Handbook of Quantitative Methods for Educational Research

Timothy Teo (Ed.)

SensePublishers

Handbook of Quantitative Methods for Educational Research

Handbook of Quantitative Methods for Educational Research

Edited by
Timothy Teo
University of Auckland, New Zealand

SENSE PUBLISHERS
ROTTERDAM/BOSTON/TAIPEI

A C.I.P. record for this book is available from the Library of Congress.

ISBN: 978-94-6209-402-4 (paperback)
ISBN: 978-94-6209-403-1 (hardback)
ISBN: 978-94-6209-404-8 (e-book)

Published by: Sense Publishers,
P.O. Box 21858,
3001 AW Rotterdam,
The Netherlands
https://www.sensepublishers.com/

Printed on acid-free paper

TABLE OF CONTENTS

FOREWORD

This is the age of "evidence" and all around are claims about the need for all to make evidence based decisions. Evidence, however, is not neutral and critically depends on appropriate interpretation and defensible actions in light of evidence. So often evidence is called for, collected, and then analysed with little impact. At other times we seem awash with data, soothed by advanced methods, and too easily impressed with the details that are extracted. Thus there seems a tension between the desire to make more meaning out of the aplenty data, and the need for interpretations that have defence and consequences.

This book shows this tension – there are many sophisticated methods now available but they require an advanced set of understandings to be able to interpret meaning and can be technically complex. With more students being less prepared in basic mathematics and statistics, taking courses in experimental design and survey methods, often these methods appear out of reach. This is notwithstanding the major advances in computer software. Not so long ago structural equation modelling required a knowledge of Greek, matrix calculus, and basic computer logic; now many programs require the facility to distinguish between boxes and circles, manipulate arrows, and read pictures. This is not a plea that only those who did it "the hard way" can appreciate the meaning of these methods – as many of these chapters in this book show how these modern methods and computer programs can advance how users think about their data and make more defensible interpretations.

The sheer number of methods outlined in the book shows the advances that have been made, and too often we can forget that many of these can be traced to some fundamental principles. The generalised regression model and the non linear factor model are two such claims for 'general models' – for example many of the item response family are variants of the non-linear factor models and understanding these relations can show the limitations and advantages of various decisions the user has to make when using these methods. For example, would a user be satisfied with a model specifying a single factor with all items loading the same on this factor – as this is what the Rasch item response model demands.

Each chapter shows some of these basic assumptions, how the methods relate to other similar methods, but most important show how the methods can be interpreted. That so many of the most commonly used methods are in one book is a major asset. The methods range from measurement models (CTT, IRT), long developed multivariate methods (regression, cluster analysis, MANOVA, factor analysis, SEM), meta-analysis, as well as newer methods include agent-based modelling, latent growth and mixture modelling.

There are many types of readers of this book, and an aim is to speak to them all. There are 'users' who read educational literature that includes these methods

and they can dip into the book to find more background, best references, and more perspective of the place and meaning of the method. There are 'bridgers" who will go beyond the users and will become more adept at using these methods and will want more detail, see how the method relates to others, and want to know how to derive more meaning and alternative perspectives on the use of the method. Then there are "clones" that will use this book to drill down into more depth about the method, use it to educate others about the method, and become more expert in their field. There are also 'lurkers", those from various disciplines who have been told to use a particular method and want a reference to know more, get an overall perspective, and begin to see how the method is meant to work. There is an art of providing "just enough" for all users, to entice them to want more, seek more, and learn more about the many aspects of the methods that can be put into a short chapter.

One of my favourite books when I was a graduate student was Amick and Walberg (1975). This book included many of the same methods in the current Handbook. I referred to it often and it became the book most often 'stolen' by colleagues and students. It became the 'go to' book, a first place to investigate the meaning of methods and begin to understand 'what to do next'. This Handbook will similarly serve these purposes. The plea, however, is to go beyond the method, to emphasise the implications and consequences. Of course, these latter depend on the appropriateness of the choice of method, the correctness in making critical decisions when using these methods, the defence in interpreting from these methods, and the quality of the data. Happy using, bridging, cloning and lurking.

John A. Hattie
University of Melbourne

REFERENCE

Amick, D., & Walberg, H. (1975). *Introductory multivariate analysis: Fro education, psychological, and social research*. Berkeley, CA: McCutchan.

SECTION 1

MEASUREMENT THEORY

MARK WILSON & PERMAN GOCHYYEV

1. PSYCHOMETRICS

Psychometrics is the study of the measurement of educational and psychological characteristics such as abilities, aptitudes, achievement, personality traits and knowledge (Everitt, 2006). Psychometric methods address challenges and problems arising in these measurements. Historically, psychometrics has been mostly associated with intelligence testing and achievement testing. In recent times, much of the work in psychometrics deals with the measurement of latent (or unobserved) traits and abilities.

In order to make our presentation both clear and accessible for those with practical interests in applying psychometrics in educational settings, this chapter is based on the *Construct Modeling* approach (Wilson, 2005): this is a "full-cycle production" measurement framework consisting of four building blocks: the *construct map*, the *items design*, the *outcome space*, and the *measurement model*. The construct modelling approach provides an explicit guiding framework for the researcher wishing to apply psychometric ideas in assessment. Activities that involve constructing and using an instrument – from hypothesizing about the construct to be measured to making interpretations and decisions – can be organised into these four building blocks. The researcher will be called the *measurer* throughout the chapter: this is the person designing and developing the measure.

For the most part, we will assume that the measurer already knows what s/he is intending to measure (at least to a certain extent). Note that this is different from the currently popular *data mining* approach (Nisbet, Elder, & Miner, 2009) where the data is expected to generate the solutions. Thus, we expect that the steps to be conducted by the measurer are *confirmatory*, rather being broadly *exploratory*. It will be helpful to note that the philosophical position of the authors is that the practice of psychometrics, and particularly the activity of *constructing measures*, is more to be considered a practical and engineering activity rather than as a basic science. Psychometricians construct measures (engineering), and build models to analyse these measures (reverse-engineering). It might not be an accident that L. L. Thurstone, a person considered to be one of the fathers of psychometrics, was a trained engineer.

MEASUREMENT

Measurement, in its broadest sense, is the process of assigning numbers to categories of observations in such a way as to represent quantities of attributes (Nunnally, 1978). Stevens (1946) noted that these numbers can be *nominal*, *ordinal*, *interval*, or *ratio.* However, simply assigning numbers at these different levels does not guarantee that

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 3–30.

the resulting measures are indeed at those corresponding levels (Michell, 1990). Instead, the level needs to be established by testing whether the *measurement model* is appropriate (van der Linden, 1992).

Corresponding to the type of measurement model that holds, measurement can be *fundamental*, *derived*, or *implicit* (van der Linden, 1992). Fundamental measurement requires that the measure has the following properties: it has an *order relation, unit arbitrariness*, and *additivity* (see Campbell, 1928). Derived measurement assumes that products of fundamental measurement are mathematically manipulated to produce a new measure (such as when density is calculated as the ratio of mass to volume). In contrast, in the implicit measurement situations in which our measurer is involved, neither of these approaches are possible: Our measurer is interested in measuring a hypothetical entity that is not directly observable, namely, the *latent variable*. Now, latent variables can only be measured indirectly via observable indicators – *manifest variables,* generically called items. For example, in the context of educational testing, if we wanted to measure the latent variable of a student's knowledge of how to add fractions, then we could consider, say, the proportion correct by each student of a set of fractions addition problems as a manifest variable indicating the student's knowledge. But note that the, the student knowledge is measured *relative to* the difficulty of the set of items. Such instances of implicit measurement can also be found in the physical sciences, such as the measure of the *hardness* of an object.

To illustrate how different fundamental measurement is from implicit measurement of a latent variable, consider the following example. If the weight of the Golden Gate Bridge is 890,000 tons and the weight of the Bay Bridge is 1,000,000 tons, then their combined weight is estimated as the sum of the two, 1,890,000 tons. However, the estimated ability of the respondent A and respondent B working together on the fractions test mentioned above would not be the sum of the performances of respondent A and respondent B separately. Implicit measurement allows quantification of latent variables provided variables are measured jointly (Luce & Tukey, 1964). For an in-depth discussion, see Michell (1990) and van der Linden (1992).

THE CONSTRUCT

Planning and debating about the purpose(s) and intended use(s) of the measures usually comes before the measurement development process itself. We will assume that the measurer has an underlying latent phenomena of interest, which we will call the *construct* (also called *propensity*, *latent variable*, *person parameter*, *random intercept*, and often symbolized by θ).

It will be assumed in this section that there is a single and definite construct that is being measured. In practice, a single test might be measuring multiple constructs. If such is the case, we will (for the purposes of this chapter) assume that each of these constructs is being considered separately. Constructs can be of various kinds: Abilities, achievement levels, skills, cognitive processes, cognitive strategies, developmental stages, motivations, attitudes, personality traits, emotional states, behavioural patterns

and inclinations are some examples of constructs. What makes it possible and attractive to measure the construct is the belief and understanding on the part of the measurer that the amount or degree of the construct varies among people. The belief should be based on a theory. Respondents to the test can be people, schools, organizations, or institutions. In some cases, subjects can be animals or other biological systems or even complex physical systems. Note that the measurer does not measure these respondents – the measurer measures the construct these respondents are believed to have.

Depending on the substantive theory underlying the construct, and one's interpretational framework, a construct could be assumed to be dimension-like or category-like. In this chapter, we will be assuming former, in which the variability in the construct implies some type of continuity, as that is most common situation in educational testing. Much of the following development (in fact virtually all of it up to the part about the "measurement model"), can be readily applied to the latter situation also—for more information on the category-like situation see Magidson & Vermunt (2002). There are many situations where the construct is readily assumed to be dimension-like: in an educational setting, we most often can see that there is a span in ability and knowledge between two extremes; in attitude surveys, we can see a span of agreement (or disagreement); in medicine, there are often different levels of a health condition or of patient satisfaction, but also a span in between. Consider the following example for better understanding of *continuity*: the variable "understanding of logarithms" can be present at many levels. In contrast, the variable "pregnancy" is clearly a dichotomy – one cannot be slightly pregnant or almost pregnant. It is possible that in some domains the construct, according to an underlying theory, has discrete categories or a set of unordered categories. A respondent might be a member of the one of the *latent classes* rather than at a point on a continuous scale. These classes can be ordered or unordered. Various models in psychometrics such as latent class models are designed to deal with constructs of that type (see Magidson & Vermunt, 2002).

The type of measurement presented in this chapter can be understood as the process of locating a respondent's location on the continuum of the latent variable. As an example, imagine a situation where one wants to find out about a respondent's wealth but cannot ask directly about it. The measurer can only ask questions about whether the respondent is able to buy a particular thing, such as "*Are you able to buy an average laptop?*" Based on the obtained responses, the measurer tries to locate the respondent on the *wealth* continuum, such as claiming that the respondent is between "*able to buy an average laptop*" and "*able to buy an average motorcycle.*"

A SURVEY OF TYPES AND PURPOSES OF MEASUREMENT

From the broadest perspective, we can distinguish two types of measurement (De Boeck & Wilson, 2006). The first type is the accurate measurement of the underlying latent variable on which the respondents are arrayed. This implies the use of the test at the level of individual respondents. Inferences regarding the individual, or perhaps groups of individuals, are of primary interest. This approach is intuitively named as

the *measurement approach.* Measurement with this purpose is also referred to as the *descriptive* measurement. In contrast, another purpose of the measurement has a different perspective. Rather than focusing on the individual, the main purpose is to seek relationships of the observations (responses to the items) to other variables. These variables can be characteristics of the respondents (gender, race, etc.), or characteristics of the items (item format, item features, etc.). This approach is referred to as the *explanatory approach*. Explanatory measurement can help in predicting behaviour in the future and can also serve to support a theory or hypothesis. As an example, a researcher might be interested in the effectiveness of the two different teaching methods. Here, the interest is in the teaching method rather than in the individual differences. A test can be designed and analysed to serve both purposes, but serving both kinds of purpose can lead to inefficiencies and challenges.

Depending on the context, the purposes of the measurement might also differ. One classification of measurement purposes in the educational context is into *norm-referenced* and *criterion-referenced* interpretations. Norm-referenced interpretations are relevant when the measurer wishes to locate a respondent's position within a well-defined group. In comparison, criterion-referenced interpretations are used in identifying a degree of proficiency in a specified content domain. College admission tests in United States (e.g., SAT, ACT) are examples of norm-referenced interpretations, as their main purpose is to rank applicants for university entrance. In contrast, criterion-referenced tests might be based on the topics in a lesson or the curriculum, or in the state standards. Some tests are designed for both types of interpretations—generally norm-referenced interpretations are always available, whereas criterion-referenced interpretations require more effort. (See below for the Construct Modeling approach to criterion-referenced measurement.)

Another perspective in looking at measurement purposes in an educational context is summative versus formative uses of tests. When a test is used to look back over what a student has learned, and summarise it, then that is a summative use. When a test is used to decide what to do next, to advance the student within a lesson, or to remediate, then that is a formative use (see Wiliam, 2011 for a broad summary of these).

From a very different perspective, the measurement, or more precisely the measurement model, can be *reflective* versus *formative*[1]. In the reflective measurement approach to modeling, which is the type of measurement model considered in this chapter and the common assumption among a majority of psychometricians, the belief is that the responses to the items are the indicators of the construct and the construct (effectively) "causes" respondents to respond to the items in such way. In contrast, in the formative measurement approach to model, which is more popular in the fields of sociology and economics, the assumption is that it is the items that influence the latent variable. For instance, returning to our example about the wealth construct above: (a) from the reflective perspective we assume that the person's location on the *wealth* construct will cause respondents to answer questions such as "*are you able to buy an average laptop?*"; but (b) from the formative perspective, the assumption is that responses to these items will "cause" the *wealth* latent variable.

(Note that we avoid using the word *construct* in the latter case, as it is discrepant to our definition of the construct. The terms *index* is often used in the formative case.)

CONSTRUCT MODELING: THE FOUR BUILDING BLOCKS APPROACH

We now outline one particular approach to developing measures—Construct Modeling. We do not claim that this is a universally optimal way to construct measures, but we do see it as a way to illustrate some of the basic ideas of measurement. Note that, although we present just a single cycle of development, one would usually iterate through the cycle several times. The Construct Modelling approach is composed of Four Building Blocks[2]: the Construct Map, the Items Design, the Outcome Space, and the Measurement Model. Note that we will label the person being measured as the "respondent" (i.e., the one who responds to the item).

The Construct Map

In order to help one think about a construct, we present the *construct map* (Wilson, 2005). Thinking in the "construct map" way prompts one to consider both sides of the measurement situation: the respondent side and the item side. A construct map is based on an ordering of both respondents and the items from a lower degree to a higher degree. A generic example of the basic form of the construct map is shown in Figure 1.[3] Respondents who possess a low degree of the construct (bottom left), and the responses that indicate this amount of the construct (bottom right) are located at

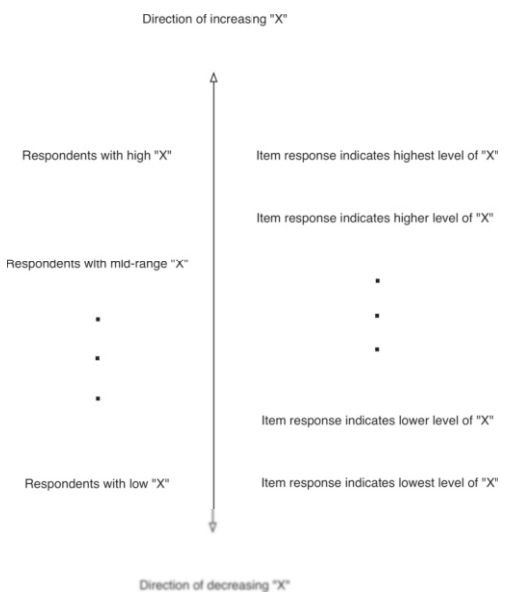


Figure 1. A generic construct map for the construct "X".

the bottom of the construct map. Similarly, respondents who possess a high degree of the construct (top left), and the responses that indicate this amount of the construct (top right) are located at the top of the construct map. In between these extremes are located qualitatively different locations of the construct, representing successively higher intensities of the construct.

Depending on the hypothesis and the setting being applied, construct maps can be connected or nested within each other and interpreted as *learning progressions*. (See Wilson, 2009 for illustrations of this.)

The construct map approach advances a coherent definition of the construct and a working assumption that it monotonically spans the range from one extreme to another – from low degree to high degree. There might be some complexities between the two extremes. We are interested in locating the respondent on the construct map, the central idea being that, between the two extremes, the respondent higher on the continuum possesses more of that construct than the respondent lower on the continuum. Thus, a respondent higher on the continuum has a better chance to be observed demonstrating the higher levels of the responses. This is called the assumption of *monotonicity.*[4]

The idea of a construct map forces the measurer to take careful consideration of the theory concerning the construct of interest. A clear definition of what is being measured should be based on the body of literature related to the construct of interest. The definition of the construct shouldn't be too vague, such as, for instance the definition of "intelligence" given by Galton (1883), as: "that faculty which the genius has and the idiot has not." It is best to support the hypothetical nature and order of the locations in the construct map from a specific theory. The coherence of the definition of the construct in the construct map requires that the hypothesized locations be clearly distinguishable. Note that the existence of these locations does not necessarily contradict the concept of an underlying continuum, as they can readily represent distinct identifiable points along a continuous span.

The advantage of laying out the construct on the construct map is that it helps the measurer make the construct explicit. Activities that are carried out in the *construct map* phase can also be described as *construct explication* (Nunnally, 1978) – a term used to describe the process of making an abstract concept explicit in terms of observable variables.

Note that each respondent has only one location on the hypothesized unidimensional (i.e., one-trait, single-factor) construct. Of course, the construct of interest might be multi-dimensional and thus the respondent might have multiple locations in the multidimensional space of several construct maps. As was noted earlier, for simplicity, we are assuming one-dimensional construct, which is believed to be recognizably distinct from other constructs. This is also called the assumption of *unidimensionality*. Note that this assumption relates to the set of items. If the construct of interest is multidimensional, such as "achievement in chemistry", which can have multiple dimensions (see, Claesgens, Scalise, Wilson & Stacy, 2009), each strand needs to be considered separately in this framework to avoid ambiguity, although the measurement models can be multidimensional (e.g., see Adams, Wilson, & Wang, 1997). For

example, consider the following two variables: (a) the wealth of a person, and (b) the cash readily available to a person. Although we would expect these two variables to be highly correlated, nevertheless, each person would have two distinct locations.

A Concrete Example: Earth and the Solar System. This example is from a test of science content, focusing in particular on earth science knowledge in the area of "Earth and the Solar System" (ESS). The items in this test are distinctive, as they are Ordered Multiple Choice (OMC) items, which attempt to make use of the cognitive differences built into the options to make for more valid and reliable measurement (Briggs, Alonzo, Schwab & Wilson, 2006). The standards and benchmarks for "Earth in the Solar System" appear in Appendix A of the Briggs et al article (2006). According to these standards and the underlying research literature, by the 8th grade, students are expected to understand three different phenomena within the ESS domain: (1) the day/night cycle, (2) the phases of the Moon, and (3) the seasons—in terms of the motion of objects in the Solar System. A complete scientific understanding of these three phenomena is the top location of our construct map. See Figure 2 for the ESS construct map. In order to define the lower locations

Location	Description
5 8th grade	Student is able to put the motions of the Earth and Moon into a complete description of motion in the Solar System which explains: • the day/night cycle • the phases of the Moon (including the illumination of the Moon by the Sun) • the seasons
4 5th grade	Student is able to coordinate apparent and actual motion of objects in the sky. Student knows that: • the Earth is both orbiting the Sun and rotating on its axis • the Earth orbits the Sun once per year • the Earth rotates on its axis once per day, causing the day/night cycle and the appearance that the Sun moves across the sky • the Moon orbits the Earth once every 28 days , producing the phases of the Moon COMMON ERROR: Seasons are caused by the changing distance between the Earth and Sun. COMMON ERROR: The phases of the Moon are caused by a shadow of the planets, the Sun, or the Earth falling on the moon.
3	Student knows that: • the Earth orbits the Sun • the Moon orbits the Earth • the Earth rotates on its axis However, student has not put this knowledge together with an understanding of apparent motion to form explanations and may not recognize that the Earth is both rotating and orbiting simultaneously. COMMON ERROR: It gets dark at night because the Earth goes around the Sun once a day.
2	Student recognizes that: • the Sun appears to move across the sky every day • the observable shape of the Moon changes every 28 days Student may believe that the Sun moves around the Earth. COMMON ERROR: All motion in the sky is due to the Earth spinning on its axis. COMMON ERROR: The Sun travels around the Earth. COMMON ERROR: It gets dark at night because the Sun goes around the Earth once a day. COMMON ERROR: The Earth is the center of the universe.
1	Student does not recognize the systematic nature of the appearance of objects in the sky. Students may not recognize that the Earth is spherical. COMMON ERROR: It gets dark at night because something (e.g., clouds, the atmosphere, "darkness") covers the Sun. COMMON ERROR: The phases of the Moon are caused by clouds covering the Moon. COMMON ERROR: The Sun goes below the Earth at night.
0	No evidence or off-track

Figure 2. Construct map for student understanding of earth in the solar system.

of our construct map, the literature on student misconceptions with respect to ESS was reviewed by Briggs and his colleagues. Documented explanations of student misconceptions with respect to the day/night cycle, the phases of the Moon, and the seasons are displayed in Appendix A of the Briggs et al article (2006).

The goal was to create a single continuum that could be used to describe typical students' understanding of three phenomena within the ESS domain. In contrast, much of the existing literature documents students' understandings about a particular ESS phenomena without connecting each understanding to their understandings about other related ESS phenomena. By examining student conceptions across the three phenomena and building on the progressions described by Vosniadou & Brewer (1994) and Baxter (1995), Briggs et al. initially established a general outline of the construct map for student understanding of ESS. This general description helped them impose at least a partial order on the variety of student ideas represented in the literature. However, the locations were not fully defined until typical student thinking at each location could be specified. This typical student understanding is represented in the ESS construct map shown in Figure 2, (a) by general descriptions of what the student understands, and (b) by limitations to that thinking in the form of misconceptions, labeled as "common errors." For example, common errors used to define category 1 include explanations for day/night and the phases of the Moon involving something covering the Sun or Moon, respectively.

In addition to defining student understanding at each location of the continuum, the notion of common errors helps to clarify the difference between locations. Misconceptions, represented as common errors at one location, are resolved at the next higher location of the construct map. For example, students at location 3 think that it gets dark at night because the Earth goes around the Sun once a day—a common error for location 3—while students at location 4 no longer believe that the Earth orbits the Sun daily but rather understand that this occurs on an annual basis.

The top location on the ESS construct map represents the understanding expected of 8th graders in national standards documents. Because students' understanding of ESS develops throughout their schooling, it was important that the same continuum be used to describe the understandings of both 5th and 8th grade students. However, the top location is not expected of 5th graders; equally, we do not expect many 8th grade students to fall among the lowest locations on of the continuum.

The Items Design

Items are the basic building blocks of the test. Each item is a stimulus and each use of it is an attempt to obtain an observation that usefully informs the construct. In order to develop these items in an orderly way, there needs to exist a procedure of designing these observations, which we call the *items design*. In a complementary sense, the construct may not be clearly and comprehensively defined until a set of items has been developed and tried out with respondents. Thus, the development of items, besides its primary purpose to obtain a useful set of items, plays an important

step in establishing that a variable is measureable, and that the ordered locations of the construct map are discernible.

The primary purpose of the items is to prompt for responses from the respondents. Items should be crafted with this in mind. Items with different purposes, such as the ones that teach the content of the test, may be costly in terms of efficiency, but, of course, may also play an important part in instruction. It is possible to see each item as a mini-test, and we will see the usefulness of this type of thinking when talking about the indicators of the instrument quality later in the chapter. Thus, a test can be seen as a set of repeated measures, since more than one observation is made for the respondent, or, put another way, a test can be considered an experiment with repeated observations—this perspective places models commonly used in psychometrics in a broader statistical framework see, for example, De Boeck & Wilson, 2004).

Item formats. Any systematic form of observation that attempts to reveal particular characteristics of a respondent can be considered as an item. Information about the construct can be revealed in many ways, in, say, a conversation, a directly asked question, or from observing respondents, in both formal and informal settings. As was mentioned above, at early stages, information revealed in any of these ways can be used to clarify the ordered locations of the construct. The item format should be appropriate to the nature of the construct. For instance, if one is interested in respondent's public speaking skills, the most appropriate format is direct observation, where the respondent speaks in public, but this is just the start of a range of *authenticity* which ranges all the way to self-report measures.

The open-ended item format is probably the most basic and the most "unrestrictive" format. In this format the responses are not limited to predefined categories (e.g., True or False), and there may be broad latitude in terms of modes of communication (e.g., written, figurative, or oral), and/or length. Open-ended items are the most common type of format that are typically observed in informal and social settings, such as within classrooms. However, due to their simplicity for evaluation, the most common item format used in formal instruments is the *fixed-response format.* Commonly, fixed-response format items will start out as in an open-ended item format—the responses to these can be used to generate a list of the types of responses, and this in turn can be used to design multiple alternatives. A fixed-response format is also very common in attitude surveys, where respondents are asked to pick the amount of intensity of the construct (i.e., Strongly Agree/Agree/etc.). This item format is also referred to as the Likert-type response format (Likert, 1932).

The list of alternative ways to give respondents a chance to reveal their place on the construct has expanded with the advances in technology and computerized testing. New types of observations such as simulations, interactive web-pages, and online collaborations require more complex performances from the respondent and allow the delineation of new locations on constructs, and sometimes new constructs altogether (Scalise & Wilson, 2011). The potential of these innovative item formats is that they might be capable of tapping constructs that were "unreachable" before.

Item development. The item development process requires a combination of art and creativity on the part of the measurer. Recall that an *item*, regardless of the format, should always *aim*[5] at the construct. Ramsden, Masters, Stephanou, Walsh, Martin, Laurillard & Marton (1993), writing about a test of achievement in high school physics noted:

> Educators are interested in how well students understand speed, distance and time, not in what they know about runners or powerboats or people walking along corridors. Paradoxically, however, there is no other way of describing and testing understanding than through such specific examples.

Sometimes it may be sufficient to simply ask for a formal "piece of knowledge"—the product of 2 and 3, or the freezing point of water in centigrade, etc.—but most often we are interested in seeing how the respondent can use their knowledge and skills.

One important aspect is the planned difficulty of the test and its respective items. One needs to consider the purpose of the instrument when selecting an appropriate difficulty level for the items. Often, items are arranged from the easiest to the most difficult one, so that respondents do not become frustrated and not get to relatively easy items. In general, the measurer needs to develop items that aim at *all* locations of the construct. (This point will be elaborated on in the validity section below.)

Another important aspect is the "grainsize" of the items. Each item, in order to provide a contribution in revealing the amount of the construct, should span at least two locations of the construct. For example, a dichotomous item will aim at *at or above* the location of the item and *below* the location of the item. A polytomous item might aim at more than two locations. Note that Likert items, by their design will generally aim at more than two locations.

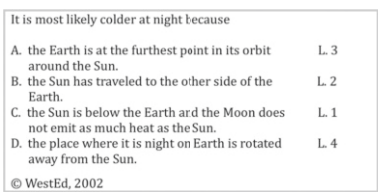
One more important activity that needs to be occuring in this phase is "listening to respondents" (AERA/APA/NCME, 1999). This activity is a very effective tool for "tuning up" the items of the instrument. Listening can either be in the form of *think alouds* or in the form of *exit interviews* (sometimes called "cognitive interviews"). In think alouds, participants are prompted to say aloud what they are thinking as they are working on the tasks. The measurer tries to take a note of everything the respondent says without any filtering. Of course, this sort of self-report has strong limitations, but at least it can indicate the sorts of issues that the respondent is working through. In exit interviews, the measurer interviews the respondent after the test is over. There should not be a long gap in time between the administration of the instrument and the exit interview. Exit interviews can be conducted over the phone, in-person, or using paper-and-pencil or a computerized survey. The findings from both think alouds and exit interviews need to be well-documented. It is recommended that the sessions be audio or video-taped, both in order to be able to return to the evidence later in the process of instrument development and to document such valuable evidence. As we will see later (in the *Validity* section), this evidence will prove to be an important one for validating the test. Also, as is the case with all steps, it is very important that the measurer stays neutral throughout the entire process.

The ESS Example Continued. Returning to the ESS example, the OMC items were written as a function of the underlying construct map, which is central to both the design and interpretation of the OMC items. Item prompts were determined by both the domain as defined in the construct map and canonical questions (i.e., those which are cited in standards documents and commonly used in research and assessment contexts). The ESS construct map focuses on students' understanding of the motion of objects in the Solar System and explanations for observable phenomena (e.g., the day/night cycle, the phases of the Moon, and the seasons) in terms of this motion. Therefore, the ESS OMC item prompts focused on students' understanding of the motion of objects in the Solar System and the associated observable phenomena. Distractors were written to represent (a) different locations on the construct map, based upon the description of both understandings and common errors expected of a student at a given location and (b) student responses that were observed from an open-ended version of the item. Each item response option is linked to a specific location on the construct map, as shown in the example item in Figure 3. Thus, instead of gathering information solely related to student understanding of the specific context described in the question, OMC items allow us to link student answers to the larger ESS domain represented in the construct map. Taken together, a student's responses to a set of OMC items permit an estimate of the student's location on the ESS construct, as well as providing diagnostic information about that specific misconception.

The Outcome Space

As has been pointed out above, an instrument can be seen as an experiment used to collect qualitative data. However, in the behavioural and social sciences, the measuring is not finished when data are collected – much needs to happen after the data are collected (van der Linden, 1992). The *outcomes space* is the building block where the responses start to be transformed into measures. The main purpose of the outcome space is to provide a standard procedure to categorize and order observations in such a way that the observed categories are informative about the locations on the construct.

The outcomes space as a term was first used and described by Marton (1981). He used students' responses to open-ended items to discover qualitatively different


It is most likely colder at night because

A. the Earth is at the furthest point in its orbit around the Sun. L. 3
B. the Sun has traveled to the other side of the Earth. L. 2
C. the Sun is below the Earth and the Moon does not emit as much heat as the Sun. L. 1
D. the place where it is night on Earth is rotated away from the Sun. L. 4

© WestEd, 2002

Figure 3. A sample OMC item based upon ESS construct map. (L indicates location on construct map.)

ways students responded to sets of tasks. Dahlgren (1984) described an outcome space as a sort of analytic map:

> It is an empirical concept which is not the product of logical or deductive analysis, but instead results from intensive examination of empirical data. Equally important, the outcome space is content-specific: the set of descriptive categories arrived at has not been determined a priori, but depends on the specific content of the [item]. (p. 26)

Within the Four Building Blocks framework, the term *outcomes space* has a somewhat broader meaning. The outcome space is an ordered, finite, and exhaustive set of well-defined, research-based, and context-specific categories (Wilson, 2005). That the categories are a finite set means that the possibly infinite number of potential responses needs to be categorized into a small (but not too small) set of categories. That the categories are exhaustive means that the categories should be inclusive—every possible response has a place (at least potentially) among the categories. That the categories are ordered means that there exists an ordering of the categories that is consistent with the ordered locations on the construct map—though the ordering might only be partial. That the categories are well-defined means that the measurer must have a way to consistently categorize the responses into the categories—this might include having: (a) definitions of the construct locations; (b) background materials explaining important concepts, etc., involved in the locations; (c) samples of the items and responses for each locations; and (d) a training procedure for raters. As was noted earlier, concerning the locations of the construct map, the categories of the outcome space need to be research-based, that is, informed by appropriate research and theory. That the categories are context-specific means that nature of the construct need to be considered when developing the categories. For example, the requirement that the alternatives to the correct prompt in multiple-choice items be superficially reasonable is one such.

Scoring. Scoring is the procedure of assigning numerical values to the ordered locations of the outcome space. Scoring should be designed so that the categories can be related back to the responses side of the construct map. The traditional procedure for multiple-choice items is to score the correct response as unity and the incorrect ones as zero. For OMC items, the ordered locations may be used as a basis for scoring. For Likert-style response items, the lowest extreme (e.g., "Strongly disagree") is often scored as zero and each subsequent category as 1, 2, 3, etc., respectively.

Open-ended items require more effort for coding and scoring. The outcome categories must be ordered into qualitatively distinct locations on the continuum, with possibly several categories within each location. Coding open-ended items can be expensive and time-consuming. With the developments of machine learning techniques, it is becoming possible to use computers to categorize and score open-ended items (Kakkonen, Myller, Sutinen, & Timonen, 2008).

Missing responses should be handled appropriately in the scoring process. If the measurer has a reasonable belief that the response is missing because the respondent was not administered the item, coding it as "missing" is an appropriate choice. If the measurer judges that the response was missing due to the high difficulty of the item (such as when a respondent fails to respond to a string of hard items at the end of the test), the missing response could be coded as zero. Although missing response indicates no information about the respondent in relation to the item, investigating potential reasons for missing responses might be a useful strategy to improve the items.

The ESS Example Continued. In the ESS example, the outcome space is simply the locations of the ESS Construct Map (see Figure 2). And the scoring guide for each item is given simply by the mapping of each item distractor to its respective location on the construct map, as exemplified for the item in Figure 3. This need not be the case, items may be developed that have much more complex relationships with the relevant construct map.

The Measurement Model

The *measurement model* phase of *Construct Modeling* closes the cycle, relating the scored outcomes back to the construct map. The measurement model predicts the probability of the response of a respondent to a particular item conditional on the respondent's location on the ability continuum and the item's location on difficulty *in relation to* the construct. The measurement model should help the measurer interpret the distance between (a) a respondent and a response on the construct map; and (b) different responses and different respondents on the construct map. The primary function of the measurement model is to bridge from the scores produced by the outcome space back to the construct map.

We will start by discussing two different approaches to the measurement model. The first approach focuses on the scores, and its relation to the construct – namely, the *instrument-focused* approach. The *instrument-focused* approach was the main driving force of *Classical Test Theory* (CTT; Spearman, 1904). The fundamental relationship in CTT is the relationship of the *true score* (T) with the *observed score* (*X*):

$$X = T + E, \tag{1}$$

where E is the error, and where the true score is understood as the average score the respondent would obtain over many hypothetical re-tests, assuming there are no "carry-over" effects.[6] In contrast, the second measurement approach focuses on each item and its relationship to the construct – thus, termed as the *item-focused* approach. The most prominent example of the item-focussed approach is the work of Guttman (1944, 1950), who based his *scalogram* approach on the idea that tests could be developed for which respondents would invariably respond according

to the (substantive) difficulty order of the items. This assumption of invariance allows a very straightforward item-wise interpretation of the respondents' scores. Although this approach was an important advancement in the conceptualization of psychometrics, the dependence of Guttman's approach on the invariant ordering has been found to be impracticable (Kofsky, 1996). The Construct Modelling approach can be seen as a synthesis of the item-focused and instrument-focused approaches.

There have been a numerous measurement models proposed within the last several decades. We will focus on one such model, namely the *Rasch* model (Rasch, 1960), due to (a) its interpretational simplicity and (b) its alignment with the measurement framework presented in this chapter[7]. The construct modelling approach is both philosophically and methodologically based on the work of Georg Rasch, a Danish mathematician, who first emphasized the features of his epynomous Rasch model. Parallel to this development by Rasch, similar approaches were also being developed, generally under the label of Item Response Theory or Latent Trait Theory (van der Linden & Hambleton, 1997; Chapter 3, this volume).

Generally, given the uncertainty inherent in sampling a respondent's relationship to a construct via items, it makes sense that one would prefer a measurement model that aligns with a probabilistic formulation. A major step forward in psychometrics occurred when the test items themselves were modelled individually using probabilistic models as opposed to deterministic models. Where the deterministic approach focuses on the responses itself, this probabilistic approach is focused on the probability of the correct response (or endorsement). In the case of the Rasch model, the probabilistic function is dependent on the item location and respondent location. Depending on the context, item location can be, for instance, interpreted as the difficulty of responding correctly or difficulty of endorsing a particular statement. The respondent location is the point where the respondent is located on the construct continuum: It can be interpreted as the respondent's ability to answer the item correctly or to endorse a particular statement. The distance between the item location and the person location is the primary focus of the model and also the feature that provides for ease of interpretation.

The Rasch model asserts that the probability of a particular response depends *only* on the person location (θ) and item location (δ). Mathematically, this statement is represented as

$$\text{Probability}(\text{correct}|\theta, \delta) = f(\theta - \delta) \tag{2}$$

The requirement for the person and item locations (person and item *parameters*) is that both are unbounded (there can always be a higher respondent or more difficult item), thus $-\infty < \theta < \infty$, and $-\infty < \delta < \infty$, but the probability is, of course, bounded between 0 and 1. The two most common probabilistic models are based on the logistic and cumulative normal functions—the Rasch model uses the logistic formulation. With a multiplicative constant of 1.7, the two are very similar, particularly in the range

of –3 and 3 (Bradlow, Wainer, & Wang, 1999). Specifically, the logistic expression for the probability of a correct response on an item (represented as: $X = 1$) is:

$$\text{Probability}(X = 1|\theta, \delta) = \exp(\theta - \delta)/\Phi, \quad (3)$$

and the probability of an incorrect response on an item (represented as: $X = 0$) is:

$$\text{Probability}(X = 0|\theta, \delta) = 1/\Phi, \quad (4)$$

where Φ is a normalizing constant, the sum of the numerators:

$$1 + \exp(\theta - \delta).$$

The *item response function* (IRF, sometimes called the item characteristic curve—ICC) summarizes the mathematical expression of the model by illustrating the relationship between the probability of the response to an item and the ability of the respondent. (See Figure 4.)

In order to calculate the probability of an observed response vector over a set of items, the probabilities for each item are multiplied together, relying on the assumption of *local independence.* Items are locally independent of each other if, once we know the respondent and item locations, there is no more information needed to calculate their joint probability. This assumption can be violated when several items have a relationship over and above what would be indicated by their respective difficulties, and the respondents' abilities. For example, if several items relate to the same stimulus material, such as in a paragraph comprehension test, then we would suspect that there might be such a relationship. In this case, understanding or misunderstanding the paragraph can improve and/or worsen performance on all items in the set, but not on other items in the test. Elaborations of basic models that account for this type of dependence have been proposed (see Wilson & Adams, 1995, Bradlow, Wainer, & Wang, 1999, and Wang & Wilson, 2005).

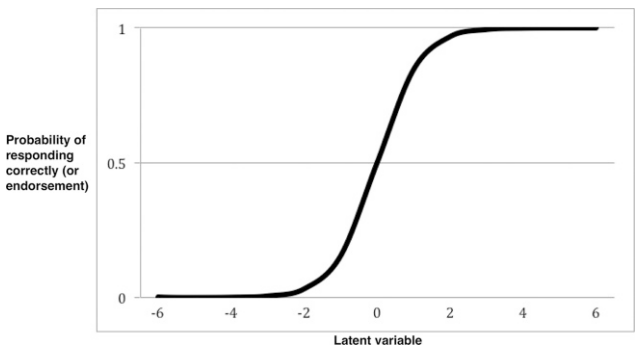


Figure 4. Item response function of the Rasch model (note, for this item, $\delta = 0.0$).

In the Rasch model, the total score of the correct (endorsed) items is monotonically (but not linearly) related to the estimated ability.[8] This property of the Rasch model will be elaborated and its implications will be described below. One fundamental property that is associated with the Rasch model is what is referred as the *sufficient statistic* – the total number of correct responses by the respondent is said to be sufficient for the person ability, which means that there is no more information available in the data that can inform the estimation of the item difficulty beyond the number correct. This concept also applies to the items – the total number of respondents responding correctly to the item is a sufficient statistic for the item difficulty. Most measurement models do not have this property.

One implication of this feature is that Rasch model is simple to interpret and explain compared to more complicated models with more complex scoring and/or parameterization. Models of the latter type might make it difficult to justify the fairness of the test to the public, such as when a respondent with a higher total score is estimated at lower location than the respondent with a lower total score.[9]

The second implication, stemming from the same argument, is that all items provide the same amount of information (all items are assumed to be equally good measures of the construct). Items differ only in difficulties. The higher the person location relative to the item location, the more likely it is that the respondent will answer correctly (endorse) the item. Thus, when this assumption is true, only two parameters (person location and item location) are needed to model achievement on the item.

A further manifestation of the uniqueness of the Rasch model is referred to as *specific objectivity* (Rasch, 1960). This can be understood in the following way: if the Rasch model holds true, then locations of two respondents on a test can be compared with each other regardless of the difficulties of the items used to measure them, and symmetrically, the locations of two items can be compared with each other regardless of the locations of the respondents answering the items.

Choosing the measurement model. Of course, all models are less complex than reality, and hence, all models are ultimately wrong—this applies to measurement models as much as any others. Some models are more suitable than others, depending on the hypothesized construct, one's beliefs, the nature of the instrument, the sample size, and the item type. Nevertheless, in the process of modelling, one must posit a sensible starting-point for model-building.

Among many criteria in choosing the model, one principle that guides the choice is the law of parsimony, also referred as Occam's razor, as Occam put it:

It is vain to do with more what can be done with fewer[10]

Thus, among the models, generally the more parsimonious models (models with fewer parameters and more degrees of freedom) will offer interpretational advantages. For example, linear models are in most instances, easier to interpret than non-linear ones. A more parsimonious model should be (and will be) a consequence

of good design, and in this context, good design includes careful development and selection of the items.

Models can be categorized according to various criteria. A model can be deterministic vs. probabilistic, linear vs. nonlinear, static vs. dynamic, discrete vs. continuous, to name several such categorizations. Some models can allow one to incorporate subjective knowledge into the model (i.e., Bayesian models), although, in truth, *any* assumption of the form of an equation is a subjective judgement. The ideal measurement model should provide a best possible basis for interpretation from the data – the central idea being to approximate ("fit") the real-world situation, at the same time having not so-many parameters as to complicate the interpretation of the results. The evaluation of the model is based on checking whether the mathematical model provides an accurate description of the observed data. For this the model "fit" is an important test whether our measurement procedure was successful. (see De Ayala, 2009 and Baker & Kim, 2004).

For the Rasch model to fit, the data should meet the relevant fit criteria. One measure of the fit of the items in the Rasch model, known as the item and respondent fit (or misfit) statistic, is obtained by comparing the observed patterns of responses to the predicted patterns of responses (See, e.g., Embretson & Reise, 2000). This type of diagnostic is an important validation step and check of the model fit. Items that are different in their measurement quality from other items (those with different slopes) need to be reconsidered and investigated. The measurer should filter out items that do not fit with the model. The idea of filtering due to the model fit has been a source of debates for many years. The approach described here might be considered a strict standard, but this standard provides for relatively straightforward interpretation via the Wright map (as described below).

The Wright Map. The Wright map provides a visual representation of the relationship between the respondent ability and the item difficulty estimates by placing them on the same logit[11] scale. This provides a comparison of respondents and items that helps to visualize how appropriately the instrument measures across the ability range. An example of a hypothetical Wright map for science literacy (including the ESS items) is shown in Figure 5. The left side of the map shows examinees and their locations on the construct: respondents estimated to have the highest ability are represented at the top, and each "X" represents a particular number of respondents (depending on the sample size). The items are represented on the right side of the map and are distributed from the most difficult at the top to the least difficult at the bottom. When the respondent and the item have the same logit (at the same location), the respondent has approximately a 50% probability of answering the item correctly (or endorsing the item). When the respondent is above the item, the probability is higher, when the respondent is below, it is lower. In this way, it is easy to see how specific items relate both to the scale itself and to the persons whose abilities are measured on the scale. The placement of persons and items in this kind

of direct linear relationship has been the genesis of an extensive methodology for interpreting the measures (Masters, Adams & Wilson, 1990; Wilson, 2005; Wright, 1968; Wright, 1977).

For example, segments of the line representing the measurement scale can be defined in terms of particular item content and particular person proficiencies. This allows the measurer to make specific descriptions of the progress of students or other test-takers whose ability estimates place them in a given segment. The set of such segments, illustrated in Figure 5 using Roman numerals II, IV and V, can be interpreted as qualitatively distinct regions that characterize the successive ordered locations on the outcome variable. Defining the boundaries of these 'criterion zones' is often referred to as standard setting. Wright Maps have proven extremely valuable in supporting and informing the decisions of content experts in the standard setting process. See Draney & Wilson (2009) and Wilson & Draney (2002) for descriptions of standard setting techniques and sessions conducted with Wright Maps in a broad range of testing contexts.

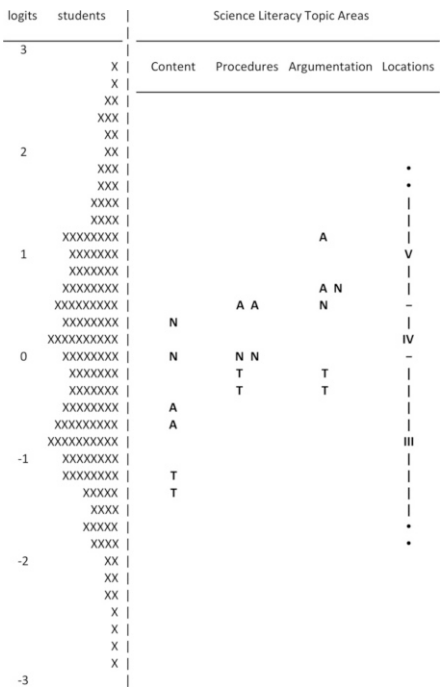


Figure 5. A Wright map of the scientific literacy variable.

Comments. (a) Each 'X' represents 5 cases; (b) "T", "N", and "A" represent different types of items; (c) Roman numerals II, IV and V represent different locations of the construct.

VALIDITY AND RELIABILITY

The two most fundamental concepts in psychometrics are test *reliability* and test *validity*. Statistical procedures exist to estimate the level of test reliability, and reasonably simple and general procedures are available to increase it to desirable levels. But statistical procedures alone are not sufficient to ensure an acceptable level of validity. Regardless of their separate consideration in much of the literature, the view of the authors is that two concepts are closely related.

Reliability

The reliability of a test is an index of how consistently a test measures whatever it is supposed to measure (i.e., the construct). It is an integral part of the validity of the test. If the instrument is sufficiently reliable, then the measurer can assume that measurement errors (as defined via Equation 1) are sufficiently small to justify using the observed score.

Thus, one can see that the closer the observed scores are to the true scores, the higher the reliability will be. Specifically, the reliability coefficient is defined as the ratio of the variance of these true scores to the variance of the observed scores. When a respondent provides an answer to the item, there are influences on the response other than the true amount of the construct, and hence, the estimated ability will differ from the true ability due to those influences. There are many potential sources for measurement error in addition to the respondents themselves, such as item ordering, the test administration conditions and the environment, or raters, to name just a few. Error is an unavoidable part of the measurement process that the measurer always tries to reduce.

The reliability coefficients described below can be seen as summaries of measurement error. The logic of most of these summary indices of measurement error is based on the logic of CTT, but this logic can readily be re-expressed in the Rasch approach. Note that the values calculated using them will be dependent on the qualities of the sample of respondents, and on the nature and number of the items used.

Internal consistency coefficients. Internal consistency coefficients inform about the proportion of variability accounted for by the estimated "true ability" of the respondent. This is equivalent to the KR-20 and KR-21 coefficients (Kuder & Richardson, 1937) for dichotomous responses and the coefficient alpha (Cronbach, 1951; Guttman, 1944) for polytomous responses. By treating the subsets of items as repeated measures (i.e., each item thought of as a mini-test), these indices apply the idea of replication to the instrument that consists of multiple items. There are no absolute standards for what is considered an adequate level of the reliability coefficient: standards should be context-specific. Internal consistency coefficients count variation due to the item sampling as error, but do not count day-to-day

variation as error (Shavelson, Webb & Rowley, 1989). The IRT equivalent of these coefficients is called the separation reliability (Wright & Stone, 1979).

Test-retest reliability. Test-retest reliability is in some respects the complement of the previous type of reliability in that it does *count day*-to-*day variation* in performance as *error* (*but not the variation due* to the *item sampling)*. The test-retest index is simply the correlation between the two administrations. As the name of the index implies, each respondent gives responses to the items twice, and the correlation of the responses on the test and the retest is calculated. This type of index is more appropriate when a relatively stable construct is of interest (in order to make sure that no significant true change in the construct is influencing the responses in the re-administration of the instrument). In addition, it is important that the respondents are not simply remembering their previous responses when they take the test the second time—the so-called "carry-over" effect (mentioned above). When calculating test-retest reliability, the time between the two administrations should not be too long in order to avoid true changes in the construct; and should not be too short in order to avoid the carry-over effect.

Alternate-forms reliability. Alternate-forms reliability counts both variation due to the item sampling and day-to-day variation as error. In calculating this index, two alternate but equivalent forms of the test are created and administered and the correlation between the results is calculated. Similarly, a single test can be split into two different but similar halves and the correlation of the scores on these two halves can be computed—the resulting index is what is referred to as the *split-halves* reliability. In this case, the effect of reducing the effective number of items needs to be taken into account using the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910) Using this formula, the measurer can estimate the reliability of the score that would be obtained by doubling the number of items, resulting in the hypothetical reliability (see Wilson, 2005, pg. 149).

Inter-rater reliability. The concept of reliability also applies to raters. Raters and judges themselves are sources of uncertainty. Even knowledgeable and experienced raters rarely are in perfect agreement, within themselves and with one another. There are four different types of errors due to raters: (a) *severity* or *leniency*, (b) *halo effect*, (c) *central tendency*, and (d) *restriction of range* (For more information, see Saal, Downey, & Lahey, 1980).

Generalizability Theory. The concept of reliability is central to a branch of psychometrics called generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Generalizability theory focuses on (a) the study of types of variation that contribute to the measurement error and (b) how accurately the observed scores allow us to generalize about the respondents' behaviour in a defined

universe of situations. “The question of reliability thus resolves into a question of accuracy of generalization, or generalizability” (Cronbach et al., 1972, p.15). For an introduction to generalizability theory see Shavelson, Webb & Rowley (1989).

Validity

A test is considered valid if it measures what it claims to be measuring. Test validity can be better understood from the causal inference perspective: for the test to be a *perfectly* valid, the degree of the construct (or presence or absence of it) should be the only cause for the observed responses—but this we know to be unattainable. This also implies that solely statistical procedures will hardly ensure validity – correlations and other forms of statistical evidence will provide only a partial support for test validity. Without a careful validation procedure, no amount of statistical methodology can provide the jump from correlation to causation.

Validity of the instrument’s usage requires evidence as to whether the instrument does indeed accomplish what it is supposed to accomplish. In general, a validity argument in testing consists of not only providing evidence that the data support the intended use and the inferences, but also showing that alternative explanations are less warranted (Messick, 1989).

Many contemporary authors endorse the view that validity is based on a holistic argument (e.g., the “Test Standards”—AERA/APA/NCME, 1999; Kane, 2006). Nevertheless, evidence for validity can be of various strands (AERA/APA/NCME, 1999). These different strands of argument will be considered next.[12]

Evidence based on the instrument content. Evidence of this kind is an attempt to answer the question: What is the relationship between the content of the test and the construct it is designed to measure? The measurer should study and confirm this relationship using whatever evidence is available[13]. This is in fact what is happening when one goes through the Four Building Blocks process described above. Going beyond a mere definition of the construct, all the steps described in the four building blocks can provide useful evidence: the development of the construct, the crafting of the set of items, the coding and scoring of responses according to the outcome space, and the technical calibration and representation of the construct through the Wright map. Evidence based on instrument content is the central and first part of the validity study – this evidence is a prerequisite for all the other strands of evidence to be useful, in the sense that all the other forms of evidence are conceptually based on this first strand.

Evidence based on the response processes. Asking respondents what they are thinking about during and after the test administration provides validity evidence based on the response processes involved in answering the items. Recall that this information should also be used during the process of item development in order to improve the items. As was mentioned above, the two major methods of

investigations of response processes are think alouds and interviews. Reaction time and eye movement studies have also been proposed as other methods to gather such evidence (Ivie & Embretson, 2010; National Research Council, 2008). With the use of computerized testing, recording the actions by the respondents such as movement of the mouse cursor and log of used functions and symbols can also serve as useful information for this strand of evidence (Cooke, 2006).

Evidence based on the internal structure. If the measurer follows the steps of the four building blocks, a hypothesized internal structure of the construct will be readily provided via the ordered locations. The agreement of the theoretical locations on the construct map to the empirical findings in the Wright map provides direct evidence of internal structure. The measurer needs to compare the hypothesized order of the items from the construct map to the order observed from the Wright maps: A *Spearman rank-order correlation* coefficient can be used to quantify this agreement (see Wilson, 2005, p. 160). The higher the correlation, the better is the match (note that there is no predetermined lowest acceptable value—this will need to be a matter of judgement). Because this analysis occurs after the procedures of the four building blocks has taken place, a negative finding implicates all four of the steps: A low correlation implies that at least one of the four building blocks needs to be re-examined.

One should also examine whether the item locations adequately "cover" the person locations in order to makes sure that respondents are being measured adequately throughout the whole continuum. For example, a small range of the difficulty of the items would look like "an attempt to find out the fastest runner in a distance of two meters".

A similar question can be asked at the item level: the behaviour of the items need to be checked for consistency with the estimates from the test. Consistency here is indexed by checking that respondents in each higher response category tend to score higher on the test as a whole. This ensures that each item and the whole test are acting in concordance.[14]

Evidence Based on Relations to Other Variables

One type of external variable is the set of results of a second instrument designed to measure the same construct. A second type arises if there is established theory that implies some type of relationship of the construct of interest with the external variable (i.e., positive, negative, or null, as the theory suggests). Then the presence or the lack of that relationship with the external variable can be used as one of the pieces of evidence. Usually the correlation coefficient is adequate to index the strength of the relationship, but, where a non-linear relationship is suspected, one should always check using a scatterplot. Examples of external variables are scores on other tests, teachers' or supervisors' ratings, the results of surveys and interviews, product reviews, and self-reports.

Just as we could apply the logic of the internal structure evidence down at the item level, the same applies to this strand of evidence. Here the evidence is referred to as differential item functioning (DIF). DIF occurs when, controlling for respondent overall ability, an item favours one group of respondents over another. Finding DIF implies that there is another latent variable (i.e., other than the construct) that is affecting the probability of responses by members of the different groups. Ideally, items should be functioning similarly across different subgroups. Respondents' background variables such as gender or race should not influence the probability of responding in different categories. One way to investigate DIF is to calibrate the data separately for each subgroup and compare the item estimates for large differences (Wilson, 2005), but another approach directly estimates DIF parameters (Meulders & Xie, 2004). DIF is clearly a threat to the validity of the test in the sense of fairness. Longford, Holland, & Thayer (1993), and Paek (2002) have recommended practical values for the sizes of DIF effects that are large enough to be worthy of specific attention.

Evidence based on the consequences of using an instrument. Since the use of the instrument may have negative consequences, this type of evidence should have a significant influence on whether to use the instrument or not. If there is a negative consequence from using the instrument, alternative instruments should be used instead, or developed if none exists. If any alternative instrument will also have the negative consequence, then perhaps the issue lies with the construct itself. Note that this issue arises when the instrument is used according to the recommendations of the measurer. If the instrument is used in ways that go beyond the recommendations of the original measurer, then there is a requirement that the new usage be validated, just as was the original use. For instance, if the instrument was designed for the use for placement purposes only, using it for selection or diagnosis will be considered as a misuse of the test and should be avoided. The cautionary message by Messick (1994) below better reflects this point:

> Validity, reliability, comparability, and fairness are not just measurement issues, but *social values* that have meaning and force outside of measurement wherever evaluative judgments and decisions are made (p. 2).

In thinking of test consequences, it is useful to think of the four-way classification of intended versus unintended use and positive versus negative consequences (Brennan, 2006). Intended use with positive consequence is seldom an issue and is considered as an ideal case. Similarly, for ethical and legal reasons, there are no questions on avoiding the intended use with negative consequences. The confusion is with unintended uses. Unintended use with a positive consequence is also a benefit. The major issue and confusion arises with unintended use with negative consequences. The measurer has a limited responsibility and a limited power in preventing this being the case once a test is broadly available. However, it is the measurer's responsibility to document the intended uses of the test.

CONCLUSION

Each use of an instrument is an experiment and hence requires a very careful design. There is no machinery or mass production for producing the instruments we need in education – each instrument and each construct requires a customized approach within a more general framework, such as that outlined above. The amount of effort you put in the design of the instrument will determine the quality of the outcomes and ease of the interpretation based on the outcome data.

In order to model real-life situations better, there have been many developments in psychometric theory that allow extensions and increased flexibility starting from the simple probability-based model we have used here. Models that allow the incorporation of item features (e.g. the linear logistic test model (Janssen, Schepers, & Peres, 2004)) and respondent characteristics (e.g. latent regression Rasch models (Adams Wilson & Wu, 1997)), and multidimensional Rasch models (Adams, Wilson & Wang, 1997) have been developed and used extensively. Recently there have been important developments introducing more general modelling frameworks and thus recognizing previously distinct models as special cases of the general model (e.g., De Boeck & Wilson, 2004; Skrondal & Rabe-Hesketh, 2004)). As a result, the range of tools that psychometricians can use is expanding. However, one should always bear in mind that no sophisticated statistical procedure will make up for weak design and/or poor items.

Psychometrics as a field, and particularly educational measurement, is growing and having an effect on every student's journey through their education. However, as these developments proceed, we need principles that act as guarantors of social values (Mislevy, Wilson, Ercikan & Chudowsky, 2003). Researchers should not be concerned about valuing what can be measured, but rather stay focused on measuring what is valued (Banta, Lund, Black & Oblander, 1996). Measurement in the educational context should be aimed squarely at finding ways to help educators and educational researchers to attain their goals (Black & Wilson, 2011).

This chapter is not an attempt to cover completely the whole range of knowledge and practice in psychometrics – rather, it is intended to outline where one might *begin*.

NOTES

1 Note, do not confuse this use of "formative" with its use in the previous paragraph.

2 These four building blocks are a close match to the 3 vertices of the NRC's Assessment Triangle (NRC, 2001)—the difference being that the last two building blocks correspond to the third vertex of the triangle.

3 Borrowed from Wilson (2005).

4 The fundamental assumption in most of the modern measurement models is monotonicity. As the ability of the person increases, the probability of answering correctly increases as well (unfolding IRT models being an exception—See Takane, (2007).

5 i.e., It should provide useful information about certain locations on the construct map.

6 The carry-over effect can be better understood with the *brainwashing* analogy. Assume that the respondent forgets his/her answers on the test items over repeated testings. Aggregating over the sufficiently large (perhaps infinite) number of hypothetical administrations gives the true location of the respondent (i.e., the True Score).

[7] In the development below, we will assume that the items in question are dichotomous, but the arguments are readily generalized to polytomous items also.

[8] Recall that instrument-focused approach of CTT is also based on the number correct. There is an important sense in which the Rasch Model can be seen as continuation and completion of the CTT perspective (Holland & Hoskens, 2003).

[9] Note that while some see this property as the advantage of the Rasch model, this has also been a point of critique of the Rasch model. The critique lies in the fact that Rasch model ignores the possibility that there is information in the different respondent response patterns with the same total. In our view, the best resolution of the debate lies the view that the instrument is an experiment that needs to be carefully designed with carefully-crafted items. This point will be elaborated later in the chapter.

[10] quote from Occam cited in , Thorburn, 1918.

[11] The natural logarithm of the odds ratio.

[12] Note that these strands should not be confused with categories from earlier editions of the "Test Standards," such as construct validity, criterion validity, face validity , etc.

[13] The simplest thing one can do is to examine the content of the items (this has been also intuitively referred to as the *face* validity), though this is far from sufficient.

[14] This information will also usually be reflected in the item fit statistics used in the Rasch model. Another indicator is the *point-biserial correlation*—the correlation of the binary score with the total score, also called as the *item-test* or *item-total correlation.*

REFERENCES

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 47–76.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council for Measurement in Education (NCME). (1999). *Standards for psychological and educational tests*. Washington D.C.: AERA, APA, and NCME.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Dekker.

Banta, T. W., Lund, J. P., Black, K. E., & Oblander, F. W. (1996). *Assessment in practice: Putting principles to work on college campuses*. San Francisco: Jossey-Bass.

Baxter, J. (1995). Children's understanding of astronomy and the earth sciences. In S. M. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 155–177). Mahwah, NJ: Lawrence Erlbaum Associates.

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives*, *9*, 1–52.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), Educational measurement (4th ed.).Westport, CT: Praeger.

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33–63.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation.* London: Longmans, Green & Co.

Claesgens, J., Scalise, K., Wilson, M., & Stacy, A. (2009). Mapping student understanding in chemistry: The perspectives of chemists. *Science Education, 93*(1), 56–85.

Cooke, L. (2006). Is the mouse a poor man's eye tracker? *Proceedings of the Society for Technical Communication Conference.* Arlington, VA: STC, 252–255.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.
Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.
Dahlgren, L. O. (1984a). Outcomes of learning. In F. Marton, D. Hounsell & N. Entwistle (Eds.), *The experience of learning*. Edinburgh: Scottish Academic Press.
De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: The Guilford Press.
De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.
Draney, K., & Wilson, M. (2009). Selecting cut scores with a composite of item types: The Construct Mapping procedure. In E. V. Smith, & G. E. Stone (Eds.), *Criterion-referenced testing: Practice analysis to score reporting using Rasch measurement* (pp. 276–293). Maple Grove, MN: JAM Press.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum & Associates.
Everitt, B. S. (2010). *Cambridge dictionary of statistics* (3rd ed.). Cambridge: Cambridge University Press.
Galton, F. (1883). *Inquiries into human faculty and its development*. AMS Press, New York.
Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150.
Guttman, L. A. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. A. Guttman, F. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), Studies in social psychology in world war two, *vol. 4*. Measurement and prediction. Princeton: Princeton University Press.
Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly-nonparallel test. *Psychometrika, 68*, 123–149.
Ivie, J. L., Embretson, S., E. (2010). Cognitive process modeling of spatial ability: The assembling objects task. *Intelligence*, *38*(3), 324–335.
Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item-group predictors. In, P. De Boeck, & M. Wilson, (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.
Kakkonen, T., Myller, N., Sutinen, E., & Timonen, J. (2008). Comparison of dimension reduction methods for automated essay grading. *Educational Technology & Society*, *11*(3), 275–288.
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.).Westport, CT: Praeger.
Kofsky, E. (1966). A scalogram study of classificatory development. *Child Development, 37*, 191–204.
Kuder, G. F., & Richardson, M. W. (1937). *The theory of the estimation of test reliability*. *Psychometrika*, *2*, 151–160.
Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140, 52.*
Longford, N. T., Holland, P. W., & Thayer, D. T.(1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.
Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1–27.
Magidson, J., & Vermunt, J. K. (2002). A nontechnical introduction to latent class models. *Statistical innovations white paper No. 1*. Available at: www.statisticalinnovations.com/articles/articles.html.
Marton, F. (1981). *Phenomenography: Describing conceptions of the world around us*. *Instructional Science*, *10*(2), 177–200.
Masters, G. N., Adams, R. J., & Wilson, M. (1990). Charting of student progress. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies. Supplementary, Volume 2* (pp. 628–634). Oxford: Pergamon Press.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education and Macmillan.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher*, *32*(2), 13–23.
Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In, P. De Boeck, & M. Wilson, (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R, J., Wilson, M., Ercikan, K., & Chudowsky, N. (2003). Psychometric principles in student assessment. In T. Kellaghan, & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation*. Dordrecht, The Netherlands: Kluwer Academic Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment* (Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser, (Eds.), Division on behavioural and social sciences and education). Washington, DC: National Academy Press.

National Research Council. (2008). *Early childhood assessment: Why, what, and how?* Committee on Developmental Outcomes and Assessments for Young Children, Catherine E. Snow & Susan B. Van Hemel, (Eds.), Board on children, youth and families, board on testing and assessment, division of behavioral and social sciences and education. Washington, DC: The National Academies Press.

Nisbet, R. J., Elder, J., & Miner, G. D. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press.

Nunnally, C. J. (1978). *Psychometric theory* (2nd ed.) New York: McGraw Hill.

Paek, I. (2002). *Investigation of differential item functioning: Comparisons among approaches, and extension to a multidimensional context.* Unpublished doctoral dissertation, University of California, Berkeley.

Ramsden, P., Masters, G., Stephanou, A., Walsh, E., Martin, E., Laurillard, D., & Marton, F. (1993). Phenomenographic research and the measurement of understanding: An investigation of students' conceptions of speed, distance, and time. *International Journal of Educational Research*, *19*(3)*,* 301–316.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogische Institut.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the Ratings: Assessing the Psychometric Quality of Rating Data. *Psychological Bulletin*. *88*(2), 413–428.

Scalise, K., & Wilson, M. (2011). The nature of assessment systems to support effective use of evidence through technology. *E-Learning and Digital Media*, *8*(2), 121–132.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922–932.

Spearman, C. C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101.

Spearman, C, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

Takane, Y. (*2007*). *Applications of multidimensional scaling in psychometrics*. In *C. R. Rao, & S. Sinharay* (Eds.), *Handbook of statistics, Vol. 26: Psychometrics*. Amsterdam: Elsevier.

Thorburn, W. M. (1918). The myth of occam's Razor. *Mind*, *27*(107), 345–353.

van der Linden, W. (1992). Fundamental measurement and the fundamentals of Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice Vol. 2*. Norwood, NJ: Ablex Publishing Corp.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory.* New York: Springer.

Vosniadou, S., & Brewer, W. F. (1994). Mental models of the day/night cycle. *Cognitive Science*, 18, 123–183.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29,*126–149.

Wiliam, D. (2011). *Embedded formative assessment*. Bloomington, IN: Solution Tree Press,

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal for Research in Science Teaching, 46*(6), 716–730.

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, *60*(2), 181–198.
Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), *Measurement and multivariate analysis* (Proceedings of the International Conference on Measurement and Multivariate Analysis, Banff, Canada, May 12–14, 2000), pp. 325–332. Tokyo: Springer-Verlag.
Wright, B. D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 invitational conference on testing* (pp. 85–101). Princeton, NJ: Educational Testing Service.
Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97–116.
Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MESA Press.

ZE WANG & STEVEN J. OSTERLIND

2. CLASSICAL TEST THEORY

GENERAL DESCRIPTION

Classical test theory (CTT) is the foundational theory of measurement of mental abilities. At its core, CTT describes the relationship between observed composite scores on a test and a presumed but unobserved "true" score for an examinee. CTT is called "classical" because it is thought to be the first operational use of mathematics to characterize this relationship (cf. Gullicksen, 1950). Modern theories of measurement, such as IRT (item response theory), do not obviate CTT or even contradict it; rather, they extend it although there are important distinctions in both the underlying philosophies and in the statistics employed for implementation.

A primary feature of CTT is its adherence to learning theories that follow notions of classical and operant conditioning (e.g., behaviorism, social learning theory, motivation). CTT presumes extant a domain of content apart from any particular examinee, although – significantly – the domain is not reified; it remains an abstraction. This perspective places CTT outside cognitivist theories of learning (e.g., information processing, constructivism). Thus, for application of the theory, the domain is defined anew in each appraisal. For example, if "reading" is the domain for an appraisal, "reading" must be defined for that specific assessment. In another assessment "reading" will have a slightly different meaning. Hence, in CTT, no two independent tests are identical, although strictly parallel forms for a given assessment may be developed. Further, in CTT the domain (whether "reading" or other) with its theoretical parameters, can be accurately sampled by a test's items or exercises. This means (to continue the reading example) that the main idea of a paragraph can be dependably deduced. The items on the test are stimuli designed to manifest observable behavior by the examinee: the response. The focus of CTT is to determine the degree to which the examinee has mastered the domain: the implied individual's true score which is inferred through responses to the test's stimuli.

Lord and Novick (1968), in their classic work *Statistical Theories of Mental Test Scores*, begin the explanation of CTT with definitions of a true score and an error score. They maintained that one must keep in mind what a true score represents and the basic assumptions about the relationships among the true score, the error score, and the observed score. In the CTT framework, an individual's observed score on a test is considered to be a random variable with some unknown distribution. The individual's true score is the expected value of this distribution, typically denoted as E (symbol for expectation; not to be confused with the error term described below)

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 31–44.

in general statistical theory. The discrepancy between the individual's observed score and true score is measurement error, which is also unobserved and stochastic. These features, then—true score, observed score, and error—compose CTT.

From these elements CTT builds two central definitions, including (1) the true score τ_{gp} of a person p on measurement g is the expected value of the observed score X_{gp}; and (2) the, error score E_{gp} which is the difference between the two elements (i.e., observed score and the true score, $X_{gp} - \tau_{gp}$). Under CTT, τ_{gp} is a constant yet unobserved value, and X_{gp} is a random variable that fluctuates over repeated sampling of measuring g. This fluctuation is reflected by a propensity distribution F_{gp} for that person p and measurement g. The expectation in definition (1) is with respect to that propensity distribution. From this stand point the mathematical model for CTT can be deduced, and consists of two equations:

$$\tau_{gp} = E(X_{gp}) \tag{1}$$

$$E_{gp} = X_{gp} - \tau_{gp} \tag{2}$$

However, in most cases, researchers are interested in the traits of a population of people rather than in the trait of a fixed person p. Therefore, any person p from that population can be considered a random sample. The notation X_g presents a random variable defined over repeated sampling of persons in a population, which takes a specific value x_g when a particular person is sampled. Similarly, Γ_g is a random variable over repeated sampling of persons in a population, which takes a specific value τ_g when a particular person is selected. Finally, E_g is random variable representing the error score. Under this construction, Lord and Novick (1968) had the theorem that $X_g = \Gamma_g + E_g$. Without loss of generality, the subscript g is omitted when only one measurement is considered. And, thus, is defined the familiar CTT equation,

$$X = \Gamma + E \tag{3}$$

It is important to remember that in equation (3), all the three elements are random variables. In CTT they are called "random variables," although in the more general probability theory they are classified as stochastic processes.

CTT as a theory requires very weak assumptions. These assumptions include: (a) the measurement is an interval scale (note: there are other types of scales such as classifications; those are not part of the CTT model although with some score transformation they can be incorporated in CTT); (b) the variance of observed scores σ_X^2 is finite; and (c) the repeated sampling of measurements is linearly, experimentally independent. Under those assumptions, the following properties have been derived (Lord & Novick, 1968):

1. The expected error score is zero;
2. The correlation between true and error scores is zero;

3. The correlation between the error score on one measurement and the true score on another measurement is zero;
4. The correlation between errors on linearly experimentally independent measurements is zero;
5. The expected value of the observed score random variable over persons is equal to the expected value of the true score random variable over persons;
6. The variance of the error score random variable over persons is equal to the expected value, over persons, of the error variance within person (i.e., $\sigma^2(X_{gp})$);
7. Sampling over persons in the subpopulation of people with any fixed true score, the expected value of the error score random variable is zero;
8. The variance of observed scores is the sum of the variance of true scores and the variance of error scores; that is:

$$\sigma_X^2 = \sigma_\Gamma^2 + \sigma_E^2. \tag{4}$$

It is important to note that the above properties are not additional assumptions of CTT; rather, they can be mathematically derived from the weak assumptions and easily met by most test data. Because of this, CTT is a test theory that provides, "a theoretical framework linking observable variables…to unobservable variables…a test theory cannot be shown to be useful or useless" (Hambleton & Jones, 1993).

From this discussion, it can be realized that with additional assumptions, CTT can be stated as a model eligible for testing against data. This empiricism is pronounced in modern test theory, especially in IRT where the model is tested against data in each new test application.

RELIABILITY

One of the most important features in CTT is reliability. The term is concerned with precision in measurement, and it is described as consistency of test scores over repeated measurements (Brennan, 2001). This definition has remained largely intact since the early days of modern measurement, although its emphasis has evolved to focus more on standard errors of measurement (cf. Brennan, 2001; Osterlind, 2010). Evolution of the term's development can be traced in each subsequent edition of the *Standards for Educational and Psychological Tests* (cf. 1966, 1974, 1985, 1999).

The mathematics of reliability is quite straightforward. Working from the formulation of CTT as given in formula (3) above (cf., $X = \Gamma + E$), Γ and E are uncorrelated

$$\rho_{\Gamma E} = 0 \tag{5}$$

This leads directly to Lord and Novick's final assumption, given above as the 8th property in the list above and expressed in Equation (4): that is, variances are

additive: $\sigma_X^2 = \sigma_\Gamma^2 + \sigma_E^2$. It follows that whenever an observed score is extant the variance of true scores and the variance of error scores is less than the variance of observed scores, or

$$\sigma_\Gamma^2 \leq \sigma_X^2 \qquad \text{and} \qquad \sigma_E^2 \leq \sigma_X^2 .$$

The ratio of these variances is expressed as:

$$\rho_X = \frac{\sigma_\Gamma^2}{\sigma_X^2} = \frac{\sigma_\Gamma^2}{\sigma_\Gamma^2 + \sigma_E^2} \tag{6}$$

This ratio quantifies the reliability of using observed scores to describe the traits of a population of individuals and ρ_X is the reliability coefficient of the measurement. As such, it is foundational to CTT. It is also obvious from equation (6) that the reliability coefficient ranges from 0 to 1.

While this coefficient is easily derived, applying it to live data in a real-world testing scenario is challenging at best, due primarily to practical considerations. From the mathematical derivation we can see that reliability requires multiple measurements. Further, in theory the measurements are presumed to be independent—even, a very large number of them would be stochastic. Practically, this is difficult to achieve even when forms of a test are strictly parallel. Using a given form and splitting it into two halves does not obviate the problem. Another practical problem concerns the attributes themselves. Attributes for educational and psychological measurements are nearly always latent constructs or proficiencies. Here is where the problem arises: as humans such latencies are labile, or changing in unpredictable and uneven ways. At some level, this makes multiple measurements even more suspect.

These two practical difficulties are not easily overcome; nonetheless, recognizing these conditions, reliability can be determined to a sufficient degree that it is useful for our purposes. Due to these problems there is not a single, universally adopted expression for the reliability coefficient. Instead, the reliability coefficient has many expressions. Generally, they are of either about the internal consistency of a test or its temporal stability. Internal consistency seeks to examine the degree to which the individual elements of a test (i.e., items or exercises) are correlated. The Cronbach's coefficient alpha (described more fully later on) is an example of gauging a tests' internal consistency. Similarly, a coefficient that indicates a test's temporal stability tries to find a similar correlational relationship between repeated measurements.

Although parallel forms are not necessary to describe relationships among quantities of interest under CTT, it is usually easier to describe those statistics with respect to parallel forms. Parallel forms are measures that have the same true score and identical propensity distribution, between the measures, for any person in the population. That is, for any given person *p* in the population, if forms *f* and *g* satisfy

that $\tau_{fp} = \tau_{gp}$, and $F_{fp} = F_{gp}$, we say forms f and g are parallel. The requirements of parallel forms can be reduced to $\tau_{fp} = \tau_{gp}$ and $\sigma^2(E_{fp}) = \sigma^2(E_{gp})$ for any given person p, if X_{fp} and X_{gp} are linearly experimentally independent, that is, the expected value of X_{fp} does not depend on any given value of x_{gp}, and that the expected value of X_{gp} does not depend on any given value of x_{fp}.

When two test forms are parallel, the distribution of any of the three random variables, X, Γ, and E, and any derived relationships (e.g., correlations, covariances) involving those random variables are identical between the two forms. In other words, the two forms are exchangeable. It matters not which test form is administered. However, those random variables do not have to follow a particular distribution, such as a normal distribution.

Then, too, there can be types of parallelism. Non-parallel forms, depending on the degree to which they differ from parallelism, can be tau-equivalent forms, essentially tau-equivalent forms, congeneric forms, and multi-factor congeneric forms. Specifically, tau-equivalent forms relax the assumption of equal error variance but the assumption of equal true scores still holds; essentially tau-equivalent forms further relax the assumption of equal true scores by requiring only that the true scores for any given person between two forms differ by a constant which depends only on the forms but not the individual; congeneric forms allows a shortening or lengthening factor of the measurement scale from one form to the other after adjusting for the constant difference in true scores at the origin of one form; multi-factor congeneric forms further breaks down the true score on either form into different components and allows each component to have a relationship similar to that exists between congeneric forms. For mathematical representations of those types of non-parallelism, see Feldt and Brennan (1989).

If X and X' are observed scores from two parallel forms for the same sample of people from the population, we have

$$\rho_{XX'} = \rho_X = \rho^2_{X,\Gamma} \tag{7}$$

where X and X' are test scores obtained from the two parallel forms.

That is, the reliability coefficient can be thought of as the correlation between two parallel forms, which is the square of the correlation between the observed scores and true scores.

Therefore, based on formula (7), if parallel forms are administered to the same sample, the reliability coefficient is the correlation coefficient squared. Sometimes, the same test form is administered twice assuming no learning has happened between the two administrations, the reliability coefficient is then based on the two administrations. This is the referred to as the test-retest reliability.

Often, a single test form is administered once and only one total test score is available for each individual. In this case, formula (6) has to be used. The challenge is that this formula provides the definition, not the calculation of reliability. Like the

true scores, the variance of true scores in the population is unknown and has to be estimated from the data. Ever since Spearman (1910) and Brown (1910), different coefficients have been proposed to estimate test reliability defined in formula (6). Those approaches are based on the thinking that each test score is a composite score that consists of multiple parts. Spearman-Brown's split half coefficient is calculated under the assumption that the full test score is the sum of two part-test scores and that the two parts are parallel:

$$_{SB}\rho_X = \frac{2\rho_{X_1X_2}}{1+\rho_{X_1X_2}} \tag{8}$$

where $\rho_{X_1X_2}$ is the correlation between the two parts. If X_1 and X_2 are two parallel forms of the same test, the above equation also serves as a corrected estimation for the reliability coefficient of the test if the test length is doubled. For more information on the relationship between test length and test reliability, see Osterlind (2010, pp. 143–146).

As parallelism between the two parts is relaxed, other formulas can be used. The applications of those formulas with degrees of parallelism can be found in Feldt and Brennan (1989). Reuterberg and Gustafsson (1992) show how confirmatory factor analysis can be used to test the assumption of tau equivalence and essentially tau equivalence.

The most popular reliability coefficient remains Cronbach's coefficient alpha (1951). This coefficient is a measure of internal consistency between multiple parts of a test and is based on the assumption that part scores (often, item scores) are essentially tau-equivalent (i.e., equal true score variance but error score variances can be different across parts). Under this assumption, coefficient alpha is:

$$_{\alpha}\rho_X = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_X^2 - \sum\sigma_{X_f}^2}{\sigma_X^2}\right) \tag{9}$$

where n is the number of parts, σ_X^2 is the variance of observed scores of the full test, and $\sigma_{X_f}^2$ is the variance of observed scores for part f.

When the parts are not essentially tau equivalent, Cronbach's alpha is the lower bound of the standard reliability coefficient. If the n parts are n items in a test that are scored dichotomously (0 or 1), Cronbach's coefficient alpha reduces to KR-20 (Kuder & Richardson, 1937):

$$_{20}\rho_X = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum\phi_f(1-\phi_f)}{\sigma_X^2}\right) \tag{10}$$

where ϕ_f is the proportion of scores of 1 on item f.

STANDARD ERROR OF MEASUREMENT

Another index is one closely related to reliability of a test: the standard error of measurement (SEM). The SEM summarizes within-person inconsistency in score-scale units. It represents the standard deviation of a hypothetical set of repeated measurements on a single individual (i.e., the standard deviation of the distribution of random variable E_{gp} in (2). In CTT models, it is usually assumed that the standard error of measurement is constant for all persons to facilitate further calculations. With this assumption,

$$\text{SEM} = \sigma_E = \sigma_X(1 - \rho_X) \tag{11}$$

where ρ_X is the reliability coefficient.

The choice of the reliability coefficient makes a difference in calculating the SEM, because different reliability coefficients capture different sources of errors. For example, a SEM based on a test-retest reliability reflects the inconsistency of test scores for an individual over time, while a SEM calculated on Cronbach's coefficient alpha reflects the inconsistency of test scores for an individual over essentially tau-equivalent test forms. Thus, when reporting or examining the SEM, one should be aware what source of error is reflected.

ESTIMATION OF TRUE SCORES UNDER CTT

One purpose of CTT is to make statistical inferences about people's true scores so that individuals can be compared to each other, or to some predefined criteria. Under CTT, the true score of each person τ_p is fixed yet unknown. In statistics, we call such a quantity a parameter. A natural following question is: Can we find an estimate for that parameter? With only one test administration, the commonly used practice to estimate a person's true score is to use the observed score x_p. This is an unbiased estimate of τ_p which is defined as the expected value of the random variable X_p, as long as the weak assumptions of CTT hold. Sometimes, an additional distributional assumption is added to a CTT model to facilitate the construction of an interval estimation of an individual's true score. A commonly used assumption is that σ_E^2 is normally distributed. With this additional assumption, the interval estimation of τ_p is $x_p \pm z\sigma_E$, where z is the value from the standard normal distribution corresponding to the probability associated with the interval.

Another less commonly used construction of a point estimation and interval estimation of τ_p depends on an additional assumption that, with a random sample of multiple persons on whom test scores are observed, the random variables Γ and X follow a bivariate normal distribution. With this assumption, a point estimate of an individual's true score is $\rho_X(x_p - \mu_X) + \mu_X$, where ρ_X is the reliability coefficient, and μ_X is the population mean of observed scores , which can be replaced by the sample mean of $\bar{X}$ in practice. The corresponding interval estimation for τ_p is

$[\rho_X(x_p - \mu_X) + \mu_X] \pm z\sigma_E\sqrt{\rho_X}$. It can be shown that this construction is consistent with confidence intervals of mean predictions in multiple linear regression.

VALIDITY

The idea that test scores are used to make inferences about people is directly related to another important concept in measurement, namely, validity. The past five decades has witnessed the evolution of the concept of validity in the measurement community, documented particularly in the five editions of the *Standards for Educational and Psychological Testing* published in 1954, 1966, 1974, 1985, and 1999, respectively (referred to as the *Standards* since different titles are used in those editions). In the first edition of the *Standards* (APA, 1954), validity is categorized into four types: content, predictive, concurrent, and construct. In the second edition of the *Standards* (AERA, APA, & NCME, 1966), validity is grouped into three aspects or concepts: content, criterion, and construct. In the third edition of the *Standards* (AERA, APA, & NCME, 1974), the three categories are called types of validity. In the fourth edition of the *Standards* (AERA, APA, & NCME, 1985), the three categories are called "types of evidence" and the central role of construct-related evidence is established. In the fifth edition of the *Standards* (AERA, APA, & NCME, 1999), the content/criterion/construct trinitarian model of validity is replaced by a discussion of sources of validity evidence.

The description of sources of validity evidence in the *Standards* is consistent with and perhaps influenced by Messick's treatment of validity as an integrated evaluative judgment. Messick (1989) wrote:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment… Broadly speaking, then, validity is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use. Hence, what is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators – inferences about score meaning or interpretation and about the implications for action that the interpretation entails… It is important to note that validity is a matter of degree, not all or none… Inevitably, then, validity is an evolving property and validation is a continuing process. (p. 13)

The process of collecting validity evidence – validation—can be carried out by examining the test content, its relationships with criteria, and the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (Messick, 1989, p. 13). More recently, Kane (2006) considers validation as "the process of evaluating the plausibility of proposed interpretations and uses" and validity as "the extent to which the evidence supports or refutes the proposed interpretations and uses" (p. 17). Importantly, he divides the validation process

into a stage of interpretative argument and a stage of evaluation of the interpretive argument (i.e., validity argument). The interpretive argument serves as the theoretical framework for the proposed interpretations and uses of test results. The validity argument evaluates the coherence, plausibility, and assumptions of the interpretive argument. Kane's (2006) treatment of validity incorporates the unitary notion of validity as an integrated judgment and also provides some guidance for validation studies. With this treatment, other previously used notions such as face validity, content validity and convergent validity can be incorporated into the two stages of validation.

Despite this evolution, the idea that construct-related evidence of validity has the central role with content- and criterion-related evidence playing a subordinate role is still prevalent in textbooks on measurement and psychological testing (e.g., McIntire & Miller, 2006; Raykov & Marcoulides, 2010). One reason may be due to the fact that it is easier to empirically collecting evidence that way.

CTT AND OTHER TECHNIQUES

Notably, CTT models have been related to other techniques as a special case and most such relationships are based on some mathematical and statistical equivalence. Before talking about those equivalences, it is important to point out that CTT is a measurement theory that bears both semantic and syntactic definitions. With a semantic definition, the more abstract constructs can be linked to observable behaviors. With a syntactic definition, those constructs and relationships between them can be stately more broadly. These two aspects together are made possible through "a particular, mathematically convenient and conceptually useful, definition of true score and on certain basic assumptions concerning the relationships among true and error scores" (Lord & Novick, 1968, p. 29).

CTT is also a theory of composite scores, with a focus on properties of intact tests. If multiple forms are available, observed scores obtained from those forms can be subject to a one-factor confirmatory factor analysis and the latent factor serve the role of true score in CTT. Parallel and non-parallel test forms correspond to constraints on parameters of factor analysis models. One the other hand, when only one test form is available, treating items (or test parts) on that test as multiple test forms, we can assess the applicability of different reliability coefficients. For example, Reuterberg and Gustafsson (1992) have shown that Cronbach's coefficient alpha assumes an equal factor loading from the latent factor to item scores but does not assume equal residual variances. In this sense, CTT is a special case of confirmatory factor analysis. However, this type of testing through factor analysis is for assumptions that are later imposed to form different CTT models, not for the weak assumptions of CTT themselves. For example, in the case of Cronbach's coefficient alpha, we can use factor analysis to test the applicability of this reliability coefficient for a particular test but it would be incorrect to claim that CTT does not apply if factor analysis results are not consistent with data.

Unlike CTT, IRT is for item-based models. Because characteristics can be examined for various items separately under IRT, items are not bound with a particular test and they are not sample dependent. In contrast, item characteristics under CTT depend on the sample and items are compared against the composite scores on the tests. However, CTT statistics can be derived using IRT with very general assumptions (Holland & Hoskens, 2003).

There are still more perspectives on CTT. For instance, CTT can also be viewed as a special case of generalizability (G) theory, first introduced by Cronbach and colleagues in response to the limitations of CTT (L. J. Cronbach, Gleser, Nanda, & Rajaratnam, 1972; L. J. Cronbach, Rajaratnam, & Gleser, 1963; Gleser, Cronbach, & Rajaratnam, 1965; Rajaratnam, Cronbach, & Gleser, 1965). In CTT, the error term E represents undifferentiated random error and does not distinguish different sources of the error. In G theory, multiple sources of error can be investigated with one design. The universe score in G theory is analogous to the true score in CTT and is the score obtained if that individual has taken all possible items that tap the proficiency/ability that the test is trying to measure under all possible conditions. Of course, since an individual cannot take all the possible items, the universe score is unknown. However, if the items on a particular test form can be considered as a random sample of all possible items and different conditions such as raters can be considered as a random sample of all possible conditions, the error term can be decomposed to reflect multiple sources, together with a source of variability of true scores across different people. In CTT, the observed scores only have the variability of true scores due to different people and the variability of scores of an agglomeration of errors.

ITEM ANALYSIS

Although the focus of CTT is usually with the total test scores, analyzing items that consist of the test is useful during the earlier stages of test development (e.g., field testing) and can be informative when examining item and test shifting. The two most important statistics for any item within the CTT framework are (a) item difficulty and (b) item discrimination. For a dichotomous item scored as correct or incorrect, item difficulty (usually denoted as p) is the percentage of individuals in the sample who answered the items correctly (that is, item difficulty measures the "easiness" of an item in the sample). For a dichotomous item, the correlation between item and total test scores is the point-biserial correlation. A large correlation suggests larger difference in the total test scores between those who answered the item correctly and those who answered the item incorrectly. That is, the correlation between item and total test score is a measure of item discrimination. When multiple score points are possible for one item, item difficulty is the average score on that item expressed as a proportion of the total possible point; and item discrimination is the Pearson product moment correlation between item and total test scores. In reality, item discrimination is usually calculated as the correlation between the item

scores and total test scores excluding the item scores for the item being evaluated. This "corrected" item discrimination eliminates the dependence of total test scores on the item being evaluated.

From the above, it is obvious that both item difficulty and item discrimination under CTT is dependent upon the sample of individuals whose responses are used for those calculations. For example, the same item may have a large p values if data are from a higher-ability group of individuals, compared to a lower-ability one. Actually, this interdependency between item and sample is the most attacked weakness of CTT, especially when it is compared to IRT.

AN ILLUSTRATIVE STUDY

Obviously—and logically—examining test items and exercises after a test has been administered to a group of examinees is the most frequent application of CTT. Such item analysis has several purposes, including interpreting the results of an assessment, understanding functioning of an item wholly, exploring parts of the item (i.e., the stem, distractors), discovering its discriminating power, and much more. While many of the statistics used for the purposes can easily be calculated by hand, it is much more convenient to use a computer. And, of course, many computer programs, both home grown and commercial, are available to do this. We explain the output from one program, called MERMAC, to illustrate typical statistical and graphical CTT output for item analysis. Figure 1 illustrates the output for one multiple-choice item, in this case Question 44.

Note in Figure 1 that the item analysis is presented in two types: tabular and graphical. In the table (left side of the figure), the results are reported for each fifth of the population, divided on the basis of their total test score (the most able group is at the top 5th; the least able is the 1st group). Such fractile groupings are common in item analysis. In addition to showing item discrimination between five ability groups, they can also be used in reliability analyses. In the table, the raw number of examinees who endorsed a given response alternative is shown. This is useful because following down the ability groups (from the top 5th to the 1st) one observes that more of the less able examinees endorsed incorrect responses, showing greater discrimination for the item. Additionally, it is instructive for both interpretation

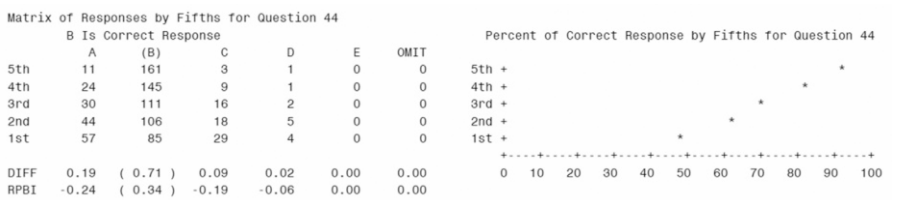


Figure 1. Graphical item analysis output from the MERMAC program.

of test results and for item improvement, to note which distractors were selected by what ability group. Below the table are two rows, labeled "DIFF" and "RPBI" meaning "difficulty" and "point bi-serial correlation." The difficulty statistic is the percent of examinees who endorsed each response alternative (both correct and incorrect). For example, overall 71 percent of examinee responded correctly to this item. The point bi-serial correlation is a theoretical conception of treating dichotomous test items (typically multiple-choice) as a true dichotomy between correct and anything not correct: as 1, 0. A correlation coefficient is then calculated between this theoretical variable and the examinee's total test score. This coefficient is interpreted as a measure of the item's discriminating power. A positive value for the coefficient indicates good discrimination; hence, one looks for a positive RPBI value for the correct alternative and negative value for the distractors, the case with the example item in Figure 1.

The right side of the MERMAC output is a graphical representation of the table, showing an asterisk for each ability group. The horizontal axis is percent endorsing the correct response; hence it is a graph of the Difficulty row.

As an illustration, suppose the same test is administered to students taking the same statistics course in four semesters. This test consists of 32 items: 4 multiple-choice items that clearly state there is only one answer, 7 multiple-choice items that ask students to choose as many (as few) correct answers, the other 21 items are constructed-response items where students are asked to conduct simple calculations or to explain and interpret results related to topics covered in the course. The 11 multiple-choice items are worth 1 point each, with partial points possible for those with multiple answers. Of those constructed-response items, 9 are worth 1 point each, 6 worth 2 points each, 2 worth 3 points each, and 4 worth 4 points each. Partial credits are possible for all constructed-response items. The total possible score for this test is 54 and there are 54 students during the four semesters who took this test. The data for four students and each item are in Table 1. Assuming the 32 items are essentially tau equivalent, the Cronbach's coefficient alpha calculated from formula (9) is .803. The corresponding SEM, calculated from formula (11), is 1.47. The 32 items can also be split in half so that the number of items and the total possible scores are the same in the two split halves. The correlation between the two split parts is .739, which results in a split-half reliability coefficient of 0.850 using equation (8). The corresponding SEM, calculated from formula (11), is 1.12.

Item difficulties and corrected item discriminations are also in Table 1. There are several very easy items. In this example, everyone answered Item 10 correctly so this item does not have any discriminating power. Item 9 is a dichotomously scored item and 4 out of the 54 students answered this item incorrectly, which renders a discrimination coefficient rounded to zero. All but one answered Item 3 correctly and the resultant item difficulty is .99 and item discrimination is −.22. This is a very easy item. In fact, it is so easy that an incorrect response is more likely given by a person with a higher total test score than one with a lower total test score. This item should be deleted.

Table 1. An example of item and test scores

Student	*I1*	*I2*	*I3*	*I4*	*I5*	*I6*	*I7*	*I8*	*I9*	*I10*	*I11*	*I12*	*I13*	*I14*	*I15*	*I16*	*I17*	*I18*	*I19*
1	1	0	1	1	1	0	1	0.5	1	1	1	1	1	2.5	1	0.5	1	1	0
2	1	0	1	1	1	1	1	1	1	1	0	1	1	2.5	1	1	1	0	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
53	1	1	1	2	2.5	1	1	0.5	1	1	1	1	3	2.5	0	0.5	0	2	3
54	1	1	1	2	2	1	1	0.5	1	1	1	0	3	2.5	1	0.5	1	1	1
Difficulty	.93	.89	.99	.79	.69	.78	.94	.91	.93	1.00	.80	.93	.80	.67	.59	.66	.69	.45	.38
Discrimination	.28	.29	-.22	.54	.68	.48	.05	.15	.00	.00	.22	.26	.51	.58	.19	.12	.14	.30	.52

Student	*I20*	*I21*	*I22*	*I23*	*I24*	*I25*	*I26*	*I27*	*I28*	*I29*	*I30*	*I31*	*I32*	*Total*	*Splithalf-1*	*Splithalf-2*
1	1	0	0	1	1.5	0.5	1	1	1	1	0	1	2	27.5	15	12.5
2	1	1	1	2	1	0.5	1	1	1	1	0	1	2.5	31.5	15	16.5
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
53	1	0	0	0.5	2	2	2	0	0	1	1	1	3.5	39	20	19
54	1	0	1	1	1.5	0	2	0	1	1	0	1	3	35	18.5	16.5
Difficulty	.98	.35	.57	.57	.61	.59	.86	.61	.68	.69	.34	.81	.74			
Discrimination	.26	.14	.12	.15	.46	.46	.56	.32	.22	.13	.22	.14	.46			

From the above, it is evident that the approach to mental measurement offered by CTT is both powerful and useful. It represents an application of the theory of true score and it has several practical applications in real-world testing situations, including developing a test, reporting a score for an examinees, item analysis, and some understanding of error in the measurement. For these reasons CTT remains a most popular approach to measuring mental processes.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association (APA). (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: Author.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*(4), 295–317. doi: 10.1111/j.1745–3984.2001.tb01129.x

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904–1920, 3*(3), 296–322.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurement: Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163.

Feldt, L. S., & Brennan, R. L. (1989). *Reliability*. In R. L. Linn (Ed.), *Educational measurement, 3rd ed.* (pp. 105–146). New York: American Council on Education and MacMillan.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influences by multiple sources of variance. *Psychometrika, 30*(4), 395–418. doi: 10.1007/bf02289531

Gullicksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47. doi: 10.1111/j.1745–3992.1993.tb00543.x

Holland, P., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika, 68*(1), 123–149. doi: 10.1007/bf02296657

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). West Port, CT: American Council on Education/Praeger.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McIntire, S. A., & Miller, L. A. (2006). *Foundations of psychological testing: A practical approach*: Sage.

Messick, S. (1989). *Validity*. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education & MacMillian Publishing Company.

Osterlind, S. J. (2010). *Modern measurement: Theory, principles and application of mental appraisal* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall/Merrill.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika, 30*(1), 39–56. doi: 10.1007/bf02289746

Raykov, T., & Marcoulides, G. A. (2010). *Introduction to psychometric theory*: Routledge.

Reuterberg, S.-E., & Gustafsson, J.-E. (1992). Confirmatory factor analysis and reliability: Testing measurement model assumptions. *Educational and Psychological Measurement, 52*(4), 795–811. doi: 10.1177/0013164492052004001

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904–1920, 3*(3), 271–295.

XITAO FAN & SHAOJING SUN

3. ITEM RESPONSE THEORY

INTRODUCTION

The past few decades have witnessed an exponential growth of applications of Item Response Theory (IRT), also known as "latent trait theory" or "item characteristic curve theory," in educational research and measurement. Simply speaking, IRT refers to a system that describes the relationship between an individual's response to an item and the underlying trait being measured (Embretson & Reise, 2000). Such a relationship is typically summarized and assessed by a family of statistical models, namely, item response models.

The major tenet of IRT modeling is that a respondent's recorded score on a test item is driven by certain unobservable, or latent, trait. In comparison to traditional test theory (i.e., classical test theory, or CTT), IRT has some unique properties and advantages for test construction, scoring, ability assessment, etc. Hambleton and Swaminathan (1985) summarized four main advantages of IRT models: (a) item parameter estimates do not depend on the particular group of examinees of the population for which the test is developed; (b) examinee trait assessment does not depend on the particular set of administered items sampled from a population pool of items; (c) statistical information is provided about the precision of the trait estimates; and (d) traditional reliability information is replaced by relevant statistics and its accompanying standard errors. The aforementioned features make IRT modeling more flexible and powerful, in contrast to CTT. For instance, when two examinees were administered with samples of items of differing difficulty, test scores based on traditional testing methods may fall short in providing information about the performance and ability of each examinee. Within the IRT framework, however, this task is easier, because estimates of examinee abilities are independent of sampled items from the same item population pool.

IRT has been applied to a wide spectrum of research settings, including, but not limited to, computer adaptive testing, test equating, identification of biased test items, and latent trait scoring. In the following sections, we will first introduce basic IRT terminologies and statistical models. We will then provide an illustrative example of applying IRT to a real data set. Finally, we discuss some research issues, and future directions for IRT modeling and application.

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 45–67.

DICHOTOMOUS ITEM RESPONSE MODELS

Assume that we are interested in assessing one's mathematics ability. The ability or trait, as a construct, is latent and not observable. As such, it can only be inferred from one's observable performance on certain measurement instruments, such as a mathematical test. Logically, the better one scores on the math test, the higher mathematical ability the respondent is judged to possess. The relationship between math ability and one's math test score, thus, can be modeled to assess one's latent trait, as well as the quality of measurement items.

Typically, item response modeling is based on three general assumptions: (a) the underlying trait is unidimensional (recently, however, more progress has been made about multi-dimensional IRT models, MIRT); (b) conditional on the respondent's level of the latent trait being measured, responses to different items are independent of each other, which is referred to as conditional independence or local independence; and (c) responses to an item can be depicted as a mathematical item response function (Ayala, 2009).

Item Response Function (IRF)

IRF describes the relationship between an examinee's underlying ability and the corresponding probability to endorse an item. The function can be succinctly presented as below (Yen & Fitzpatrick, 2006):

$$p_i(\theta) \equiv p_i(X_i = x_i | \{\theta\}, \{\delta_i\}) \quad (1)$$

In Equation 1, θ denotes the latent trait being measured, p denotes the probability for endorsing an item, X_i represents the score for item i, and δ_i represents the parameters of that particular item. The IRF function expresses the probability for one examinee to score x_i on that item, given that examinee's level on the latent trait and item parameters. Put differently, one's response to an item is predicated on both person parameter (e.g., latent trait level) and item parameters (e.g., item difficulty). A graphical presentation of the IRF is usually called item response curve.

Another relevant graphical technique is called item characteristic curve (*ICC*) or item characteristic function, which plots the expected response score in relation to the trait being measured. For a binary item, the *ICC* can be expressed as (Hambleton & Swaminathan, 1985):

$$f_i(\theta) \equiv P_i(\theta)^{ui} Q_i(\theta)^{1-ui} \quad (2)$$

In Equation 2, P represents the probability to correctly answer the item, whereas $Q = 1 - P$. In addition, U represents the dichotomous responses where correct response coded as 1 and incorrect response coded as 0. In short, the *ICC* expresses the expected probability for an examinee to select response 1, for the given level of the examinee's ability or trait.

Response options, however, can be more than two, and can be of different relationships (e.g., ranking order). Further, the *IRF* does not follow a linear relationship; instead, *IRF* has two major forms. One is called *normal ogive model*, which is the integral of normal distributions. The other one is based on logistic regression function distribution for a dichotomous outcome, and this is the more widely used form. It should be noted that normal ogive models and logistic regression models are comparable in many respects, and they yield similar results with simple transformations (Hambleton, Swaminathan, & Rogers, 1991). In this chapter, we focus on logistic regression models for various item response functions. We begin with the basic one-parameter *IRT* model with dichotomous response options.

One-Parameter Logistic Model

The one-parameter logistic model (1PL), also known as *Rasch* model, involves only one item parameter for estimation. The following equation for 1PL, in addition to the item parameter, entails a person parameter θ which refers to the respondent's ability or trait level to endorse the item. Again, the response option 1 means endorsing the item, or answering the item correctly, and response option 0 means not endorsing the item, or answering the item incorrectly.

$$p(X_{is}=1|\theta_s,\beta_i)=\frac{\exp(\theta_s-\beta_i)}{1+\exp(\theta_s-\beta_i)},\ i=1,2,\cdots,n \tag{3}$$

where, X_{is} is the response of person s to item i (response options 0 or 1),
θ_s is the latent trait level of person s, and
β_i is difficulty level of item i.

In the model, $p(X_{is}=1|\theta_s,\beta_i)$ denotes the probability of one individual with trail level θ_s to endorse that item in the trait-consistent direction. The only parameter, item difficulty, represents the required trait level for an individual to have 50% chance to respond to an item correctly, i.e., in the expected direction. So, the higher the value of the parameter β, the more difficult the item is for examinees to endorse.

Figure 1 presents the item characteristic curves of three items with different difficulty levels ($b=-2$, 0, and 2, respectively). The β parameter for an item is the point on the trait axis where the probability of a correct response is 0.5. It should be noted that the underlying trait and the item difficulty are projected to the same coordinate, x axis. The basic assumption is that the higher the item difficulty value is, the higher ability level the item requires for endorsement. Therefore, along the trait continuum, from the left to the right, the ability level goes from lower to higher levels. The lower the β value, the more the item is located to the left end of the trait continuum. Therefore, a characteristic curve denoting difficulty value –2 is located to the left of the curve of difficulty of 0, which then is to the left of an item with

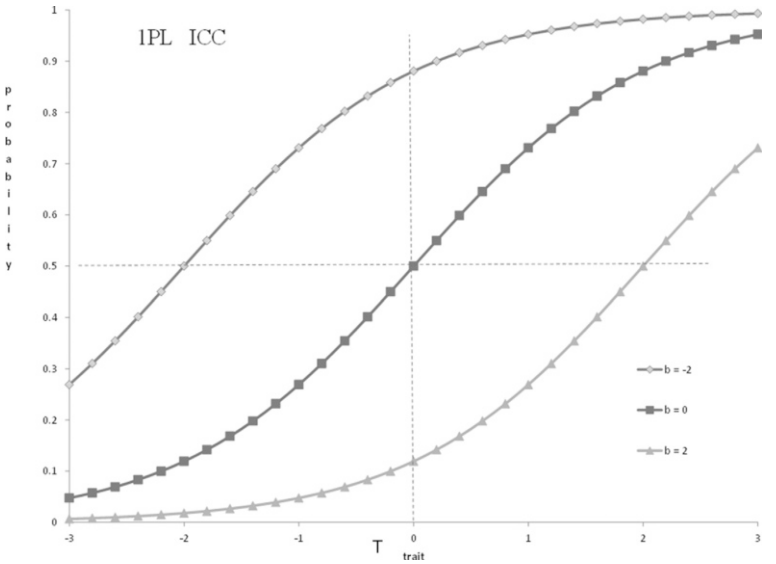


Figure 1. ICCs for three items with different difficulty levels.

difficulty of 2. From another perspective, in the middle of the graph is a horizontal line which intersects with the three curves. The line represents $p = .5$ probability to endorse an item. In fact, when an examinee's ability matches the item difficulty (i.e., $\theta = \beta$), the probability of endorsing the item is 0.5. Also, at this point, the slope of the *ICC* (i.e., the first derivative of the function) reaches its maximum of 0.25 when 1PL is utilized for modeling the function (Hambleton & Swaminathan, 1985; Hambleton et al., 1991).

Two-Parameter (2PL) Logistic Model

In addition to item difficulty parameter, the 2PL model involves another item parameter α, known as item discrimination. The discrimination parameter is proportional to the *ICC* slope at the difficulty level b along the trait continuum (Hambleton et al., 1991). The larger the discrimination parameter, the more powerful the item is in separating lower-ability from higher-ability examinees. Theoretically, the discrimination parameter can range from negative infinity to positive infinity. A negative value of discrimination, however, is counterintuitive because it means the decrease of probability of endorsing an item with the increase of ability. In practice, item discrimination parameter α typically is within the range between 0 and 2 (Hambleton et al.). Other researchers recommended that reasonably good values for item discrimination parameter range from .8 to 2.5 (e.g., Ayala, 2009). The equation representing 2PL model is as below:

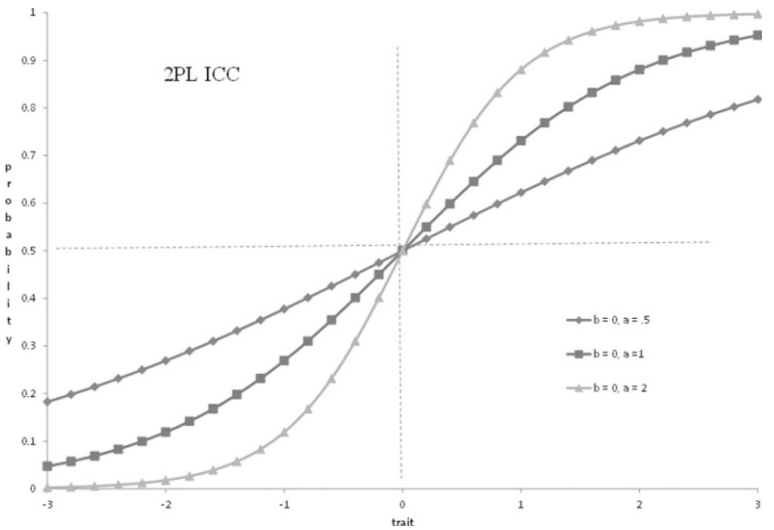


Figure 2. ICCs for three items with different discrimination parameters.

$$p(X_{is}=1|\theta_s,\beta_i,\alpha_i)=\frac{\exp[\alpha_i(\theta_s-\beta_i)]}{1+\exp[\alpha_i(\theta_s-\beta_i)]},\ i=1,2,\cdots,n \tag{4}$$

where, X_{is} = response of person s to item i (with response options 0 or 1)
θ_s = latent trait level for person s
β_i = difficulty level for item i
α_i = discrimination power for item i

Similar to the 1PL model, the 2PL involves both person parameter and item parameters, but with one more parameter α. Figure 2 presents the *ICC*s of three items of 2PL model. The three items possess the same item difficulty level ($\beta = 0$). Therefore, we can see that the three ICC cross at the point which corresponds to 0.5 endorsement probability. As explained earlier, at the probability 0.5, the item difficulty matches the measured ability perfectly. However, because of the different values of discrimination parameter ($\alpha = 0.5$, 1.0, and 2.0, respectively), the three curves show different "steepness." The steepest curve corresponds to the highest discrimination power ($\alpha = 2$), whereas the most flat curve has the lowest discrimination power ($\alpha = .5$).

Three-Parameter (3PL) Logistic Model

Compared with 1PL and 2PL IRT models, the 3PL model incorporates one more item parameter *c*, which represents the guessing parameter or pseudo-chance-level

parameter. This additional parameter represents the probability for a low-ability examinee to answer the item correctly, but the correct response is not the result of the examinee's ability, but of some other random errors such as guessing. The mathematical expression of the 3PL is presented below:

$$p(X_{is}=1|\theta_s,\beta_i,a_i,c_i)=c_i+(1-c_i)\frac{\exp[a_i(\theta_s-\beta_i)]}{1+\exp[a_i(\theta_s-\beta_i)]},\ i=1,2,\cdots,n \qquad (5)$$

where, X_{is} = response of person s to item i (with response options 0 or 1)
θ_s = latent trait level for person s
β_i = difficulty level for item i
a_i = discrimination power for item i
c_i = random guessing factor for item i

Figure 3 presents three *ICC*s for three items, with the same item difficulty ($\beta = 0$) and item discrimination ($\alpha = 1$), but with different guessing parameter values ($c = 0.0$, 0.1, and 0.2, respectively). On the graph, the guessing parameter value is reflected by the asymptote on the left end of the trait continuum. As the graph shows, for low-ability examinees, even with a trait value of −3, the persons have some probability of endorsing the items, or answering the items correctly, depending on the guessing parameter values.

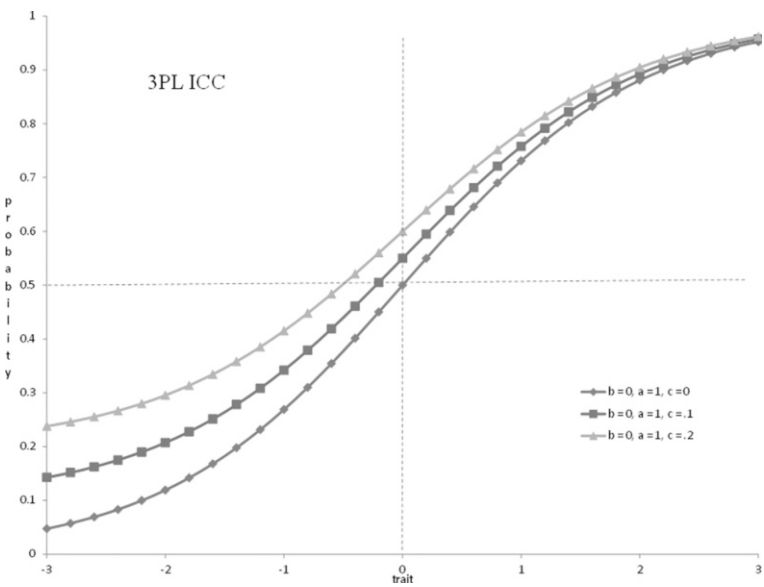


Figure 3. ICCs for three items with different guessing parameter values.

Up to this time, we have focused on binary response items. In practice, however, an item may involve three or more response options. For instance, an item with Likert-scale response options could have five response categories ranging from 1 (strongly disagree) to 5 (strongly agree). As such, binary response models do not apply, and polytomous response models should be utilized.

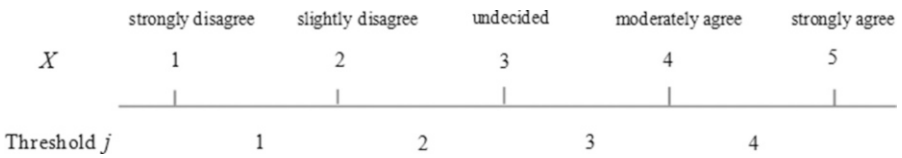
POLYTOMOUS ITEM RESPONSE MODELS

There are three major types of polytomous IRT models: graded response model, partial credit model, and nominal response model. Because of space constraints, we will discuss the graded response model with some details, but only provide brief descriptions of other two models.

The Graded Response Model (GRM)

The GRM is an extension of the 2PL binary model. Assuming we have an item with five response categories, we will have the following response dichotomies: (a) category 1 vs. categories 2, 3, 4, and 5; (b) categories 1 and 2 vs. categories 3, 4, and 5; (c) categories 1, 2, and 3 vs. categories 4 and 5; and (d) categories 1, 2, 3, and 4 vs. category 5. Suppose we attempt to measure students' self-esteem with the following exemplar item:

On the whole, I am satisfied with myself.



Equation 6 below is the mathematical presentation of the GRM, where $p^*_{ix}(\theta)$ denotes the probability of endorsing each response option category x or higher as a function of the latent trait θ, whereas $p_{i1}(\theta)$ denotes the probability of responding in the first category. By the same token, $p_{i2}(\theta)$ represents the probability of responding in the second category, and so on.

$$p^*_{ix}(\theta) = \frac{\exp[a_i(\theta - \beta_{ij})]}{1 + \exp[a_i(\theta - \beta_{ij})]} \quad i = 1, 2, \cdots, n \quad j = 1, 2, \cdots, n-1 \tag{6}$$

$$p_{i1}(\theta) = 1.0 - p^*_{i2}(\theta)$$
$$p_{i2}(\theta) = p^*_{i2}(\theta) - p^*_{i3}(\theta)$$
$$p_{i3}(\theta) = p^*_{i3}(\theta) - p^*_{i4}(\theta)$$
$$p_{i4}(\theta) = p^*_{i4}(\theta) - p^*_{i5}(\theta)$$
$$p_{i5}(\theta) = p^*_{i5}(\theta) - 0$$

Category Response Curve (CRC). In the GRM, category response curves (CRC) are used to describe the probability of endorsing a particular category option as a function of the latent trait. In general, each *CRC* peaks in the middle of two adjacent threshold parameters β_{ij}, and the more peaked or narrow the *CRC*, the more item discrimination power it has (Embretson & Reise, 2000). Figure 4 below is a graphical presentation of CRC of a measurement item with seven graded response options:

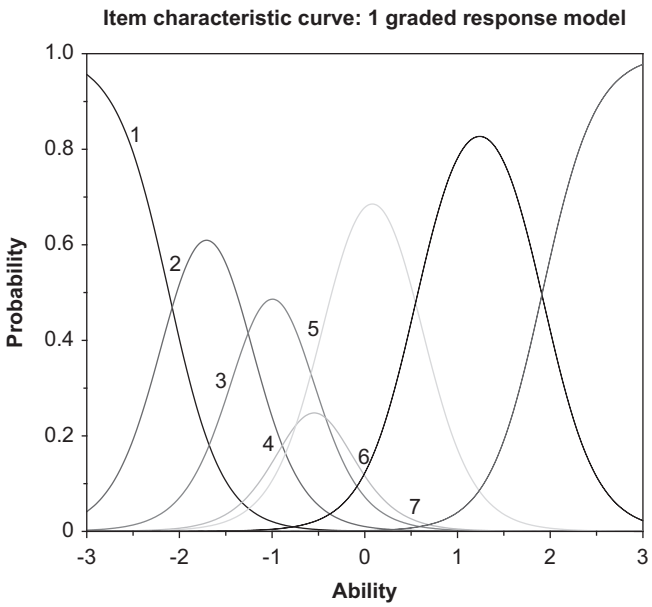


Figure 4. CRCs for an item with seven graded response options.

Partial Credit Model (PCM)

As an extension of the 1PL model, PCM is originally developed to score test items which require multiple steps, and hence entailing assigning examinees partial credits for their response. Consider a simple example. An examinee needs to solve a math problem: (5*4) + 6 =? To reach the final solution, the examinee has to go through a step-by-step process, where the first step requires multiplication and the second step requires summation. That said, an examinee who only gets the first step right will be awarded partial credit, whereas one who gets both steps right will receive full credit. In PCM, all items are assumed to have the same slope or discrimination power. The only parameter is termed step difficulty or transition location parameter, often denoted as δ. Compared to the GRM, the equation for the PCM is relatively unwieldy and hence is not presented here. Interested readers can consult any book discussing polytomous IRT models for more details.

Nominal Response Model (NRM)

Nominal response models do not involve ordered response categories. Instead, all response options are parallel to each other. Multiple-choice test is a case of applying nominal models. The model models the probability for an examinee with certain trait level to select a particular response category. Bock's nominal response model is presented below:

$$p_{ix}(\theta) = \frac{\exp(a_{ix}\theta + c_{ix})}{\sum_{x=0}^{m} \exp(a_{ix}\theta + c_{ix})} \tag{7}$$

In Equation 7 above, *i* represents items, and *x* represents the response categories. By adding identification constraints, the two parameters of the model can be estimated. It should be pointed out, however, the graded response model discussed previously can be considered as a special case of nominal response models, with all the response categories being ordered.

SOME MAJOR CONSIDERATIONS IN IRT APPLICATIONS

Model Specification and Fit Assessment

If the earlier-introduced unidimensionality and local independence assumptions are met, an inevitable question is how to select the best model among the wide range of IRT models including 1 PL, 2 PL, 3 PL, and polytomous models. In other words, with what procedure and against what criteria do we judge whether an IRT model captures the sample data well? Consider a simple example: the actual data follow an IRT model with varying slopes, but we fit a 1PL model which assumes a uniform slope value (i.e., the same discrimination power) across items. Then, it is reasonable to expect the 1PL model would fit poorly or biasedly because it involves model misspecification. In practice, often, different items of a test are represented by different IRT models. A pertaining case is when a test is comprised of both dichotomous and polytomous response items. In such circumstances, model fit assessment usually unfolds on an item-by-item basis (DeMars, 2010). Specifically, researchers rely on various residual statistics by computing the difference between observed and model-implied (expected) proportion-correct (or proportion endorsed) on a particular item. Large residuals typically indicate poor item fit which may result from a variety of reasons. For example, violation of unidimensionality, a non-monotonous relationship between item response and the underlying trait, unspecified item parameters are only a few possible cases in point (Embretson & Reise. 2000).

It is also possible to judge model fit at the test level. The typical approach is similar to procedure of model evaluation in structural equation modeling. By comparing two different models, often one nested within another, researchers can examine

the likelihood ratio comparison indices (e.g., Chi-square) to determine whether statistically significant differences exist between the complex model and the more parsimonious one. Later, we will present an example applying this procedure to compare 1PL and 2PL model fit with the same set of data.

Item and Test Information

With an instrument, each item contributes to our understanding of an examinee's position on the ability continuum, and reduces our uncertainty about one's ability location. Correspondingly, *item information function*, denoted as $I_i(\theta)$, serves as an index to evaluate the amount of information that one particular item contributes to ability assessment. $I_i(\theta)$ is related to the previously discussed item parameters. In general, item information is higher under the conditions: (a) when the difficulty value β is closer to the trait value θ, than when the two values are far different from each other; (b) when the discrimination value α is high; and (c) when the value of guessing parameter approaches zero.

For dichotomous item response models, $I_i(\theta)$ can be derived from the following:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)(1 - P_i(\theta))} \tag{8}$$

In the above, $P_i(\theta)$ denotes the conditional probability of endorsing a binary item, given the item response function, and $P_i'(\theta)$ refers to the first derivative of item response function, given the estimated trait level. Further mathematical reduction can lead us to a simplified form for item information function of 2PL models (Embretson & Reise, 2000; Hambleton et al., 1991):

$$I_i(\theta) = a_i^2 P_i(\theta)(1 - P_i(\theta)) \tag{9}$$

Figure 5 presents the item information functions of three items under 2PL response model. The three items possess the same difficulty values but different discrimination parameter values (0.5, 1.0, and 2.0, respectively). Clearly, the item with the largest discrimination value demonstrates the highest information, whereas the less discriminative item shows less item information. Moreover, as mentioned earlier, an item would convey more information when item difficulty matches an examinees' trait level. In the present example, when trait level is close to the difficulty value 0, the item conveys more information in differentiating examinees' ability levels. For examinees with trait levels far from $\theta = 0$ (e.g., $1.5 < \theta < 2$), the items, even the one with the highest level of item information function, will provide much less information to help us in differentiating examinees with different ability levels.

An important property of item information curves is that they are additive. When item information from all the items on a test is added, it leads to the test information

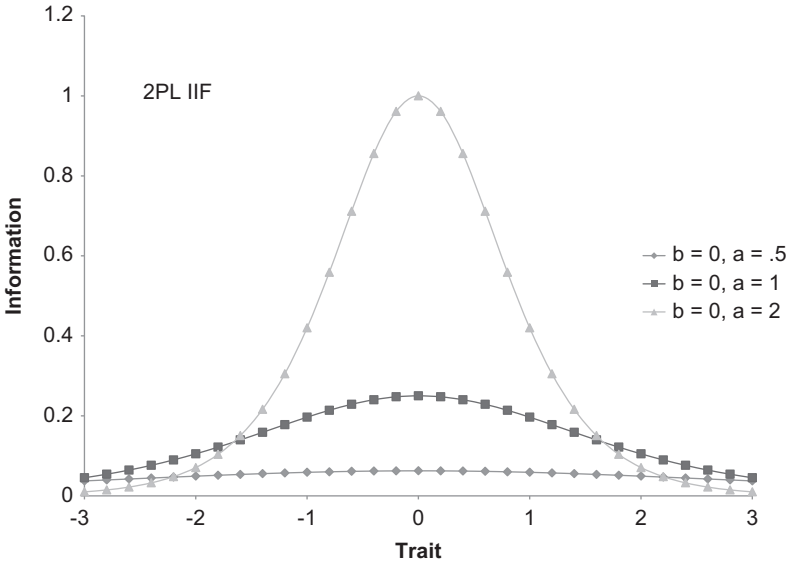


Figure 5. Item information functions of three items.

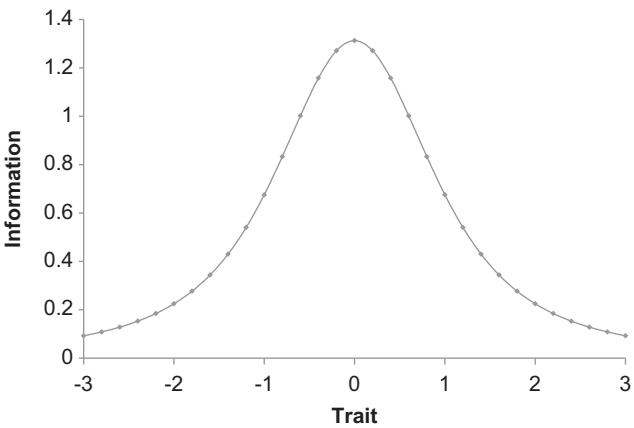


Figure 6. Test information function of a three-item test.

curve. Figure 6 is the plot of test information function by combining the three items in Figure 5 above. This test information curve informs us that this test of three items would be most effective in differentiating examinees with trait level θ in the range of -1 to $+1$. For examinees with trait level θ below -1, or above $+1$, this test would not be effective in differentiating the examinees.

Item information in IRT modeling analysis plays a role similar to what reliability does in classical test theory. Each item response function can be transformed to

corresponding item information function, which provides insight about the precision of ability assessment along the trait range. As item information [$I_i(\theta)$] of different items is additive, the resultant sum across all the items on a test is the test information [$I(\theta)$], which reflects the information contribution of the whole instrument to ability assessment. In addition, the standard error of estimation, or standard error of measurement, which reflects the variance of latent trait estimation, is the reciprocal of test information, as shown in Equation 10. As such, the higher the test information, the smaller the standard error of estimation is, and the less error there is in ability assessment.

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \tag{10}$$

In Equation 10, $I(\theta)$ represents the test information, and $SE(\hat{\theta})$ represents the standard error of estimation of the whole test or instrument. In classical test theory, the standard error of measurement is constant for a designated test, regardless of examinee's ability level. In *IRT*, however, the standard error varies with the trait level θ, and hence conveys more precise information with respect to a specified trait level θ. Also, because measurement error is mapped to the same metric as the latent trait, confidence intervals can be easily constructed.

ADVANCED TOPICS OF IRT

Multidimensional Model

As mentioned earlier, two important assumptions of IRT model estimation are local independence and unidimentionality of the underlying trait being measured. Nonetheless, the latter assumption does not always hold in practical settings. It is very likely that an examinee's response to a test is driven by more than one latent trait. For instance, an examinee's performance on a math test depends on his/her math ability. On the other hand, understanding the wording of the math problems is a prerequisite for tackling the question. Thus, the examinee's response or final score could be a reflection of both math ability, and reading ability, although these two types of ability may play different roles in the hypothesized situation. As a result, IRT model based on unidimentionality assumption is not the most applicable in this or other similar situations. Further, if one arbitrarily ignores the situation of multidimensionality and continue to apply a model with unidimensionality constraints, calibrated scores could be misleading and difficult to interpret (Yen & Fitzpatrick, 2006).

Test Score Linking and Equating

When examinees take different tests measuring the same latent construct, are those examinees still comparable in terms of their test scores? The aforementioned question directly speaks to test score equating and scale linking. Simply speaking,

linking is the process of aligning different metrics so that parameter estimates from different samples or models can be compared. Equating refers to procedures of adjusting or converting, to a common metric, the scores of different examinees on different tests so as to better compare individuals (Ayala, 2009). Generally speaking, the goal of linking is to adjust item parameter estimates, whereas the goal of equating is to adjust person location estimates (Ayala, 2009). Equating test scores with IRT models usually entail four steps: (a) select the suitable equating design, (b) decide the appropriate item response model, (c) build a common metric for item or trait parameters, and (d) determine the scale for test score reporting (Hambleton & Swaminathan, 1985). Researchers who are more interested in test equating can consult the more comprehensive and detailed work by Kolen and Brennan (2010).

Differential Item Functioning (DIF)

DIF refers to such a situation where respondents from different groups (e.g., gender groups, cultural groups) have the same level of the measured ability (θ), but show different probability for endorsing an item. Alternatively stated, the item is said to be biased against a particular group of respondents as opposed to other groups. When *DIF* occurs, for the group that the test is biased against, the test scores fails to represent the true levels of examinees' ability or trait that is being measured. A typical approach for detecting DIF is to compare item response functions. The logic is straightforward: an item with *DIF* will not show identical response function across different groups. Conversely, if an item does show identical response functions across groups, then no *DIF* exists (Hambleton et al., 1991).

ILLUSTRATIVE EXAMPLE FOR BASIC *IRT* MODELING ANALYSIS

Data Sample

The data used for this illustration came from the Texas Assessment of Academic Skills (TAAS) tests administered to 11th-grade students in the early 1990s (Fan, 1998). The original dataset is very large, so we randomly selected 1000 observations for the current illustrative example. The test contained 48 reading items and 60 math items. For illustrative purpose, we only selected 10 math items. All items are dichotomously coded, with 0 denoting incorrect answer and 1 denoting correct answer.

Assessment of Data Unidimensionality

As mentioned earlier, unidimensionality is an important assumption for most *IRT* models. Therefore, we conducted categorical factor analysis with the weighted least squares estimation in *Mplus* (Version 5.0). The overall χ^2 test of one-factor model is not statistically significant ($\chi^2_{(35)} = 47.574$, $p > .05$). Other fit indices also showed

that the one-factor model fits the data well (*CFI* = .990; *TLI* = .987; *RMSEA* = .019; *SRMR* = .049). The "scree" plot of eigenvalues for the estimated polychoric correlation matrix was shown in Figure 7, which suggests that one-factor solution is viable for the data.

We also examined the two-factor solution, and the results indicated that this would be overfactoring the data (e.g., only one item loaded high on the second factor (see Cole et al., 2011). Factor loadings of both the one-factor solution and the two-factor solution were presented in Table 1. It should be noted, in the two-factor solution as shown in Table 1, Item 10's loading on the second factor was beyond the value of

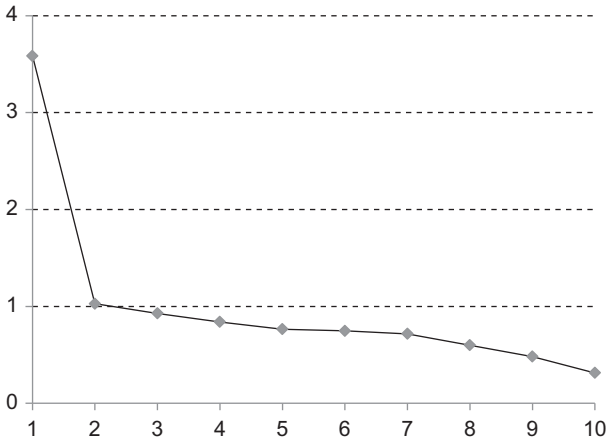


Figure 7. Eigenvalue "scree" plot for the 10-item "mini-test".

Table 1. Factor loadings of one- and two-factor solutions

		Two Factors	
Items	*One Factor*	*Factor 1*	*Factor 2*
1	0.487	0.564	–0.070
2	0.499	0.548	–0.037
3	0.385	0.449	–0.061
4	0.576	0.541	0.049
5	0.687	0.574	0.116
6	0.282	0.214	0.082
7	0.507	0.540	–0.018
8	0.570	0.487	0.094
9	0.512	0.578	–0.051
10	0.782	0.001	1.681

typical expectation, and it is very different from those of the rest of the items. Further inspection revealed a negative estimate of the residual variance for that item. Such a Haywood case further supports the one-factor solution.

IRT Analysis

A host of specialized software has been developed for *IRT* modeling analysis. Among them, BILOG-MG is designed mainly for binary item analysis. PARSCALE can perform analysis of both binary and polytomous items. Another popular software is MULTILOG, which can be used for implementing most IRT models, including rating scale, graded response, multiple choice, partial credit, etc. More recently, IRTPRO has come out (Cai, Thissen, & du Toit, 2011) to replace MULTILOG. The new software incorporates almost all the functions that MULTILOG can provide, and is more powerful and promising because it deals with both unidimensional and multidimensional *IRT* models. It should be pointed out that most IRT software is specialized, and hence has limited size of users. On the other hand, some general statistical software such as *Mplus* or *SAS* also offers some IRT modeling analysis. But, if a researcher is interested in more comprehensive IRT analysis, specialized software typically is the better option.

For the present analysis, we utilized the beta version of IRTPOR software (Cai, Thissen, & du Toit, 2011). Cronbach's coefficient α for the "mini-test" of ten items is 0.63, with more detailed item statistics and other information in Table 2.

Table 2. Item statistics and related information of the "mini-test"

Item	*p-value (std.)*	*Corrected*[a] *Item-Total r*	*Item-Deleted*[b] *Coefficient* α
1	0.910 (0.286)	0.2435	0.6202
2	0.857 (0.350)	0.2786	0.6131
3	0.907 (0.291)	0.1881	0.6285
4	0.799 (0.401)	0.3398	0.5997
5	0.841 (0.366)	0.3671	0.5953
6	0.782 (0.413)	0.159	0.6394
7	0.632 (0.483)	0.3097	0.6077
8	0.764 (0.425)	0.3312	0.6013
9	0.618 (0.486)	0.326	0.6034
10	0.725 (0.447)	0.4621	0.5678

[a] For calculation of these correlation coefficients, the "total" was obtained without the contribution of a particular item in question.

[b] This is the coefficient *α* of the remaining nine items, without the contribution from a particular item in question.

We first fitted a 2PL model to the data. Estimates of item difficulty parameter β and those of item discrimination parameter a were presented in Table 3.

By comparing Table 2 and 3, we can see that the two tables provided consistent information about item discrimination and difficulty (see also, Fan, 1998). In Table 2, the corrected item-total correlation represents item discrimination. Values in Table 2 showed that Item 10 has the highest discrimination power (.4621), whereas Item 6 possesses the lowest discrimination power (.1590). In Table 3, corresponding discrimination parameter values for Item 10 is 2.13, and for Item 6 is 0.53. For item difficulty, item *p*-value in Table 2 represents the percentage of endorsement for each item. Lower item *p*-value means smaller proportion of respondents endorsing an item, or answering an item correctly, and thus the more difficult the item is. For example, In Table 2, Item 1, with an endorsement percentage of .91, is the least difficult among the ten items, and while Item 9, with endorsement percentage of .618, is the most difficult on this "mini-test". In Table 3, Column *c* represents item difficulty information. It is shown that Item 1 has the lowest difficulty value of –2.75 (reverse sign to the tabled value), and item 9 has the highest difficulty value of –0.58.

It should be noted that IRTPRO outputs two different forms of parameter estimates. Correspondingly, the response function also takes two forms (Equation 11; see IRTPTO user's guide). The former function (first part of Equation 11) is called slope-intercept model, where α is the slope or discrimination parameter and c is the intercept. In the latter equation (second part of Equation 12), β is the threshold parameter.

$$P = \frac{1}{1+\exp[-(\alpha_i\theta_s + C_i)]} = \frac{1}{1+\exp[-\alpha_i(\theta_s - \beta_i)]}, i = 1, 2, \cdots, n \qquad (11)$$

Table 3. 2PL model item parameter estimates [logit: aθ + c or a(θ – b)]

Item	*α (s.e.)*	*c (s.e.)*	*b (s.e.)*
1	1.09 (0.17)	2.75 (0.17)	–2.53 (0.31)
2	1.05 (0.15)	2.14 (0.13)	–2.05 (0.23)
3	0.80 (0.15)	2.53 (0.14)	–3.14 (0.51)
4	1.25 (0.15)	1.77 (0.12)	–1.42 (0.13)
5	1.75 (0.21)	2.47 (0.19)	–1.41 (0.11)
6	0.53 (0.11)	1.35 (0.08)	–2.57 (0.48)
7	0.96 (0.12)	0.65 (0.08)	–0.68 (0.10)
8	1.23 (0.14)	1.51 (0.11)	–1.23 (0.11)
9	0.98 (0.12)	0.58 (0.08)	–0.59 (0.09)
10	2.13 (0.25)	1.66 (0.17)	–0.78 (0.06)

If we compare Equation 11 above with Equation 4 introduced earlier, simple algebraic re-arrangement leads to the following:

$$P = \frac{1}{1+\exp[-a_i(\theta_s - \beta_i)]} = \frac{\exp[a_i(\theta_s - \beta_i)]}{1+\exp[-a_i(\theta_s - \beta_i)]} \quad i = 1, 2, \cdots, n \qquad (12)$$

Attention should be paid to the positive sign before c, and negative sign before β. So, to interpret β and c comparably, we need to add a negative sign to the c values presented in Table 3.

Figure 8 provides the graphs of trace lines and item information curves for three selected items on this "mini-test": Item 6, Item 7, and Item 10. The two "trace lines" (the two solid curve lines in each graph) represent the respective probabilities of endorsing one of the two response categories (0, 1) for a given ability level θ, which is represented by the x-axis of the graph. The item information curve is represented by the dashed line in each graph. As the graph indicates, each item provides the maximum amount of information around the point where the two trace lines for

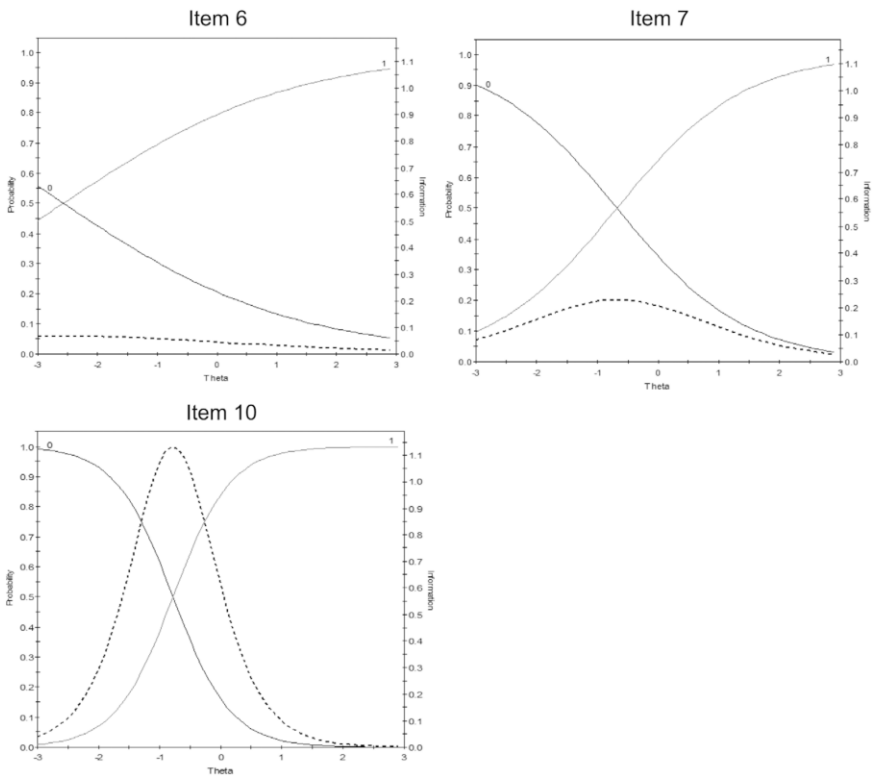


Figure 8. Trace lines and item information curves of three selected items.

the two response categories (1, 0) intersect with each other. In other words, an item provides most information about an examinee's ability when an examinee's θ level is at the point where exist approximately equal probabilities for the examinee to endorse either of the two categories.

Among the three selected items, Item 6 provides very little information along the continuum of latent trait θ, and the item information curve is both very low and essentially flat. This means that this item contributes very little to our knowledge about examinees' ability or performance, as it does not differentiate between examinees with lower- vs. higher-ability for any level of θ. The main reason that Item 6 provides very little item information at any ability level is that, Item 6 has very low discrimination value (Table 2: corrected item-total $r = 0.1590$; Table 3: $a = 0.53$). Consequently, this item is not useful in discriminating or separating examinees with lower vs. higher ability levels for any given θ level.

Item 7 and Item 10 have very different item information curves. Item 7 has somewhat symmetrical, but very low and flat item information curve, with the lowest and highest point of the curve being approximately 0.05 and 0.2, respectively. This means that, Item 7 is not really that informative in differentiating examinees' ability levels, and it contributes little to our understanding about which examinee has higher trait level compared with other examinees. On the other hand, Item 10 also has somewhat symmetrical information curve, but its curve is very steep, with the lowest and highest point of the curve being 0.00 and 1.10, respectively. In addition, the steep curve peaks approximately at the point of $\theta = -0.75$. All this informs us that, (a) this is a relatively informative item with respect to an examinee's trait level; (b) this item is the most informative for examinees with trait level θ at approximately –0.75; (c) for examinees with trait level θ being considerably lower or higher than –0.75, this item will be much less informative. For example, this item would be almost useless in differentiating among examinees with trait level $\theta > 1$. The difference between Items 7 and 10 as discussed above is reflected by their discrimination index information presented in Table 2 (corrected item-total correlations of .0.3097 vs. 0.4621, respectively) and Table 3 (item discrimination parameter estimates of 0.96 vs. 2.13, respectively). It is obvious item information function provides much richer information than item discrimination index alone.

The brief discussion above reveals one crucial difference between item discrimination information (e.g., corrected item-total correlation) in classical test theory and item discrimination and item information function in IRT. In classical test theory, we talk about item discrimination as if it were applicable for examinees at any trait (θ) level. In IRT, however, item information function is operationalized and quantified relative to a given trait level θ. Within this framework, it may not be an issue of whether or not an item is informative or useful in general, but whether an item is informative relative to certain θ level range. Because of this, an item may or may not be informative, depending on what trait level is being considered. This characteristic of item information naturally leads to the discussion about test information curve below.

Figure 9 presents the test information curve of this 10-item test. As discussed above, test information curve has similar meaning as item information curve, but this is about the whole test. In this graph, there are two curve lines. One (solid curve line) is the test information curve line, and the other (dashed line) is the standard error line. The horizontal axis represents the continuum of the trait level (θ) being measured. The left vertical axis represents the amount of test information relative to the trait level θ. The right vertical axis represents the magnitude of standard error of estimation relative to the trait level θ. As defined in Equation 10, the standard error of information in IRT modeling is the inverse of the square root of test information.

Figure 9 shows that the test provides much more information for the trait level range of $-2 < \theta < 0$. That is, in the lower ability range, this "mini-test" provides more information in differentiating examinees with different levels of the trait. In contrast, this "mini-test" provides relatively little information in the higher ability range (i.e., $\theta > 0$). The reason for the low information for high-ability examinees is simple: the items on this "mini-test" were easy, as shown in Table 2. It should be noted that, unlike classical test theory in which standard error of measurement is a constant for all examinees with different levels of trait measured by a test, the magnitude of standard error of estimation in IRT framework has an inverse relationship with the test information. As a result, the magnitude of standard error of

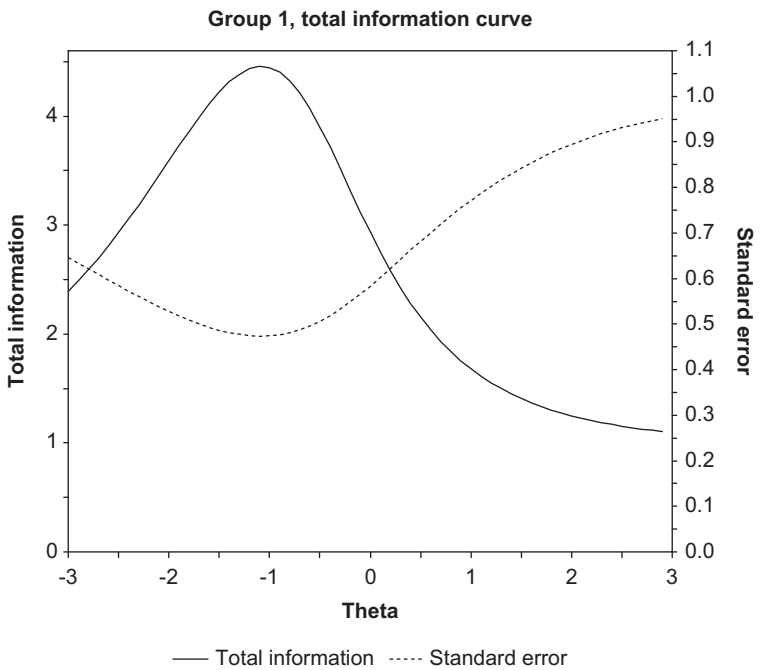


Figure 9. Test information curve (solid line) of the "mini-test".

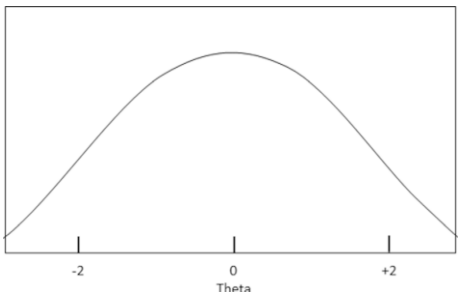


Figure 10. Hypothetical test information function for a test designed for a broad range of trait level.

estimation is not constant across the ability range; instead, the magnitude of standard error of estimation is related to the trait level θ. In Figure 9, the magnitude of standard error of estimation is considerably greater for the trait level range of $\theta > 1$, and our measurement based on this "mini-test" is most precise around trait level of $\theta = -1$ where test information curve peaks. So this "mini-test" provides much more measurement precision for examinees at or around the level of $\theta = -1.00$. But for examinees with higher performance level, the scores of this "mini-test" are much more crude with considerably larger measurement error.

Test information curve provides extremely useful information for test improvement considerations. Whether or not the test information curve as shown above is good enough depends on the purpose of this "mini-test." For example, if this "mini-test" is designed to separate students into two groups, one group with very low math performance such that they will need remedial courses before they can take regular math classes, and the other group who are ready to take regular math classes now. If the cut-off point separating these two groups is set at approximately $\theta = -1$, then this test information curve is perfect for this intended purpose. Lack of test information above $\theta = 0$ would not be our concern, as we are only interested in separating those at or around $\theta = -1$, and we have no interest in differentiating those with math trait level above $\theta = 0$.

On the other hand, if the purpose is to provide measurement to cover a broad range of trait level in math performance (e.g., $-2 < \theta < 2$), the test information curve in Figure 9 would be considered deficient, primarily because it provides very little information about higher range of the trait level (e.g., $\theta > 0$). To serve such a purpose, we will need a test information curve similar to Figure 10, which has relatively high level of test information over the range of $-2 < \theta < 2$.

The fit indices showed that, in general, the 2PL model fits the data well ($M_2 = 53.83$, $df = 35$, $p = .02$, $RMSEA = .02$). The log likelihood is 9256.19, but this information is only informative when this model is being compared to another nested model. Therefore, we proceeded to test a more constrained and nested 1PL model. By constraining the item discrimination parameter to be equal across

Table 4. 1PL model item parameter estimates [logit: aθ + c, or a(θ – b)]

Item	*α (s.e.)*	*c (s.e.)*	*b (s.e.)*
1	1.12 (0.06)	2.78 (0.13)	–2.48 (0.15)
2	1.12 (0.06)	2.19 (0.11)	–1.95 (0.12)
3	1.12 (0.06)	2.74 (0.13)	–2.44 (0.15)
4	1.12 (0.06)	1.70 (0.09)	–1.52 (0.10)
5	1.12 (0.06)	2.04 (0.10)	–1.82 (0.11)
6	1.12 (0.06)	1.58 (0.09)	–1.41 (0.10)
7	1.12 (0.06)	0.68 (0.08)	–0.61 (0.07)
8	1.12 (0.06)	1.46 (0.10)	–1.30 (0.09)
9	1.12 (0.06)	0.61 (0.08)	–0.54 (0.07)
10	1.12 (0.06)	1.21 (0.09)	–1.08 (0.08)

all items, the previous 2PL model is reduced to 1PL model. As shown in Table 4, the discrimination parameter for the ten items is the same ($a = 1.12$), and only the item difficulty parameter c varies across the items. Because this 1PL model is more constrained, we expect that the model would not fit the data as well as the 2PL does. The fit indices do show a worse fit ($M_2 = 120.72$, $df = 44$, $p = .0001$, $RMSEA = .04$). The corresponding log likelihood is 9325.75. The difference of log likelihood values of nested models approaches chi-square distribution, which would provide a statistical test for testing which model fits better. As such, the difference of log likelihood between 1PL and 2PL is 69.56 (9325.75 – 9256.19 = 69.56), with df difference (df_Δ) being 9 (i.e., 44 – 35 = 9). This test on the difference of log likelihood values of nested models is statistically highly significant. In other words, the difference of model fit between the two models is not trivial, and it is more than what we would expect from sampling error or sampling fluctuation. Based on this evidence, we would conclude that the 2 PL model is preferable to the 1PL model for this measurement data set.

RESEARCH ISSUES

Item response theory, as a measurement framework that is still developing, holds great promise for applications in educational measurement and research, as it offers many advantages over the framework of classical test theory. Here, we briefly discuss a few directions wherein IRT may have important research applications.

First, test construction and scale development is an area where IRT can have significant influence. In contrast to traditional methods, IRT can more readily identify biased test items, thus enhancing measurement validity for examinees from different populations. Measurement invariance is always an important issue

in cross-cultural and cross-group research. As mentioned before, identifying items biased against certain groups (e.g., an ethnic or marginalized group) can greatly improve the validity of measurement and assessment. Some scales may be unbiased in one cultural setting, but may turn out to be biased in another cultural setting. For example, a scale can be more readily endorsed by males than females in Asian culture, but not necessarily in the Western culture. To learn more about applying IRT to scale construction, interested readers may consult, for example, Wilson (2005).

Second, for assessment in the areas of personality, cognition, and attitudes, IRT applications may provide information not readily available before. For example, application of the mixed-measurement IRT model, incorporating both latent-trait and latent-class analysis, can help detect both qualitative and quantitative individual differences (Embretson & Reise, 2000). Recently, Huang and Mislevy (2010) integrated evidence-based design and polytomous Rasch model to assess students' problem-solving ability. As the authors discussed, combining cognitive psychology, task design, and psychometric analysis will open new avenues for educational measurement. Moreover, in addition to providing the standard descriptive information about items, IRT also can be very useful for explanatory purposes. Interested readers are encouraged to read Boeck and Wilson (2004) for more information.

Third, Computerized Adaptive Testing (CAT) and item banking are another area to which IRT can contribute significantly. With the rapid diffusion of new computing technologies and psychometric modeling, CAT has clearly become a trend for the future. As compared to traditional paper-and-pencil tests, CAT possesses a number of advantages, such as practicality of automatically creating tests tailored to individual examinees, and the possibility of shortening the test length, not to mention the time and cost saving. For CAT applications, a viable item bank should consist of a sufficient number of items with good discrimination power and difficulty level across the latent trait range (Thissen & Wainer, 2001). IRT is fundamental for item selection, item bank construction, and for scoring examinees in any CAT applications.

Fourth, IRT is a promising technique for assessing reliability and validity. For example, IRT can help address the construct validity of cognitive and developmental assessment. Specifically, the technique helps assess dimensionality, decompose cognitive process, detect qualitative differences among respondents, and facilitate interpretation of measured ability (Embretson & Reise, 2000). In this regard, the classic work by Wainer and Braun (1988) devotes several chapters explicating the linkage between IRT and various aspects of measurement validity.

Fifth, many sustaining and challenging research topics are related to test equating and scale calibration. These topics include, but not limited to, scale drift, equating strains, scale shrinkage, as well as nonparametric IRT models (Kim, Harris, & Kolen, 2010). Tackling some of these issues depends on advances of IRT, computational statistics, statistical theories, and other related methodology areas.

For future research, more work is needed in applying IRT in the context of multidimensional latent traits. Indeed, in recent years, there have been significant development and advancement of MIRT. Nonetheless, software application of MIRT

is still rare, and research applications of MIRT are still lacking. On a different note, bridging the gap between IRT and other latent variable modeling techniques is another fascinating area. Factor analysis, multilevel modeling, structural equation modeling and IRT have differences, but also share commonalities (Skrondal & Rabe-Hesketh, 2004). More research is warranted to link these techniques and apply them in broader research and application contexts.

REFERENCES

Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Cai, L., Thissen, D., & du Toit, S. H. (2011). IRTPRO: *Flexible, multidimensional, multiple categorical IRT modeling* [Computer software]. Chicago, IL: Scientific Software International.

Cole, D. A., Cai, L., Martin, N., Findling, R. L., Youngstrom, E. A., Garber, J., & Forehand, R. (2011). Structure and measure of depression in youths: Applying item response theory to clinical data. *Psychological Assessment, 23*, 819–833.

DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational & Psychological Measurement, 58*, 357–381.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Norwwell, MA: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Huang, C., & Mislevy, R. J. (2010). An application of the polytomous Rasch model to mixed strategies. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 211–228). New York: Routledge.

Kim, S., Harris, D. J., & Kolen, M. J. (2010). Equating with polytomous item response models. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 257–291). New York: Routledge.

Kolen, M. J., & Brennan, R. L. (2010). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman & Hall.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum.

Wainer, H., & Braun, H. I. (1988). *Test validity*. Hillsdale, NJ: Lawrence Erlbaum.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: Praeger.

SECTION 2

METHODS OF ANALYSIS

KEN KELLEY & JOCELYN H. BOLIN

4. MULTIPLE REGRESSION

Multiple regression is a commonly used analytic method in the behavioral, educational, and social sciences because it provides a way to model a quantitative outcome variable from regressor variables.[1] Multiple regression is an especially important statistical model to understand because special cases and generalizations of multiple regression are many of the most commonly used models in empirical research. Correspondingly, multiple regression occupies a core position in the analytic architecture of behavioral, educational, and social science research.

In this chapter we (a) provide an overview of multiple regression, (b) emphasize the meaning and interpretation of the various regression model parameters, (c) discuss inference based on the regression model, and (d) briefly discuss selected important topics in an effort for readers to be in a better position to understand and use the multiple regression model. Throughout the chapter we use an illustrative data set to motivate and demonstrate an application of the multiple regression model. After a delineation of the model and presentation of the analytic details, we turn to a "big picture" perspective in the discussion section on what we see as the three primary purposes of multiple regression. In particular, we discuss the primary purposes of the multiple regression being (a) description, (b) prediction, and (c) explanation, which may not be mutually exclusive.[2] Being able to effectively interpret, contribute to, critique, or use results of the research literature requires a fundamental understanding of multiple regression. We hope this chapter provides such a fundamental understanding of multiple regression.

ILLUSTRATIVE DATA

Throughout the chapter we will refer to a data set from Cassady and Holden (2012) consisting of 486 undergraduate students (304 females and 182 males) from a midsized Midwestern university. The sample was obtained from a psychology participant pool. The majority of the participants were majoring in teacher education. Two females did not report their age. The mean (standard deviation) of the age for the 302 females that reported their age was 20.7550 years (.2024) and for the 182 males was 21.3352 years (.2276). The data consist of measures of college academic performance, study skills, test anxiety (emotional and cognitive), and feelings of tests as threats, among others.

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 71–101.

College academic performance is operationalized by current college grade point average (GPA). Study skills are operationalized by the Study Skills and Habits (SS&H) scale (Cassady, 2004), which measures typical patterns of study behaviors and abilities. Emotional Test Anxiety (ETA) is operationalized by the Sarason Bodily Symptoms scale (taken from the Reactions To Tests Scale from Sarason, 1984), which measures physical responses to stress and is used as an indicator of the emotionality/affective component of test anxiety. Cognitive Test Anxiety (CTA) is operationalized by the CTA scale (Cassady & Johnson 2002), which measures distractibility, worry over tests, and self-deprecating ruminations during both test preparation and test performance. The feeling of tests as threats is operationalized by the Perceived Test Threat (PTT) scale (Cassady, 2004), which measures students' perceptions of a specific upcoming test as threatening to their academic or personal status. In addition, other variables were measured (e.g., age, race, SAT math and verbal scores), but we do not discuss them here because they are not the focus of our discussion of the multiple regression model (see Cassady, 2001, for related details).

The descriptive statistics for the full sample, which contain missing data, are given in Table 1. Table 2 contains the descriptive statistics for the data after listwise deletion was performed. Listwise deletion is when all individuals are deleted from the analysis when those individuals have any missing data on the relevant variables.[3] After listwise deletion considering the five variables, the sample size was reduced to 411. Table 3 shows the correlation matrix of the variables. In addition to the values of each of the correlations, the *p* value for the two-sided test of the null hypothesis is provided, along with an asterisk, which denotes statistical significance at the .05 level, or two asterisks, which denotes statistical significant at the .01 level.

In addition to the tabular summaries, Figure 1 is a scatterplot matrix, which shows a plot of the bivariate relationship between each pair of variables. Scatterplot matrices can be helpful to visually gauge the bivariate patterns in the data, such as the strength of linear and nonlinear relationships, and to check for possible outliers or miscoded data. Notice that the principal diagonal is blank because it represents the location where each variable would be plotted against itself. Also notice that

Table 1. Descriptive statistics for the observed data

	Descriptive Statistics				
	N	*Minimum*	*Maximum*	*Mean*	*Standard Deviation*
Current College GPA	466	1.80	4.00	3.1192	.48941
Study Skills and Habits	472	8	32	21.97	4.850
Emotional Test Anxiety	472	10	40	15.86	6.486
Cognitive Test Anxiety	458	17	68	35.11	10.828
Perceived Test Threat	464	27.00	81.00	48.6272	10.24374
Valid N (listwise)	411				

Table 2. Descriptive statistics after listwise deletion

	Descriptive Statistics After Listwise Deletion				
	N	*Minimum*	*Maximum*	*Mean*	*Standard Deviation*
Current College GPA	411	1.80	4.00	3.1447	.48873
Study Skills and Habits	411	8.00	32.00	21.94	4.928
Emotional Test Anxiety	411	10.00	40.00	16.01	6.593
Cognitive Test Anxiety	411	17.00	68.00	35.04	10.853
Perceived Test Threat	411	27.00	81.00	48.5328	10.32122
Valid N (listwise)	411				

*Table 3. Correlation table with the two-tailed significance level (*p*-value) for the correlation coefficient*

		Correlations				
		Current College GPA	*Study Skills and Habits*	*Emotional Test Anxiety*	*Cognitive Test Anxiety*	*Perceived Test Threat*
Current College GPA	Pearson Correlation	1	.186**	–.106*	–.301**	–.056
	Sig. (2-tailed)		.000	.031	.000	.256
	N	411	411	411	411	411
Study Skills and Habits	Pearson Correlation	.186**	1	–.293**	–.383**	–.270**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	411	411	411	411	411
Emotional Test Anxiety	Pearson Correlation	–.106*	–.293**	1	.719**	.329**
	Sig. (2-tailed)	.031	.000		.000	.000
	N	411	411	411	411	411
Cognitive Test Anxiety	Pearson Correlation	–.301**	–.383**	.719**	1	.469**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	411	411	411	411	411
Perceived Test Threat	Pearson Correlation	–.056	–.270**	.329**	.469**	1
	Sig. (2-tailed)	.256	.000	.000	.000	
	N	411	411	411	411	411

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

the plots below and above the principal diagonal are redundant, as they are the transposition of one another. Now that a description of the data has been provided, we begin our discussion of multiple regression model.

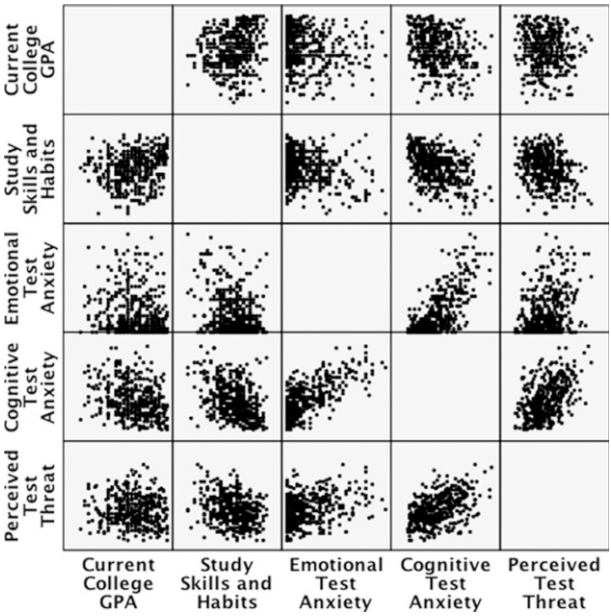


Figure 1. Scatterplot matrix showing the bivariate relationship of each of the variables. Note that the principal diagonal is blank because that represents the location where each variable would be plotted against itself. The plots above the principal diagonal are the transposition of the corresponding plots below the principal diagonal.

THE MULTIPLE REGRESSION MODEL

Multiple regression can be described as a general data analytic system due to its flexibility in handing different types of data and research questions (e.g., Cohen, 1968).

Multiple regression attempts to model the variation in an outcome variable as a linear function of a set of regressors. This process is accomplished through a linear equation that quantifies, via regression coefficients, the contribution of each regressor variable on the outcome variable.

The population multiple regression model linking the set of regressors to the outcome variable can be written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \varepsilon_i, \tag{1}$$

where Y_i is the observed value of the outcome variable for the ith individual ($i = 1, \ldots, N$), β_0 is the population value of the intercept, β_k is the population value of the regression coefficient for the kth regressor ($k = 1, \ldots, K$), and ε_i is the population value of the error for the ith individual. The error is the part of an individual's score

that cannot be accounted for by the particular regression model (i.e., the model with the K regressors). Notice that the multiple regression model is a linear model because Y_i is a sum of an intercept, K coefficients multiplied by the corresponding variables, and an error term.[4]

Estimated Regression Model Based on Data

Although the population regression coefficients (i.e., the β_k values) from Equation 1 are of interest, they are generally unknowable. However, the regression coefficients can be estimated based on data. The sample analog to Equation 1 is

$$Y_i = b_0 + b_1 X_{1i} + \cdots + b_K X_{Ki} + e_i, \tag{2}$$

where b_0 is the estimated intercept, b_K is the estimated regression coefficient for the kth regressor, and e_i is the error for the ith individual. The errors (i.e., the e_i values) are the difference between the model-implied value[5] of the outcome and the observed value of the outcome. The model-implied value, denoted $\hat{Y}_i$ for the ith individual, is obtained by using each individual's set of regressors in the estimated regression equation:

$$\hat{Y}_i = b_0 + b_1 X_{1i} + \cdots + b_K X_{Ki}. \tag{3}$$

The value of $\hat{Y}_i$ obtained by this equation is the model-implied conditional mean of the outcome variable for the particular combination of the ith individual's K regressors.

Using the illustrative data, we will model GPA (our outcome variable) as a linear function of SS&H, ETA, CTA, and PTT (our four regressors). The multiple regression equation that models GPA (i.e., the analog of Equation 3) for our example data is

$$\widehat{\text{GPA}}_i = b_0 + b_1 \text{SS\&H}_i + b_2 \text{ETA}_i + b_3 \text{CTA}_i + b_4 \text{PTT}_i. \tag{4}$$

The realized values of the regression equation or, in other words, the model with the regression coefficients that have been estimated, is

$$\widehat{\text{GPA}}_i = 3.135 + .01 \times \text{SS\&H}_i + .017 \times \text{ETA}_i - .022 \times \text{CTA}_i + .006 \times \text{PTT}_i. \tag{5}$$

We say more about this later, but we will note now that the intercept and the four regression coefficients are statistically significant. Three regressors have positive effects (SS&H, ETA, & PTT) and one regressor has a negative effect (namely, CTA). The regression coefficients should be used beyond simply saying there is a positive or a negative effect. The value of each regression coefficient conveys the expected change in GPA for a one-unit change in the corresponding regressor,

holding constant the other regressors. For example, the conditional mean of GPA is expected to increase by .01 for every unit increase of SS&H, holding everything else constant. The negative coefficient for CTA conveys that the conditional mean for GPA decreases .022 units for every unit increase of CTA. We will return to this regression model later.

In multiple regression, much concerns the errors. The error is formally defined as the difference between the observed value and the predicted value,

$$e_i = Y_i - \hat{Y}_i, \tag{6}$$

which is the difference between the observed value of the outcome variable and the model-implied value of the outcome variable. These errors are often termed residuals.

The way in which the regression coefficients are estimated in traditional multiple regression is with the least squares estimation procedure, which is why multiple regression is sometimes termed ordinary least squares regression or OLS regression. The least squares estimation method estimates coefficients such that the sum of the squared errors are minimized, that is,

$$\begin{aligned} min\left(\sum_{i=1}^{N} e_i^2\right) &= min\left(\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2\right) \\ &= min\left(\sum_{i=1}^{N} (Y_i - [b_0 + b_1 X_{1i} + \cdots + b_K X_{Ki}])^2\right). \end{aligned} \tag{7}$$

Thus, for a particular model, the estimated regression coefficients based on a particular set of data are those coefficients that minimize the sum of the squared errors, which is generally an advantageous method of estimation.[5] The variance of the error (residuals) is termed the error (residual) variance. In the regression model of GPA, the variance of the residuals is .209 (SD of errors is .457). The standard deviation of the errors plays an important role in null hypothesis significance testing and confidence interval formation by way of the standard error of the estimated regression coefficients.

Although the estimated regression coefficients provide the best point estimates of their corresponding population values, those estimates are fallible, meaning in almost all cases the estimated regression coefficients differ from the population values they estimate. This same issue arises with essentially any estimate of a population quantity. Correspondingly, users must always be aware that estimates have sampling error. An implication of sampling error is that, hypothetically speaking, if the same regression model were fitted using a random sample of the same size from the same population, the estimates would change for each of the random samples. Now, such replication is not generally done, but if it were, then the variability of the estimates could be easily seen. The degree of variability of the estimates is quantified by the standard error of each of the regression coefficients. The standard error of the regression coefficients

plays a central role in hypothesis testing and confidence interval formation, which we discuss formally later.

INTERPRETATION OF THE ESTIMATED REGRESSION MODEL COEFFICIENTS

The Intercept

The estimated intercept (i.e., b_0, which was 3.135 from our example data) is interpreted as the conditional mean of Y for a set of K regressors that are all zero. More formally, $\mathrm{E}[Y|(X_1 = \cdots X_K = 0)] = b_0$, where $E[\cdot]$ is the expectation of the expression in the brackets, with "|" representing a conditional statement. In other words, when all K regressors are zero, the best estimate for the outcome variable, or what would be predicted, is the intercept.

Depending on the particular situation, the intercept may or may not be a useful quantity from a practical perspective. In particular, it could be the case that (a) the set of K regressors can never all be zero, (b) there is no data consistent with the set of regressors all being zero, or (c) the intercept represents a value from an uninteresting research question. In such situations, the intercept is part of the model and serves as a scalar of the regression equation, but it may not, by itself, provide interpretational value of the phenomena under study. For example, from Table 2, the example data set does not have any regressors that have a value of zero (the minimum value for each regressor is above zero). Thus, in our model, the intercept represents a quantity that is outside the scope of our data. It is, however, our best estimate for an individual's GPA that has scores of zero for SS&H, ETA, CTA, and PTT, yet such a combination of regressors is absent from our data. Thus, the intercept has little direct interpretational value in this situation, though it does serve as an important scalar (in the sense that the intercept adjusts the regression equation such that the model-implied values reduce the squared error). So, in that sense, it is a necessary quantity yet it does not provide much interpretational value.

Although not necessary, it can be advantageous to rescale data so that the intercept has a more useful interpretation. For regressors that have a value added or subtracted (such as the mean), the value of the regression coefficients are left unchanged, only the intercept will change. One useful way to rescale regressors is to center each regressor. By centering, we mean that the data are put in deviation form. In other words, the mean of a regressor is subtracted from the individual values of the corresponding regressor. The deviations (i.e., the centered scores) are then used instead of the regressor itself in the regression model.

To illustrate centering on the example data, we return to the situation of modeling GPA from the four regressors, with the caveat that the four regressors have now been centered. Now, the model-implied regression equation is

$$\widehat{\mathrm{GPA}}_i = 3.145 + .01 \times \mathrm{ss\&h}_i + .017 \times \mathrm{eta}_i - .022 \times \mathrm{cta}_i + .006 \times \mathrm{ptt}_i, \qquad (8)$$

where lowercase letters are used for the regressors to denote that the regressors have been centered (note that GPA was not centered). Now, with centered regressors, the intercept changes from 3.135 to 3.145. Although in this particular instance the intercept did not change by much, centering *can* have a much more dramatic effect on the value of the intercept. While the intercept is still interpreted as the conditional mean of the outcome variable when all of the regressors are 0, the interpretation now of a regressor being 0 is when that regressor is at its mean. Thus, for the mean on all four regressors (and thus ss&h, eta, cta, and ptt=0), the conditional mean of GPA (i.e., the model-implied value) is 3.145. What may not be obvious initially is that if the original regression equation (i.e., with the intercept of 3.135) had been used and the values of the regressors were their respective means, the conditional mean of GPA would also be 3.145:

$$\begin{aligned}\widehat{\text{GPA}}_i &= 3.135 + .01 \times 21.94 + .017 \times 15.86 - .022 \times 35.11 + .006 \times 48.63 \\ &= 3.145, \end{aligned} \tag{9}$$

which is the same model-implied value of the intercept when the regressors were all centered. Thus, even though the intercept is different for the two models, the model-implied values can be recovered (regardless of the type of linear transformation performed) for equivalent model and data specifications.

For another example of rescaling to facilitate the interpretation of the intercept, suppose grade-level (Grade) for high school students as well as Sex are used as regressor variables in a multiple regression model. For the Grade variable, it would be perfectly fine to use 9, 10, 11, and 12 to represent freshmen, sophomores, juniors, and seniors, respectively. For the Sex variable, 0 could be used to represent female and 1 male. In such a situation, the intercept would not have a meaningful interpretation beyond that of a necessary scaling parameter in the model, because while one variable (Sex) could be zero, the other variable (Grade) could not be zero for the data at hand. One could argue that a value of 0 for grade-level would represent kindergarten, but that is an extreme extrapolation and nonsensical in most situations. However, it would be perfectly fine to scale Grade so that 0, 1, 2, and 3 represented freshmen, sophomore, junior, and senior, respectively. In such a case, the intercept would represent the model-implied (i.e., conditional mean) value for a female freshman (i.e., when all regressors are 0). Regardless of Grade being scaled as 9, 10, 11, and 12 or 0, 1, 2, and 3, the E[*Y*|(Sex, Grade)] would be the same for equivalent situations, as illustrated for the GPA example with and without centering. Thus, unlike the GPA example (in which rescaling was done by mean centering), here a different type of rescaling provided a more useful interpretation of the intercept (namely subtracting 9 from each regressor). Depending on the specific situation, if there is a desire to rescale regressors to make the intercept more interpretable, there will usually be multiple ways to proceed.

Regression Coefficients

In some situations, holding constant other variables is built into the design of the study by randomly assigning participants to groups, as in traditional applications of analysis of variance. In such situations, by randomly assigning participants to groups, in the population at least, it is known that there are no spurious variables that may be responsible for any differences that exist between the groups, other than the treatment(s).[6] However, when interest concerns the relationship between X_1 and Y, if random assignment to the level of X_1 was not done, any relation between X_1 and Y may be due to a some other variable, say X_2. However, by including both X_1 and X_2 in the multiple regression model, the effect of X_1 on Y can be evaluated while statistically holding constant X_2. This is an important aspect of the multiple regression model, as many times regressor variables are of interest but cannot be controlled by the researcher.

The interpretation of regression coefficients (i.e., the b_k values) is that they quantify the expected change of Y for a one-unit increase in X_k while controlling for the other $K - 1$ regressors. Controlling in this context refers to a statistical control in which the effect of one regressor is evaluated holding constant all other regressors, not a direct manipulation, which would be a control built into the study itself. Correspondingly, the regression coefficient can be thought of as the unique contribution a regressor has on the outcome variable. In other words, regression coefficients quantify the unique linear effect that each regressor has on the outcome variable while controlling for the other $K - 1$ regressors in the model. In this respect, the regression coefficients are technically partial regression coefficients.

For the example data, recall that the estimated regression equation is

$$\widehat{\text{GPA}}_i = 3.135 + .01 \times \text{SS\&H}_i + .017 \times \text{ETA}_i - .022 \times \text{CTA}_i + .006 \times \text{PTT}_i . \quad \text{(5, repeated)}$$

The value of .01 for SS&H is the estimated impact on GPA of a one-unit increase in SS&H, controlling for (holding constant) ETA, CTA, and PTT. The idea of controlling (holding constant) other variables when interpreting a regression coefficient has a precise statistical meaning and is not intended to imply that the researcher has, or even could, manipulate (i.e., directly control) the level of a particular regressor of an individual (e.g., the level of emotional text anxiety).

Regression coefficients are scaled in terms of both the outcome variable as well as the particular regressor variable. Provided that the rescaling is in the form of a linear transformation, the value of regression coefficients can be easily converted from the original unit into the new units (e.g., standardized units). The regression coefficients are a type of effect size because they convey the magnitude of effect that each regressor has on the outcome variable while holding constant the remaining regressors.

MODEL FIT STATISTICS

In assessing the overall fit of the model, the most common way is to consider the squared multiple correlation coefficient, denoted R^2 for the sample value, which is often termed the *coefficient of determination*. The squared multiple correlation coefficient quantifies the proportion of the variance in the outcome variable that can be explained by the set of regressors. Said another way, the variance of the outcome variable can be partitioned into that which can be accounted for and that which cannot be accounted for by the particular model in a particular data set.

An estimate of the population squared multiple correlation coefficient is the ratio of the sum of squares due to the regression model to the total sum of squares as

$$R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}} = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}. \tag{10}$$

This ratio conveys the amount of the variance in Y that can be accounted for by the model to the overall amount of variance in Y. Thus, Equation 10 is a ratio of variances. As such, the proportion of the variance of Y (the denominator of Equation 10) that can be accounted for by the model (the numerator of Equation 10) is a useful way of quantifying a model's effectiveness. A different way of conceptualizing R^2 is that it represents the squared correlation between Y and the model-implied values of Y:

$$R^2 = (r_{Y\hat{Y}})^2. \tag{11}$$

Correspondingly, when $R = r_{Y\hat{Y}} = 0$, it signifies a perfect lack of linear association between Y and $\hat{Y}$, whereas when $R = r_{Y\hat{Y}} = 1$, it signifies a perfect linear association between Y and $\hat{Y}$.

Although R^2 is the correlation between Y and $\hat{Y}$ for the sample values, R^2 as an estimator of the population squared multiple correlation coefficient is positively biased. A better estimate, namely one that is more unbiased, is the adjusted squared multiple correlation coefficient, which is given as

$$R^2_{Adj} = 1 - (1 - R^2)\frac{N-1}{N-K-1}. \tag{12}$$

This adjustment to the squared multiple correlation coefficient corrects for inflation due to sample size and number of predictors included in the model. In large samples with a moderate number of predictors, the adjusted and unadjusted squared multiple correlation coefficients will be very similar. However, in small samples or with large numbers of regressors, the adjusted squared multiple correlation coefficient can decrease substantially.

For the example data, in which GPA is modeled with SS&H, ETA, CTA, and PTT, $R^2 = .133$. Thus, for the particular data, 13.3% of the variance in GPA was accounted for by the four regressors. However, $R^2 = .133$ is a positively biased estimate of the proportion of variance accounted for in the population. The adjusted value R^2 is $R^2_{Adj} = .125$. Thus, in the population, 12.5% of the variance being accounted for by the four regressors is a better estimate, in the sense that it is (nearly) unbiased.

INFERENCE IN REGRESSION

In general, data are collected to make inferences about what is true in a population. For example, in our data set, we are not literally interested in the 411 participants who took part in the study, but rather in how those 411 participants allow us to make inferences about their corresponding population. Hypothesis tests and confidence intervals are inferential procedures because they use sample data to draw conclusions (i.e., make inferences) about what is true in the population.

Inference for the Squared Multiple Correlation Coefficient

In order to evaluate if the model has accounted for more variance in the outcome variable than would be expected by chance alone, a null hypothesis significance test of the squared multiple correlation coefficient can be performed. The specific test is an *F*-test and it tests an omnibus (i.e., overarching) effect size that evaluates that all of the K regressors are 0 in the population. That is, the *F*-test tests the null hypothesis that $\beta_1 = \cdots = \beta_K = 0,$ which is equivalent to the population squared multiple correlation coefficient is 0. The *F*-test is similar to the *F*-test in an analysis of variance because it evaluates the amount of variance accounted for by the regressors to the amount of unaccounted for variance. In fact, this test is a type of analysis of variance in that the ratio of variances is examined (namely the variance due to regression model is compared to the variance due to the error). The *F*-test for the overall model fit is given as

$$F = \frac{MS_{\text{Regression}}}{MS_{\text{Residual}}} = \frac{R^2/K}{(1-R^2)/(N-K-1)}, \tag{13}$$

with K and $N - K - 1$ degrees of freedom and *MS* denoting *mean square*. Under the multiple regression assumptions discussed, when this test is statistically significant (i.e., when the *p*-value is less than the Type I error rate, usually .05), the null hypothesis that the population squared multiple correlation coefficient is 0 is rejected.

For the example data, the value of the *F*-statistic is 15.633 with 4 and 406 degrees of freedom. The corresponding *p*-value is <.001. Thus, the null hypothesis can be rejected and the claim made that the model is able to account for more variance in the outcome variable than would be expected from chance alone. That is, the value

of the observed R^2 of .133 would be exceedingly unlikely if the population squared multiple correlation coefficient were in fact 0.

Confidence intervals for the population squared multiple correlation coefficient are useful, but complicated to obtain. At present, there is no way to obtain these confidence intervals using SPSS point-and-click menu options. There is no closed form solution for such confidence intervals, but they can be obtained easily with the MBESS *R* package (Kelley & Lai, 2010) among other programs with specialized scripts (e.g., see Kelley, 2008; Algina & Olejnik, 2000; Mendoza & Stafford, 2001). In addition to the assumptions for inference previously discussed, forming confidence intervals for the population squared multiple correlation coefficient assumes multivariate normality, which is a much more stringent assumption than normality of errors. Multivariate normality implies that the K regressors and the outcome variable have a $K + 1$ dimensional multivariate normal distribution in the population.

For the example data, the 95% confidence interval for the population squared multiple correlation coefficient is [.07, .19]. Although the observed value of the squared multiple correlation coefficient is .133, the population value could conceivably be as low as .07 or as high as .19, with 95% confidence. This confidence interval assumes that the regressors are random, which is generally the case in application of the multiple regression model in empirical research, although other methods for fixed regressors also exist.[7]

Inference for the Intercept and Regression Coefficients

In order to evaluate the individual regressors uniquely contribute to the modeling of the outcome, a null hypothesis significance test of the regression coefficients can be performed. Under the multiple regression assumptions discussed, when the null hypothesis is true, a regression coefficient divided by its standard error follows a t-distribution with $N - K - 1$ degrees of freedom. Correspondingly, p-values can be determined to test the null hypothesis that the population value of the regression coefficient is some specified value, such as 0.

The t-test for testing the kth regression coefficient is

$$t = \frac{b_k - \beta_{k_0}}{s_{b_k}}, \tag{14}$$

where β_{k_0} is the specified null value for the kth population regression coefficient with $N - K - 1$ degrees of freedom. Most often, $\beta_{k_0} = 0$, which then leads to the simpler and more common way of writing the t-test:

$$t = \frac{b_k}{s_{b_k}}. \tag{15}$$

In the example data, the regression coefficients, as previously noted, are .01, .017, –.022, and .006 for SS&H, ETA, CTA, and PTT, respectively, each with 406 degrees

of freedom. The standard errors for the regression coefficients are .005, .005, .003, and .002, respectively. In each case, interest concerns evaluating the regression coefficient against a population value of 0. This is the case because we are interested in detecting a linear relationship between each of the regressors and the outcome variable, while holding each of the other regressors constant. Thus, the *t*-statistics for the four regression coefficients are 2.00, 3.50, −6.70, and 2.34, respectively. Each of these *t*-statistics is statistically significant at the .05 level, with two-sided *p*-values of .046, .001, <.001, and .02, respectively.

In addition to the null hypotheses that each of the regression coefficients equals zero, which provides a directionality of the relationship, estimating the size of the contribution each regressor has on the outcome variable is important. That is, we seek to understand the degree to which each regressor has an impact on the outcome. Although the null hypothesis was rejected for each regression coefficient, confidence intervals are important in order to convey the uncertainty of the estimate with regards to the plausible values of the population parameter. Two-sided confidence intervals for regression coefficients are formally given by

$$probability[b_k - t_{(1-a/2,N-K-1)}s_{b_k} \leq \beta_k \leq b_k + t_{(1-a/2,N-K-1)}s_{b_k}] = 1 - a, \qquad (16)$$

where $t_{(1-a/2,N-K-1)}$ is the critical value. Alternatively, the confidence interval limits can be written as

$$b_k \pm t_{(1-a/2,N-K-1)}s_{b_k},$$

which is simply the estimate plus or minus the margin of error (the margin of error is $t_{(1-a/2,N-K-1)}s_{b_k}$). This provides a less formal way of conveying the confidence interval limits.

For the example data, the confidence intervals for the population regression coefficients for SS&H, ETA, CTA, and PTT are [.0002, .020], [.008, .027], [−.028, −.015], and [.001, .011], respectively. Confidence intervals for the intercept and regression coefficients are available from SPSS via the "Statistics" option in the linear regression analysis.[8]

ASSUMPTIONS FOR INFERENCE IN MULTIPLE REGRESSION

The estimation methods discussed above, namely least squares estimation, do not, in and of themselves, depend on assumptions. However, like all other statistical procedures, inference in multiple regression is based on a set of assumptions about the population from which the sample was collected. By inference, we mean the null hypothesis significance tests and confidence intervals for the squared multiple correlation coefficient, the intercept, and the *K* regression coefficients.

Inference for the regression coefficients depends on four assumptions: (a) linearity, (b) normality of errors, (c) homoscedasticity, and (d) independence.

The linearity assumption is that each of the *K* regressors is linearly related to the outcome variable. The assumption of linearity between the outcome variable and the regressors is arguably the "most important mathematical assumption of the regression model" (Gelman & Hill, 2007, p. 46). If the linearity assumption does not hold, then using a linear model is necessarily a flawed way of modeling the relationship of the outcome with the regressors. Correspondingly, nonlinear models may be more appropriate when linearity does not hold. For example, a learning curve is typically sigmoidal ("S" shaped) in nature. A nonlinear regression model with a sigmoidal functional form (e.g., asymptotic regression, Gompertz, or logistic curves) may be more appropriate than a multiple regression model (e.g., Seber & Wild, 1989). Alternatively, as previously noted (footnote 4), functions of the regressors can be used, rather than the values of the regressors themselves, as a way to satisfy the linear assumption of regression. However, in many situations, transformations are difficult to interpret and provide a poor substitute for an inherently nonlinear model.

The normality of errors assumption means that the distribution of the e_i values follows a normal distribution. When inference concerns regression coefficients, this normality assumption is for the errors only, not the distribution of regressors. However, confidence intervals for the population squared multiple correlation coefficient, at least for the most common approach to confidence interval formation, requires multivariate normality among the regressor and outcome variables.

The homoscedasticity assumption is that the conditional variance of the outcome variable for any combination of the regressors is the same in the population. The reason that this assumption is necessary for inference in the case of least squares regression is because there is only a single error term that estimates the population error variance. If the population error variance depends on/changes with the particular set of regressors, then using only a single value to estimate the population error variance would be problematic. In our example data, the estimated error variance is .209. Thus, homoscedasticity implies that the variance of the errors, regardless of the combination of regressors, is .209.

The independence assumption is that the unit of analysis (i.e., whatever the *i* represents in the multiple regression equation, such as individuals, schools, or students) are all independent of one another. That is, the independence assumption stipulates that there is no correlation among any subset, or clustering, of the units of analysis. This is best handled with an appropriately designed study. For example, if schools are the unit of analysis, having multiple schools from the same school district/corporation (if there are multiple districts/corporations) would be problematic, as schools within the same district/corporation would tend to be more similar than schools from different districts/corporations due to the common effects of the district/corporation. When dependencies are built into the data by a common grouping/nesting structure, such as multiple students from the same class in which

multiple classes are included, other models, such as hierarchical linear models (HLM), may be more appropriate (See Osborne & Neupert, this volume).

Checking the Assumptions

The assumptions of linearity, normality of errors, and homoscedasticity are generally assessed by graphical means, but more formal assessments can also be made. In graphical assessment of the linearity assumption, scatterplot matrices (such as Figure 1) can be useful in order to assess if the relationship among variables seems linear. Cohen et al. (2003, chapter 4) recommend scatterplots of the residuals from the regression model of interest plotted against each regressor variable and against the model-implied outcome values, along with lowess regression lines (e.g., at 0, −1, and 1 standard deviation from the mean residual). Lowess regression, and thus a lowess regression line, is a nonparametric approach to obtaining a smooth regression line that does not presuppose that relationships between variables are linear. Thus, if the lowess regression line differs to a non-trivial degree from a horizontal line (recall it is the residuals that are being plotted, not the outcome values themselves), then there may be cause for concern that a linear model is not appropriate and adjustments to

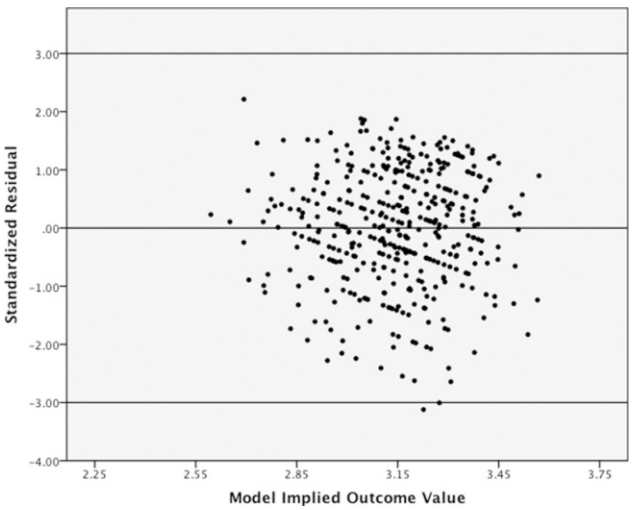


Figure 2. Scatterplot of standardized residuals plotted against the model-implied outcome values for the model in which GPA is modeled from study skills and habits, emotional test anxiety, cognitive test anxiety, and perceived test threat.
Note that the three horizontal lines are at the mean (which is 0), 3 standard deviations above the mean, and 3 standard deviations below the mean for the standardized residual. The reason the horizontal lines are provided at 3 standard deviations above and below the mean is to help identify possible outliers.

the model should be considered. Additionally, a scatterplot of standardized residuals plotted against the model-implied outcome values can be useful. Figure 2 provides such a plot (due to space limitations, we do not provide plots of the residuals against each regressor and the model-implied outcome values). In such plots, obvious patterns would be a concern because there may be important variables missing from the model (or homoscedasticity does not hold, which is an assumption we discuss in a moment).

The normality of errors assumption involves assessing the normality of a variable, residuals in our case, in the way one would typically evaluate normality. Regarding visual approaches, we recommend assessing normality with a normal Q-Q (quantile-quantile) or P-P (percentile-percentile) plot, which is a plot of the expected cumulative quantiles/probabilities of the residuals given they are normally distributed against the observed cumulative quantiles/probabilities of the residuals. If the points do not differ in a non-trivial way from the equiangular line (i.e., the line of slope 1), then the assumption of normality of the residuals may be satisfied. Figure 3 provides a P-P plot of the residuals. Formal assessment with statistical tests or by testing the skew and kurtosis are also possible.

The homoscedasticity assumption implies that the variance of the errors is the same across all model-implied values and across all values of the regressor variables. From the same plots for assessing linearity discussed above (e.g., Figure 2), the residuals should not differ from a rectangular shape if, in fact, homoscedasticity holds. For example, if residuals were small for small values of X_1 but began to spread as X_1 increased, a violation of the homoscedasticity assumption may have occured.

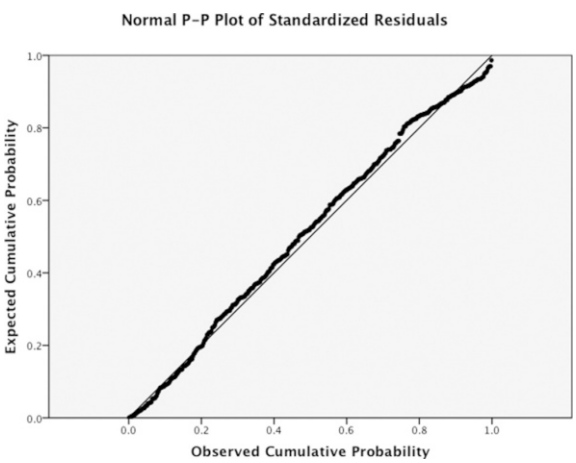


Figure 3. Normal probability-probability plot (P-P Plot) of the residuals for the model Residuals for the model in which GPA is modeled from study skills and habits, emotional test anxiety, cognitive test anxiety, and perceived test threat.

More formal assessment with statistical tests for the homoscedasticity assumption also exist, but visual methods can be very enlightening.

The fourth assumption of independence has two conditions that need to be considered: nested data structures and autocorrelation. When the units are not a simple random sample from a population, but rather are collected based on clusters of individuals, the observations will not likely be independent. Generally, issues of nonindependent observations are best considered at the design stage of the research. As mentioned previously, if a clustering/nesting structure exists, more appropriate models such as HLM can be used. For example, if multiple classrooms with multiple students are included in a study, the students in the same classrooms will tend to be more similar than students from different classrooms. When such a nesting structure (students within classrooms) is part of the design, it should be explicitly dealt with as part of the model. Failure to consider such a nested data structure results in violations of the independent observation assumption, and the multiple regression model is not robust to violations.

In addition to nesting type structures, correlation among residuals is also a violation of the independence assumption. For example, if residuals for values of X_1 that are close together are more similar than residuals for values of X_1 that are farther apart, such a situation would illustrate serially correlated errors. For example, when X_1 represents time, time values close together will tend to have outcome variables that are more similar than if the time values were farther apart, which tends to produces adjacent errors that are more similar than if errors were random. Such a situation would then likely involve errors that have an autocorrelation. The Durbin-Watson statistic is a statistical procedure that measures the degree of correlation of residuals with the immediately preceding residuals (1st order autoregressive correlation). The Durbin Watson statistic ranges from 0–4, with values at 2 indicating perfect lack of first order autocorrelation. Values near 2 are thus not considered problematic, but as the values move close to 0 or 4, evidence of 1st order autocorrelation exists. In our example data, the Durbin-Watson statistic is 1.804. Estimated critical values for the Durbin-Watson statistic are discussed in more technical works on regression and time series.

In addition to the assumptions we have discussed, an issue of concern is the measurement of the regressors used in the regression model. In particular, it is ideal for the regressors to be measured without much error. Some sources state regressors being measured without error as an assumption of inference in multiple regression. We do not regard regressors being measured without error as an assumption per se, but results obtained using regressors measured with error may differ substantially from results obtained when regressors are measured without error. That is, there will be a bias in the estimated regression coefficients, standard errors, and model fit statistics (e.g., R^2) when regressors are measured with error. Correspondingly, measurement error in the regressors is an important consideration. Of course, the less measurement error, the better the conclusions. When a nontrivial amount of error exists in the regressors, latent variable models (e.g., confirmatory factor and structural equation models) should be considered (e.g., Mulaik, 2009). Such models

require multiple measures of the same construct (e.g., well-being, motivation, conscientiousness) rather than a single measure as typically included in regression (e.g., either based on a single measure or a composite score of multiple measures).

EXAMPLE IN SPSS

To solidify the information just presented in the previous section on the results from the example data, we now show output from the SPSS linear regression procedure in Table 4. Table 4 consists of three types of output: (a) Model Summary, (b) ANOVA, and (c) Coefficients.

The linear regression procedure is available from the point-and-click SPSS interface via the *Analyze* menu on the toolbar. From the *Analyze* menu, the *Regression* menu is selected, and then the *Linear* procedure is selected. Within the *Linear Regression* procedure, the outcome variable of interest is selected as *Dependent* and the regressors of interest are selected as *Independent(s)*. Additional information and output are available in the *Statistics, Plots, Save,* and *Options* menu buttons. Figures 1–3 were created using the *Plots* options. We chose SPSS to illustrate the multiple regression model we have been discussing because it seems to be the most widely used software package in behavioral, educational, and social science research.

Table 4a. Summary of overall multiple regression model fit

Model Summary[b]				
Model	*R*	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*
1	.365[a]	.133	.125	.45718

[a] Predictors: (Constant), Perceived Test Threat, Study Skills and Habits, Emotional Test Anxiety, Cognitive Test Anxiety

[b] Dependent Variable: Current College GPA

Table 4b. Analysis of variance source table testing the overall fit of the model to infer if the collection of regressors accounts for a statistically significant amount of variation in the dependent variable (college GPA)

ANOVA[b]						
Model	*Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*	
1	Regression	13.070	4	3.267	15.633	.000[a]
	Residual	84.860	406	.209		
	Total	97.929	410			

[a] Predictors: (Constant), Perceived Test Threat, Study Skills and Habits, Emotional Test Anxiety, Cognitive Test Anxiety

[b] Dependent Variable: Current College GPA

Table 4c. Estimated regression coefficients, tests of statistical significance, and confidence intervals for the fitted multiple regression model

Coefficients[a]							
	Unstandardized Coefficients		*Standardized Coefficients*			*95.0% Confidence Interval for B*	
Model	*B*	*Std. Error*	*Beta*	*t*	*Sig.*	*Lower Bound*	*Upper Bound*
1 (Constant)	3.135	.186		16.821	.000	2.768	3.501
Study Skills and Habits	.010	.005	.101	2.002	.046	.000	.020
Emotional Test Anxiety	.017	.005	.233	3.501	.001	.008	.027
Cognitive Test Anxiety	–.022	.003	–.487	–6.702	.000	–.028	–.015
Perceived Test Threat	.006	.002	.123	2.336	.020	.001	.011

[a] Dependent Variable: Current College GPA

EXTENSIONS OF THE BASIC MULTIPLE REGRESSION MODEL

We have presented the basic multiple regression model. However, there are many extensions and special uses of the multiple regression model in applied research in the behavioral, educational, and social sciences that we would be remiss if we did not discuss. We discuss six important extensions and special uses of the multiple regression model in the subsections that follow. However, due to space limitations, we can only briefly discuss each of these six extensions and special uses. Thus, our treatment is necessarily limited and additional sources should be consulted before using the extensions and special uses of the multiple regression model.

Moderation Models

The basic multiple regression model as presented is additive because in that each of the regressors enters the model as a main effect only. This implies that the effect of a regressor on the outcome variable does not change at different levels of other regressors. When additivity does not hold, a moderation model may be appropriate. A moderation model is one in which there are one or more interaction terms in the regression model (in addition to the main effects). An interaction term is a regressor that is the product of two (or more) other regressors. Such a model allows not only for effects to be additive, but also to be multiplicative. The following equation shows a moderated multiple regression model for two regressors (i.e., a multiple regression model with an interaction):

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{1i} X_{2i}. \tag{17}$$

Moderators are important in many areas because they answer questions about how the level of one variable *moderates* the relationship between another regressor and the outcome of interest. Moderations are realized via interactions, which are multiplicative effects. This means that the model is more complex than additive terms, but that multiplicative terms are also necessary. Interactions in multiple regression have the same interpretation as in factorial ANOVAs. However, whereas factorial ANOVAs can only incorporate categorical independent variables, multiple regression may include interactions between categorical as well as continuous independent variables. Interactions between any number of independent variables can be incorporated into a model. However, the interpretation of interactions involving more than two regressors can be difficult.

Interpreting the results of a moderated regression model (i.e., a model with one or more interactions) is more involved than interpreting an additive model, such as those previously discussed. In particular, by the very definition of an interaction, the main effects can no longer be interpreted as having a constant effect on the outcome variable. Rather, the main effect of X_1 (i.e., b_1 in Equation 17) provides an unambiguous interpretation itself at only one value of X_2, namely, when X_2 is 0. When X_2 is 0, the values of b_2 and b_3 are not a concern because they cancel from the equation; b_1 is then the slope of the effect of X_1 on Y. As explained in Cohen et al., "in general, in a regression equation containing an interaction, the first-order regression coefficient [i. e., the main effect] for each predictor involved in the interaction represents the regression of Y on that predictor, *only at the value of zero on all other individual predictors with which the predictor interacts*" (2003, p. 260).

There is increased complexity when interpreting a regression model that contains one or more interactions. In many situations, the interpretation of the regression model can be improved by using centered regressors. Centered regressors set the zero value of the regressors to their respective means. Thus, the main effect of X_1 is interpreted at the mean of X_2, which is the now zero due to centering. Additionally, because of the increased complexity in interpreting regression models with interactions, it can oftentimes be beneficial to plot the model-implied regression equations for selected combinations of regressors. In particular, the model-implied relationship between Y and X_1 at the mean of X_2, one standard deviation above the mean of X_2, and one standard deviation below the mean of X_2 can be plotted to visualize the effect of an interaction. (e.g., see Aiken & West, 1991, for details).

Mediation Models

Mediation models are important in the context of causal modeling because they attempt to disentangle the causal pathways of how one (or more) variables cause one (or more) other variables, which in turn cause one (or more) other variables. For example, it might be theorized that X_1 causes Y, but it does so through X_2. That is, X_1 causes X_2 and then X_2 causes Y. There may be what is termed "complete mediation,"

when the entire effect of X_1 on Y is explained through X_2, or what is termed "partial mediation," when there is some effect of X_1 on Y above that which goes through X_2. The notions of complete (or full) mediation and partial mediation, although widely used, are qualitative descriptions of what is inherently a quantitative process. Preacher and Kelley (2011) review such issues and discuss various effect sizes in the context of mediation analysis.

A widely used framework for showing support for mediation uses regression and is known as the "causal steps approach," which was popularized by Baron and Kenny (1986; see also Judd & Kenny, 1981). This framework can be interpreted as consisting of four conditions that must hold in order to support a mediation hypothesis. These four conditions are as follows:

The exogenous regressor (i.e., the independent variable) must be related to the outcome variable. This condition, in the population, requires β_1^* in

$$Y_i = \beta_0^* + \beta_1^* X_{1i} + \varepsilon_i^* \tag{18}$$

to be nonzero.

The exogenous regressor must be related to the endogenous regressor (i.e., the mediating variable). This condition, in the population, requires β_1^{**} in

$$X_{2i} = \beta_0^{**} + \beta_1^{**} X_{1i} + \varepsilon_i^{**} \tag{19}$$

to be nonzero. Note that we use asterisks to distinguish the parameter values from Equations 18 and 19 (above) from Equation 20 (below).

The endogenous regressor must be related to the outcome variable (i.e., the dependent variable) after controlling for the exogenous regressor. This condition, in the population, requires that β_2 in the equation below to be nonzero

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i. \tag{20}$$

When both the regressor and the mediating variable are used simultaneously to predict the outcome variable, the impact of the regressor is reduced. Conditions 1–3 can be evaluated with correlation coefficients or regression analysis (we have presented them in the regression form), but Condition 4 can be evaluated with the use of regression analysis or a more general procedure, such as path analysis or structural equation modeling.

An alternative conceptualization of mediation is that the product of β_1 and β_2 from Equation 20 does not equal zero (i.e., mediation holds if $\beta_1 \times \beta_2 \neq 0$). The β_1 and β_2 regression coefficients are equivalent to the causal paths from the independent variable to the mediator to the dependent variable from a path analytic, or generalized via a structural equation model, framework. Thus, if mediation holds, the causal path must be non-zero (e.g., MacKinnon et al., 2002).

Implicit in the mediation framework is the passage of time. That is, if X_1 causes X_2, X_1 must precede X_2 in time. Similarly, if X_2 causes Y, X_2 must precede Y in time. Issues of simultaneous causality, while theoretically possible, may not be reasonable in many situations. Cole and Maxwell (2003) discuss mediation models in the context of longitudinal data (see also Gallob & Reichardt, 1985). We recommend interested readers consult MacKinnon (2008), who discusses mediation models and the underlying assumptions in detail.

Hierarchical Regression

Traditional applications of the multiple regression model examine the contributions of regressors simultaneously. In other words, all variables are considered at the same time. However, it can be advantageous to examine sets of regressors in a prespecified sequence or in a defined priority order. The order that the regressors enter the model should be theoretically driven. When regressors are added, the model is referred to as *hierarchical regression.*[9] Hierarchical regression is a model comparison approach in which richer models (i.e., with more regressors) are compared to simpler models. Such comparison is used to infer if additional regressors account for a statistically significant amount of the variance of the outcome variable that was previously unexplained. In particular, the change in R^2 from model 1 (a simpler model with q regressors) to model 2 (richer model with $q + r$ regressors) is tested to infer if the model with $q + r$ regressors has a larger population squared multiple correlation coefficient than the model with only q regressors. The number of "blocks" of variances that enter into a model depends on the number of regressors available. The way in which such models are tested is with the following F-statistic,

$$F = \frac{(R^2_{q+r} - R^2_q)/r}{(1 - R^2_{q+r})/(N - q - r - 1)}, \tag{21}$$

where the numerator and denominator degrees of freedom are r and $N - q - r - 1$, respectively.

In hierarchical regression models, there are often regressors that researchers would like to control for before assessing the effects of the regressors of primary importance. These regressors are used for controls (i.e., control variables) and are often not of theoretical importance, but rather are important to control for as they may explain a large portion of variance. A common approach to this situation is to include these variables in the first block of a hierarchical regression analysis. For example, a researcher may choose to include demographic characteristics (e.g., Sex, Age, SES) in block one of a sequential hierarchical model. The R^2 for this model will provide the variance accounted for by the collection of demographic regressor variables. Then, as subsequent variables of theoretical interest or blocks of regressors are added to the model, the change in R^2 will provide information on how

much variance the regressors account for–and this is key–above what the control variables accounted for.

Stepwise Regression

Stepwise regression is a procedure in which a variety of algorithms can be used to mechanically select which regressors (potentially many) should be included in a model based on statistical, not theoretical, criteria. Stepwise regression can proceed from forward selection (fewer to more regressors) or backwards selection (more to fewer regressors) methods. When there are more than a few available regressors, the number of models fitted by stepwise procedures can be large. A large number of fitted models can have false positives due to the sheer number of models fitted in order to arrive at a final model. That is, in the final model from the stepwise procedure, there will be a higher rate of false positives than in a prespecified model.

One way forward selection may begin is by entering the regressor into the regression model with the strongest correlation with the outcome variable. Then, the second variable is entered into the model that has the biggest impact on the model (e.g., highest change in R^2), and so on. This process will continue until the addition of new regressors does not add enough to the variance accounted for (e.g., a statistically significant change) in the outcome variable.

In contrast to forward selection, backward selection may begin with all regressors included in the model. Then the regressor that has the least impact on the model (e.g., smallest change in R^2) is removed, an so on. This process can continue until the removal of regressors impacts the variance accounted for (e.g., a statistically significant decrease) in the outcome variable.

Stepwise regression is completely mechanical/machine driven. Stepwise regression is thus a completely atheoretical way of modeling the relationship between an outcome variable and a set of regressors. When working from theory, stepwise regression is not recommended. If research is completely exploratory, stepwise regression may shed some light on regressor variables that may be effective at modeling the outcome variable. However, we generally recommend against stepwise regression because, in the vast majority of situations, there is some theory available to suggest what variables are, for example, best used as control variables versus those that are more theoretically interesting.

Categorical Regressors

Although the dependent variable for a multiple regression model needs to be continuous, or nearly so, the regressors can be continuous or categorical. However, categorical variables must be treated differently than continuous variables when entered into a regression model. In order to use a categorical regressor in a regression analysis, we generally recommend a process called *dummy coding*.[10] Dummy coding represents the different levels of a categorical variable (group

membership) as 1 or 0, depending on whether or not the variable represents a particular group. This procedure codes the levels of a categorical variable with J levels into $J - 1$ dummy variables. One level, denoted the reference level, is represented by all of the other levels being 0. When a categorical variable only has two levels, only one dummy coded variable is necessary. For example, the Cassady data includes the participants' sex, in which the variable Sex is 1 for male and 0 for female. Correspondingly, female is the reference group. The value of the regression coefficient, holding everything else constant, represents the difference in the conditional mean of the outcome variable for males.

When a categorical variable has more than two levels, multiple dummy variables can be used to include the categorical information into a regression analysis. For example, the self-identified race of the participants is included in the example data file. Initially, the variable Race was coded as a single variable in which 1 = Caucasian, 2 = African-American, 3 = Asian, and 4 = Other. However, such a coding scheme would not be used in applications of multiple regression, as the numbers are not meaningful; they simply represent a category rather than any sort of continuum. However, the variables can easily be recoded into 3 dummy codes (recall $J - 1$ dummy codes are needed, which is why, with four levels, only 3 dummy codes are necessary.). Any of the four levels of Race can be used as the reference category, but we use Caucasian as the reference category because it represents the majority in this sample. We form three dummy coded variables in which AA represents African-American, Asian represents Asian, and Other represents a self-identified other categorization. Thus, for a participant who has a 0 for each of the three race variables, that participant would be Caucasian.

When interpreting dummy variables, we can learn several different pieces of information. First, we can infer if the conditional mean for the outcome variable in the population differs, holding everything else in the model constant, for the particular group as compared to the reference group (in this case Caucasian). Such an inference is made by the p-value from the corresponding null significance test. If there are no other variables in the model, then the t-value obtained for the test of two-levels of the variable (i.e., in which only a single dummy variable is necessary) is exactly equal to that obtained in the context of a two independent groups t-test. Second, an estimate of the conditional mean difference, holding everything else in the model constant, is available by way of the regression coefficient. Third, the confidence interval for the population regression coefficient of the particular group provides the range of plausible parameter values. The wider this confidence interval, the more uncertainty there is of the population conditional mean difference on the outcome variable.

Cross-Validation

Often in studies that use multiple regression, especially when prediction is of interest, it is advantageous to provide evidence of the effectiveness of the obtained

multiple regression model estimates from one sample as they would apply to another. As mentioned previously, it is generally not possible to obtain the true population regression values. In an application of a multiple regression model, the estimated regression coefficients are idiosyncratic to the characteristics of the particular sample drawn from the population. Correspondingly, estimating how well those values would predict in a future sample can be very useful.

As discussed, R^2 is the squared correlation between Y and $\hat{Y}$. Let Y^* be the observed values of the outcome variable for a second sample and $\hat{Y}^*$ be the model-implied (i.e. predicted) values of the outcome variable from the second sample when the estimated regression coefficients from the first sample are applied to the regressors from the second sample. The squared correlation between $\hat{Y}$ and $\hat{Y}^*$ can be calculated, denoted $R^2_{Y^*\hat{Y}^*}$. This value of $R^2_{Y^*\hat{Y}^*}$ will tend to be smaller than R^2, and is often termed the *shrunken* R^2. The difference between R^2 and $R^2_{Y^*\hat{Y}^*}$ is known, therefore, as shrinkage. If the shrinkage is small, then evidence suggests that the first regression equation obtained in the first sample cross-validates well in future samples. However, if the shrinkage is large, then there is not strong evidence that the model obtained in the first sample will be good at predicting values of the outcome variable. Pedhazur recommends that when the shrinkage is small, the samples be combined and the regression coefficients estimated as a way to improve prediction in future samples (1997; see also Mosier, 1951). Darlington discusses several methods of estimating the shrunken R^2 from single samples (1990).

SPECIAL CASES AND EXTENSIONS OF THE MULTIPLE REGRESSION MODEL

The multiple regression model is a special case of the general linear model. In its most general form, the general linear model allows multiple continuous outcome variables to be modeled from multiple regressor variables. These regressor variables might be grouping variables or continuous variables from either observational work or randomized experiments, and any combination thereof. Correspondingly, some general linear model special cases can be conceptualized as a multiple regression model (e.g., a correlation, single sample, paired sample, and independent-samples *t*-test or an analysis of (co)variance). The multiple regression model extends to other statistical models that have multiple regression as a special case (e.g., path analysis, confirmatory factor analysis, structural equation modeling, discriminant function analysis, canonical correlation, and multivariate analysis of (co)variance). The multiple regression model can also be extended to situations in which there are nesting structures, such as students nested within classrooms (with HLM/multilevel modeling).

In addition, generalizations of the general linear model to situations of categorical and limited outcome variables are termed *generalized linear model*. Generalized linear models use the exponential family of distributions (e.g., logistic, probit, tobit, Poisson) to link a function of a linear model to the outcome variable. For example, the proportion of 3rd grade students who pass a state-wide assessment within different

schools in the same district/corporation has a limited dependent variable that has a range of 0 to 1. A linear model may have model-implied values of the proportion of students passing outside the range of 0 to 1 that require homoscedasticity, which would not be reasonable in general. A generalized linear model with a logistic link function would provide a more appropriate way to model such data.

Connecting multiple regression models to other models can be done, yet we are restricted on space here. Our point in mentioning special cases and generalizations of the multiple regression model is to illustrate how multiple regression plays a core role in the analytic architecture in behavioral, educational, and social science research.

SUGGESTIONS FOR FURTHER READING

Multiple regression is often the focus of an entire graduate level course, and many book-length treatments range from very applied to very theoretical. Correspondingly, we are unable to cover the full scope of the multiple regression model and its various uses. However, we offer several suggestions for further reading on the richness of the multiple regression model. For a general introduction to multiple regression, we suggest Kahane (2008), which provides a nontechnical introduction that is useful for understanding the fundamentals of regression. For treatments appropriate for applied researchers and users of research, we recommend Cohen, Cohen, West, and Aiken (2002) and Pedhazur (1997). For a more advanced treatment of regression from a general linear model perspective, we suggest Rencher and Schaalje (2008).

In addition to sources that discuss the multiple regression model, sources that discuss the design aspects of a study that will use multiple regression are of great importance. When designing a study that will use multiple regression, among other things, sample size planning is important. Sample size planning can be done from (at least) two different perspectives: statistical power and accuracy in parameter estimation. Statistical power concerns correctly rejecting a false null hypothesis of interest (e.g., for the test of the squared multiple correlation coefficient or a specific regression coefficient). Accuracy in parameter estimation involves obtaining sufficiently narrow confidence intervals for population effect sizes of interest (e.g., squared multiple correlation coefficient or a specific regression coefficient). Cohen (1988) details sample size planning for statistical power and Kelley and Maxwell (2008) detail sample size planning for accuracy in parameter estimation, both of which are written in the multiple regression context.

DISCUSSION

Having now discussed the regression model itself, showed examples of its use, and connected it with other models widely used in the behavioral, educational, and social sciences, we now take a big picture view of the purpose of the model. We regard multiple regression as having three primary purposes: (a) description, (b) prediction, or (b) explanation, which may not be mutually exclusive.[11]

Although there may be a conceptual distinction between using multiple regression for description, prediction, or explanation, there are no differences in the multiple regression model itself. We briefly discuss these three potential uses of multiple regression to not only help clarify the generality of multiple regression, but also to shed light on its limitations.

Descriptive uses of multiple regression seek to identify ways in which a set of regressors can be used to model the outcome variable. Such a use of regression serves to identify regressors that have some correlation with the outcome after controlling for the other regressors in the model. No stringent philosophical underpinnings are necessary. Rather, the outcome is only a description of the relationship among variables. Using multiple regression for description fails to capitalize on the model for making predictions or explaining relationships, two purposes we discuss momentarily. Using multiple for regression for description can be considered less sophisticated than using it for prediction or explanation. The conclusions that can be legitimately drawn from such a use of regression as a descriptive method are rather weak, unless additional assumptions are made. Nevertheless, as a purely statistical tool, regression can be used to partition the variance in the outcome variable into that which can be modeled by each regressor and that which remains unexplained. The regression coefficients themselves identify the extent to which each regressor has a relation to the outcome variable when controlling for the other regressors.

Rather than saying "the extent to which each regressor has a relation to the outcome variable" as we just did when referring to descriptive uses of multiple regression, it is tempting to say that each regressor "predicts" or "impacts" or "influences" the outcome variable. However, those terms should be reserved for predictive or explanation purposes. That is, for descriptive regression, prediction does not take place. For example, the full data set may be used to estimate the regression coefficients and not used on a future sample for prediction. Alternatively, terms such as "impacts" or "influences" conjure more causal relationships, such as "changes in the kth regressor leads to a b_k amount of change in Y. However, such casual-like statements are generally not warranted.

In the prediction context, a model is formed based on one set of data (training data) but used on data where the outcome variable is unknown. Regression as a predictive model provides an estimated value for outcome variables based on the regression coefficients obtained in the training data sample and the values of a set of predictor variables. For example, one could predict students' first year of college GPA with various individual difference measures, performance in high school (e.g., high school GPA at the end of junior year), and measures of academic achievement (e.g., ACT or SAT scores) based on training data. Then, with the information obtained from the training data, the regression model could be applied to high school seniors to predict their college success, given the relevant regressors (i.e., those used in the model developed from the training data). The purpose of the multiple regression model in this case is not to say what *causes* college GPA, but rather to form a prediction equation that might be useful for predicting academic success as

operationalized by college GPA. Such prediction models can be important because they can help identify those high school students who will likely be successful in college. Of course, no prediction model is perfect. However, statistical prediction (e.g., by using multiple regression) has been shown again and again to outperform expert judgments (e.g., see Grove & Meehl, 1996).

Multiple regression, when used in the explanation context, is ultimately interested in identifying causal variables as well as estimating how much of an impact those variables have on the outcome variable, while holding constant the other regressors. In many cases, it is not possible to unambiguously show causality, but under the appropriate conditions, causal relationships can sometimes be discerned. The situations in which unambiguous causes can be identified require that a random sample of individuals from some population be randomly assigned different levels of the regressor variables of interest. In the vast majority of applications of multiple regression in education, levels of regressors are not randomly assigned but will differ across individuals. In such instances, there is necessarily a limitation of the multiple regression model to be used to infer causation. However, even for models without randomization to level of the regressors, such a model *may* shed light on causal relationships or causal pathways (e.g., via mediation models). Assuming the assumptions of the regression model hold, showing that a particular regressor accounts for some of the variance in the outcome variable (i.e., a non-zero regression coefficient) in a nonrandomized situation (e.g., an observational study) is a necessary, but not a sufficient, condition for causal inference. In such cases, the effect of one regressor on the outcome variable, after including in the model the other regressors, could be a causal agent, but it may not be. Realizing the limitations of multiple regression in making causal inferences is important and has many real-world consequences.

Although understanding what variables are associated with the outcome variable of interest in the context of a set of regressors can be useful in its own right (e.g., for descriptive purposes), the lack of randomization of the levels of regressors does not denigrate the multiple regression model. Any suggestion that a regressor causes (or similarly impacts, influences, effects, acts upon, is an antecedent to, etc.) an outcome variable necessitates a discussion that is above and beyond the regression model itself.

Multiple regression is such a key model in the behavioral, educational, and social sciences that a single chapter cannot replace the need for more detailed study by those that the will use the model directly (e.g., primary researchers) or use it indirectly (e.g., policy makers). Being able to effectively interpret, contribute to, critique, or use the results of the research literature essentially requires a fundamental understanding of multiple regression. We hope this chapter has clearly articulated the multiple regression model for applied researchers and has provided a solid fundamental understanding. Additionally, we hope our chapter has been thought-provoking and that it instills confidence in the presentation and interpretation of results from the multiple regression model.

NOTES

[1] The outcome variable can be termed a "criterion" or "dependent variable", whereas the regressor can be termed a "predictor", "explanatory", or "independent variable". Without loss of generality, we use the terms outcome and regressor throughout the chapter.

[2] Many times multiple regression is said to have two distinct purposes: prediction and explanation (e.g., Pedhazur, 1973). However, we regard description as potentially distinct from prediction or explanation.

[3] Using listwise deletion for missing data in multiple regression is not necessarily optimal. Other methods are available, such as using maximum likelihood or multiple imputation, but they are beyond the scope of this chapter. In general, we would suggest not using listwise deletion in multiple regression, but do it here for simplicity.

[4] Although the model is linear in its parameters, that does not prevent arbitrary functions of the variables from being used, such as taking the logarithm of X_2 or squaring X_3. In fact, the population multiple regression model of Equation 1 can be written as $Y_i = \beta_0 + \beta_1 f_h(X_{1i}) + \cdots + \beta_K f_H(X_{Ki})$, where $f_h(\ )$ is some arbitrary function of the variable in parentheses ($h = 1, \cdots, H$). In most applications of multiple regression no transformation is made, and thus the function would simply be the identity function (i.e., the variable itself is used).

[5] Recalling that Y is in fact a mean, namely a conditional mean, there is a direct parallel to the sum of squares in the context of the estimated variance. In particular, to minimize $\sum(X_i - C)^2$, where C can be any constant real value, the mean of X is the minimizer. That is, $\sum(X_i - \bar{X})^2$ is the minimum value, which is the numerator of the variance. Thus, the regression least squares criterion of minimizing $\sum(Y_i - \hat{Y})^2$ is analogous to why the mean is often regarded as a superior measure of central tendency. It produces the most efficiency (i.e., least variance) compared to any other estimate of central tendency.

[6] Spurious variables are also known as "lurking", "confounding" or, "third variables".

[7] After *R* is installed and then the MBESS package installed within *R*, the way in which a confidence interval for the population squared multiple correlation coefficient can be constructed is as follows: "require(MBESS)" (to load the package) followed by "ci.R2(R2 = .133, K = 4, N = 411, conf.level = .95)" (to implement the ci.R2() function with the appropriate values for the model of interest). See Kelley (2007a; 2007b) for more information on MBESS.

[8] Only confidence intervals for the population unstandardized regression coefficients are available via the SPSS point-and-click interface. Confidence intervals for the population standardized regression coefficients, which is when regression is performed for standardized scores, can be obtained indirectly with the use of the noncentral *t*-distributions. See Kelley (2007b) for a discussion of such confidence intervals and information on how they can be implemented easily via the MBESS R package.

[9] Hierarchical regression should not be confused with the similar sounding hierarchical linear model, usually denoted HLM, as they are completely separate models.

[10] Other coding schemes exist, such as *effect coding* and *orthogonal coding*. In effect coding, a "–1" is used to represent the reference category and a 1 or a 0 is used to represent the other category of interest. In orthogonal coding, coefficients are used that form a set of orthogonal comparisons. Orthogonal comparisons are such that each comparison provides independent information from other comparisons.

[11] Many times multiple regression is said to have (only) two distinct purposes: prediction and explanation (e.g., Pedhazur, 1973). However, we regard description as potentially distinct from prediction or explanation.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Algina, J., & Olejnik, S. (2000). Determining sample size for accurate estimation of the Squared multiple correlation coefficient. *Multivariate Behavioral Research*, *35*, 119–136.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *The Journal of Personality and Social Psychology*, *51*, 1173–1182.

Cassady, J. C. (2001). The stability of undergraduate students' cognitive test anxiety levels. *Practical Assessment, Research & Evaluation*, 7(20).

Cassady, J. C. (2004). The influence of cognitive test anxiety across the learning-testing cycle. Learning and Instruction, *14*(6), 569–592.

Cassady, J. C., & Holden, J. E. (2012). Manuscript currently being written.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic procrastination. *Contemporary Educational Psychology*, *27*, 270–295.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Mahwah, NJ: Erlbaum.

Cohen, J, Cohen, P, West, S, Aiken, Leona S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, *112*, 558–577.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge.

Gollob, H. F., & Reichardt, C. S. (1991). Interpreting and estimating indirect effects assuming time lags really matter. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 243–259). Washington, DC: American Psychological Association.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–Statistical controversy. *Psychology, Public Policy, and Law*, *2*, 293–323.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment interventions. *Evaluation Review*, *5*, 602–619.

Kahane, L. H. (2008). *Regression basics* (2nd ed.). Sage: Thousand Oaks, CA.

Kelley, K. (2007a). Methods for the behavioral, educational, and social science: An R package. *Behavior Research Methods*, *39*, 979–984.

Kelley, K. (2007b). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*(8), 1–24.

Kelley, K. (2008). Sample size planning for the squared multiple correlation coefficient: Accuracy in parameter estimation via narrow confidence intervals. *Multivariate Behavioral Research*, *43*, 524–555.

Kelley, K., & Lai, K. (2010). MBESS (Version 3.0 and greater) [computer software and manual]. Accessible from http://cran.r-project.org/web/packages/MBESS/.

Kelley, K., & Maxwell, S. E. (2008). Power and accuracy for omnibus and targeted effects: Issues of sample size planning with applications to multiple regression. In P. Alasuuta, L. Bickman, & J. Brannen (Eds.), Handbook of social research methods (pp. 166–192). Newbury Park, CA: Sage.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Earlbaum.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheet, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*, 83–104.

Mendoza, J. L., & Stafford, K. L. (2001). Confidence intervals, power calculations, and sample size estimation for the squared multiple correlation coefficient under the fixed and random regression models: A computer program and useful standard tables. *Educational and Psychological Measurement*, *61*, 650–667.

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement*, *11*, 5–11.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. New York, NY. CRC Press.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Harcourt Brace.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*, 93–115.

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons.

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology, 46,* 929–938.

Seber, G. A. F., & Wild, C. J. (1989). *Nonlinear regression*. New York, NY: John Wiley & Sons.

CHRISTINE DISTEFANO & DIANA MINDRILA

5. CLUSTER ANALYSIS

INTRODUCTION

Large multivariate datasets may provide a wealth of information, but often prove difficult to comprehend as a whole; therefore, methods to summarize and extract relevant information are essential. Such methods are the multivariate classification procedures, which use multiple variables to identify characteristics that groups of individuals have in common. By definition, classification refers to the process of dividing a large, heterogeneous group into smaller, homogeneous groups where members are similar to each other while different from cases in other groups (Gordon, 1981; Clogg, 1995; Heinen, 1993; Muthén & Muthén, 2000). The objective is to identify groups underlying a larger set of data, where the number of groups is unknown at the onset. Once created, groups can then be thought of as possessing like patterns of characteristics and cases within the same group may be treated similarly.

Procedures to identify clusters focus on creating smaller groups of cases using the responses to a set of variables. This scenario is conceptually similar to exploratory factor analysis methods, but differs as exploratory factor analysis aims to create smaller groups of variables using responses from a set of cases. Gordon (1981) describes two general reasons why classification may be useful:

Data simplification. Given that large quantities of data can hinder understanding, classification can be useful to detect important relationships and patterns within a larger set of data. If meaningful groups can be identified, groups can be named and properties of the group summarized to allow for more efficient organization and retrieval of information.

Prediction. If a larger set of data can be summarized and patterns within the data to be observed more clearly, it may be of interest to predict how these relationships develop. On a simple level, prediction could be used to predict properties not yet measured, such as inferring about the similarity of cases within a group on variables other than those used to identify the grouping structure. On a deeper level, prediction could be used to posit hypotheses that may account for the groups. Prediction could be conducted in a two step approach where first, an exploratory analysis is used to identify an initial classification system; second, hypotheses of antecedents which contribute to the group structure are tested on an independent sample drawn from the same population.

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 103–122.

Classification methods are well-known and well-used in the social sciences. For example, marketing researchers may group people by spending patterns and store preferences, researchers in education may group students based upon ability or interest in a subject area, anthropologists may group indigenous cultures based upon customs and rituals. Classification systems may be especially useful in educational research, where the goal is often to explain, and provide information that helps assist, intervene, instruct, etc. individuals with a variety of needs.

Much of the research conducted in the social sciences has utilized a variable-oriented approach to data analysis where the focus is on identification of relationships among variables (e.g., multiple regression or correlational procedures) or investigation of mean differences (e.g., ANOVA). This approach is useful for studying inter-individual differences but less so for understanding intra-individual dynamics (Bergman & Magnusson, 1997). In order to address this concern about the study of dynamics, relatively more attention has been devoted to the use of person-oriented analyses.

Bergman and Trost (2006) made the distinction between the theoretical and methodological aspects of person-oriented and variable-oriented approaches. In variable-oriented approaches, basic concepts are considered as variables, and the importance of these concepts is derived from their relationships with other variables, which are investigated using linear statistical models. In contrast, person-oriented theories consider all the variables simultaneously as interrelated components of an indivisible entity, and studies them "as an undivided whole", by employing pattern-oriented approaches (Bergman & Trost, 2006 pp. 604). Such approaches include cluster analytic techniques (Bergman & Magnusson, 1997), which focus upon classification of individuals in order to consider intra-individual variation.

DESCRIPTION AND PURPOSE OF THE METHOD

Cluster analysis refers to a family of procedures which group cases to uncover homogeneous groups underlying a data set (Anderberg, 1973; Aldenderfer & Blashfield, 1984; Blashfield & Aldenderfer, 1988; Everitt, 1993; Hartigan, 1975; Milligan & Cooper, 1987). The researcher has many choices to make when clustering. This discussion will provide an overview of selected procedures and considerations for educational researchers interested in using cluster analysis for classification.

Starting Cluster Analysis

Assumptions and variable considerations. Each case's set of scores across of many variables is evaluated with a cluster analysis. The collection of scores creates a multivariate profile for each case, which is used in analyses to identify like cases. For example (note: this scale will be discussed later in the chapter), Figure 1 illustrates the profiles for two cases from across a set of 14 variables. Profiles can be plotted to provide information about the "height" (the magnitude of the scores on the variable's scale) and the "shape" (the pattern of peaks and troughs for a case) across the set of variables.

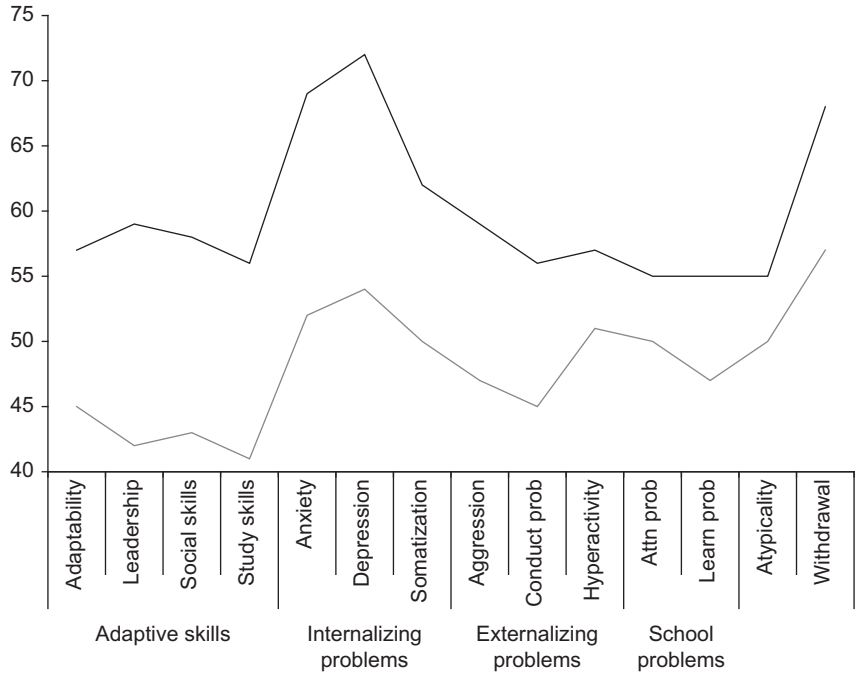


Figure 1. Sample profile of scores for two cases selected from the BASC TRS-C norm dataset.

As with any statistical method, there are assumptions and considerations underlying cluster analysis. First, the choice of variables to include is of primary importance for cluster analysis. Because the multivariate profile is used to create the groups, the variables included in the cluster analysis should be the most relevant to the situation under study. The discussion of how many variables to include for each case is likely to be related to factors of the research situation, such as time, expense, and ease of collecting data information (Everitt, 1993). While researchers in the social sciences often err on the side of collecting more variables than fewer, it is important to note that the groups identified from a cluster analysis may differ markedly if different numbers of variables are included. Finally, both cases and variables selected for a cluster analysis are assumed to be independent and variables are assumed to be uncorrelated.

A second consideration is that variables used in cluster analyses are thought to be at the observed, rather than latent, level. Therefore, variables in cluster analysis may be considered as directly measured and do not necessarily refer to underlying latent variables. The data are also scale-dependent, meaning that variables with both large mean differences and/or standard deviations may suppress the influence of other variables (Everitt, 1993).

Another consideration is the metric level of the data. Typically, cluster analysis requires that data are on the same metric level. For cluster analysis, data may be of any scale type (e.g., nominal, ordinal, interval, or ratio); however, the metric level of the data will impact the choice of proximity measure (described in following section) used to describe relationships between the cases. If the variables have large standard deviations or are measured on different scales, the variables may be standardized to put all values on a common metric before clustering (Aldenderfer & Blashfield, 1984; Everitt, 1993; Milligan, 1996). Standardizing will remove the undue influence due to problems of metric or variability (Milligan, 1996); however, it may also have the disadvantage of masking differences on variables which best differentiate the groups (Duda & Hart, 1973; Everitt, 1993). Results concerning the need to standardize are conflicting. For example, Milligan (1980) found that standardization of variables produced only minor differences in cluster analysis results versus the use of non-standardized data, while other researchers have found that standardization did impact results (e.g., Stoddard, 1979). The decision of whether or not to standardize variables should be made by considering the problem at hand, the metric level of the variables, and the amount of variability in the data. Overall, researchers should be aware that clustering results may differ if standardization is, or is not, carried out.

When data are measured on varying metric levels, there are other transformations that may be useful. For example, principal components factor analysis may be conducted first to reduce the variables into related components, which are then used as input for the cluster analysis (Aldenderfer & Blashfield, 1984). This procedure may be attractive if there is significant multicollinearity between variables, because components are clustered instead of scores from many single variables. However, principal components has been criticized because it may merge modes present in the data set, resulting in data that are normally distributed and may not reflect the original nature of the data (Aldenderfer & Blashfield, 1984).

Another problem encountered with cluster analysis is when it is of interest to group cases by variables of different type. (Everitt, 1993). While it may be of interest to include all variables together to create groups, using a mixed set of variables poses problems in cluster analysis. Researchers have offered suggestions, including categorizing all interval level data to ordinal or nominal level data before clustering (Everitt, 1993). An obvious disadvantage to this option is the loss of potentially important information in the transformation process. A second possibility would be to cluster cases separately, by type of variable, and to try to synthesize results across the different studies (Everitt, 1993). This, too, may not be optimal because information is lost when all profiles of scores are not considered together as one multivariate set.

Sample size requirements for cluster analysis have not been specifically stated. The number of cases needed will be related to the number of variables included, where more cases are needed as the number of variables used to create the groups increases. Cluster analysis is typically conducted with large samples (e.g., >200). However, a rule of thumb is to follow recommendations presented from multiple

regression or factor analysis and use a minimum of 10–15 cases per variable (e.g., Pedhauzer, 1997), with 200 cases as a minimum.

Finally, it is noted that cluster analysis as a methodology is not without criticism. A major criticism is that cluster analytic evaluation criteria are heuristic in nature and a researcher's subjectivity may bias the choice of a solution (Aldenderfer & Blashfield, 1984). Additional criticisms include the lack of statistical indices to assist in the choice of a final solution, and the sensitivity of the clustering algorithm upon the results (Bergman & Magnusson, 1997, Steinley, 2003).

Proximity measures. After deciding on the variables to include, clustering requires an index to use to group cases. The proximity measure transforms the multivariate raw data, via a mathematical formula, into a matrix which is used to evaluate how alike cases are (Romesburg, 1984). There are two general types of proximity measures, similarity indices and dissimilarity indices, where elements in the data matrix vary based on direction of the relationship between the cases. For dissimilarity indices, *smaller* values indicate that two cases are more alike; for similarity indices, *larger* values indicate that two cases are more alike.

While variables for cluster analysis can be measured on any metric level, in social sciences, clustering often takes place with data that is at least ordinal in nature. Examples of data which may be of interest to educational researchers include grades, test scores, or standardized test scores. While proximity indices can be used with nominal or ordinal data, focus will be given to proximity measures used with interval data. Researchers interested in using nominal data have options to create a proximity matrix, such as the simple matching coefficient or Jaccard's coefficient (c.f. Aldenderfer & Blashfield, 1984, pp. 28–29; Everitt, 1993, pp. 40–41), or with Sneath and Sokal's method (c.f. Gordon, 1981, p. 24) when ordinal data used. For data that are on at least interval level of measurement, or even an ordinal level, but treated as continuous data[1], correlation and distance measures may be used. These measures are the two types that commonly used with cluster analysis in the social sciences (Aldenderfer & Blashfield, 1984).

A popular similarity index metric used in cluster analysis is the correlation coefficient (Aldenderfer & Blashfield, 1984; Everitt, 1993). This value summarizes the amount of relationship between cases as:

$$r_{jk} = \frac{\sum (x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k})}{\sqrt{\sum (x_{ij} - \overline{x_j})^2 (x_{ik} - \overline{x_k})^2}}, \quad (1)$$

Where x_{ij} is the value of variable i for case j and $\overline{x_j}$ is the mean of all values of the variable for case j (Aldenderfer & Blashfield, 1984).

Correlations are not scale-dependent, and the values are bounded from –1 to +1, making interpretation relatively easy. While the correlation coefficient has some attractive qualities, this index is often criticized. It has been suggested that the correlation coefficient is a useful measure of similarity in those situations where

absolute "size" of the differences alone is seen as less important than the "shape" of the profile (Everitt, 1993). Thus, the correlation similarity index is sensitive to the shape of a profile (i.e., the pattern of dips and rises across the set of variables). For example, consider two cases where the profiles are of different magnitude (i.e., elevation) but of similar shape. These cases would have a high correlation value, meaning a high degree of similarity. Figure 1 shows two profiles of scores, where cases differ in elevation but are similar shape. Note that the collection of variable means for a set of variables used to define a cluster is called a centroid. Another criticism noted is that computing a correlation coefficient for use in clustering requires that the mean value must be obtained across different variables rather than across cases. Researchers have argued that this type of summarizing does not make statistical "sense" (Aldenderfer & Blashfield, 1984).

A second type of proximity measures, *dissimilarity* indices illustrate how different two cases are from each other across the set of variables. Two highly dissimilar cases would receive a higher value, or greater distance, between cases, while highly similar cases receive a low value, showing greater similarity (Everitt, 1993). Two cases with identical scores on all variables would receive a distance measure of zero, showing perfect agreement across the two profiles. While these measures have a minimum of zero, there is no upper bound, making distance scores themselves hard to interpret. Distance measures are also scale-dependent (Everitt, 1993) and sensitive to fluctuations in variability across variables used in the clustering.

A very popular distance measure used with cluster analysis is the Euclidean distance. From Aldenderfer and Blashfield (1984), the Euclidean distance between two cases, *i* and j, is described as:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}, \tag{2}$$

where d_{ij} is the distance between case *i* and case *j,* x_{ik} is the value of the k^{th} variable for the i^{th} case, x_{jk} is the value of the k^{th} variable for the j^{th} case. For a multivariate profile, x_{ik} is represented as a vector, and differences between variables are summed over all variables used in the clustering procedure. When calculating a Euclidean distance, two case profiles, two group centroids, or an individual case profile and a group centroid can be used in the formula.

To eliminate the square root symbol, the Euclidean distance value is often squared, and the squared value (d_{ij}^2) is reported as the squared Euclidean distance.

A final important distance is the Mahalanobis D^2, which is defined as:

$$d_{ij} = (X_i - X_j)' \Sigma^{-1} (X_i - X_j), \tag{3}$$

Where Σ is the pooled within-groups variance-covariance matrix and X_i and X_j are vectors of the values for cases *i* and *j*. Unlike the Euclidean distance, this metric incorporates relationships among variables into the equation. When the relationship

between variables is zero, the Mahalanobis D^2 is equivalent to the squared Euclidean distance. Given that the groups underlying a dataset are typically unknown at the start of a cluster analysis, the entire dataset would need to be used as the choice of Σ (Everitt, 1993).

Other measures of dissimilarity that can be used with continuous data are (a) the city block distance, (b) the Minkowski distance, (c), the Canberra distance, (d) the Pearson correlation, and (e) the angular separation (Everitt, Landau, Leese & Stahl, 2011). Although distance measures may have some advantages, they are often criticized because a computed index may be affected by the elevation or "height" of the profiles. In other words, while two cases have a similar shape across the set of variables, the level of the scores impacts the dissimilarity index (Aldenderfer & Blashfield, 1984). Considering the cases shown in Figure 1, these two cases may have a large correlation index, showing similarity, but the same two cases could have a large distance measure, showing dissimilarity. This discussion illustrates the importance of selecting a proximity index based on what considerations are most important for a researcher's purpose for clustering. Also, with distance measures, cases with large standard deviations and size differences can overpower the effects of variables with smaller size differences or standard deviations. Finally, distance indices are affected by transformations of scale, such as standardizing variables. Even given these caveats, distance measures are among the most often used with cluster analysis.

Clustering Algorithms

There are many choices of clustering algorithms to join cases into groups or clusters. When choosing an algorithm to join cases, the method needs to be compatible with the purpose for clustering, the level of the variables, and the choice of similarity matrix. Also, each method represents a different perspective and could produce different results, even when applied to the same dataset.

At its most basic level, there are different families of procedures that cluster cases according to a general method. There are seven major families of clustering methods (cf. Aldenderfer & Blashfield, 1984, pp. 35–53); however, the three most popular families used in social sciences are discussed: hierarchical agglomerative methods, iterative partitioning methods, and factor analytic variants. Several different clustering techniques underlie each of these families of clustering methods. These selected options, and additional clustering algorithms (not discussed here) are readily available to researchers through software packages often used in educational research (e.g., SAS, R, SPSS, etc.).

Hierarchical algorithms. Hierarchical algorithms join cases into groups using a series of merger rules. These techniques can be subdivided into two types: (1) agglomerative techniques, which successively group single cases to arrive at one group of size N, and (2) divisive methods, which separate the N cases into

smaller subsets (Everitt, 1993). Divisive methods are far less popular in the social sciences than hierarchical techniques (Everitt, 1993). Therefore, focus will be on agglomerative methods; discussions of divisive methods may be found in clustering texts (cf. Everitt, 1993; Hartigan, 1975; Lorr, 1983).

Hierarchical agglomerative methods have been the most popular procedure of linking cases used with clustering (Aldenderfer & Blashfield, 1984). These methods examine the proximity matrix and sequentially join two cases (or cases to cluster) that are the most similar. After cases are joined, the similarity matrix is re-examined to join the two cases/clusters with the next smallest distance to another case/cluster. A total of $N - 1$ steps are made through a dataset, grouping cases from singletons to one large set of N cases, where N is the number of cases in the dataset (Lorr, 1983).

Different ways to join the data underlie hierarchical agglomerative methods. The single linkage (or nearest neighbor) method joins like cases in terms of similarity index. Here, new cases are joined to groups on the basis of a high level of similarity to any member of the group. Therefore, only a "single link" is required between two cases to merge the group(s). A drawback to this linking process is that it may produce long "chains" of cases, where cases are linked one-by-one to create one large cluster.

The complete linkage (or furthest neighbor) method is the counterpart to the previous techniques, in that cases are considered to be included into an existing cluster must be within a specified level of similarity to all members of the group. This is a much more rigorous rule than imposed by the single linkage method. As a result, the complete linkage method tends to create smaller, tighter elliptical-shaped groups (Aldenderfer & Blashfield, 1984). As a middle ground, the average linkage method essentially computes an average of the similarity index for a case with all cases in an existing cluster and cases are joined to the group based on the average similarity with members. Other methods, (e.g., Mean vector [or centroid] clustering and median clustering) work similarly to group cases.

The most popular hierarchical agglomerative method used in the social sciences (Aldenderfer & Blashfield, 1984) is Ward's method (Ward, 1963). This procedure creates groups which are highly homogeneous by optimizing the minimum variance, or error sum of squares (ESS), within clusters. The ESS formula, as stated in Everitt (1980) is:

$$ESS = \sum_{i=1}^{n} (X_i - \bar{X})^2, \tag{4}$$

where X_i is the case (or group) in question and $\bar{X}$ is the cluster centroid. The n may refer to the total number of cases (at the start of the process) or the number of groups (as the clustering process proceeds). At the first step of the process, each case is its own cluster, and the ESS among groups is 0. Cases are joined into clusters which result the smallest increase of ESS, computed as a sum over all clusters. Although

Ward's method is very popular, it is also sensitive to the elevation of the profiles and may group cases into different clusters based on elevation, even if the shapes of the profiles are similar.

While hierarchical agglomerative methods are useful, they do suffer from problems. One major problem is cases joined at an earlier step cannot be reassigned– even if the case has a closer association with a different group by the end of the assignment process. In other words, a poor start to the clustering process can not be overcome because only one pass to assign cases is made through the dataset. Second, hierarchical agglomerative methods do not have "stopping" rules which state the number of clusters underlying a dataset. The researcher can plot the union of cases through the clustering process by using a dendrogram (Everitt, 1993). Dendrograms visually represent the merging of cases at each step, from the lowest level (where all *N* cases are their own group) to the highest level (with all cases forming one large group of *N* cases). A researcher can examine the plot for suggestions of the number of clusters underlying the dataset by looking for the large divisions or "steps" in the graph.

For example, Figure 2 shows a dendrogram plot for a subset of 50 random cases from the example dataset (described later in this chapter). For the clustering, Ward's method with squared Euclidean distances was used to group cases. At the bottom of the plot, each case is its own group, and similar cases are joined in a hierarchal manner. The plot shows that there may be four groups underlying the dataset. Three clusters have multiple cases and one cluster consists only of one case (case ID number 36). This case may be examined to determine if there were measurement/scoring problems or to see if it is an outlier. If it is of interest to keep this case, it may be of interest to try differing number of cluster solutions (e.g., 3, 4, 5 clusters) to evaluate the placement of this case in different cluster solutions.

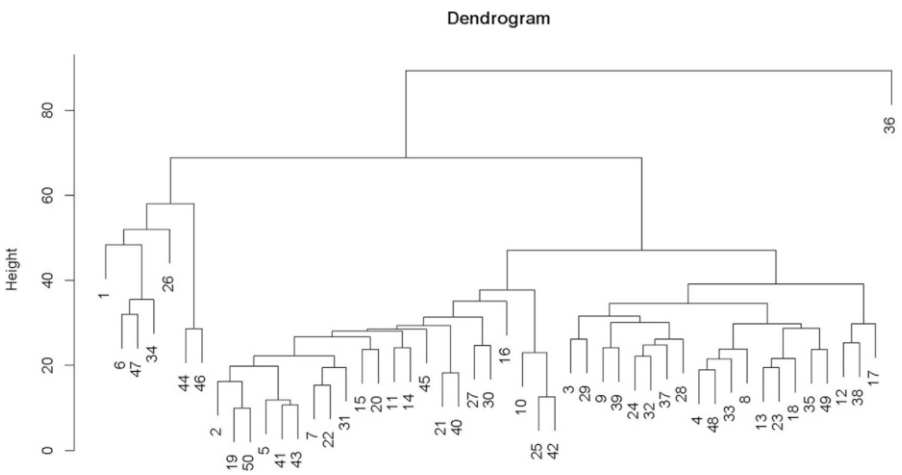


Figure 2. Dendrogram for a selection of 50 cases from the BASC TRS-C norm dataset.

A third consideration is that groups created using hierarchical agglomerative methods are nested. This means the clusters do not overlap and each cluster can be subsumed as a member of a larger, more inclusive group at later steps in the process. Finally, hierarchical methods may not provide stable groupings. A researcher may obtain different results if a dataset is simply reordered, reshuffled, and re-clustered (Blashfield & Aldenderfer, 1988).

Iterative partitioning methods. Iterative partitioning methods are another choice of clustering algorithm. As with hierarchical methods, there are many procedures and choices underlying this family of methods. Approaches within this family use a similar set of general procedures, where clustering is conducted by completing a series of steps (Aldenderfer & Blashfield, 1984; Everitt, 1993):

1. At first, an initial cut or partition of the data set into *k* clusters is made, where *k* is specified by the researcher. The centroid, or the arithmetic mean values across the set of variables, is computed for each of the *k* clusters. If group information is unknown, the initial partition may be conducted arbitrarily. For example, one choice may be to allow the first *k* cases to serve as the cluster centroids for the *k* groups or randomly choosing *k* cases to serve as the initial group means.
2. Next, individual cases are assigned to the cluster that has the nearest centroid. This is typically conducted using distance measures (e.g., squared Euclidean) in the proximity matrix and assigning cases to the cluster with the smallest distance to a given centroid.
3. Once all cases in the dataset are assigned, the centroids of the *k* clusters are re-computed. The dataset is re-examined to see if any cases have a smaller distance to the cluster centroid from the initial assignment to a different cluster centroid. Cases with smaller distances to different groups are re-assigned.
4. Steps 2 and 3 are repeated by making "passes" or iterations through the dataset until no cases change cluster assignment.

With iterative clustering procedures, the type of pass used to group the data refers to how cases are assigned after each iteration. There are two basic types of passes: *k*-means and hill climbing (Aldenderfer & Blashfield, 1984). Hill climbing passes assign cases to a cluster if the proposed assignment optimizes the value of a statistical criterion, which is concerned with cluster homogeneity. Alternatives for the criterion can be based on the within-group variation, **W**, the pooled within cluster covariance matrix or in combination with the between-group variation, **B**, the pooled between cluster covariance matrix (Everitt, 1993). Using these matrices, optimization criteria focus on minimizing **trW**, minimizing the determinant of **W**, and maximizing the trace of $\mathbf{BW}^{-1}$.

The *k*-means pass involves assigning cases to the cluster with the nearest centroid. There are many options to assign cases: the process may be combinatorial or noncombinatorial, inclusive or exclusive. Combinatorial methods allow for recalculation of a centroid after each membership change, while noncombinatorial

methods recalculate cluster centroids after the entire dataset has been classified. When computing centroid values, individual cases can be included in the calculations (inclusive) or a case in question may be removed from the centroid calculations (exclusive method).

While not explicitly stated, the *k*-means procedure tries to minimize the trace of the pooled within covariance matrix (**trW**) (Aldenderfer & Blashfield, 1984). This criterion is similar to Ward's method, as minimizing the trace of the within group sum of squares matrix is equivalent to minimizing the sum of squared Euclidean distances between cases and the cluster centroid (Everitt, 1993); however, the hierarchal process is optimized within k-means to identify the "best" (global) solution underlying a dataset (Steinley, 2006).

While iterative partitioning methods do allow cases to switch clusters, these methods are not without problems. Iterative partitioning procedures are sensitive to the initial cut in the data and may not be able to fully overcome a poor initial partition (Steinley, 2003). As with hierarchal methods, *k*-means may produce a suboptimal solution if the initial cut is poor. This has been referred to as the problem of finding a local optimal solution, rather than a global solution for the entire dataset (Aldenderfer & Blashfield, 1984). To avoid the problem of a poor starting point, centroid values, or "seed" values may be used, where the input for the initial partition are *k* centroids based upon prior knowledge or previous analyses. Another strategy that has been recommended is to use the final solution from the Ward's hierarchical agglomerative method as the starting point, or seed, for the iterative partitioning procedure (Huberty et al., 1997; Ward, 1963). By using the final Ward's solution as the initial starting point for the *k*-means procedure, the researcher gains the benefits of both clustering algorithms. As with hierarchal methods, random shuffling of a dataset and re-clustering can help determine if identified clusters are stable entities.

As with other clustering methods, it is noted that a researcher may achieve different results if different choices are made and applied to the same iterative partitioning method. For example, results may differ if a *k*-means versus hill climbing procedures is used or different optimization criteria are considered, even with the same dataset (Aldenderfer & Blashfield, 1984; Everitt, 1993). Finally, the most well used method, minimization of **tr(W)**, is scale dependent and may produce different solutions even if raw data and standardized data from the same dataset are clustered.

Factor analytic variants. Factor analytic approaches to clustering have been used in more in psychological research than in the other social sciences (Aldenderfer & Blashfield, 1984). These methods are often termed *Q*-analyses and focus on using factor analytic techniques (e.g., principal components) to reduce a correlation matrix of relationships between cases (i.e., rows). This method is similar to more traditional exploratory factor analysis, which examines the relationships between variables (i.e., columns) of a dataset.

While Q-analysis techniques have been used to group cases, distinctions have been identified between this procedure and other clustering methods (Lorr, 1983).

For example, factor analytic methods have been termed as dimensional approaches, which simplify a dataset by identifying a fewer extracted factors to explain the relationship between cases (Weiss, 1970 as cited in Lorr, 1983, p.123). Where hierarchical and iterative partitioning procedures aim to reduce a dataset into smaller, discrete groups, the groups obtained from these analyses may not be mutually exclusive (Lorr, 1983). Because factor analytic approaches to clustering cases are not as common as iterative partitioning or hierarchical clustering methods, but, may be used in social science research, these methods will be briefly discussed. More detail on factor analytic variants may be found in clustering texts (cf., Aldenderfer & Blashfield, 1984; Lorr, 1983).

Like traditional factor analysis scenarios, Q-analysis procedures typically use a similarity index to compare cases. While correlations are typically used to group similar profiles, it is known that this choice of proximity index does not consider the level of the profile(s). Alternatives to use of the correlation matrix in a Q-analysis include using covariances between variables or the sum of score cross-products. An advantage of using covariances as the index of similarity is that the variables remain in their original metrics – this may aid interpretation if variables possess similar amounts of variability. Similarly, using the sum of score cross-products uses raw scores; results from raw score cross products provide results that approximate results from using distance coefficients (Lorr, 1983).

Once a proximity matrix is constructed, the relationships between cases are factor analyzed using common methods, such as principal components or principal axis factoring. Lorr (1983) noted principal components method is the most popular procedure used to identify a smaller number of groups. However, when assigning groups to clusters, researchers will encounter similar problems as with traditional factor analysis —where criteria need to be used to classify cases to groups. A common procedure is to assign a case into a group is to use a cutoff value, such as a minimum correlation (e.g., value of at least .5) between the case and the extracted group (Lorr, 1983). Also, to create mutually exclusive groups, cases should not "cross-load" with other groups above a maximum value (e.g. no higher than a .35 association with other groups). While guidelines are provided, these considerations are subjective and may vary among researchers; care is needed when choosing outcomes under Q-cluster analysis.

Finally, some considerations apply when conducting Q-analysis. For example, if variables are on different metrics interpretation problems may arise. In this situation it is recommended that the variables are centered and then standardized across the cases before calculating correlations or sums of raw cross-products (Lorr, 1983). Second, because the cutoff values to assign cases to clusters are chosen by the researcher, the homogeneity of the clusters may be influenced by the cutoff value chosen. Lower cutoff points would allow for a greater number of cases to be classified into a group, but the resulting cluster would be relatively heterogeneous. Higher cutoff values for association would result in more homogeneous groups, but lower coverage of the dataset.

In addition to the clustering algorithms described above, there are some clustering procedures that have unique features and cannot be assigned to any of the clustering families described in the literature. Such clustering techniques are: (a) methods based on density search and mode analysis; (b) methods that allow overlapping clusters; (c) methods that cluster data matrices instead of proximity matrices and, therefore, group both variables and individuals at the same time; (d) methods that constrain the clustering process by using, in part, external information to determine the cluster memberships; (e) fuzzy methods, where individuals have fractional memberships to several groups and a membership function indicates the strength of membership to each cluster; and (f) neural networks, where groups are identified by modeling pattern recognition algorithms employed by highly connected networks such as neurons in the human nervous system (Everitt et al., 2011, pp. 215–255).

Conducting Cluster Analysis

Determining a starting point for cluster analysis. Given that the goal of clustering is to determine the number of cases from an ungrouped set of data, a natural question when beginning the process is: "How many groups underlie the dataset?" Heuristic criteria may be used to suggest the optimal number of clusters. Three statistics, Cubic Clustering Criterion, Pseudo F, and Pseudo t-square, can be plotted by the number of possible clusters (maximum of *N* clusters) to judge the number of groups underlying a data set (Aldenderfer & Blashfield, 1984; Sarle, 1983). The plots are analogous to a scree plot in factor analysis. Here, graphs are examined to determine large changes in level of the plot, where the drop suggests the number of clusters underlying the dataset (Everitt, 1993). Additionally, with hierarchical methods, dendrogram plots can be examined to identify where "steps" or breaks in the graph are, denoting different groups. If factor analytic methods are used, scree plots can be used to determine the number of groups which may underlie the data. Using dendrogram or other plots are subjective methods to determine the number of clusters. As with exploratory factor analysis, when conducting cluster analysis, researchers should use the suggested number of clusters as a starting point and evaluate a range of cluster solutions above and below this point.

Nevertheless, the cluster analysis literature provides a variety of statistical tests, indices, and procedures that may be used to obtain additional information and help researchers identify the optimal number of clusters in a data set. Such criteria are: (a) the Calinski and Harabsz's index (Calinski & Harabasz, 1974), (b) Hartigan's rule (Hartigan, 1975), (c) the Krzanowski and Lai test (Krzanovski & Lai, 1985), (d) the silhouette statistic (Kaufman and Rousseeuw, 1990), (e) approximate Bayes factors (Kass & Raftery, 1995; Frayley & Raftery, 1998), (f) the gap statistic (Tibshirani, Walther, & Hastie, 2001), (g) nonparametric measures of distortion (Sugar & James, 2003), (h) stability analysis (Steinley, 2008), or (i) bootstrapping (Fang & Wang, 2012). Many of these techniques were developed to address specific problems and, therefore, do not have a general applicability. Furthermore, some of the methods

that have a wider applicability are computationally intensive or "require strong parametric assumptions" (Sugar & James, 2003, p. 750). Thus, such statistical tests are not widely used with cluster analysis.

Choosing a cluster solution. Once a researcher has obtained a solution, interpreting cluster analysis results involves two main components. First, the centroid information is evaluated for each cluster (Aldenderfer & Blashfield, 1984; Everitt, 1993). Through examining cluster centroids, one may determine if a cluster's centroid values identify a subgroup of the population. Second, supporting information about each cluster's demographic characteristics may be considered. Within educational research, demographic characteristics may include gender distributions, racial/ethnic membership, family socioeconomic status, and cluster size relative to the total sample. A cluster is "named" by comparing the centroid information and demographic characteristics to existing theoretical perspectives and prior research. This information can be evaluated and compared for a range of solutions to determine which one fits the data best.

After the final cluster solution is agreed upon, additional investigations and use of the solution can be made. These may be internal validation and external validation procedures. Internal validation procedures center on using the same dataset. This may be conducted by shuffling the dataset and reclustering to see which cluster definitions are stable. Another method which is useful if the dataset is sufficiently large is to split the dataset into half samples. One half-sample may be used to build a classification rule using predictive discriminant analysis techniques, and applied to the second half sample (Huberty et al., 1997). This has the effect of treating the second sample as ungrouped cases, where the cases are assigned into the cluster with the closest association (Huberty, 1994). Concordance between the two classification methods may be assessed.

External classification procedures focus on using independent datasets. Validation of a cluster solution is paramount to illustrating that it is an optimal solution (Aldenderfer & Blashfield, 1984). Replication is an important criterion, not only to determine the appropriate number of clusters, but to ensure that the agreed upon solution holds the same meaning in independent samples from the same population (Aldenderfer & Blashfield, 1984). Validation procedures may also be conducted by determining if there are differences between groups (i.e., Analysis of Variance – ANOVA) on important variables which were *not* used to group cases into clusters.

ILLUSTRATIVE STUDY

As an example to illustrate different classification methods, the Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992) Teacher Rating Scales–Child (TRS-C) norming dataset was utilized. The TRS-C includes 148 items that are rated by a child's teacher. For each child, teachers rate the frequency of behaviors

exhibited during the last several months, using a four-point scale of "Never", "Sometimes", "Often" and "Almost Always." This form is appropriate for children aged 6 to 11 years old.

Items on the TRS-C are organized into 14 subscales of behavior measure both adaptive and maladaptive behaviors. Teacher ratings for each subscale can be transformed to a T-score (mean of 50, standard deviation of 10), and, for each child, values across the set of 14 variables may be used to evaluate a child's emotional and behavioral health in the school setting. Generally, higher scores represent greater levels of problematic behavior; however, for scales measuring adaptive skills (e.g., Study Skills), lower scores represent more maladaptive behaviors.

While the BASC TRS-C (1st edition) dataset was selected for a number of reasons, the primary reason for inclusion is that it represents a situation which is often encountered with classification: where it is of interest to create subgroups of cases from a large, ungrouped dataset. A realistic objective may be to create a classification system for describing child behavior in the school setting – including both good behavior and problematic behavior/emotional competency – as rated by teachers. Knowledge of subgroups of children who behave similarly to other children within the same group may be of interest for academic remediation, referral for counseling or special education, or differentiated instruction.

For the TRS-C norm sample, the average age was reported as 8½ years old and consisted of 598 (49%) female and 630 (51%) males. The sample of children was primarily Caucasian (*n = 820*, 67%), with 33% (n = 408) classified as minorities. Of the children included in the norm database, the majority had not been diagnosed with a behavioral or emotional disorder (*n = 1131,* 92%); however, 91 (7.5%) of the children had received one clinical diagnosis, and six children (0.5%) received two prior diagnoses.

To begin clustering, both CCC plots and dendrograms were run with SAS software (version 9.2). The plots suggested that 6–8 clusters were underlying the dataset; however, 4 through 9 cluster solutions were run and interpreted. For clustering, Ward's method was used with the Squared Euclidean distance as the proximity measure. To evaluate the cluster solutions, group centroids for the solutions were examined and matched to theoretical knowledge of child behavior in school settings as well as prior research solutions. The size of the cluster relative to the total norm sample and the gender membership in the cluster was used to help identify characteristics of the groups. To validate the final solution, an ANOVA was run on the Behavioral Symptoms Index (BSI) which is a measure of a child's overall level of risk. BSI values are measured on a T-score metric and are comprised from a collection of information on the TRS-C form. While it is recognized that the information is not truly unique and would not be an optimal choice of a variable for validation, it is used to illustrate how validation procedures may be conducted.

After evaluating and comparing multiple solutions, a seven cluster solution was interpreted. The seven groups uncovered by the Ward's clustering procedure were named by examining the centroids across the set of 14 variables (listed in Table 1)

Table 1. BASC teacher rating scale–child norming data: mean t-scores by scale for the seven cluster solution under ward's clustering algorithm

Ward's Method	*Cl.1*	*Cl.2*	*Cl.3*	*Cl.4*	*Cl.5*	*Cl.6*	*Cl.7*
N	*463*	*160*	*277*	*89*	*38*	*25*	*176*
Externalizing Problems							
Aggression	44.15	51.04	45.71	**67.33**	**68.95**	44.52	57.76
Conduct Problems	45.39	48.13	47.16	**67.22**	**71.00**	52.76	54.50
Hyperactivity	43.68	49.80	47.67	**66.51**	**68.34**	47.56	59.31
Internalizing Problems							
Anxiety	45.42	54.39	45.99	55.72	**72.03**	48.64	55.03
Depression	44.48	52.41	45.82	**60.39**	**79.11**	50.76	57.16
Somatization	46.10	54.96	46.89	49.26	**64.00**	46.08	58.23
Other Scales							
Atypicality	45.14	48.95	47.71	**64.79**	**81.79**	55.36	54.19
Withdrawal	44.99	50.71	48.42	56.54	**71.58**	**76.52**	53.63
School Problems							
Attention Problems	41.46	47.35	52.97	**65.47**	**68.34**	**63.04**	57.99
Learning Problems	42.52	47.49	52.99	**64.76**	**66.13**	**63.80**	56.85
Adaptive Skills							
Adaptability	58.15	50.21	48.04	**36.33**	**31.76**	41.40	41.59
Leadership	57.73	52.78	42.17	**40.11**	**40.58**	**33.16**	45.57
Social Skills	57.52	52.16	43.02	**39.57**	42.21	**33.32**	45.30
Study Skills	58.79	53.73	43.55	**36.75**	**38.61**	**34.44**	42.69
Percentage of Total (Cluster Size)	38	13	23	7	3	2	14
Percentage Male / Female	39/61	46/54	57/42	82/18	58/42	48/52	63/37
Cluster Name	Well Adapted	Average	Low Adaptive	DBP	GP-S	Acad. Prob.	Mildly Disrup.

Notes. Values that differ from the mean by one standard deviation or more (regardless of direction) are printed in boldface. Cl= Cluster, DBP = Disruptive Behavior Problems, GP-S = General Problems—Severe; Internal. Problems = Internalizing Problems, Acad. Prob. = Academic Problems.

and matching the descriptions to prior research. The groups identified were named: (1) Well Adapted, (2) Average, (3) Low Adaptive, (4) Disruptive Behavior Problems, (5) General Problems—Severe, and (6) Mildly Disruptive and (7) School Aversion. Each cluster is briefly described to illustrate the naming process.

The Well Adapted cluster ($n = 417$) was named because of its significant elevations on adaptive scales and absence of behavioral problems. There were more girls (60%) reported in this group than boys (39%). The second cluster was labeled Average. With 160 members, this cluster reported all 14 variables close to expected mean values of 50 and had slightly more girls in the cluster.

A third cluster of 277 members was identified as Low Adaptive Skills. This group looked similar to the Average cluster, with the exception of low scores on three of the four Adaptive Skills scales. This cluster had a higher percentage of boys as members.

A fourth cluster of 89 members was identified as Disruptive Behavior Problems. Significant adaptive behavior deficits and elevation on externalizing scales mark this cluster. As expected, males dominated this cluster (82%). The Disruptive Behavior Problems group accounted for seven percent of the total norm sample.

The Mildly Disruptive group had 176 members, was predominantly male (63%) and accounted for 14% of the norm sample. This cluster is differentiated from the Disruptive Behavior Problems cluster by comparatively mild elevations on the Aggression, Hyperactivity, and Adaptability scales.

The cluster General Problems – Severe is the most behaviorally impaired of all the cluster types. This small cluster ($n = 38$) is predominantly male (58%) and the group exhibited a diverse array of problems including psychotic thought processes (significant Atypicality scores) and impaired adaptive skills. Additionally, children in this cluster exhibited high levels of externalizing behaviors. General Problems–Severe children comprised only a small percent (3%) of the norm sample.

A small cluster of children ($n = 25$) was found with scores within one half standard deviation of the mean on Internalizing Problems and Externalizing Problem scales. However, this scale had significantly high levels of School Problems scales, very low Adaptive Skills, and the highest Withdrawal T-scores across the set of clusters. The group was roughly equally split across genders. This group was named School Aversion because it shares similarities with the Academic Problems cluster identified in previous studies (e.g., Kamphaus et al., 1997), but the levels seen here are much more extreme.

An ANOVA was run to see if the groups illustrated mean differences on BSI. The ANOVA test reported significant mean differences across BSI values for the different clusters ($F_{6,1221} = 815.47$, $p < .001$). The lowest T-scores, illustrating lower 'at-risk' status were seen for students in the Well-Adaptive group; highest BSI values were reported for the General Problems—Severe students. Bonferroni post-hoc tests were conducted to determine which group scores were significantly different. With the exception of scores for students in the Average group and the Academic Problems group, there were significant differences among the BSI mean scores. From lowest T-score to highest, the groups were ordered as: Well-Adaptive ($M = 42.6$), Low Adaptive ($M = 47.1$), Average/Academic Problems ($M = 50.1/51.9$), Mildly Disruptive ($M = 58.7$), Disruptive Behavior Problems ($M = 66.7$), and General Problems –Severe ($M = 78.8$).

SUMMARY AND CONCLUSIONS

Cluster analysis has a long history in the social sciences. The goal of this chapter was to introduce cluster analysis as a classification procedure. Additionally, the example was presented to illustrate how different choices in the classification process can produce different results. While limited procedures were presented, the figures and tables provide information and may be used as a resource for interpretation of various classification techniques.

One recognized limitation of the current chapter is that the viability of cluster solution in the presence of missing data was not discussed. This was omitted to concentrate on an introduction of the procedures and assumptions underlying cluster analysis. However, it is recognized that missing data is often encountered, especially in a field such as the educational research which often uses self-report or test data in investigations. Briefly, there are many ways that missing data may be handled, some of which vary depending on the type of missing data, and some of which vary based upon the classification procedure in use. While these issues are very involved and complex, it is beyond the scope of the chapter to discuss different types of missing data (e.g., missing at random, missing completely at random, and missing not at random). Readers are referred to texts on missing data, such as Little and Ruben (2002) or Enders (2010), for more detailed information about various types of and treatments for missing data.

If cluster analysis is of interest and the percent of missing data is not too high, data could be imputed or the dataset reduced through listwise deletion methods (or pairwise deletion in the case of factor analytic variants of cluster analysis). Again, these methods to treat missing data come with known caveats (e.g., Enders, 2009), and if too much data are lost through listwise deletion, the accuracy of the groups is questionable. Further, if too much data is imputed for the variables without taking into consideration characteristics of the case (e.g., mean imputation), the variability within the dataset will be reduced. If possible, it is recommended to impute mean scores using information from an individual's pattern of scores (e.g., mean imputation for an individual on a missing item based on other items within the same subscale). On a whole, cases with a lot of missing data may be investigated to see if there is enough information to include these cases in the cluster analysis. Researchers may want to create an arbitrary cut-off value (e.g., 25% missing data) and include cases with less missing data and exclude those cases with missing data above the cut-off value. Other, more sophisticated methods of treating missing data in cluster analysis include estimating a missing data point using regression procedures (Gordon, 1981).

We also note that there are many possible numbers of combinations of cluster applications which may be used. Hopefully the presentation of the algorithms and proximity values along with the example can give researchers an idea of the magnitude of choices available when conducting cluster analysis. Researchers are encouraged to apply more than one technique when conducting classification work (e.g., two different clustering algorithms) to determine which groups consistently re-emerge. Further,

while validation procedures were briefly discussed, the importance of validation in the context of clustering cannot be stated loudly enough. Validation is crucial to ensure that the groups identified are not artifacts of just one sample of data.

CONCLUSION

In summary, cluster analysis can be valuable tools in the exploration of large sets of multivariate data. By organizing data into subgroups, these smaller groups can help researchers identify patterns present in the data and uncover the unique characteristics of the group structures. Application of cluster analysis in practice requires care because, as shown in the chapter, there are many areas where choices need to be made, and criteria to evaluate which are subjective and open to different interpretations by different researchers. As Everitt, Landau, Leese and Stahl (2011, p. 287) state "Simply applying a particular method of cluster analysis to a dataset and accepting the solution at face value is in general not adequate." As classification methods rearrange the "facts" of a dataset for pattern recognition and group identification, validation work is crucial to conduct before trusting that a solution represents an underlying taxonomy. Careful analysis and execution of the all the decisions underlying classification will help the methodology fulfill its potential as an efficient and useful tool for applied researchers.

NOTE

[1] In the social sciences, data that are ordinal are often treated as interval level data. A common example is data from self-report questionnaires where data arise from the use of a Likert scale.

REFERNENCES

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis. Beverly Hills, CA: Sage Publications.*

Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press, Inc.

Bergman, L. R., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development & Psychopathology*, *9*, 291–319.

Bergman, L. R., & Trost, K. (2006). The person oriented versus the variable-oriented approach: Are they complementary, opposites, or exploring different worlds?, *Merrill-Palmer Quarterly*, *3*, 601–632.

Blashfield, R. K., & Aldenderfer, M. S. (1988). The methods and problems of cluster analysis. In J. R Nesselroade, R. B Cattell (Eds.). *International handbook of multivariate experimental psychology* (pp. 311–359). New York: Plenum Press.

Calinski, R. B., & Harabasz, J. (1974). A denrite method for cluster analysis. Communications in Statistics, 3, 1–27.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences*.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: John Wiley & Sons, Inc.

Enders, C. K. (2010). *Applied missing data analysis*. New York: The Guilford Press.

Everitt, B. S. (1993). *Cluster analysis* (3rd ed.). New York: John Wiley & Sons, Inc.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.), Chichester, UK: John Wiley & Sons, Ltd.

Everitt, B. S. (1980). *Cluster analysis* (2nd ed.). London: Heineman Educational Books Ltd.
Fang, Y., & Wang, J. (2012). Selection of the number of clusters via the bootstrap method, *Computational Statistics and Data Analysis*, *56*, 468–477.
Frayley, C., & Raftery, A. (1998). *How many clusters? Which clustering method? Answers via model-based cluster analysis*, Technical report 329, University of Washington, Department of statistics.
Gordon, A. D. (1981). *Classification*. New York: Chapman and Hall.
Hartigan, J. A. (1975). *Clustering algorithms*. New York: John Wiley & Sons, Inc.
Heinen, T. (1993). *Discrete latent variable models*. Tilburg University Press, Tilburg.
Heinen, T. (1996). *Latent class and discrete latent trait models*. Thousand Oaks, CA: Sage Publications.
Huberty, C. J. (1994). *Applied discriminant analysis*. New York: John Wiley & Sons, Inc.
Huberty, C. J., DiStefano, C., & Kamphaus, R. W. (1997). Behavioral clustering of school children. *Multivariate Behavioral Research*, *32*, 105–134.
Kamphaus, R. W., Huberty, C. J., DiStefano, C., & Petoskey, M. D. (1997). A typology of teacher rated child behavior for a national U. S. sample. *Journal of Abnormal Child Psychology*, *25*, 453–463.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors, Journal of the American Statistical Association, 90, 773–795.
Kaufman, L., & Rousseeuw, P. (1990). *Finding groups in data: An introduction to cluster analysis*, New York: Wiley.
Krzanowski, W. J., & Lai, Y. T. (1985). A criterion for determining he number of clusters in a data set, *Biometrics*, *44*, 23–34.
Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*, (2nd ed.). New York: John Wiley & Sons.
Lorr, M. (1983). *Cluster analysis for social scientists: Techniques for analyzing and simplifying complex blocks of data*. San Francisco: Jossey-Bass.
Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45*, 325–342.
Milligan, G. W. (1996). Clustering validation: Results and implications for applied analyses. In P. Arabie, L. J. Hubert, & G. De Soete (eds.) *Clustering and classification* , River Edge, NJ: World Scientific.
Milligan, G. W., & Cooper, M. C. (1987). Methodology review: Clustering methods. Applied Psychological Measurement, 11, 329–354.
Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, *24*, 882–891.
Pedhauzer, E. J. (1997). *Multiple regression in behavioral rsearch: Explanation and prediction* (3rd Ed). New York: Wadsworth Publishers.
Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior assessment system for children*. Circle Pines, MN: American Guidance Service, Inc.
Romesburg, H. C. (1984). *Cluster analysis for researchers*. Belmont, CA: Lifetime Learning Publications.
Sarle, W. S. (1983). *Cubic clustering criterion*, SAS Technical Report A-108. Cary, NC: SAS Institute.
Steinley, D. (2003). Local optima in K-means clustering: What you don't know may hurt you. *Psychological Methods*, *8*, 294–304.
Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology, 59*, 1–34.
Steinely, D. (2008). Stability analysis in K-means clustering. *British Journal of Mathematical and Statistical Psychology*, *61*, 255–273.
Stoddard, A. M. (1979). Standardization of measures prior to cluster analysis. *Biometrics*, *35*(4), 765–773.
Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a data set: An information-theoretic approach. *Journal of the American Statistical Association*, *98*, 750–763.
Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set using the gap statistic, *Journal of the Royal Statistical Society*, *63*, 411–423.
Ward, J. H. (1963). Hierarchial grouping to optimize an objective function. J*ournal of the American Statistical Association, 58*, 236–244.

LISA L. HARLOW & SUNNY R. DUERR

6. MULTIVARIATE ANALYSIS OF VARIANCE

With Discriminant Function Analysis Follow-up

INTRODUCTION TO MANOVA

Multivariate analysis of variance (MANOVA) allows an examination of potential mean differences between groups of one or more categorical independent variables (IVs), extending analysis of variance (ANOVA) to include *several* continuous dependent variables (DVs) (e.g., Grimm & Yarnold, 1995; Harlow, 2005; Maxwell & Delaney, 2004; Tabachnick & Fidell, 2013). As with ANOVA, MANOVA is a useful procedure whenever there are limited resources or when it is important to identify which groups may need specific treatments, interventions, or note. MANOVA can illuminate whether and how groups differ, and on which DVs.

In ANOVA a researcher can posit one or more categorical IVs, each with two or more groups, and one continuous DV. With MANOVA, the same structure of IVs can be considered except that *two or more DVs* are analysed. Hence, MANOVA allows for a much more realistic appraisal of group differences than does ANOVA. MANOVA can also be extended to incorporate one or more covariates, becoming a multivariate analysis of covariance (MANCOVA) that allows for one or more categorical grouping variables, one or more continuous covariates, and two or more continuous dependent variables. As will be seen later, MANOVA is closely related to the multivariate procedure of discriminant function analysis (DFA), which is mathematically equivalent but switches the roles of the independent and dependent variables. That is, DFA allows two or more continuous *IVs* and a categorical *DV.* Thus, in MANOVA researchers start with a focus on the categorical groups and ask how the means of several DVs differ. In contrast, with DFA researchers start with several (usually) continuous IVs and ask how these variables can help discriminate between the categorical groups of the DV. To preview analyses for an example later in the chapter, DFA is sometimes used as a follow-up procedure to a significant MANOVA, in order to investigate which of the continuous variables is differentiating among the groups. In what follows, we describe the basic purposes for MANOVA, along with the main equations needed, and how to assess the overall analysis with significance tests, effect sizes, confidence intervals, and a follow-up DFA. An example further illuminates the use of MANOVA and DFA.

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 123–143.

Description and Purpose of MANOVA

Several main purposes for which MANOVA is used are briefly described below.

MANOVA for an experimental design. The best use of MANOVA is when an IV is experimentally manipulated, and participants are randomly selected from a relevant population and then randomly assigned to groups. In this case, the goal is to assess whether the manipulated IV brought about or caused significant group differences between groups on a set of meaningful DVs. For example, a researcher could randomly assign students to an innovative phoneme training reading group or a standard educational reading group. At the end of the study, the researcher could examine the mean scores on reading comprehension, reading interest, and vocabulary between the two groups. A MANOVA would reveal whether there were any significant differences between the groups on a linear combination of the three DVs. Follow-up analyses (e.g., a set of ANOVAs, or a single DFA) could be conducted to determine which of the DVs were most clearly showing differences across reading groups. Differences could be attributed to the phoneme training if scores from that group were significantly higher.

MANOVA for a repeated measures design. MANOVA can be used to assess whether there are mean differences across time on a set of DVs. In this case, the IV is time and the groups are the various time points in which the data are collected on the set of DVs. For example, the reading researcher in the previous study may want to assess mean scores on reading comprehension, reading interest, and vocabulary across three time points (e.g., pre-test at the beginning of the school year, post-test at the end of the first semester, and follow-up at the end of the school year). In this example, time is the IV with levels representing the three separate time points, and the set of DVs is measured "k" (i.e., the number of levels or groups in the IV) times. This is also called a within-groups design as the analysis is assessed within a same group of participants, across time. It could also be referred to as a dependent MANOVA since the scores at each time point are dependent on the previous time point with the same (within-group) sample providing repeated measures across time. Regardless of how this design is named, a researcher could assess whether there were significant differences across time. If the phoneme training were successful, there should be significant differences between the pre- and post-test administered at the beginning and end of the first semester, respectively. If changes were long-term, there would be significant differences between the post-test and follow-up scores; and even possibly between the pre-test and follow-up scores collected at the beginning and end of the academic school year, respectively.

MANOVA for a non-experimental design. Although it is not ideal, MANOVA can be used to assess differences between two or more intact groups, on two or more DVs. For example, a reading researcher could examine whether there are differences

between two classrooms, one of which used phoneme training and the other that used standard reading training, on a set of DVs (i.e., reading comprehension, reading interest, and vocabulary). However, even if significant differences were found between the two classrooms, it would be impossible to attribute causality to the type of training, especially since the IV was not manipulated and participants were not randomly assigned to classrooms. In this design, it would be very difficult to control for all possible confounds that could be explaining differences in the DVs. For instance, classrooms may have differed as to initial reading level, basic intelligence, socioeconomic status, and amount of reading in the home, to name a few. If a researcher was fairly sure that this set of four potentially confounding variables were the most important considerations outside of the type of training in comparing across classrooms, these variables could be assessed as covariates with a MANCOVA. This would provide some degree of control and probably elevate the study to a quasi-experimental design, although results would still not be as definitive as in an experimental design with random assignment to groups. Nonetheless, this form of non- or quasi-experimental MANOVA is sometimes used, with results then interpreted more descriptively than inferentially.

The Main Equations for MANOVA

The main equation to describe the nature of a *DV* score, *Y*, for MANOVA is:

$$Y_i = \mu_{yi} + \tau + E \tag{1}$$

where Y_i is a continuous DV, μ_{yi} is the grand mean of the *i*th DV, τ is the treatment effect or group mean, and E is error.

For MANOVA, another equation is needed to reflect that linear combinations of the continuous DVs are being formed before examining group differences. Thus,

$$V_i = b_1Y_1 + b_2Y_2 + \ldots + b_pY_p \tag{2}$$

where V_i is the *i*th linear combination, b_i is the *i*th unstandardized weight, and Y_i is the i^{th} DV.

When there are more than two groups in a MANOVA more than one linear combination can be formed. The number is determined by:

$$\#\text{ of } V_i\text{'s} = \text{minimum } (p, k - 1), \tag{3}$$

where p is the number of continuous variables, and k is the number of groups or levels of the IV. When there are only two groups, only one linear combination can be formed (i.e., $k - 1 = 2 - 1 = 1$) no matter how many dependent variables are included in a design. This will be the case in the example later in the chapter.

In MANOVA, even though there may be one or more linear combinations, each with a specific set of weights (i.e., the "*b*" values in equation 2), the weights

and linear combination(s) are not a point of focus until conducting DFA, which is discussed later in the context of an example. To preview, the linear combinations in equation 2 are called *discriminant functions* in DFA, and for that analysis the weights are of prime importance. For now, know that in MANOVA, the focus is on modelling mean differences in the DVs, across the groups of the IV. Similar to what occurs in ANOVA, a ratio of between-group variance over within-group variance is formed in MANOVA, except that now the ratio involves variance-covariance matrices. The between-group variance-covariance matrix could be labelled B, although to distinguish it from the unstandardized b weights in the linear combination, this matrix is often labelled as $\boldsymbol{H}$ for the "Hypothesis" matrix. The within-group variance-covariance matrix is often labelled as $\boldsymbol{E}$ to represent error variance (Harris, 2001). Whereas in ANOVA there is just a single dependent variable in which to delineate between- and within-group variance, in MANOVA we need to focus on "p" sets of variances, one for each DV, as well as $p(p - 1)/2$ covariances among the p DVs. We store the between-group variances and covariances in the $\boldsymbol{H}$ matrix and the pooled within-group variances and covariances in the $\boldsymbol{E}$ matrix.

Thus, in MANOVA, another equation of interest is the ratio of the between-groups variance-covariance matrix over the error variance-covariance matrix:

$$\boldsymbol{H} / \boldsymbol{E} = \boldsymbol{E}^{-1}\,\boldsymbol{H} \tag{4}$$

Those familiar with matrix operations will realize that $\boldsymbol{E}^{-1}$ refers to the inverse of the divisor matrix, $\boldsymbol{E}$, which is multiplied by the dividend matrix, $\boldsymbol{H}$. Subsequently, it will become apparent that one of the challenges in conducting a MANOVA is considering different ways of summarizing this ratio of matrices with a single number that can be assessed for significance with an F-test. With ANOVA, where there is only one DV, there is just a single number to indicate the ratio of between- over within-group variances. In MANOVA, however, this ratio involves two matrices, which after multiplying the inverse of the $\boldsymbol{E}$ matrix by the $\boldsymbol{H}$ matrix still results in a matrix, and not a single number such as an F in ANOVA. Drawing on matrix operations and features, several methods are suggested shortly to summarize a matrix (e.g., $\boldsymbol{E}^{-1}\,\boldsymbol{H}$) with a single number. One method involves finding a *determinant*, which is a generalized variance of a matrix that can summarize how different the information is in a matrix. If all of the variables are essentially the same, the determinant will be very small, indicating that there is very little variation to assess within the matrix. Thus, it is important to choose dependent variables that are at least somewhat different, and IV groups that are fairly different from each other in order to provide a reasonable size determinant. Another matrix method for summarizing a matrix is a trace that represents the sum of the diagonal elements in a matrix. For the $\boldsymbol{E}^{-1}\,\boldsymbol{H}$ matrix mentioned earlier, the sum of the diagonals will refers to the sum of variances for this product matrix. Still another way to summarize a matrix is to calculate eigenvalues, which are the variances of the linear combinations of a matrix. Referring back to equation 3, there will be as many eigenvalues as there are linear combinations in

MANOVA (or DFA). To review, there will only be one linear combination of the continuous variables, and thus, one eigenvalue when there are just two groups for the categorical IV in MANOVA (or the categorical DV in DFA). For those interested in more information about matrices, see a 96-page book by Namboodiri (1984); Chapter 6 in Harlow (2005); or Appendix A in Tabachnick & Fidell (2013).

For now, it important to realize that there are various ways to summarize the matrix formed by the ratio of between- over within-matrices in equation 4 for MANOVA. Just as with ANOVA, it is important to focus on this ratio of the variance between means over the variance within scores in MANOVA. If this ratio is large, the null hypothesis of no significant differences between means can be rejected. Let's see more about how this is done by considering various ways of specifically summarizing between- and within-group information in MANOVA.

Overall Assessment for MANOVA

Just as with ANOVA, MANOVA results should be interpreted first at a macro or omnibus level. At this level, the first focus is on determining whether there is a significant macro-level group-difference. In addition, MANOVA is concerned with an overall shared variance effect size, as well as with which DVs are showing significant differences across groups, both of which are presented shortly.

Several macro-assessment summary indices have been offered to summarize the matrix results for MANOVA, borrowing on the matrix summary values suggested earlier. Wilks' (1932) Lambda, which uses determinants to summarize the variance in the ratio of matrices formed in MANOVA, is probably the most widely used macro-assessment summary index. Wilks found it difficult to calculate the between-groups matrix, specifically, due to computational limitations at that time. Instead, he suggested that the determinant of the within-groups variance-covariance matrix over the determinant of the total (i.e., within plus between) variance-covariance matrix indicates how much of the variation and covariation between the grouping variable(s) and the continuous variables was *unexplained*. Thus, one minus Wilks' Lambda is a measure of the shared or explained variance between grouping and continuous variables. Two other macro-assessment summary indices incorporate the trace of a variance-covariance matrix to summarize group difference matrices. Hotelling's trace is simply the sum of the diagonal elements of the matrix formed from the ratio of the between-groups variance-covariance matrix over the within-groups variance-covariance matrix. Pillai's trace is the sum of the diagonal elements of the between-groups variance-covariance matrix over the total (i.e., between plus within) variance-covariance matrix. A fourth macro-assessment summary is Roy's Greatest Characteristic Root (GCR: Harris, 2001). The GCR is actually the largest eigenvalue from the between over within variance-covariance matrix, providing a single number that gives the variance of the largest linear combination from this matrix.

Below, we delineate further how to assess the initial MANOVA macro-level information, focusing on suggested criteria (e.g., determinant, trace, or eigenvalue)

for summarizing the ratio of some form of the between-groups matrix (i.e., ***H***) over within-groups matrix (i.e., ***E***), along with a significance test.

A Significance Test. Each of the four main macro summary indices just briefly introduced has an associated *F*-test for assessing whether group differences are significantly different from chance in MANOVA.

For *Wilks' Lambda*, showing the amount of variance in the linear combination of DVs that is not explained by the IVs, low values (closer to zero than 1) are best. However, the associated *F*-statistic should be large and significant in order to conclude that there are significant differences between at least two groups on the linear combination of DVs. Wilks' Lambda can be calculated as the determinant of the ***E*** matrix, divided by the determinant of the sum of the ***H*** and ***E*** matrices:

$$\Lambda = | \boldsymbol{E} | / | \boldsymbol{H} + \boldsymbol{E} | \tag{5}$$

where | | stands for the determinant of the matrix inside the parallel lines (See a matrix algebra book or computer program to find the determinant of a matrix).

The second macro summary index, the *Hotelling-Lawley trace*, is formed by summing the diagonal elements in the $\boldsymbol{E}^{-1}\boldsymbol{H}$ matrix as given below.

$$\text{Hotelling-Lawley trace} = tr\,[\boldsymbol{E}^{-1}\boldsymbol{H}] \tag{6}$$

The Hotelling-Lawley trace can also be calculated as the sum of eigenvalues of the $\boldsymbol{E}^{-1}\boldsymbol{H}$ matrix. The reason these are equivalent is because the sum of the eigenvalues of a matrix is equal to the sum of the diagonal values of the original matrix. For both methods, which summarize the essence of the variance of the $\boldsymbol{E}^{-1}\boldsymbol{H}$ matrix, an associated *F*-test indicates whether there is a significant difference between means on the linear combination(s); and thus large *F*-values are preferred.

Pillai's trace, the third macro summary index, is the sum of the diagonal values of the matrix product of ***H*** times the inverse of ***E*** + ***H*** matrices, as given below.

$$\text{Pillai's trace} = tr\,[(\boldsymbol{H} + \boldsymbol{E})^{-1}\boldsymbol{H}] \tag{7}$$

Similar to the Hotelling-Lawley trace, Pillai's trace can also be formed from the sum of the eigenvalues of the $[(\boldsymbol{H} + \boldsymbol{E})^{-1}\boldsymbol{H}]$ matrix. As with the other indices, the associated *F*-test for Pillai's trace should be large and significant to indicate that there are significant differences on the means for the linear combination(s). An advantage of Pillai's trace is that it is the most robust of the four summary indices when there are less than ideal conditions, such as when there is unequal sample size across groups or heterogeneity of variances. In an example presented later, Pillai's trace will be preferred due to unequal sizes in the IV groups, and an indication of significant heterogeneity for variance-covariance matrices. Another benefit of Pillai's trace is that it can be interpreted as the proportion of variance in the linear combination of DVs that is explained by the IV(s). Thus, it is intuitively meaningful.

The fourth macro summary index, *Roy's largest root or the greatest characteristic root (GCR)*, is a single value simply represented as given below.

$$\text{GCR} = \text{the largest eigenvalue of } \boldsymbol{E}^{-1}\boldsymbol{H} \tag{8}$$

As with the other indices, a large and significant *F*-test is preferred for the GCR, again indicating that there are significant differences across groups on the means of the linear combination(s). Aside from recommendations to use GCR by Harris (2001), the GCR is not used as often as other indices, particularly Wilks' lambda and Pillai's trace, the former used most often, probably due to being introduced before the others, and the latter due to its robustness with non-standard conditions.

Effect Size. A common multivariate effect size for MANOVA is Eta-squared:

$$(\eta^2) = (1 - \Lambda), \tag{9}$$

where η^2 represents the proportion of variance in the best linear combination(s) of DVs that is explained by the grouping IVs, and Λ represents Wilks' Lambda (see equation 5). Eta-squared (i.e., η^2) can be interpreted with multivariate guidelines for shared variance effect sizes (e.g., Cohen, 1992). Thus, a small multivariate shared variance effect size would be equal to about .02, a medium effect size would equal .13 or better, and a large effect size would be greater than or equal to about .26 or more.

If the macro-level *F*-test is significant in MANOVA and there is a reasonable effect size, there are one or two more layers to interpret. Just as with ANOVA, this could involve micro-level significance tests of specific group differences if there are more than two IV groups, and effect sizes for group means. But first, it is important to conduct a "mid-level" evaluation of the dependent variables.

Follow-up Analyses after a Significant MANOVA

After finding a significant macro-level *F*-test in MANOVA and summary criteria (e.g., Wilks' lambda, Pillai's trace, etc.), it is important to assess which DVs are significantly showing mean differences. Follow-up analyses can take one of several forms, described below.

Separate ANOVAs for each DV. Probably the most common follow-up to a significant MANOVA is to conduct a separate ANOVA for each DV. Researchers would hope to find a significant *F*-test for each DV, indicating that these variables each show significant differences across two or more groups. Although ANOVAs are widely conducted after finding significant MANOVA results, a drawback is that separate ANOVAs do not take into account whether the DVs are related in any way. That is, analysing the DVs separately could mislead researchers into thinking there is a large cumulative effect across the DVs and groups, which is most likely not an accurate picture if DVs are related. Thus, it may be preferable to consider

other follow-up procedures that recognize any overlap among the DVs. A set of ANCOVAs, one for each DV, or a single DFA, offer alternative follow-up options.

Separate Analyses of covariance (ANCOVA)s for each DV. ANCOVA is just like ANOVA, except that it allows for the inclusion of one or more continuous covariates. With ANCOVA, mean differences across groups on an outcome variable are assessed after partialling out the relationship between covariates and the DV. This allows for a more fine-tuned assessment of group differences. Thus, a better follow-up than ANOVA is to conduct a separate ANCOVA for each dependent variable, using the remaining dependent variables as covariates in each analysis. This has been suggested by Bock (1966; Bock & Haggard, 1968) and is considered a step-down procedure. If these analyses revealed significant *F*-tests, it would suggest that there were significant group differences on a dependent variable after partialling out any overlapping variance among the remaining continuous dependent variables used in the MANOVA. Thus, group differences would be revealed for the unique portion of each dependent variable that is distinct from any relationship with other dependent variables. This provides a rigorous assessment of group differences although it is not often seen in the literature, possibly due to unfamiliarity with this option, and the difficulty in finding significant differences with ANCOVA on such small, unique portions of the dependent variables.

Discriminant function analysis follow-up. Another possible follow-up procedure after a significant macro-level F-test with MANOVA, is to conduct a single DFA with the same variables that were used in the MANOVA except that the roles (independent or dependent) are reversed. Thus, a DFA would use each of the continuous (*dependent*) variables from a MANOVA as the continuous *independent* variables. The categorical (*independent*) grouping variable from MANOVA would now become the categorical *dependent* variable in DFA. The goal would be to assess how each of the continuous variables discriminated among the groups of the DFA outcome variable. The standardized weights or the structure coefficients would be the focus in DFA, such that continuous variables with large standardized weights or structure coefficients would also be the variables that have notable group differences on the categorical variable. In this way, we could assess which of the continuous variables are showing the clearest differences across groups without having to conduct separate (ANOVA or ANCOVA) analyses for each dependent variable. Thus, the overall error rate is most likely smaller with a single DFA follow-up than with *p* follow-up ANOVAs or ANCOVAs, especially if the error rate was not adjusted (as with a Bonferroni approach). Moreover, the multivariate nature of DFA would take into account any relationship among the continuous variables, providing a more precise depiction of group differences than is portrayed when conducting a set of individual ANOVAs that do not correct for shared variance between the set of variables.

Follow-up planned comparisons. When there is a significant effect of a DV and there are more than two groups in the IV(s), it is advisable to assess which pair(s) of

groups showed significant differences on a DV. Tukey (1953) Honestly Significant Difference (HSD) tests between pairs of means would provide some protection for overall Type I error (i.e., rejecting Ho when it is true), particularly if there were several groups and a large number of paired comparisons were conducted. Another alternative, a Bonferroni approach, could be adopted whereby the total alpha is split among the number of pair-wise group tests (e.g., 4 tests could each use an alpha of .0125 to maintain an overall .05 alpha). Still another possibility is to increase statistical power and reduce the probability of a Type II error (i.e., retaining Ho when it is false) by using an alpha level of .05 for all comparisons. Researchers need to decide for themselves which error is more important to protect, Type I or Type II, when assessing between group differences.

Follow-up effect sizes. If ANOVAs or ANCOVAs are conducted for each DV, following a significant MANOVA, an η^2 or omega-squared (i.e., ω^2) univariate effect size could be calculated for each DV to assess how much variance was shared between that specific continuous variable and the grouping variable(s). Computer packages sometimes refer to η^2 values as R^2, which is the same value. Cohen's (1992) guidelines for univariate effects would apply for any of these: .01 for a small effect, .06 for a medium effect, and about .13 or more for a large effect. For MANOVA, Cohen's *d* can also provide a micro-level effect size for the difference between a pair of means (e.g., Cohen, 1988, 1992), just as with ANOVA. This is easily calculated by a difference between means in the numerator and a pooled or average standard deviation in the denominator. By Cohen's guidelines, a standardized *d* or difference of .20 is a small effect, a *d* of .50 is a medium effect, and .80 or more represents a large effect.

Additionally, just as with univariate ANOVAs, group means on DVs can be graphed after a significant MANOVA; alternatively, boxplots can be provided that pictorially display what is called the "five-number summary" (i.e., maximum, 75th percentile or 3rd quartile, the median called the 50th percentile or Q2, the 25th percentile called Q1, and the minimum). Most computer programs easily allow for these. As boxplots convey a clear visual depiction of a set of specific indices for each dependent variable, across groups, these are presented later in the MANOVA example introduced below.

AN ILLUSTRATIVE STUDY WITH MANOVA

An example is provided to illustrate how to conduct a MANOVA and follow-up DFA, along with supplemental analyses. The example draws on data collected from 265 faculty at a New England university to assess work environment (Silver, Prochaska, Mederer, Harlow, & Sherman, 2007), with a National Science Foundation institutional transformation grant (No. 0245039: PI: Barbara Silver; CO-PIs: Lisa Harlow, Helen Mederer, Joan Peckham, and Karen Wishner) to enhance careers of all faculty, particularly women in the sciences.

For the analyses presented, the independent grouping variable is gender, which is somewhat evenly split (i.e., 55% men and 45% women). Three continuous and reliable variables (with coefficient alpha internal consistency reliability given in parentheses for each) – Career Influence (coefficient alpha = .83), Work Respect (coefficient alpha = .90), and Work Climate (coefficient alpha = .93) – allow examination of a three-tier conceptual structure of individual, interactional, and institutional variables, respectively (Risman, 2004). These three variables, all averaged composite scores on a 1 to 5 Likert scale, serve as DVs in the MANOVA example, and conversely as IVs in the follow-up DFA.

In addition, scatterplots, and separate boxplots are presented for each of these three variables, across the two gender groups, in order to further explore relationships and reveal group differences for these three variables. Moreover, computer set-up for three packages – SPSS, SAS, and R – is presented to provide researchers with several options for conducting MANOVA and related analyses.

Preliminary Analyses before Conducting MANOVA

Before conducting a MANOVA, it is important to assess basic descriptive statistics (e.g., mean, standard deviation, five-number summary, skewness, kurtosis, correlations), as well as scatter plots in order to evaluate assumptions of normality, homoscedasticity, and linearity, and any possible collinearity (i.e., high correlation or redundancy) among variables. Table 1 shows descriptive statistics on the three continuous variables, Career Influence, Work Respect, and Work Climate.

Table 1. Descriptive statistics on three continuous variables

Statistic		*Career Influence*	*Work Respect*	*Work Climate*
N	Valid	265	265	265
	Missing	0	0	0
Mean		2.599	3.873	3.814
Standard Deviation		.702	.836	.881
Skewness		–.085	–.698	–.754
Std. Error of Skewness		.150	.150	.150
Kurtosis		–.267	–.161	.220
Std. Error of Kurtosis		.298	.298	.298
Minimum		1.000	1.270	1.000
Percentiles	25 = Q1	2.162	3.333	3.292
	50 = Q2	2.556	4.037	3.944
	75 = Q3	3.056	4.540	4.486
Maximum		4.050	5.000	5.000

Means are higher than the mid-point (i.e., 3) of the 5-point scales for Work Respect and Work Climate, indicating relatively high overall scores for these two variables. The mean for Career Influence (i.e., 2.599) is lower; suggesting that appraisal of one's individual influence was somewhat less than the interactional, as well as the institutional appraisal of Work Respect and Work Climate, respectively. Standard deviations were similar, and much smaller than the respective means, suggesting fairly consistent scores within this sample, for each of these variables. Skewness and kurtosis, which should be around zero in normally distributed data, indicate that the normality assumption appears reasonable for these data. In the bottom portion of Table 1 are five-number summaries for the three variables. Notice that the lower 50% of the scores (i.e., from the minimum to the median or 50th percentile) cover a broader range of scores (i.e., 1.27 to 4. 037; and 1.0 to 3.944) than the top range of scores for Work Respect and Work Climate, respectively. This pattern suggests somewhat uneven distributions for these variables. Five-number summaries are depicted in boxplots, later, for men and women, separately, to illuminate potential group differences on these variables.

Figures 1 to 3 show scatterplots for the three variables. These allow further examination of how well assumptions are met. When data meet assumptions of normality, linearity and homoscedasticity, scatterplots should reveal a fairly even elliptical pattern of points. When data are nonnormal, the pattern of points may be bunched up in one end or the other indicating some evidence for skewness or kurtosis. If data are not completely linear, the pattern shows some curve indicating that after a certain point, the relationship between a pair of variables changes from linear to non-linear, thereby reducing the linear correlation. Similarly, if the pattern showed a wider range of points at either end of the scatterplot, heteroscedasticity would be present, indicating that individuals who have low scores on one variable may tend to

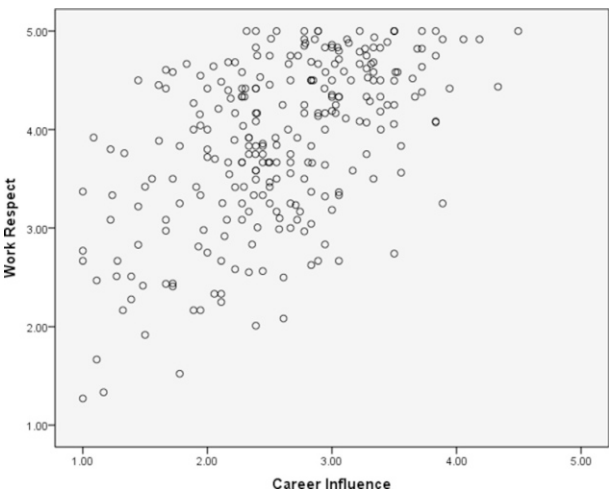


Figure 1. Scatterplot of career influence and work respect.

have a wide range of scores on the other variable. None of these patterns would be optimal as MANOVA, similar to many statistical methods, is more accurate when the data follow a normal, linear and homoscedastic pattern. Examining Figure 1, the scatterplot appears to follow rather closely the preferred elliptical pattern, with no obvious deviations from normality, linearity or homoscedasticity for the relationship between Career Influence and Work Respect.

Figures 2 and 3 scatterplots for Work Climate, with Work Respect and Career Influence, respectively, are reasonable but do not seem to follow an elliptical pattern

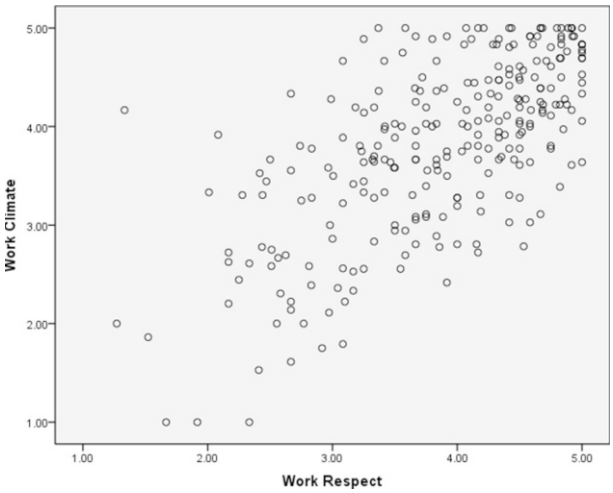


Figure 2. Scatterplot of work climate and work respect.

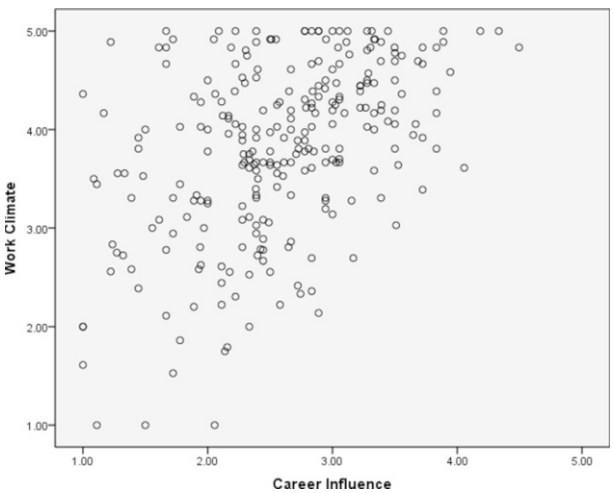


Figure 3. Scatterplot of work climate and career influence.

quite as clearly. Points are more densely located near the upper right-hand corner of both figures, with some possible evidence of outliers in the bottom left, and possibly the upper left corners of both figures. It is noteworthy that both figures 2 and 3 involve the variable,Work Climate, which may have scores that are not as consistent as for the other two variables. This speculation is later confirmed when examining boxplots for men and women, separately, for the three variables; as well as statistical tests of homoscedasticiy. As can be seen shortly, there are a few outliers for Work Climate, among the men, yielding some heterscedasticity. When conducting the MANOVA and DFA, it would probably be advisable to examine Pillai's trace, instead of Wilks' lambda, for these data as Pillai's trace is more robust to assumption violations than the other methods. Pillai's trace would also be preferred due to the slightly unequal Gender groups for this example.

Correlations among the four variables are also examined to assess any possible collinearity among the variables. Table 2 shows that none of the variables are correlated extremely highly (i.e., greater than .90, or even .70). Thus, there is no concern that collinearity is present for these variables.

Overall Results for MANOVA

A MANOVA was conducted to examine whether there were significant differences between Gender groups on a linear combination of the three-tier set of variables: Career Influence, Work Respect, and Work Climate. Analysis set-up for SPSS, SAS, and R are provided in the Appendix for the major analyses. As part of a MANOVA or DFA, researchers can request Box's test of equality of covariance matrices. If the data were to meet the assumption of homoscedasticity, this test result would be non-significant, indicating that there was no indication of significant heteroscedasticity. Unfortunately, however, the F-test in this case was significant [$F(6, 419736.37) = 2.61$, $p = .016$], suggesting some degree of violation of this assumption. This is further confirmed with results from Levene's test of equality of error variances showing significant results for Work Respect [$F(1, 255) = 8.147$, $p = .005$], and Work Climate [$F(1, 255) = 6.396$, $p = .012$], indicating some heterogeneity for these two variables. Based on findings from these two sets of tests, as well as those from the scatterplots shown earlier, Pillai's trace will be evaluated for the F-test for the overall MANOVA as it is more robust to violations.

Table 2. Correlation among the four variables

	Gender (1 = f, 2 = m)	*Career Influence*	*Work Respect*	*Work Climate*
Gender	1.000	.206	.234	.163
Career Influence	.206	1.000	.546	.460
Work Respect	.234	.546	1.000	.668
Work Climate	.163	.460	.668	1.000

Pillai's trace was .063 (so that $\eta^2 = .063$), with $F(3, 252) = 5.71$, $p = .001$ for the MANOVA analysis on these data. This finding indicates that the means of the linear combination of the three continuous variables are significantly different across Gender groups (with scores for men being somewhat higher). Using Steiger and Fouladi's (2009) R^2 program (freely available on the web), confidence intervals for a shared variance effect can be calculated. Results revealed a 95% confidence interval of [.012, .123], indicating a small-to-medium, and significant, shared variance effect between Gender and the set of three continuous DVs.

As part of the MANOVA output, most computer programs provide follow-up ANOVAs, one for each DV. Although our focus is largely on a follow-up DFA, it is worthwhile to briefly examine ANOVA results for these data (see Table 3).

Notice that there are relatively small and significant group difference effects for each of the three dependent variables, with Work Respect having the largest effect.

Further analysis, with DFA, will reveal whether this pattern of results is verified.

Follow-up Results with DFA

Macro-level SAS and R results for DFA are virtually identical to those for MANOVA, with $F(3, 253) = 5.71$, Pillai's trace = $\eta^2 = .063$, $p = .0009$. SPSS gives a chi-square test with comparable results: χ^2 (3) = 16.05, (1-Wilks' Lambda) = $\eta^2 = .063$, $p = .001$. The structure coefficients, which are within-group correlations between the standardized discriminant function (a form of equation 2) and the three continuous variables, reveal values of .807, .927, and .637 for Career Influence, Work Respect and Work Climate, respectively. These results parallel those from conducting individual ANOVAs, with Work Respect showing the largest, and Work Climate showing the smallest effect with Gender. Thus, it is to be expected that there are somewhat larger group differences for Work Respect, followed by those for Career

Table 3. Tests of ANOVA for each of the three dependent variables

Source	*Dependent Variable*	*Type III SS*	*df*	*Mean Square*	*F*	*Sig.*	R^2	*95% CI For* R^2
Corrected	Influence	5.305	1	5.305	11.251	0.001	.042	[.007, .102]
Model	Respect	9.899	1	9.899	14.819	<.001	.055	[.013, .120]
	Climate	5.417	1	5.417	6.995	0.009	.027	[.002, .079]
Error	Influence	120.241	255	.472				
	Respect	170.335	255	.668				
	Climate	197.490	255	.774				
Corrected	Influence	125.546	256					
Total	Respect	180.234	256					
	Climate	202.908	256					

Influence, and lastly those for Work Climate having the smallest difference between Gender groups. DFA provides additional information from MANOVA, showing that the discriminant function formed from these three continuous variables was able to correctly classify participants into their respective Gender groups 59.5% of the time, which is greater than the 50% chance level. Next, let's examine boxplots to further investigate Gender group differences on the three continuous variables (i.e., Career Influence, Work Respect & Work Climate).

Boxplots as Further Follow-up to MANOVA

Boxplots were constructed to visually depict the five-number summary by Gender, plus any outliers for the three continuous variables (i.e., Career Influence, Work Respect, Work Climate; depicted in dark to light gray, respectively in Figure 4).

The upper and lower most points are the maximum and minimum estimated scores, with Work Climate showing several outliers below the minimum of most scores for the men. The boxes delineate the 75th, 50th and 25th percentiles, respectively, with slightly more distinct differences between Gender groups for Work Respect than for the other variables. Notice that the scores are more spread out for the women, particularly for Work Climate and Work Respect; with scores for Career Influence also showing some spread for both men and women. For all three variables, men faculty scored somewhat higher than the women faculty.

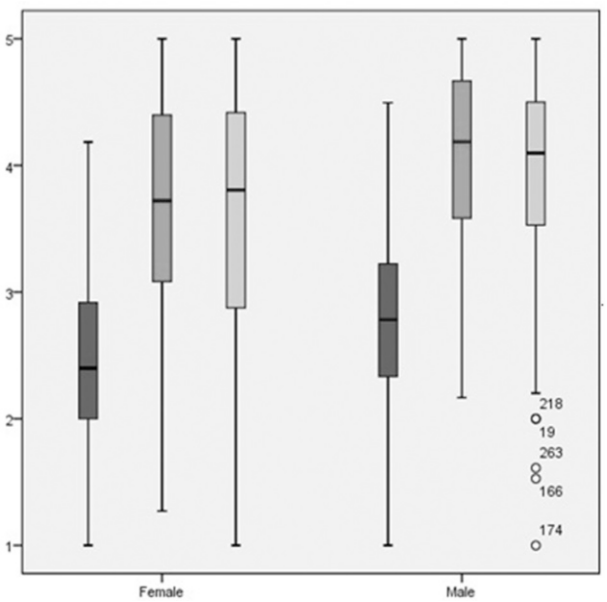


Figure 4. Boxplots for three variables, by gender.

SUMMARY

In conclusion, MANOVA and a follow-up DFA were described, and then applied to a relevant example to investigate group differences on a set of relevant variables. Whenever possible, significance tests, effect sizes, confidence intervals and figures were presented to provide a fuller picture of the findings. Results revealed small- to medium-size significant group-difference effects, with slightly higher means for men compared to the women faculty on a set of work environment variables. Other analyses, including descriptive statistics, correlations, scatterplots, boxplots, and ANOVAs helped to convey the nature of the data and group differences. The reader should recognize that analyses were conducted on data from intact groups, and thus cannot warrant the causal conclusions allowed for an experimental design in which participants are randomly assigned to treatment and control groups. Still, the example presented here provides a useful illustration of how to examine group differences on a set of relevant dependent variables, with interpretation based more descriptively, than inferentially. It should also be noted that although group differences were significant overall, and for each dependent variable, none of the effects were very large. This is actually good news, suggesting that gender differences are not very consequential in this sample of faculty, regarding individual career influence, an interactional sense of work respect, and an institutional evaluation of work climate. It would be useful to verify results on larger and more diverse, independent samples. The Appendix briefly describes syntax that can be used to conduct MANOVA, DFA, and related analyses; using SPSS, SAS, and R computer packages.

REFERENCES

Bock, R. D. (1966). Contributions of multivariate experimental designs to educational research. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology.* Chicago: Rand McNally.

Bock, R. D., & Haggard, E. A. (1968). The use of multivariate analysis of variance in behavioural research. In D. K. Whitla (Ed.), *Handbook of measurement and assessment in behavioral sciences.* Reading, MA: Addison-Wesley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* San Diego, CA: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Grimm, L. G., & Yarnold, P. R. (1995). *Reading and understanding multivariate statistics.* Washington, DC: APA.

Harlow, L. L. (2005). *The essence of multivariate thinking: Basic themes and methods.* Mahwah, NJ: Erlbaum.

Harris, R. J. (2001). *A primer of multivariate statistics* (3rd ed.). Mahwah, NJ: Erlbaum.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

Namboodiri, K. (1984). *Matrix algebra: An introduction.* Beverly Hills: Sage.

Risman, B. J. (2004). Gender as a social structure: Theory wrestling with activism. *Gender & Society, 18,* 429–450.

Silver, B., Prochaska, J. M., Mederer, H., Harlow, L. L., & Sherman, K. (2007). Advancing women scientists: Exploring a theoretically-grounded climate change workshop model. *Journal of Women and Minorities in Science and Engineering, 13,* 207–230.

Steiger, J. H., & Fouladi, R. T. (2009). *R2 user's guide, version 1.1.* University of British Columbia.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.
Tukey, J. (1953). *Honestly significant difference (HSD) test*. Unpublished manuscript, Princeton University.
Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika, 24*, 471–494.

SUGGESTIONS FOR FURTHER READING

Der, G., & Everitt, B. S. (2008). *A handbook of statistical analyses using SAS* (3rd ed.). Boca Raton, FL: Chapman & Hall/ CRC.
Everitt, B. S. (2007). *An R and S-Plus companion to multivariate analysis.* London: Springer.
Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage.
Hancock, G. R., & Mueller, R. O. (Eds.) (2010). *The reviewer's guide to quantitative methods in the social sciences.* New York: Routledge.
Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (2nd ed.). New York: Wiley.
Kabacoff, R. I. (2011). *R in action: Data analysis and graphics with R.* Shelter Island, NY: Manning.
Leech, N. L., Barrett, K. C., Morgan, G. A. (2011). *IBM SPSS for intermediate statistics*, (4th ed.). New York: Routledge.
Pallant, J. (2011). *SPSS survival manual*, Berkshire, England: McGraw Hill.
Revelle, W. (in preparation). *An introduction to psychometric theory with applications in R.* New York: Springer.
Slaughter, S. J., & Delviche, L. D. (2010). *The little SAS book for enterprise guide (4.2).* Cary, NC: SAS Institute.
Stevens. J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York: Routledge.
Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S,* (4th ed.*).* New York: Springer.

APPENDIX

Syntax for SPSS, SAS and R for Conducting

Descriptive Statistics, Correlations, Scatterplots, Boxplots, MANOVA, and DFA

For the following computer set-ups, the variable names are abbreviated as follows: Influenc = Career Influence, Respect = Work Respect, Climate = Work Climate, and sex1f2m = Gender (where 1 = female and 2 = male). Note also that although there were 265 participants in the sample, analyses that included the variable Gender (i.e., sex1f2m) only had 257 participants as gender was not given for 8 individuals. The data set used in analyses was labelled: Adv04.sav in SPSS, Adv04 in SAS, and Adv04dat in R. It should also be noted that different statistical analysis programs may produce slightly different solutions due to program-oriented differences in calculation procedures and rounding (e.g., values may differ at the 2nd or 3rd decimal). The output provided by the syntax below should yield similar inferences regardless of the software used, despite minor differences in reported values. Readers may also need to check with Google for more up-to-date syntax.

SPSS Syntax

```
GET FILE='C:\Users\User\Desktop\Adv04.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.

FREQUENCIES VARIABLES=Influenc Respect Climate
/FORMAT=NOTABLE
/NTILES=4
/STATISTICS=STDDEV MINIMUM MAXIMUM MEAN SKEWNESS SESKEW
KURTOSIS SEKURT
/ORDER=ANALYSIS.

CORRELATIONS
/VARIABLES=sex1f2m Influenc Respect Climate
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE.

GRAPH /SCATTERPLOT(BIVAR)=Influenc WITH Respect.
GRAPH /SCATTERPLOT(BIVAR)=Influenc WITH Climate.
GRAPH /SCATTERPLOT(BIVAR)=Respect WITH Climate.

EXAMINE VARIABLES=Influenc Respect Climate BY sex1f2m
/COMPARE VARIABLE
/PLOT=BOXPLOT
/NOTOTAL
/MISSING=LISTWISE.

GLM Influenc Respect Climate BY sex1f2m
/METHOD=SSTYPE(3)
/INTERCEPT=INCLUDE
/EMMEANS=TABLES(sex1f2m)
/PRINT=ETASQ HOMOGENEITY
/CRITERIA=ALPHA(.05)
/DESIGN= sex1f2m.

DISCRIMINANT
/GROUPS=sex1f2m(1 2)
/VARIABLES=Influenc Respect Climate
/ANALYSIS ALL
/PRIORS EQUAL
/STATISTICS=MEAN STDDEV UNIVF BOXM CORR TABLE
/CLASSIFY=NONMISSING POOLED.
```

SAS Syntax

DATA Adv04; INFILE 'SASUSER.Adv04'; /*Gets datafile Adv04 in Sasuser */

PROC MEANS N MEAN STD SKEWNESS KURTOSIS MIN Q1 Median Q3 MAX; VAR Influenc Respect Climate; **RUN**;

PROC CORR; VAR Sex1f2m Influenc Respect Climate; **RUN**;

PROC GPLOT; PLOT Influenc*Respect; /* This runs a scatterplot */
PLOT Influenc*Climate; PLOT Respect*Climate; **RUN**;

PROC Sort; By sex1f2m; /* Sort data by gender before conducting boxplot */

PROC Boxplot data = SASUSER.Adv04b; /* This runs a boxplot */
Plot (Influenc Respect Climate) * sex1f2m / BOXSTYLE=SCHEMATIC; **RUN**;

PROC GLM; /* This runs MANOVA: IV after class and DVs after Model */
CLASS sex1f2m; MODEL Influenc Respect Climate= sex1f2m;
LSMEANS sex1f2m /PDIFF CL; MANOVA H = _ALL_/SHORT; **RUN**;

PROC DISCRIM LIST CANONICAL MANOVA POOL=TEST WCOV;
CLASS sex1f2m; VAR Influenc Respect Climate; **RUN**;

R Syntax

Note that the MASS package in R (used for the DFA analysis) does not produce tests of significance for DFA. Because DFA is mathematically identical to MANOVA, the omnibus fit statistics for DFA must be obtained from the MANOVA procedures. It is also worth mentioning that R has multiple methods for producing similar results, often depending on which package is used (e.g., *psych* vs. *psychometric* for descriptive statistics).

```
# Load the following packages: car, psych, MASS
library(car)
library(psych)
library(MASS)
library(candisc)
# Read in the data and select variables from the larger data set
adv04dat = read.table("c:/Data/adv04na.txt", sep="","", header=TRUE)
myvars = c("sex1f2m", "Influenc", "Respect", "Climate")
work=adv04dat[myvars]
```

```
# Designate sex1f2m as a categorical variable called "gender"
work$gender=factor(work$sex1f2m, levels = c(1,2), labels = c("Female", "Male"))
attach(work)

# Produce descriptive statistics
describe(work)

# Produce correlation matrix
corr.test(work[1:4], use="pairwise.complete.obs")

# Produce a scatterplot matrix for the independent variables
scatterplotMatrix(~Influenc+Respect+Climate, diagonal="histogram",
data=work)

# Produce boxplots for the data
par(mfrow=c(2,2))
boxplot(Influenc~sex1f2m, main="Boxplot of Influence by Gender",
xlab="Gender", col="aquamarine")
boxplot(Respect~sex1f2m, main="Boxplot of Respect by Gender", xlab="Gender",
col="lightgreen")
boxplot(Climate~sex1f2m, main="Boxplot of Climate by Gender", xlab="Gender",
col="khaki ")

# MANOVA for 3 dependent variables and 1 independent variable
Y = cbind(work$Influenc, work$Respect, work$Climate)
faculty.mod = lm(Y~gender, data=work)
faculty.can1 = candisc(faculty.mod, term="gender", type="III")
Anova(faculty.mod, test="Wilks", type="III")
Anova(faculty.mod, test="Pillai", type="III")
Anova(faculty.mod, test="Hotelling-Lawley", type="III")
Anova(faculty.mod, test="Roy", type="III")
summary(faculty.can1, means = FALSE, coef="structure")

# Follow-up ANOVAs with R^2 reported
mydata.aov = Anova(aov(Influenc~gender), type="III"); mydata.aov
r <- summary.lm(aov(Influenc~gender)); r$"r.squared"
mydata.aov = Anova(aov(Respect~gender), type="III"); mydata.aov
r <- summary.lm(aov(Respect~gender)); r$"r.squared"
mydata.aov = Anova(aov(Climate~gender), type="III"); mydata.aov
r <- summary.lm(aov(Climate~gender)); r$"r.squared"

# Follow-up Discriminant Function Analysis (Requires the MASS package)
work.2 = na.omit(work)
```

```
dfa = lda(gender ~ Influenc + Respect + Climate, data=work); dfa
dfa = lda(gender ~ Influenc + Climate + Respect, data=work.2, CV=TRUE)

# Assess the predictive accuracy of the DFA
pred = table(work.2$gender, dfa$class)
diag(prop.table(pred, 1))
sum(diag(prop.table(pred)))
```

BRIAN F. FRENCH, JASON C. IMMEKUS & HSIAO-JU YEN

7. LOGISTIC REGRESSION

INTRODUCTION TO THE METHOD

Logistic regression (LR) is a statistical procedure used to investigate research questions that focus on the prediction of a discrete, categorical outcome variable from one or more explanatory variables. LR was developed within the field of epidemiology to examine the association between risk factors and dichotomous and continuous outcomes (Kleinbaum, Kupper, & Morgenstern, 1982; Tripepi, Jager, Stel, Dekker, Zoccali, 2011). Subsequently, the model has received extensive use across disciplines. In the medical domain, for example, LR has been used to identify predictors of Alzheimer's disease. In business settings, it has been employed to determine the most important factors (e.g., ease of use) for internet banking usage (Hassanuddin, Abdullah, Mansor, & Hassan, 2012). In education research, the method has been used to investigate predictors of college student persistence in engineering (French, Immekus, & Oakes, 2004). The purposes of this chapter are to describe the LR model in the context of education research and provide a real data illustration of its use to obtain results with theoretical and practical implications. The information is presented to promote the technical and practical understanding of the method.

LR is distinguishable from multiple linear regression analysis due to the fact that the (a) dependent variable is categorical in nature (e.g., group membership), not continuous, and (b) the model assumes a nonlinear relationship between the outcome and explanatory variables. Within educational research, examples of discrete outcome variables include: presence or absence of a learning disability, exceeding or not exceeding minimum proficiency requirements on an end-of-grade English Language Arts test, or being accepted or not to an institution of higher education. Whereas the outcome variable in a LR analysis can be either dichotomous (e.g., pass/fail) or ordinal (e.g., *Far Below Basic*, *Basic*, *Proficient*), for didactic purposes, this chapter focuses exclusively on instances in which the dependent, or outcome variable is binary or dichotomous.

The application of LR in educational research can be exemplified by the following research questions:

1. What student and institutional factors can be used by colleges and universities to determine the likelihood of a student earning a college degree?
2. To what extent are different academic counselling strategies effective for promoting at-risk students' attainment of a high school diploma?

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 145–165.

3. What factors are associated with exceeding minimum state requirements on an end-of-grade test among English Language Learners?

Research utilizing LR focuses on the prediction of a dichotomous dependent variable based on a set of independent variables. Potential outcomes could include: earning a college degree (0 = no degree, 1 = degree), attainment of a high school diploma (0 = no diploma, 1 = diploma), or exceeding minimum state requirements on an end-of-grade test (0 = did not exceed requirement, 1 = exceeded requirement). Within the LR model, predictor variables can be continuous, dichotomous, categorical, or a combination (Hosmer & Lemeshow, 1989, 2000; Tabachnick & Fidell, 2007). Consequently, LR results yield empirical evidence that can have relevant and substantive implications to research, practice, and policy.

The LR procedure seeks to gather empirical evidence on the predictive nature of a set of explanatory variables to account for the variance of an outcome variable much like multiple linear regression (MLR) and discriminant analysis (DA; Cizek & Fitzgerald, 1999; Davis & Offord, 1997). However, LR differs from these two procedures in important ways. Specifically, compared to MLR, (a) the independent variables specified in LR can be dichotomous, categorical, and/or continuous, (b) the relationship between the explanatory variables and the outcome is nonlinear, and (c) parameter estimation is based on maximum likelihood (ML) procedures, not ordinary least squares (OLS). Furthermore, applying MLR to predict a binary outcome also violates basic MLR model assumptions.

LR and DA can both be used to predict a categorical outcome and have been compared in terms of classification accuracy (Cleary & Angel, 1984; Fan & Wang, 1999). Comparatively, Cleary and Angel (1984) noted that "discriminant analysis yields results quite similar to logistic regression except when the probability of the event being predicted is near zero or one" (p. 341). In terms of explanatory predictors, DA is restricted to the use of continuous variables that are multivariate normal (Tabachnick & Fidell, 2007). Also, like MLR, OLS is used for parameter estimation which can result in biased estimates when the data do not meet model assumptions (e.g., multivariate normality). Fan and Wang (1999) compared the performances of LR and DA for two-group classification and, in general, found that the methods performed similarly across simulated conditions (i.e., unequal proportions, unequal group covariances, and sample size). Regardless of the approach to data analysis, researchers should understand the characteristics of the data at hand to guide the selection of an appropriate statistical analysis to address their research question(s). That said, LR has been recognized as an alternative to DA (Fan & Wang, 1999; Tabachnick & Fidell, 2007).

The aim of LR, as in many statistical models, is to identify a set of theoretically and empirically relevant explanatory variables that can be used to develop a parsimonious model to predict a dichotomous outcome. The effectiveness of LR results to address a given research question depends on many important substantive factors (e.g., group proportions, variables in the model, sample) and methodological

issue (e.g., variable selection, assumptions) that should be considered throughout data analysis. To encourage the application of LR, the chapter begins with an overview of the LR model, corresponding model assumptions, and interpreting model data fit and parameter estimates. This is followed by an illustrative study based on real data. The chapter concludes with a discussion of research issues and areas for research within LR.

The Logistic Regression Model

LR represents a model-based approach to predict an individual's or intervention group's (e.g., after-school program participation) standing on a binary outcome, such as: exceeded minimum passing score, did not exceed minimum passing score. In this case, the outcome variable Y can be assigned a value of 0 if the individual or group did not exceed minimum passing score, whereas a value of 1 would indicate otherwise (i.e., exceeded minimum passing score). In this instance, the quantitative values of 0 and 1 serve as dummy coded variables to represent the qualitative outcome variable of interest in the analysis. In this case, the outcome variable was whether or not students exceeded the minimum passing score. The outcome variable can be on a nominal (e.g., likelihood to participate in an after-school program vs. non-program participation) or ordinal (e.g., mathematics knowledge based on exceeding passing score [pass] vs. not exceeding score [fail]) level of measurement. Therefore, the specific intent of the analysis is to accurately classify individuals on the outcome of interest within the research question.

The classification of individuals on the outcome variable is addressed by determining the probability of the outcome occurring for each individual and group conditional on their standing across model predictors. Mathematically, this is represented by the following equation characterizing the nonlinear relationship between the outcome and explanatory variables:

$$P(Y = 1) = \frac{e^u}{1 + e^u}. \quad (1)$$

The left-hand side of the equation, $P(Y = 1)$, indicates that the outcome Y is operationalized in terms of the probability (P) of its occurrence (value equals 1.00). Because the likelihood of the event occurring is expressed as a probability, its value will range between 0 and 1.00. Therefore, to predict the probability of the occurrence of an outcome, LR uses ML for parameter estimation that maximizes the function that relates the observed responses on the independent variables to the predicted probabilities likelihood estimation. The use of ML leads to more accurate conclusions when a nonlinear relationship exists between the binary outcome and explanatory variables compared to OLS regression (Lottes, Adler, & DeMaris, 1996). The use of OLS under these conditions is inappropriate because the assumptions of the linear

model are not met. Specifically, the use of OLS in the prediction of probabilities is problematic because the values are not constrained to the range of zero and one. The right-hand side of the equation expresses this probability in terms of taking the base *e* of the natural logarithm *u*, which includes the linear set of explanatory variables. (Notably, *e* is approximately equal to 2.718.) Let's explore this in the context of an applied example.

In the model, *u* is defined as:

$$u = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k, \tag{2}$$

where *u* is the predicted outcome, β_0 is the intercept, and β_1, β_2, ... β_k are slope parameters indicating the linear relationship between the outcome and explanatory variables. This linear regression equation shows the direct correspondence of LR to MLR. This also includes the use of hierarchical or stepwise (e.g., backward, forward) selection procedures to identify statistically significant predictor variables. The endorsement of the hierarchical model building approach is offered as best practice as models are built up or trimmed down based on theory in comparison to statistical results (Thompson, 1995).

Notably, this linear equation creates the logit, *u*, or log odds of the event occurring, expressed as:

$$\text{logit}(u) \equiv \ln\left(\frac{\hat{u}}{1-\hat{u}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k. \tag{3}$$

As indicated in Equation (3), *u* is the natural log (log_e) of the probability of being assigned to the group with a value equal to one divided by the probability of being in the group equal to zero (Tabachnick & Fidell, 2007). This value is also referred to as the *logit*, and provides a continuous indicator of the linear association between the outcome and explanatory predictors. Its values can range from $-\infty$ to ∞, depending on the measure of *X* (Hosmer & Lemeshow, 2000). As such, no matter how it is expressed, the probability, the odds ratio, or the log odds ratio, the same information is obtained.

An example is used to situate the discussion on the use of LR in applied educational research. Specifically, one of the research questions investigated by French, Immekus, and Oakes (2005) was the extent to which the following explanatory variables predicted persistence in one's major across two cohorts (Cohort 1, Cohort 2) of undergraduate engineering students following eight and six semesters in the program: gender (males = 0, female = 1); high school rank; SAT Mathematics and Verbal scores; cumulative grade point average (CGPA); motivation; and, institutional integration; and, participation in a freshman orientation course (no = 0, yes = 1)

Hierarchical LR (HLR) was used to examine the extent to which variable clusters added to the explained variance in persistence in one's major. HLR is the preferred modelling approach when the research question seeks to address the extent to which a group of variables contribute to the explained variance of the outcome after

accounting for the variance attributed to covariates (e.g., student demographics, prior achievement). Study variables were added into the model in the following order: Background variables (i.e., gender; high school rank; SAT Mathematics [SAT_{Math}]; SAT Verbal [SAT_{Verbal}]; CGPA), motivation and institutional integration, and participation in freshman orientation course. Ordering of the variables was based on occurrence in time and practical reasons (e.g., covariates). Specification of variables in theoretically derived "blocks" assists with controlling Type I error rates (Cohen & Cohen, 1983). Within the study, Cohort 1 data was used to identify significant model predictors to obtained parameters to cross-validate results using Cohort 2 data. We use these results to walk the reader through the basics of the model.

Logistic Regression Model Assumptions

The goal of LR is to correctly classify individuals on the outcome variables based on one or more explanatory variables. As with most statistical procedures, there are several assumptions associated with the use of LR in educational research. Tabachnick and Fidell (2007) identify several important factors (e.g., *linearity in logit*) to consider in the application of LR. Consideration of such assumptions is critical to effectively use the model to address one's research question to obtain results that can be used for meaningful decision-making (e.g., program planning/ evaluation).

To begin, the LR model is used to predict an individual or groups' categorical membership on the outcome variable, and thus the primary model assumption is that the independent variables occurred before the outcome. That is, data collected on the explanatory variables was completed before data were gathered on the outcome variable *Y*. Or, more simply, the independent variables represent attributes of the individual or group (e.g., gender, prior academic achievement) that would be hypothesized to impact one's categorical membership on the outcome (e.g., obtain high school degree). Among the sample of undergraduate engineering students, information on their demographics, noncognitive beliefs (e.g., motivation), and participation in a freshman orientation course was obtained prior to their decision to persistent within their major.

Tabachnick and Fidell (2007) identify several technical assumptions of the LR model. As indicated in Equation (2), it is assumed that that there is a linear relationship between the continuous explanatory and outcome variables, while no assumption is made on nature of the relationship (linear) among the explanatory variables. To contrast with DA, LR does not assume that the predictor variables follow a multivariate normal distribution with equal covariance matrix across the dependent variable at all levels. In contrast, LR assumes that the binominal distribution describes the distribution of errors that is equal to the difference in the observed and predicted responses on the dependent variable. Second, like MLR, the predictor variables should not be strongly correlated. Correlation coefficients can be used to examine the direction and strength among continuous variables,

whereas multiway frequency tables can be used to examine relationships among categorical variables. Third, there should an adequate number of observations across combinations of discrete variables, and group membership on the outcome variable should not be perfect based on discrete explanatory variables. The latter could occur in the example above if all students who participated in the freshman orientation course persisted in their engineering major, whereas all of those who did not enroll in the orientation course matriculated to non-engineering majors.

There are other notable assumptions that researchers should be aware. Specifically, the joint expected cell frequencies for all pairs of variables exceed a value of one, whereas no more than 20% of the variable pairs contain less than five (Tabachnick & Fidell, 2007). The consequence of limited frequencies across joint pairs of cells between discrete variables is reduced statistical power. Thus, statistical power is increased with larger sample sizes. Each individual's group membership on the binary outcome is independent from one another is a key assumption. This results in the assumption of independent errors, or that discrepancy in students' actual and predicted persistence in an engineering major does not depend on the outcomes of other students in which data has been collected. Lastly, initial data screening should be conducted to identify potential outliers that may influence results. Observations that fall outside of range of typical values can be identified through inspection of the descriptive statistics of the variables included in the model or residuals between the observed and predicted values of *Y*. This follows the typical data screening procedures for many statistical methods.

Collectively, these considerations address the primary assumptions of the LR model. The extent to which one's obtained data meets model assumptions is based on careful inspection of results based on initial data screening to obtain descriptive statistics (e.g., frequency distributions), as well as results based on the analysis of the data. There are many useful resources that can be used to assist with decisions regarding the extent to which model assumptions have been met (e.g., Hosmer & Lemeshow, 1989, 2000; Manard, 1995; Tabachnick & Fidell, 2007).

Interpreting Logistic Regression Model Data Fit and Parameter Estimates

The use of LR in educational research requires consideration of the assessment of model-data fit and interpretation of model parameters. The fit of a specified LR model is evaluated in terms of the log-likelihood statistic. Parameter coefficients (β) are estimated using ML to determine the value that "most likely" produces the observed outcome (e.g., group membership). These values must be interpreted and used accurately to avoid erroneous conclusions based on the data.

The log-likelihood statistic provides a measure of model-data fit by adding the probabilities of the observed and predicted of each individual included in the analysis. The statistic is estimated iteratively based on the parameters included in the model until a convergence criterion has been obtained (e.g., <.0001). Therefore, the number of log-likelihood values reported in the output of an analysis will equal the number

of model parameters. Multiplying the log-likelihood statistic by –2 provides a basis to compare two competing LR models that differ by the number of model predictors. This is referred to as the likelihood ratio chi-square difference statistic ($x^2_{Difference}$), which is distributed as chi-square with degrees of freedom (df) equal to difference in the number of parameters between the compared models ($df_{difference}$).

The comparison of two competing LR models is advantageous in educational research to identify a parsimonious set of explanatory variables that most accurately predict group membership on the outcome. This requires the estimation of two models: constrained and free. The *constrained model* represents a restricted LR model that includes a limited set of predictors (e.g., constant-only model, one predictor variable). The *free model* is one in which additional parameters (or predictors) have been included in the model, such as the addition of a covariate in the model (e.g., prior academic achievement).

The chi-square difference value is estimated by:

$$X^2_{Difference} = -2log-likelihood_{Constrained} - (-2log-likelihood_{Free}) \quad (4)$$

Based on the $df_{Difference}$, the statistical significance of can be determined by comparing it to the critical values of the chi-square distribution associated with a pre-determined level of significance (e.g., $p > .05$) for hypothesis testing. Within the context of comparing competing LR models using the likelihood ratio chi-square difference test ($x^2_{Difference}$), the null hypothesis is that the two models do not differ statistically in predicting group membership on the outcome variable. That is, the constrained model ($x^2_{Constrained}$) with a reduced number of explanatory variables is equally effective for predicting one's standing on the outcome as a model that includes one or more predictors. This is concluded if the probability value associated with exceeds the significance level (e.g., $p > .05$). Contrary, the alternative hypothesis is that the models differ statistically and the additional model parameters included in the free model (x^2_{Free}) improve the predictive utility of the model above and beyond the constrained model. One would accept the alternative hypothesis is the probability value corresponding to the statistic is less than the determined significance level (e.g., $p < .05$).

The log-likelihood and the likelihood ratio chi-square difference statistics were used to evaluate the model data fit of competing models for predicting undergraduate students' retention in an engineering major. Before presenting results, it should be noted that multiple imputation (Enders, 2010) was used to estimate five scores for the missing Cohort 1 and 2 data. See information below on the importance of appropriately handling missing data. This resulted in five independent regression analyses on each data set and averaging parameter estimates.

For the model predicting engineering students' retention in their major, the first model included the Block 1 predictors of academic achievement and gender, which was statistically reliable, $X^2(5) = 96.31$, $p < 0.05$. The second model included the Block 2 variables of motivation and integration which resulted in a statistically significant model across the imputed data sets, $X^2_{Range}(7) = 103.73 - 105.79, ps < 0.05$. The third step in

the hierarchical LR model included the Block 3 variable of orientation course participation, which was statistically significant, $X^2_{Range}(8) = 104.54 - 106.18, ps < 0.05$.

Based on the model-data fit of each of the hierarchical regression analysis models, chi-square difference tests were used to empirically test whether the inclusion of the Block 2 and 3 variables resulted in a reliable improvement in prediction accuracy. The difference between models 1 and 2 was statistically significant at the .05 level, $X^2_{Difference}(2) = 8.80, p < 0.05$. This provides empirical evidence to support the rejection of the null hypothesis that students' levels of motivation and integration are collectively significant predictors of persistence in an engineering major. Subsequently, there was no statistical improvement with the addition of the final model with the Block 3 predictor of freshman orientation course, $X^2_{Difference}(1) = 0.74, p > 0.05$. Therefore, the null hypothesis that after accounting for student background variables (achievement, gender) and noncognitive self-beliefs (e.g., motivation), participation in a freshman orientation course did not contribute to predicting whether a student would persist in an engineering major.

Inspection of the statistical significance of model parameters provides a basis to determine the association between explanatory and outcome variables. As indicated, ML is used for model parameter estimation. Model parameters provide a basis to determine the (a) individual influence of explanatory variables on the outcome, and (b) probability of an individual being classified on the outcome variable (e.g., persisting in engineering major).

The relationship of each explanatory variable to the outcome variable is determined by its corresponding beta coefficient (β). The coefficient is interpreted as the log odds of the outcome occurring based on a one-unit change in the explanatory variable. More technically, "the slope coefficient represents the change in the logit for a change of one unit in the independent variable x" (Hosmer & Lemeshow, 1989, p. 39). Importantly, the curvilinear relationship of the predictors and outcome variables results in different likelihood of an individual being categorized on the outcome variable based on standing on the predictor variables. In the provided example, the significant (unstandardized) model predictors ($p < .05$, with coefficients), were: CGPA ($\beta = .788$); SAT_{Math} ($\beta = .005$); HS Rank ($\beta = .017$); and, motivation ($\beta = .447$). Non-significant parameters were: SAT_{Verbal} ($\beta = -.001$); Gender ($\beta = -.138$); Integration ($\beta = .187$) and, Orientation Class ($\beta = .146$). In consideration of the impact of CGPA on students' persistence in an engineering major, one would conclude that a one-unit increase in CGPA is associated with a .79 change in the log odds of persisting in an engineering major. Positive parameter coefficients are associated with a positive change in the log odds of persisting, whereas a negative coefficient would suggest a decrease in the log odds of persistence with a higher value on the predictor variable (e.g., CGPA).

The statistical significance of parameter coefficients is estimated using the Wald statistic (Hosmer & Lemeshow, 2000). The statistic is estimated as:

$$Wald = \frac{B_j}{SE(B_j)} \tag{5}$$

where, β is the estimated parameter coefficient and $SE(\beta)$ is the parameter's standard error. Based on a two-tailed p-value, the statistic is used to identify significant predictor variables.

Whereas parameter coefficients are reported by the log odds, it is more common to use the odds ratio to communicate the association between the explanatory and outcome variables. The odds ratio is estimated by taking the exponential of the log odds estimate, Exp(β). The odds ratio is interpreted as the odds of the outcome occurring based on the unit change in the predictor variable. The odds ratio is centered around 1.00, which indicates that there is no association or odds of the outcome occurring (e.g., persisting in an engineering major) based on changes in the explanatory variable. Thus, odds ratios greater than 1.00 indicate the odds of the outcome's occurrence given on a one-unit change in the predictor variable, whereas a value less than 1.00 being indicative of the decreased chance of the outcome occurring.

The odds ratio of the significant model predictors of persistence in an engineering major can be readily estimated. For instance, the odds ratio of persisting in an engineering major, based on a one-unit increase in one's CGPA is 2.19 (median value based on imputation). This indicates that undergraduate students are 2.19 times more likely to persist within an engineering major based on a one-unit increase in their cumulative GPA. Although SAT_{Math} was a significant model predictor, the odds ratio was 1.00, indicating no increase in the odds of persisting in one's engineering major with a one-point scale score increase. A similar finding was reported for HS Rank. However, a one-unit increase in motivation was associated with being 1.52 times more likely to persist in one's engineering major.

Another utility of LR model parameters is estimating each individual's predicted probability of group membership. This is conducted using Equation 1 by inserting estimated model parameters and the individual's scores. The equation for a hypothetical student with a set of predictor scores would be:

$$P(Y=1)=\begin{cases}\dfrac{e^{-4.09+(.788)3.45+(-.001)600+(.005)675+(.017).91+(-.138)0+(.447)4.25+(.187)3.80+(.146)0}}{1+e^{-4.09+(.788)3.45+(-.001)600+(.005)675+(.017).91+(-.138)0+(.447)4.25+(.187)3.80+(.146)0}} \\ \dfrac{e^{4.03}}{1+e^{4.03}}\end{cases} \quad (6)$$

The predicted probability is .98, indicating that an undergraduate student with this particular set of predictor values has an extremely high (almost 1.00) probability of persisting in an engineering major. Contrary, a student with the following set of scores would be identified as having a .91 probability of persisting in an engineering major: CGPA = 2.75; SAT_{Verbal} = 550; SAT_{Math} = 600; HS Rank = .85; Gender = 0; Motivation = 3; Integration = 2.67; and, Orientation = 0.

An important aspect of predicted probabilities is that they can be compared to a classification criterion (probability .50 or greater) to classify individuals on the dependent variable. That is, an individual with a predicted probability that exceeds

a designated criterion is assigned to the outcome with value of 1.00 (e.g., pass), otherwise to the group equal to 0.0 (e.g., failed). The predicted and observed classifications can be used to estimate a host of descriptive statistics pertaining to classification accuracy (e.g., false negative, sensitivity). Based on the hierarchical LR model used to predict persistence in an engineering major, the classification rate of Cohort 1 was 65%, with a 64% cross-classification rate for Cohort 2.

Preliminary Considerations

LR is a model-based approach to determining the extent to which a set of explanatory variables predict membership dichotomous outcome variable. As such, the aim is to specify a parsimonious model that yields results relevant to guide subsequent decision-making endeavors (e.g., research, policy). This and other factors support the need to conduct preliminary data screening procedures prior to testing LR models, and assess indicators of model-data fit (e.g., Pearson residuals).

Data screening routinely begins with using descriptive statistics to understand the characteristics of the data at hand. In general, this entails inspecting the features of the variables that will be included in the LR model. Depending on the level of measurement of the variables, frequency distributions (e.g., histograms) can be examined for distributional characteristics (e.g., skewness, kurtosis), extreme cases, and potential data entry errors. Measures of central tendency (e.g., mean) and variability (e.g., standard deviation) can be used to summative information on the continuous variables. It is surprising of the amount of information that can be obtained about the data at hand based on the thoughtful inspection of the descriptive statistics of variables.

Sample size is a concern for statistical modeling in general. Parameter estimates in logistic regression are general more stable with larger samples. Long (1997) has recommended samples greater than 100 to accurately conduct significance tests for the coefficients. That said, some research areas employing LR for a specific purpose have provide guidelines about sample sizes required for adequate power given the outcome. Simulation research on the use of LR to detect differential item functioning, for example, has suggested the need for approximately 250 persons per group to have adequate statistical power (e.g., French & Maller, 2007). It may be wise to conduct a power analysis to be certain you have adequate power and be familiar with the standards for sample size for power in your relative field. Power analysis can be conducted in SAS or via a freeware program (G*Power) available for free (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3). In our illustrative example we have over 400 participants with only 6 predictors so we can have some confidence in the parameter estimates.

There are many diagnostic statistics available to evaluate the fit of a LR model. Largely, these measures of model-data fit are based on the difference between the observed and predicted values. Two indices used to identify individual observations not accounted for by the model include the Pearson residuals and deviance residuals.

The Pearson residuals are estimated for each individual and used to calculate the Pearson chi-square statistic (sum of squares of Pearson residuals). The Pearson chi-square statistics is distributed as a chi-square distribution, based on *df* equal to $J - (p + 1)$, where J = number of unique values of the independent variables x (Hosmer & Lemeshow, 1989). A statistically significant Pearson chi-square statistics indicates an observation that is not well accounted for by the model. The deviance residual is another index that can be used to identify unique observations not explained by the model. Like Pearson residuals, deviance residuals are calculated for each individual, and summed to yield a summary statistic, distributed chi-square.

The Hosmer-Lemeshow goodness-of-fit- statistic (Hosmer & Lemshow, 1989, 2000) is another statistic that can be used to evaluate model-data fit. The statistic is based on the grouping of individuals in a contingency table based on their predicted probabilities. Each row corresponds to one of the binary outcomes (e.g., fail = 0, pass = 1), whereas columns represent the predicted probabilities of group member. Two grouping strategies have been proposed. The first approach is grouping individuals according to their predicted probabilities by percentiles. This results in 10 groups that are ranked by the classification of individuals by their predicted probabilities. The second approach classifies individuals based on cut points along the probability scale. Thus, dividing the probability continuum by .20 to establish groups would result in five groupings, with group one comprised of all individuals with predicted probabilities below .20 and the highest group being those with values above .80. A Pearson chi-square statistic is then used to estimate the statistical difference between the observed and expected cell counts. Hosmer and Lemeshow (1989) discuss the ways in which the statistics function under varying conditions.

Other diagnostic statistics can be used to examine the impact of observations on model parameter estimates and the general LR model. For instance, the $\Delta\hat{B}_j$ statistic can be used to inspect the standardized discrepancy between model parameter estimates, $\hat{B}_j$, based on the inclusive and exclusive of a particular observation. Values can be used to identify observation(s) suspected of influencing the resultant ML estimate. The impact of observations on the overall fit of a specified model can be inspected using the $\Delta_j\hat{D}$ statistic. The value reports the difference in the Pearson chi-square statistic based on models with and without the observation included in the analysis. Despite the availability of diagnostic statistics to evaluate LR models, these statistics should be used with caution. Specifically, Hosmer and Lemeshow (1989, 2000) report that the evaluation of model-data fit is largely subjective. That is, interpretation of these values is commonly done using plots with the statistics reported on the *y*-axis and predicted probabilities on the *x*-axis (see Hosmer & Lemeshow, 1989, 2000). These steps of basic data screening and assumption checking in LR that have been just described do parallel what the reader has likely experienced in standard regression models. We do not spend time reviewing these in the example due to this reason. Instead we focus on understanding the model and output for interpretation. See Lomax and Hahs-Vaughn (2012) for a clear example of LR model assumption and diagnostic checking.

AN ILLUSTRATIVE STUDY

For illustrative purposes we will use an example drawn from the student achievement literature. We provide the example in SAS 9.2 but the same analysis could be conducted in many statistical software programs including SPSS, R, and Stata. Explanations are provided for the SAS code and the shorter data set below can be used to replicate analysis even though results will differ due to the smaller dataset . The reader is encouraged to create the SAS code and attempt to replicate the results presented in this chapter to increase proficiency in software use and understanding of the method. We provide the first 15 cases in the dataset here to allow the reader to work replicate the analysis even though the estimates will not be exact.

Data

The data used for the illustrative example are drawn from grade 6 students attending a middle school in the western part of the United States. The outcome, or dependent, variable of interest is the end-of-grade performance on a state achievement test for English Language Arts. This outcome variable was coded for each student as being "Proficient" (coded as 1) or being "Not Proficient" (coded as 0). The independent, or predictor, variables included: three interim English Language Arts test scores from assessments that are administered periodically throughout the school year (e.g., fall, winter, and spring); sex (female = 0, male = 1); socio-economic status measured by free or reduced lunch status of the student (SES, 0 = No, 1 = Yes); and, fluency in the English language (LngPrt, coded 1–4 for the 4 classification levels). The levels of LngPrt included: 1 = English Language Learner, 2 = English Language, 3 = Initially

Table 1. Descriptive statistics for the 6th grade achievement data

Variable Name	*Percentage*			
Proficient	*50.53(Yes)*	*49.47(No)*		
Free/reduced lunch(SES)	*17.80(Yes)*	*82.80(No)*		
Sex	*49.47(Male)*	*50.53(Female)*		
LngPrt[b]	*15.03%(ELL)*	*66.13(EO)*	*5.86(I-FEP)*	*12.79(R-IEP)*
	M	*SD*	*Minimum*	*Maximum*
End of Course Exam[a]	347.25	47.20	224	485
Interim Assess 1 (18 items)	10.57	3.35	1	18
Interim Assess 2 (25 items)	15.71	4.25	2	25
Interim Assess 3 (20 items)	11.80	3.68	2	19

[a] This variable was transformed into the Proficient variable based on the state cutscore for proficiency.

[b] ELL = English Language Learner, EO = English Language, I-FEP = Initially Fluent English Proficient, R-FEP = Reclassified Fluent English Proficient

Fluent English Proficient, 4 = Reclassified Fluent English Proficient. Table 1 contains descriptive information for each of these variables. Table 2 contains 15 rows of data on these variables reflecting 15 students in the dataset.

Dealing with Missing data

At the onset of data analysis, attention should be paid to the assumptions as mentioned in the previous section. However, attention to robustly handling missing data is given here as it is a topic that is often overlooked. Standard statistical analyses are designed for data sets with all variables having no missing values. However, in practice it is common to have missing values regardless of the effort that was placed on data collection processes and accuracy. For example, participants completing a survey on parenting practices may inadvertently skip questions by not noticing questions printed on the back side of a page or choose not to respond to certain sensitive questions. The common and easy solution to this problem is to ignore the missing data by simply removing the complete case or row which contains any missing variable (i.e., listwise deletion). This practice, complete case analysis, is the default for many statistical software programs (e.g., SAS). However, such missing data techniques rely on critical assumptions (e.g., data missing completely at random

Table 2. Sample data on 15 students on variables analysed

Student	*Proficient*[a]	*SES*[a]	*Sex*[b]	*Language Proficiency*[c]	*Interim 1*	*Interim 2*	*Interim 3*
1	0	1	0	1	8	14	14
2	0	1	1	1	5	10	4
3	1	1	0	4	5	11	10
4	0	1	0	1	10	13	11
5	0	0	0	2	6	11	8
6	0	1	1	1	9	12	16
7	1	0	0	2	12	18	12
8	0	1	0	1	5	8	2
9	1	0	1	2	6	15	10
10	1	0	0	2	8	16	16
11	1	0	0	4	8	16	14
12	1	0	1	2	14	14	16
13	1	0	0	2	12	16	12
14	1	0	1	2	13	20	16
15	1	0	1	2	11	10	11

[a] 0 = No, 1 = Yes; [b]1 = Male, 0 = Female; [c]1 = English Language Learner, 2 = English Language, 3 = Initially Fluent English Proficient, 4 = Reclassified Fluent English Proficient

(MCAR). Violation of these assumptions can invalidate the results of corresponding analysis (Allison, 2001).

In educational research, a review of published articles in major journals (e.g., *Child Development, Educational and Psychology Measurement*) that indicated having missing data, only 2.6% employed maximum likelihood (ML) or multiple imputation (MI) procedures (Peugh & Enders, 2004), which are considered state-of-the-art methods. There are many excellent sources for reviews of missing data (e.g., Enders, 2010; Graham, Cumsille, & Elek-Fisk, 2003; Little & Rubin, 2002; Rubin 1987; Schafer & Graham, 2002). The point made here is that older methods (e.g., mean replacement, regression imputation) do not function well especially when compared to new procedures (e.g., MI and ML). For instance, mean imputation retains means, but distorts marginal distributions and measures of covariation (Little & Rubin, 2002; Schafer & Graham, 2002).

The trend of older method use is changing as the general acceptance of newer missing data methods are being widely accepted (Graham, Taylor, Olchowski, & Cumsille, 2006). Specifically, ML and MI techniques have grown in popularity due to (a) support demonstrating the production of accurate and efficient parameter estimates under various assumptions (e.g., MAR, MCAR; Allison, 2003; Schafer & Graham, 2002) and (b) having worked their way into many software programs. We encourage the reader to view the sources in the suggested readings and carefully consider the missing data in your dataset as well as the options that are available when the software package used for analysis. It may be the case that additional programs or add-ons to programs will be needed to implement imputation.

Cross-validation

The idea with predictive models is to develop a model in a sample that can be used to predict the outcome at a later point in time with a new sample without having to rebuild the model that has been established. For instance, if we construct a model that predicts proficiency levels of students with desired accuracy (e.g., accurate classification rates), then one should be able to use those regression weights in a new independent and representative sample and obtain similar results. This answers the question of real interest: "How well does the model work in the population or on other samples from this population?" (Cohen & Cohen, 1983). This gathers evidence that the model is generalizable. Thus, in this example we construct a model on a random sample of half of the data. Then, we use the final model parameter estimates from that model to conduct the LR analysis with the second half of the sample to provide classification results based on obtained model parameter estimates to evaluate generalizability of the model.

Running the LR analysis in SAS

The below SAS code is example of code for running a hierarchal logistic regression where the first block of variables included in the model are sex, SES, and language

proficiency. This allows for an evaluation of background variables that may be important in predicting the outcome that would not allow for the building of an accurate model. The second logistic statement adds the interim assessment variables to the background variables to estimate and evaluate the complete model.

```
Data achieve;
Infile 'c:\pathtodatafile\achievement.dat';
Input ID 1–4 proficient 5 SES 6 Sex 7 LngPrt 8 @ 9 (Interim1–3) (2.);

        proc logistic data= achieve descending;
model Proficient = sex SES LngPrt / ctable pprob=.50 lackfit risklimits rsq;
run;

        proc logistic data= achieve descending;
model Proficient = sex SES LngPrt Interim1 Interim2 Interim3 / ctable pprob=.50
lackfit risklimits rsq;

run;
```

Formally, the full model being estimated is:

$$\begin{aligned} \textit{predicted logit}(\text{Proficient} = 1) = \beta_0 + \beta_1 Sex + \beta_2 SES + \beta_3 LngPrt \\ + \beta_4 interim1 + \beta_5 interim2 + \beta_5 interim3 \end{aligned} \quad (7)$$

The SAS code in the box is easy to follow with this model presented with a few short definitions of terms. The data statement establishes a temporary work file in SAS that is called achieve. It could be called anything (e.g., a1, LR). The infile and input statement tells SAS where to find the data file and what the variable names are and what column to find those variables in the data file. If the reader is unfamiliar with SAS, the help files can assist in establishing this or see a many SAS sources readily available on line or in print (e.g., Allison, 1999). The Proc LOGISTIC statement indicates to SAS to run a logistic regression analysis. The descending option tells SAS to predict the log odds of being proficient (1). If this option is not used SAS will predict the log odds of being not proficient (0) as 0 comes before 1. If your data are based on complex survey data, the Proc Survey Logistic statement can be invoked.

The SAS statement also request information on the (a) classification table (CTABLE) for probabilities at 0.50; (b) Hosmer and Lemeshow Goodness-of-Fit test (LACKFIT); (c) odds ratio for each variable in the equation with the 95% confidence limits (RISKLIMITS); and, (d) the value of r-squared (RSQ). We note there is little agreement on which measure of association to use. We report Nagelkerke R^2 which SAS labels Max-rescaled R^2. See Menard (2000) for a complete discussion of LR coefficients of determination. Due to a lack of clarity of the measure, the value can

be supplementary to other, more useful indices (e.g., overall evaluation of the model, tests of regression coefficients, and the goodness-of-fit test statistic).

Reporting the Results

In the first step (Step 1), a test of the model with the student background characteristics as predictors against a constant-only model was statistically reliable, X^2 (3, N = 480) = 66.47, $p > .05$, R^2 = 17.2%. At Step 2, the model with the three interim assessments included was statistically reliable, X^2 (6, N = 480) = 287.69, $p < .05$, R^2 = 60.1%. Beyond examination of each step, interest was on evaluating if the addition of variables improved the prediction or classification of proficiency status (0 = did not meet minimal standards, 1 = exceeded minimum standards). The statistical difference between Steps 1 and 2 was significant, $X^2_{difference}$ (3) = 221.22, $p < .05$. This indicates a reliable improvement with the addition of the interim assessment variables. All variables were significant predictors with the exception of sex, as seen in Table 3. Prediction using these variables resulted in a correct classification rate of 82%.

We also evaluated the Hosmer and Lemeshow Goodness-of-Fit test to help assess overall model fit. Recall, we do not want this test to be significant. We want to *not* reject the null hypothesis that these data fit the model. This is, in fact, what we observed, X^2 (8, N = 480) = 5.99, $p > .05$. The parameter estimates from the model with the significant variables from sample 1 were used to predict proficiency for Sample 2 (n = 458) and correct classification was examined for cross-validation purposes. Comparisons of predicted proficiency and observed proficiency for Sample 2 resulted in a correct classification rate of 81%. Classification rates of incorrect and correct decisions appear in Tables 4 and 5 for Sample 1 and 2, respectively.

Table 3. Summary of regression analysis for variables predicting proficiency for sample 1 (N = 480)

				95% Confidence Interval	
Measures	*B*	*SE B*	*Odds Ratio*[1]	*Lower*	*Upper*
1. Sex	–0.077	0.258	0.926	0.588	1.536
2. SES	–0.745*	0.363	0.474	0.233	0.968
3. LngPrt	0.332*	0.166	1.394	1.006	1.931
4. Interim 1	0.113*	0.048	1.120	1.018	1.232
5. Interim 2	0.276*	0.046	1.139	1.204	1.446
6. Interim 3	0.298*	0.051	1.348	1.220	1.490

[1]the odds ratio is the increase in the odds of the occurrence of an event with a one unit change in the independent variable.

*$p < .05$

Table 4. Classification table results for predicting student proficiency for sample 1 (N = 480)

	Result of Initial Model	
Proficient	*Positive*	*Negative*
Proficient	45%	10%
Not Proficient	37%	8%

Table 5. Classification table results for predicting student proficiency for sample 2 for cross-validation (N = 458)

	Result of Cross-Validation of Model	
Proficient	*Positive*	*Negative*
Proficient	38%	10%
Not Proficient	43%	9%

To aid understanding of the model, we build the probability of being proficient for student 1 in Table 2. Note the intercept for the model was −9.676. The equation would be:

$$\ln\left(\frac{\hat{u}}{1-\hat{u}}\right) = -9.9676 + (-.077)(0) + (-0.745)(1) + (.113)(8) + (.276)(14) + (.298)(14) \quad (8)$$

Therefore, taking the absolute value of this log odds, the odds of student 1 in Table 2 being proficient is:

$$\text{Odds(of being proficient)} = \exp(1.14) = 0.319.$$

And the probability of being proficient is, P (proficient) = 0.319 / 1 + 0.319 = 0.24. This low probability coincides with the observation that this student was actually labelled as not being proficient. We can also look at the odds ratios in Table 3 to understand how each variable influences the probability of being proficient. For example, it is clear that the odds of being classified as proficient increase by over 1 for every point a student gains on any of the interim assessments. Additionally, as SES is less than 1, it indicates that students receiving free and reduced lunch (e.g., lower SES compared to their counterparts) reduces the odds of being classified as proficient or at least having a score above the proficiency level.

The results from this illustrative study demonstrate that LR can be useful in educational research. The simple model we constructed here reveals that accounting for interim progress, language proficiency, and poverty do aid to predicting a student's probability of meeting a state standard in terms of being classified as proficient on this particular skill. The proposed model demonstrated fair accuracy with a correct classification rate of 82% which was validated using a cross-validation step in the analysis. Of course, as with any statistical modelling procedure, the model is only as good, both practically and theoretically, as are the variables employed to construct the model.

Final model evaluation depends on more that the statistical criteria that that have been mentioned and discussed to this point. The strength, credibility, and usability of the model depend on the theory and research questions that drove the model construction. In reporting the results and discussing the model, the researcher has the task of making explicit how the results address the proposed research question(s) situated in within a clear theoretical framework. Discussing results should include not only how the model adds to the area of focus understanding of the variables under investigation but also out of the limitations of the modelling strategy given the data (e.g., sample, design) at hand. Such information will allow for the reader and consumer of the work to evaluate the results with a critical eye with such information.

RESEARCH ISSUES/CONCLUSIONS

Logistic regression has received growing acceptance in many areas, especially in social sciences research (Hosmer & Lemeshow, 2000) and higher education, specifically. Since the 1990s, the application of LR has appeared in many higher education articles as well as a popular method in conference presentations at such venues as the American Educational Researcher Association conference (Peng, So, Frances, & St. John, 2002). A cursory search of popular databases (i.e., PsyInfo) using the term "logistic regression" resulted in 3,416 hits on written material before 2002 and 18,677 within the last 10 years. Clearly, from this one database search there has been an increase in the use and discussion of LR in the social sciences. This is most likely due to that fact that, as an alternative to linear regression models, logistic regression provides flexibility in examining relationships of a categorical outcome variable and combination of continuous or categorical predictors. In fact, many educational datasets can make use of logistic regression to investigate categorical outcome measures (i.e., pass/fail course, retention, diagnostic accuracy) in educational research.

In spite of the popularity of logistic regression, various problems associated with the application and interpretations have been identified (Peng et al., 2002). These concerns include sample size, transformation of the scale, label of dependent variable and reference category, diagnostic analysis, and underreported statistical software, parameters of estimates, and justification of model selections. These problems affect accuracy and implication of logistic regression model across studies

(Peng et al., 2002). These technically challenging areas will continue to receive attention from a methodological angle to bring clarity or at least provide guidance to practitioners applying the method to explore various research questions. Moreover, as new uses of LR arise (e.g., multilevel analysis, person-fit LR models), the demand for methodological evaluation will wax to ensure proper statistical modeling.

The current article provides guideline and fundamental principles of how LR should be applied to promote learning about this method. However, in recent years, there is an increasing number of applied and methodological studies discussing the extension of the LR model to multilevel because individuals are indeed nested in and influenced by the context (i.e., culture, geography, school type). This type of sampling structure is common in educational literature as well as other areas such as organizational research or medical research. In general, the resulting data have a hierarchical structure; making inferences at the individual (e.g., student) level problematic or inaccurate when the structure is ignored. Essentially, bias in the standard errors is introduced and results in underestimating the standard error of the regression coefficient. This can lead to inaccurate statistically significant results (i.e., inflate Type I error). As a result, multilevel modelling takes the hierarchical structure of the data (e.g., students nested within schools) into account to accurately estimate the standard error by partitioning variance into individual level (Level 1) and contextual or cluster level (Level 2).

Examples of such multilevel work have included building multilevel logistic regression models to account for context in how items function across groups in the area of measurement invariance (e.g., French & Finch, 2010; Swanson, et al., 2002). Multilevel regression models are being used to develop explanatory person-fit analysis models in the realm of person response function (PRF) models where the probability of a correct response to an item is model as a function of the item locations (e.g., Conijin, Emons, van Assen, & Sijtsma, 2011). In addition, multilevel logistic regression models have been used to investigate neighborhood effects (high or low education levels) on individuals' visits to physicians and their chances of being hospitalized for ischemic heart disease (Larsen & Merlo, 2005). Thus, it is expected that in the years to come there will be more useful extensions of LR with educational related data to increase the accuracy of modeling the complexities of the world in which we live, work, play, and learn. The extensions of LR to multilevel data to address various outcomes from medical visits to item functioning reflect both the applied and methodological trends that will continue over the next decade. We look forward to seeing the new applications as well as model extensions with LR.

REFERENCES

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*, 545–557.

Allison, P. D.(2001). *Missing data*. Thousand Oaks, CA: Sage.

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine

handicapped groups. *Journal of Educational Measurement, 24*, 41–55.
Cizek, G. J., & Fitzgerald, S. M. (1999). An introduction to logistic regression. *Measurement and Evaluation in Counselling and Development*, *31*, 223–244.
Cleary, P. D., & Angel, R. (1984). The analysis of relationships involving dichotomous dependent variables. *Journal of Health and Social Behavior*, *25*, 334–348.
Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences,* (2nd ed.), Hillsdale, N. J.: Lawrence Erlbaum Associates.
Conijn, J. M., Emons, W. H. M., van Assen, Marcel A. L. M., & Sijtsma, K. (2011). On the usefulness of a multilevel logistic regression approach to person-fit analysis. *Multivariate Behavioral Research*, *46*(2), 365–388. doi:10.1080/00273171.2010.546733
Davis, L. J., & Offord, K. P. (1997). Logistic regression. *Journal of Personality Assessment, 68*(3), 497–507.
Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.
Fan, X., & Wang, L. (1999). Comparing linear discriminant function with logistic regression for the two-group classification problem. *Journal of Experimental Education, 67*.
French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement, 47*, 299–317.
French, B. F., Immekus, J. C., & Oakes, W. C. (2005). An examination of indicators of Engineering students' success and persistence. *Journal of Engineering Education*, *94*, 419–425.
French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for DIF detection. *Educational and Psychological Measurement*, *67*, 373–393.
Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Research methods in psychology* (pp. 87–114). Volume 2 of the Handbook of Psychology (I. B. Weiner, Editor-in-Chief). New York: John Wiley & Sons.
Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*, 323–343.
Hassanuddin, N. A., Abdullah, Z., Mansor, N., & Hassan, N. H. (2012). Acceptance towards the use of internet banking services of cooperative bank. *International Journal of Academic Research in Business and Social Sciences*, *2*, 135–148.
Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.
Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.
Kleinbaum, D., Kupper, L., & Morgenstern, H. (1982). *Epidemiologic research: Principles and quantitative methods*, Chap 1–19. John Wiley and Sons Publishers, New York.
Larsen, K., & Merlo, J. (2005). Appropriate assessment of neighborhood effects on individual health: Integrating random and fixed effects in multilevel logistic regression. *American Journal of Epidemiology*, *161*, 81–88.
Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.
Long. J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
Lottes, I. L., Adler, M. A., & DeMaris, A. (1996). Using and interpreting logistic regression: A guide for teachers and students. *Teaching Sociology*, *24*, 284–298.
Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*, 17–24.
Menard, S. (1995). *Applied logistic regression*. Thousand Oakes, CA: Sage.
Peng, C. J., So, T, H., Frances, F. K., & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988–1999. *Research in Higher Education*, *43*, 259–293.
Peugh, J. L., & Enders, C. K. (2004). Missing data in education research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*, 525–556.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. J. Wiley & Sons, New York.
Schafer, J. L., & Graham, J. W.(2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.
Schmand, B., Eilelenboom, P., & vanGool, W. A. (2012). Value of diagnostic tests to predict conversion to Alzheimer's disease in young and old patients with amnestic mild cognitive impairment. *Journal of Alzheimer's Disease*, *29*(3), 641–648.

Swanson, d. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. J*ournal of Educational and Behavioral Statistics, 27*, 53–75.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (4th ed.). Boston: Allyn and Bacon.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55, 525–534.

Tripepi, G., Jager, K. J., Stel, V. S., Dekker, F. W., & Zoccali, C. (2011). How to deal with continuous and dichotomous outcomes in epidemiologic research: Linear and logistic regression analyses. *Nephron Clinical Practice*, *118*, 399–406.

SUGGESTIONS FOR FURTHER READINGS

Allison, P. D. (1999). *Logistic regression using the SAS system: Theory and application*. Cary, NC: SAS Institute.

Agresti, A. (2007). *An introduction to categorical data analysis*(2nd ed.). New York: John Wiley & Sons.

Cizek, G. J., & Fitzgerald, S. M. (1999). An introduction to logistic regression. *Measurement and Evaluation in Counseling and Development, 31*, 223–244.

Hosmer, D. W., & Lemeshow, S. (2000). Applied logistic regression (2nd ed.). New York: Wiley.

Menard, S. (1995). *Applied logistic regression*. Thousand Oakes, CA: Sage.

W. HOLMES FINCH

8. EXPLORATORY FACTOR ANALYSIS

Exploratory factor analysis (EFA) is a very popular statistical tool that is used throughout the social sciences. It has proven useful for assessing theories of learning, cognition, and personality (Aluja, García, & García, 2004), for exploring scale validity (Manos, Rachel C.; Kanter, Jonathan W.; Luo, Wen;), and for reducing the dimensionality in a set of variables so that they can be used more easily in further statistical analyses (Mashal & Kasirer, 2012). EFA expresses the relationship between variables that can be directly measured, or observed, and those that cannot, typically referred to as latent variables. The model parameter estimation is based upon the covariance matrix among a set of the observed variables. This relative simplicity in the basic design of the method makes it very flexible and adaptable to a large number of research problems. In the following pages, we will explore the basic EFA model and examine how it can be applied in practice. We will put special focus on the various alternatives for conducting factor analysis, discussing the relative merits of the more common approaches. Finally, we will provide an extended example regarding the conduct of EFA and interpretation of results from an analysis.

Prior to discussing the intricacies of EFA, it is important to say a few words about how it fits in the broader latent model framework. Factor analysis in general is typically divided into two different but complementary analyses: EFA and confirmatory factor analysis (CFA). From a mathematical perspective these two models are very closely linked, however they have very different purposes in application. Perhaps the most distinctive difference between the two is the degree to which the underlying factor model is constrained. In EFA very few constraints are placed on the structure of the model in terms of the number of latent variables or how the observed indicators relate to their latent counterparts. In contrast, researchers using CFA constrain the model to take a very specific form, indicating precisely with which latent variables each of the observed indicators is associated, and how many such indicators exits. This statistical distinction manifests itself in practice through the different manner in which each method is typically used. EFA is most often employed in scenarios where a researcher does not have fully developed and well grounded hypotheses regarding the latent structure underlying a set of variables, or where those hypotheses have not been thoroughly examined with empirical research (Brown, 2006). CFA is typically used to explicitly test and compare theories about such latent structure by altering of the constraints described above. Thus, while the basic model may be the same for these two approaches to factor analysis, the actual analyses are conducted in a very

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 167–186.

different manner. The focus of this chapter is on EFA, and so no further discussion of CFA is presented. However, researchers should always keep the distinction between the two approaches to factor analysis in mind as they consider which would be most appropriate for their specific research problem.

Exploratory Factor Analysis Model

As discussed briefly above, factor analysis expresses the relationship between a set of observed, or directly measured, variables, and a set of unobserved, or latent variables. Typically, the latent variables are those of greatest interest to the researcher, as they might represent the true construct of interest. For example, a researcher might be particularly interested in assessing the latent structure underlying a set of items intended to measure reasons why college undergraduates consume alcohol. The researcher might have some idea based on substantive theory regarding the number and nature of these latent variables. However, this theory might be relatively untested with empirical evidence. In order to gain insights into the nature of the underlying construct(s) EFA can be employed. The basic model takes the form:

$$x = LF + u \tag{1}$$

In this matrix representation of the model, x is simply a vector of observed variables, L is a matrix of factor pattern coefficients (often referred to as factor loadings), F is a vector of common factors and u is a vector of unique variables. In the context of our example, x represents responses to the individual items asking students why they drink, F is the set of latent variables that underlie these item responses. These might be thought of as the real reasons that students consume alcohol, which cannot be directly measured. The L, or factor loadings values, express the relationship between each of the observed and latent variables, while the unique variables, u, represent all influences on the observed variables other than the factors themselves. Often, these values are referred to as uniquenesses or error terms, and indeed they are similar in spirit to the error terms in standard linear models such as regression.

The primary objective in factor analysis is to identify the smallest number of factors that provides adequate explanation of the covariance matrix of the set of observed variables (Thompson, 2004). We will discuss how one might define adequate explanation forthwith. First, however, it is worth briefly describing the underlying mechanics of how the factor model described above is optimized for a specific research scenario. The model presented in (1) can be linked directly to the covariance matrix (S) among the observed indicator variables using the following

$$\mathrm{S} = \mathrm{LFL} + \Psi \tag{2}$$

The factor loading matrix, L is as defined previously. The factor covariance matrix, F, contains the factor variances and covariances, or relationships among the factors

themselves. The term Ψ is a diagonal matrix containing the unique variances. This equation expresses the relationship between the factor loadings and the observed correlation matrix. In practice, the goal of EFA is to define each of these values in such a way that the predicted correlation matrix, $\hat{\Sigma}$, is as similar as possible to the observed correlation matrix, S, among the observed variables. Often, statisticians discuss these covariance matrices in their standardized forms, the predicted and observed correlation matrices, $\hat{R}$ and R, respectively.

Factor Extraction

The process of obtaining initial estimates of EFA model parameters, including the factor loadings, is known as factor extraction. As discussed previously, the primary goal of factor extraction is to identify factor loadings that can reproduce as closely as possible the observed correlation matrix, while maintaining the smallest number of factors possible. If the only goal were to accurately reproduce this matrix, we would simply assign each observed variable to its own factor, thus replicating the observed data (and the observed correlation matrix) exactly. However, when the additional goal of reducing the size of the data set from the total number of observed variables to a smaller number of factors, this approach would not be helpful. Therefore, there is created friction between the goal of accurately reproducing R while keeping the factor model as simple as possible.

There are a number of methods available for extracting the initial set of factor loadings. These various approaches differ in terms of how they express the optimizing function; i.e. the comparison between R and $\hat{R}$. However, despite the fairly large number of approaches for extraction, only a few are actually used routinely in practice. Only these methods will be described here, though it is useful for the researcher to be aware of the availability of a broader range of extraction techniques.

One of the most common such factor extraction approaches is principal components analysis (PCA). PCA differs from the other extraction methods in that it is designed to extract total variance from the correlation matrix, rather than only shared variance, which is the case for the other extraction approaches. In technical terms, the diagonal of R contains 1's in the case of PCA, while the off diagonal elements are the correlations among the observed variables. Thus, when the parameters in (1) are estimated in PCA, it is with the goal of accurately reproducing the total variance of each variable (represented by the diagonal 1 elements) as well as correlations among the observed variables. The latent variables in this model are referred to as components, rather than factors, and likewise the loadings in PCA are referred to as component rather than factor loadings. One interesting point to note is that when researchers use PCA with a set of scale items and thereby set the diagonal of R to 1, they make a tacit assumption that the items are perfectly reliable (consistent) measures of the latent trait (Thompson, 2004).

An alternative approach to initial factor extraction involves the replacement of the 1's in the diagonal of *R* with an estimate of shared variance only, typically the squared multiple correlation (SMC) for the variable. The SMC values, which are estimated by regressing each observed variable onto all of the others, represent only the variation that is shared among the observed variables, as opposed to the total variation used in PCA. Thus, when the factor model parameters are estimated, it is with the goal of most closely approximating the variability that is shared among the observed variables and ignoring that which is unique to each one alone. Perhaps the most popular of this type of extraction method is principal axis factoring (PAF). A third popular approach for estimating factor model parameters is maximum likelihood estimation (MLE). MLE is an extraction method based in the larger statistics literature, where this approach to parameter estimation is quite popular and widely used in many contexts. For factor analysis, the goal is to find estimates of the factor loadings that maximize the probability of obtaining the observed data. This approach to extraction is the only one that requires an assumption of multivariate normality of the data (Lawley & Maxwell, 1963). The fourth method of extraction that we will mention here, alpha factoring, was designed specifically for use in the social sciences, in particular with psychological and educational measures (Kaiser & Caffrey, 1965). Alpha factoring has as its goal the maximization of Cronbach's alpha (a very common measure of scale reliability) within each of the retained factors. Therefore, the goal of this extraction approach is the creation of factors that correspond to maximally reliable subscales on a psychological assessment. While there are a number of other extraction methods, including image factoring, unweighted least squares, and weighted least squares, those highlighted here are the most commonly used and generally considered preferred in many social science applications (Tabachnick & Fidell, 2007).

Factor Rotation

In the second step of EFA, the initial factor loadings described above are transformed, or rotated, in order to make them more meaningful in terms of (ideally) clearly associating an indicator variable with a single factor with what is typically referred to as simple structure (Sass & Schmitt, 2010). Rotation does not impact the overall fit of the factor model to a set of data, but it does change the values of the loadings, and thus the interpretation of the nature of the factors. The notion of simple structure has been discussed repeatedly over the years by researchers, and while there is a general sense as to its meaning, there is not agreement regarding exact details. From a relatively nontechnical perspective, simple structure refers to the case where each observed variable is clearly associated with only one of the latent variables, and perfect simple structure means that each observed variable is associated with only one factor; i.e. all other factor loadings are 0. From a more technical perspective,

Thurstone (1947) first described simple structure as occurring when each row (corresponding to an individual observed variable) in the factor loading matrix has at least one zero. He also included 4 other rules that were initially intended to yield the over determination and stability of the factor loading matrix, but which were subsequently used by others to define methods of rotation (Browne, 2001). Jennrich (2007) defined perfect simple structure as occurring when each indicator has only one nonzero factor loading and compared it to Thurstone simple structure in which there are a "fair number of zeros" in the factor loading matrix, but not as many as in perfect simple structure. Conversely, Browne (2001) defined the complexity of a factor pattern as the number of nonzero elements in the rows of the loading matrix. In short, a more complex solution is one in which the observed variables have multiple nonzero factor loadings. Although the results from different rotations cannot be considered good or bad, or better or worse, the goal of rotations in EFA is to obtain the most interpretable solution possible for a set of data, so that a relatively better solution is one that is more theoretically sound (Asparouhov & Muthén, 2009). With this goal in mind, a researcher will want to settle on a factor solution that is most in line with existing theory and/or which can be most readily explained given literature in the field under investigation. In short, we want the solution to "make sense".

Factor rotations can be broadly classified into two types: (1) Orthogonal, in which the factors are constrained to be uncorrelated and (2) Oblique, in which this constraint is relaxed and factors are allowed to correlate. Within each of these classes, there are a number of methodological options available, each of which differs in terms of the criterion used to minimize factor complexity and approximate some form of simple structure (Jennrich, 2007). Browne (2001) provides an excellent review of a number of rotational strategies, and the reader interested in the more technical details is encouraged to refer to this manuscript. He concluded that when the factor pattern conformed to what is termed above as pure simple structure most methods produce acceptable solutions. However, when there was greater complexity in the factor pattern, the rotational methods did not perform equally well, and indeed in some cases the great majority of them produced unacceptable results. For this reason, Browne argued for the need of educated human judgment in the selection of the best factor rotation solution. In a similar regard, Yates (1987) found that some rotations are designed to find perfect (or nearly) simple structure solution in all cases, even when this may not be appropriate for the data at hand. Based on their findings, Browne and Yates encouraged researchers to use their subject area knowledge when deciding on the optimal solution for a factor analysis. While the statistical tools described here can prove useful for this work, they cannot replace expert judgment in terms of deciding on the most appropriate factor model.

There are a number of rotations available to the applied researcher in commonly used software packages such as SPSS, SAS, R, and MPlus. Some of the most

common of these rotations fall under the Crawford-Ferguson family of rotations (Browne, 2001), all of which are based on the following equation:

$$f(\Lambda) = (1-k)\sum_{i=1}^{p}\sum_{j=1}^{m}\sum_{l\neq j,l=1}^{m}\lambda_{ij}^2\lambda_{ij}^2 + k\sum_{j=1}^{m}\sum_{i=1}^{p}\sum_{l\neq i,l=1}^{p}\lambda_{ij}^2\lambda_{ij}^2 \tag{3}$$

where
m = the number of factors
p = the number of observed indicator variables
l_{ij} = unrotated factor loading linking variable i with factor j

The various members of the Crawford-Ferguson family differ from one another in the value of k. As Sass and Schmitt (2010) note, larger values of k place greater emphasis on factor (column) complexity while smaller values place greater emphasis on variable (row) complexity. Popular members of the Crawford-Ferguson family include Direct QUARTIMIN ($k = 0$), EQUAMAX ($k = m/2p$), PARSIMAX ($k = (m - 1)/(p + m - 2)$), VARIMAX ($k = 1/p$), and the Factor Parsimony (FACPARSIM) ($k = 1$).

In addition to the Crawford-Ferguson family, there exist a number of other rotations, including orthogonal QUARTIMAX, which has the rotational criterion

$$f(\Lambda) = -\frac{1}{4}\sum_{i=1}^{p}\sum_{j=1}^{m}\lambda_{ij}^4, \tag{4}$$

GEOMIN with the rotational criterion

$$f(\Lambda) = \sum_{i=1}^{p}\left[\prod_{j=1}^{m}(\lambda_{ij}^2 + \varepsilon)\right]^{\frac{1}{m}}. \tag{5}$$

and PROMAX. The PROMAX rotation, which is particularly popular in practice, is a two-stage procedure that begins with a VARIMAX rotation. In the second step, the VARIMAX rotated factor loadings are themselves rotated through application of the target matrix

$$T_1 = (\Lambda_v'\Lambda_v)^{-1}\Lambda_v'B \tag{6}$$

where
Λ_v = Varimax rotated loding matrix
B = Matrix containing elements $\frac{\lambda_{ij}^{b+1}}{\lambda_{ij}}$
b = Power to which the loading is raised (4 is the default in most software)

This target matrix is then rescaled to T based on the square root of the diagonals of $(T_1'T_1)^{-1}$ and the Promax rotated loading matrix is defined as

$$\Lambda_P = \Lambda_v T \tag{7}$$

The interested reader can find more technical descriptions of these rotational methods in the literature (Browne, 2001; Asparouhov & Muthen, 2009; Mulaik, 2010; Sass & Schmitt, 2010).

One issue of some import when differentiating orthogonal and oblique rotations is the difference between Pattern and Structure matrices. In the case of oblique rotations, the Pattern matrix refers to the set of factor loadings that reflects the unique relationship between individual observed and latent variables, excluding any contribution from the other factors in the model. The structure matrix includes loadings that reflect the total relationship between the observed and latent variables, including that which is shared across factors. In general practice, researchers often use the Pattern matrix values because they do reflect the unique relationship and are thus perhaps more informative regarding the unique factor structure (Tabachnick & Fidell, 2007). Because orthogonal rotations by definition set the correlations among factors to 0, the Pattern and Structure matrices are identical.

In practice, VARIMAX and PROMAX are probably the two most widely used methods of factor rotation, as revealed by a search of the Psycinfo database in February, 2012. This popularity is not due to any inherent advantages in these approaches, as statistical research has identified other approaches that would be more optimal in some circumstances (Finch, in press). However, these methods are widely available in software, have been shown to be reasonably effective in statistical simulation studies, and are generally well understood in terms of their performance under a variety of conditions. This does not mean, however, that they should be the sole tools in the factor analysts rotational arsenal. Indeed, many authors (e.g., Asparouhov & Muthen, 2009) argue that because the goal of factor rotation is to produce meaningful and interpretable results, it is recommended that multiple approaches be used and the results compared with one another, particularly in terms of their theoretical soundness. At the very least, we would recommend that the researcher consider both an orthogonal and an oblique rotation, examining the factor correlations estimated in the latter. If these correlations are nontrivial, then the final rotational strategy should be oblique, so that the loadings incorporate the correlations among the factors.

Communalities

One measure of the overall quality of a factor solution is the individual communality value for each of the observed variables. Conceptually, communalities can be interpreted as the proportion of variation in the observed variables that is accounted for by the set of factors. They typically range between 0 and 1, though in certain

(problematic) circumstances this will not be the case. A relatively large communality for an individual variable suggests that most of its variability can be accounted for by the latent variables. For orthogonal rotations, the communality is simply the sum of the squared factor loadings. Thus, if a three factor solution is settled upon and the loadings for variable 1 are 0.123, 0.114, and 0.542, the communality would be $0.123^2 + 0.114^2 + 0.542^2$, or 0.322. We would conclude that together the three factors accounted for approximately 32% of the variation in this variable. It is important to note that a large communality does not necessarily indicate that the factor solution is interpretable or matches with theory. Indeed, for the previous example, the loadings 0.417, 0.019, and 0.384 would yield an identical communality to that calculated previously. Yet, this second solution would not be particularly useful given that the variable loads equally on factors 1 and 3. Therefore, although communalities are certainly useful tools for understanding the quality of a factor solution, by themselves they do not reveal very much about the interpretability of the solution.

Determining the Number of Factors

As with factor extraction and rotation, there are a number of statistical approaches for identifying the optimal number of factors. It should be stated up front that the optimal solution is the one that best matches with theory and can be defended to experts in the field, regardless of what the statistical indicators would suggest. Having said that, there are statistical tools available that can assist the researcher in, at the very least, narrowing down the likely number of factors that need to be considered. Most of these approaches are descriptive in nature, although some inferential tests are available. We will begin with the more descriptive and generally somewhat older methods for determining the number of factors, and then turn our attention to more sophisticated and newer techniques.

Perhaps one of the earliest approaches for determining the likely number of factors was described by Guttman (1954), and is commonly referred to as the eigenvalue greater than 1 rule. This rule is quite simple to apply in that a factor is deemed to be important, or worthy of retaining if the eigenvalue associated with it is greater than 1. The logic underlying this technique is equally straightforward. If we assume that each observed variable is standardized to the normal distribution with a mean of 0 and variance of 1, then for a factor to be meaningful it should account for more variation in the data than does a single observed variable. While this rule is simple and remains in common use, it is not without problems, chief of which is that it has a tendency to overestimate the number of factors underlying a set of data (Patil, McPherson, & Friesner, 2010). Nonetheless, it is one of the default methods used by many software programs for identifying the number of factors.

Another approach for determining the number of factors based on the eigenvalues is the Scree plot. Scree is rubble at the base of a cliff, giving this plot its name. It was introduced by Cattell (1966), and plots the eigenvalues on the Y axis, with the factors on the X axis. The researcher using this plot looks for the point at which the plot

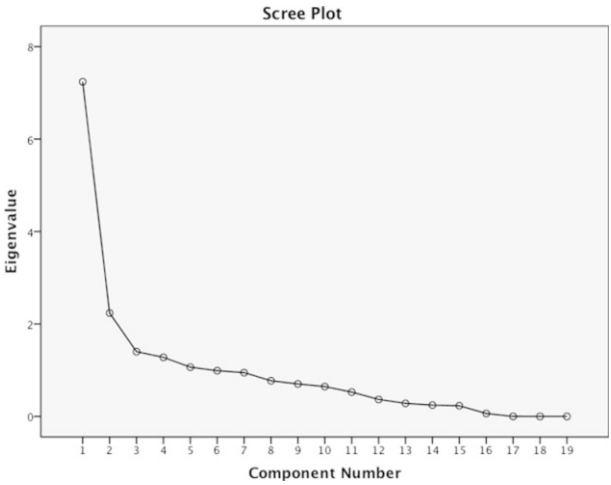


Figure 1. Example scree plot.

bends, or flattens out. Figure 1 contains an example of a Scree plot. It would appear that the line bends, or flattens out at 3 factors, thus we might retain 2. It is important to note that the interpretation of the Scree plot is subjective, so that researchers may not always agree on the optimal number of factors to retain when using it. Prior research on the effectiveness of this method has found that much as with the eigenvalue greater than 1 rule, the Scree plot tends to encourage the retention of too many factors (Patil, McPherson, & Friesner, 2010).

In addition to examining the eigenvalues themselves, researchers often will also consider the proportion of variation in the observed data that is accounted for by a particular factor solution. The total variance contained in the data is equal to the sum of the eigenvalues. Therefore, the proportion of variability accounted for by an individual factor is simply the ratio of its eigenvalue to the sum of the eigenvalues (which will be equal to the number of observed variables). While there are no rules regarding what constitutes an acceptable proportion of observed indicator variance accounted for by the latent variables, clearly more is better, while maintaining a goal of factor parsimony.

As discussed above, mathematically speaking the goal of factor analysis is to reproduce as closely as possible the correlation matrix among the observed variables, R, with the smallest number of latent variables. This predicted correlation matrix $\hat{R}$, can then be compared with the actual matrix in order to determine how well the factor solution worked. This is typically done by calculating residual correlation values (the difference between the observed and predicted correlations) for each pair of observed variables. If a given factor solution is working well, we would expect the residual correlation values to be fairly small; i.e. the factor model has done an accurate job of reproducing the correlations. A common rule of thumb (Thompson,

2004) is that the absolute value of the residual correlations should not be greater than 0.05. This cut-off is completely arbitrary, and a researcher may elect to use another, such as 0.10. While the residual correlation matrix is a reasonably useful tool for ascertaining the optimal number of factors, it can be very cumbersome to use when there are many observed variables. Some software packages, such as SPSS, provide the user with the number and proportion of residual correlations that are greater than 0.05, eliminating the need for the tedious job of counting them individually.

In addition to these purely descriptive assessments of a factor solution, there exist some inferential tools. For example, parallel analysis (PA; Horn, 1965) has proven to be an increasingly popular and reasonably dependable hypothesis testing method for determining the number of factors. The PA methodology is drawn from the literature on permutation tests in the field of statistics. Specifically, the goal of this technique is to create a distribution of data that corresponds to what would be expected were there no latent variables present in the data; i.e. if the observed variables were uncorrelated with one another. This is done by generating random data that retains the same sample size, means and variances as the observed data, but being random, has correlation coefficients among the observed variables centered on 0. When such a random dataset is created, factor analysis is then conducted and the resulting eigenvalues are retained. In order to create a sampling distribution of these eigenvalues, this random data generation and factor analysis is replicated a large number of times (e.g. 1000). Once the distribution of egenvalues from random data are created, the actual eigenvalues obtained by running factor analysis with the observed data are then compared to the sampling distributions from the random data. The random data distributions are essentially those for the case when the null hypothesis of no factor structure is true, so that the comparison of the observed eigenvalues to these random distributions provides a hypothesis test for the null hypothesis of no factor structure. Therefore, if we set $a = 0.05$, we can conclude that an observed eigenvalue is significant when it is larger than the 95th percentile of the random data distribution. This method will be used in the example below, providing the reader with an example of its use in practice.

Another alternative approach for assessing factor solutions is Velicer's minimum average partial (MAP) approach (Velicer, 1976). This method involves first estimating multiple factor solutions (i.e. different numbers of factors). For each such factor solution, the correlations among the observed variables are estimated, partialing out the factors. For example, initially one factor is retained, and the correlations among all of the observed variables are calculated after removing the effect of this factor. Subsequently, 2 factors, 3 factors, and so on are fit to the data, and for each of these models the partial correlations are calculated. These partial correlations are then squared and averaged in order to obtain an average partial correlation for each model. The optimal factor solution is the one corresponding to the minimum average partial correlation. The logic underlying MAP is fairly straight forward. A good factor solution is one that accounts for most of the correlation among a set of observed variables. Therefore, when the factor(s) are partialed out of the correlation

matrix, very little relationship is left among the variables; i.e. the partial correlations will be very small. By this logic, the solution with the minimum average squared partial correlation is the one that optimally accounts for the relationships among the observed variables.

Example

We will now consider an extended example involving the conduct of factor analysis from the initial extraction through the determination of the number of factors. For this example, we will examine the responses to a 12 item questionnaire designed to elicit information from college students regarding their reasons for drinking alcohol. The Items appear below in Table 1, and are all answered on a 7 point likert scale where a 1 indicates this is nothing like the respondent and 7 indicates this is exactly like the respondent. The researcher believes that the items measure 3 distinct latent constructs: drinking as a social activity, drinking as a way to cope with stress, and drinking as an enhancement to other activities. Data were collected on a total of 500 undergraduate students at a large university (52% female). The goal of this EFA is to determine the extent to which the underlying theory of the scale matches with the observed data collected from the college students. In other words, do the items group together into the three coherent factors envisioned by the researcher?

The researcher first conducts an EFA with 3 factors (matching the theory) using MLE extraction and PROMAX rotation. The latter choice is made in order to obtain a correlation matrix for the factors, which in turn will inform the final decision regarding the type of rotation to use (orthogonal or oblique). This correlation matrix appears in Table 2.

Table 1. Drinking scale items

Item 1: Because you like the feeling
Item 2: Because it's exciting
Item 3: Because it give you a pleasant feeling
Item 4: Because it's fun
Item 5: It helps me enjoy a party
Item 6: To be sociable
Item 7: It makes social gatherings more fun
Item 8: To celebrate special occasions
Item 9: To forget worries
Item 10: It helps when I feel depressed
Item 11: Helps cheer me up
Item 12: Improves a bad mood

Table 2. Interfactor correlation matrix

Factor	*1*	*2*	*3*
Dimension	1.000	.266	.633
	.266	1.000	.363
	.633	.363	1.000

Table 3. Eigenvalues and percent of variance accounted for by each factor

Factor	*Eigenvalue*	*Percent*	*Cumulative percent*
1	3.876	32.297	32.297
2	1.906	15.880	48.178
3	1.150	9.587	57.765
4	.837	6.975	64.740
5	.722	6.013	70.753
6	.669	5.576	76.328
7	.576	4.802	81.131
8	.557	4.643	85.774
9	.487	4.061	89.834
10	.471	3.923	93.758
11	.426	3.552	97.309
12	.323	2.691	100.000

All of the factor pairs exhibit a non-zero correlation, and factors 1 and 3 are highly correlated with one another, with $r = 0.633$. This result would suggest that an oblique rotation is likely more appropriate than orthogonal.

After determining the general rotational approach, we will next want to consider the appropriate number of factors to retain. As described above, this is not an issue with a simple answer. There are a number of statistical tools at our disposal to help in this regard, but they may provide somewhat different answers to the question of the optimal number of factors to be retained. Of course, the final determination as to factor retention is the conceptual quality of the factors themselves. First, however, we can examine some of the statistical indicators. Table 3 contains the eigenvalue for each factor, along with the proportion of variance accounted for by each individually, as well as by the set cumulatively.

An examination of the results reveals that the eigenvalue greater than 1 rule would yield a three factor solution. The first three factors explain approximately 58% of the total variation in item responses, with the first factor explaining a full third of the variance by itself. After three factors, the change in additional variance explained for each additional factor is always less than 1%, indicating that these factors do not provide markedly greater explanation of the observed data individually. The scree

Table 4. MAP results for the drinking scale data

Factor	*MAP value*
0.000000	0.083579
1.000000	0.033096
2.000000	0.026197
3.000000	0.034009
4.000000	0.053208
5.000000	0.077821
6.000000	0.108430
7.000000	0.159454
8.000000	0.224068
9.000000	0.312326
10.000000	0.504657
11.000000	1.000000

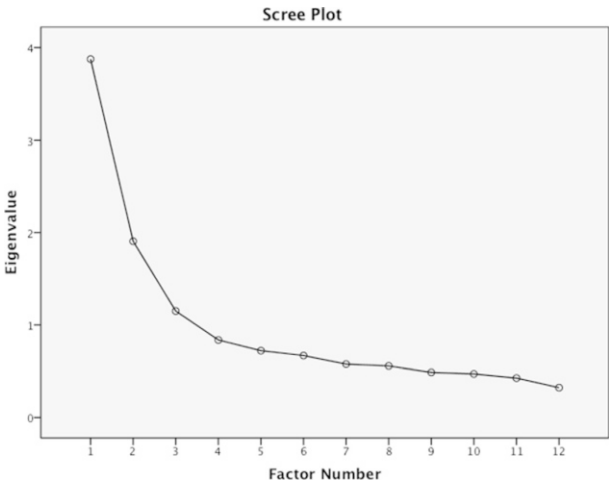


Figure 2. Scree plot for drinking scale items.

plot (Figure 2), which provides a graphical display of the eigenvalues by factor number suggests that perhaps three or four factors would be appropriate, given that the line begins to flatten out for eigenvalues between those numbers of factors.

In addition to these approaches for determining the number of factors, which are each based on the eigenvalues in some fashion, other approaches may also be used for this purpose, including MAP, PA, and the chi-square goodness of fit test from MLE extraction. The MAP results for the drinking data appear below in Table 4.

These results show that the lowest average squared correlation value was associated with the two factor solution. Thus, based on MAP we would conclude that there are 2 factors present in the data.

Another method for ascertaining the number of factors is PA. In this case, we will ask for 1000 permutations of the original datasets, and set the level of *a* at 0.05 (using the 95th percentile). Results of PA appearing in Table 5 below, suggest the presence of 3 factors. We conclude this based upon the fact that the eigenvalues from the actual data are larger than the 95th percentile values for the first three factors, but not the fourth.

Finally, because we used the MLE method of factor extraction, a chi-square goodness of fit test was also a part of the final results. This statistic tests the null hypothesis that the factor solution fits the data. More specifically, it tests the hypothesis that the reproduced correlation matrix (based upon the factor solution) is equivalent to the observed correlation matrix. It is important to note that in order to use MLE extraction, we must assume that the observed data follow the multivariate normal distribution (Brown, 2006). We can assess this assumption using Mardia's test for multivariate normality (Mardia, 1970). In this example, MLE extraction yielded *p*-values of 0.00004, 0.0102, and 0.482 two, three, and four factors, respectively. Thus, based on this test, we would conclude that four factors is the optimal solution.

In considering how to proceed next, we can examine the results of the various analyses just discussed in order to narrow down the range of options for which we should obtain factor loadings matrices. It would appear that the least number of factors that might be present in the data would be two (MAP), while the largest reasonable number would be 4 (chi-square goodness of fit test). For this reason,

Table 5. Eigenvalues for raw data and parallel analysis distribution

Factor	*Raw Data*	*Means*	*95th Percentile*
1	3.272626	0.288555	0.357124
2	1.234191	0.217483	0.267662
3	0.431697	0.162343	0.205637
4	0.110090	0.114928	0.153813
5	–0.014979	0.071712	0.107747
6	–0.034157	0.031778	0.063371
7	–0.073058	–0.005970	0.022749
8	–0.111261	–0.043912	–0.015727
9	–0.137628	–0.081220	–0.053983
10	–0.139254	–0.119163	–0.090826
11	–0.200550	–0.160150	–0.129607
12	–0.229260	–0.208650	0.172491

we will examine factor loading values for each of these three solutions. As noted previously, given that there appear to be nontrivial correlations among the factors, we will rely on PROMAX rotation, and will use MLE extraction. Pattern matrix values for the two, three, and four factor solutions appear in Table 6.

When interpreting the factor loadings in order to identify the optimal solution, it is important to remember the expected number of factors based on theory, which in this case is three. Furthermore, the items are ordered so that items 1 through 4 are theoretically associated with a common factor, items 5 through 8 are associated with a separate factor, and finally items 9 through 12 are associated with a third factor. In examining the two factor results, it appears that the 4 items theoretically associated with a common latent construct do in fact group together, while the other 8 items are grouped together in a single factor. Based on theory, it appears that factor 1 corresponds to the Enhancement construct, while factor 2 appears to conflate the Coping and Social constructs. With respect to the three factor solution, we can see that items 1 through 3 load together on factor 3, while items 5 through 8 load together on factor 1 and items 9 through 12 load on factor 2. Item 4 (drinking because it's fun) is cross-loaded with factors 1 and 3, and thus cannot be said to be associated clearly with either one. Considering these results in conjunction with the underlying theory, it would appear that factor 1 corresponds to Social reasons for drinking, factor 2 corresponds to Coping reasons for drinking and factor 3 (minus item 4) corresponds to Enhancement. We might consider whether the cross-loading

Table 6. Pattern matrices for PROMAX rotation of two, three, and four factor solutions for the drinking scale data

	Two Factors		*Three Factors*			*Four Factors*			
Item	*F1*	*F2*	*F1*	*F2*	*F3*	*F1*	*F2*	*F3*	*F4*
1	0.35	0.13	–0.11	0.01	0.63	–0.13	0.65	–0.06	0.08
2	0.40	0.08	0.01	–0.02	0.54	0.00	0.55	–0.07	0.05
3	0.39	0.08	0.02	–0.01	0.51	0.06	0.47	0.14	–0.16
4	0.81	0.06	0.48	–0.00	0.48	0.50	0.45	0.00	0.01
5	0.64	–0.07	0.63	–0.03	0.01	0.62	0.02	–0.08	0.05
6	0.72	0.01	0.72	0.05	0.01	0.74	–0.01	0.07	–0.02
7	0.69	–0.07	0.69	–0.02	0.01	0.71	–0.02	0.05	–0.09
8	0.71	–0.06	0.83	0.02	–0.14	0.81	–0.12	–0.07	0.08
9	0.04	0.57	0.04	0.58	–0.02	–0.02	0.05	0.13	0.54
10	0.08	0.54	0.12	0.58	–0.07	0.05	–0.02	0.04	0.67
11	–0.05	0.72	–0.17	0.69	0.13	–0.13	0.10	0.68	0.06
12	0.01	0.66	0.03	0.69	–0.06	–0.12	–0.12	0.69	0.06

of item 4 makes sense from a theoretical perspective. Finally, an examination of the four factor solution reveals that factor 1 corresponds to the Social construct along with the cross-loaded item 4 and factor 2 corresponds to the Enhancement construct, again considering the cross-loaded item. Factors 3 and 4 appear to be associated with the Coping construct, which has been split between items 9 (Forget worries) and 10 (Helps when depressed) on factor 3 and items 11 (Cheer me up) and 12 (Improves bad mood) on factor 4. Again, we must consider how this factor solution matches with the theory underlying the scale.

Following is a brief summary of the analyses described above. In order to decide on the final factor solution, we must consider all of the evidence described above. As mentioned previously, in the final analysis the optimal solution is the one that is theoretically most viable. Based upon the various statistical indices, it would appear that a solution between 2 and 4 factors would be most appropriate. For this reason, we used MLE extraction with PROMAX rotation and produced factor pattern matrices for 2, 3, and 4 factors. An examination of these results would appear to suggest that the 3 factor solution corresponds most closely to the theoretically derived constructs of Enhancement, Social, and Coping reasons for drinking. It is important, however, to note two caveats regarding such interpretation. First of all, item 4 (drinking because it's fun) cross-loads with two factors, which does not match the theory underlying the scale. Therefore, further examination of this item is warranted in order to determine why it might be cross-loading. Secondly, interpretation of the factor loading matrices is inherently subjective. For this reason, the researcher must be careful both in deciding on a final solution and on the weight which they place it. In short, while the factor solution might seem very reasonable to the researcher, it is always provisional in EFA, and must be further investigated using other samples from the population and confirmatory factor analysis (Brown, 2006).

Factor Scores

One possibly useful artifact of EFA is the possibility of calculating factor scores, which represent the level of the latent variable(s) for individuals in the sample. These scores are somewhat controversial within the statistics community, and are not universally well regarded (see Grice, 2001 and DiStefano, Zhu, & Mindrila, 2009, for excellent discussion of these issues). They are used in practice not infrequently, however, so that the knowledgeable researcher should have a general idea of how they are calculated and what they represent. There are multiple ways in which factor scores can be estimated once a factor solution has been decided upon. By far the most popular approach to estimating these scores is known as the regression method. This technique involves first standardizing the observed variables to the Normal (0,1) distribution; i.e. making them z scores. The factor scores can then be calculated as

$$F = ZR^{-1}l \tag{6}$$

where F is the vector of factor scores for the sample, Z is the set of standardized observed variable values, R is the observed variable correlation matrix, and l is the matrix of factor loadings. These factor scores are on the standard normal distribution with a mean of 0.

Researchers can then make use of these factor scores in subsequent analyses, such as regression or analysis of variance. However, as noted previously such practice is not without some problems and is not always recommended. Among the issues that must be considered when using such scores is the fact that the scores were obtained using a single factor extraction technique. Given that no one extraction method can be identified as optimal, and that the solutions might vary depending upon the extraction method used, the resultant factor scores cannot be viewed as the absolute best representation of the underlying construct for an individual or for a sample. In short, these values are provisional and must be interpreted as such. This indeterminacy of solutions means that another researcher using the same sample but a different method of extraction could obtain different factor scores, and thus a different result for the subsequent analyses. Neither of these outcomes could be viewed as more appropriate than the other, leading to possible confusion in terms of any substantive findings. A second concern with respect to the use of factor scores obtained using EFA is whether the factor solutions are equivalent across subgroups of individuals within the samples. Finch and French (2012) found that when factor invariance does not hold (factor loading values differ across groups), the resultant factor scores will not be accurate for all members of the sample, leading to incorrect results for subsequent analyses such as analysis of variance. With these caveats in mind, researchers should consider carefully whether derived factor scores are appropriate for their research scenario. If they find multiple extraction and rotation strategies result in very similar solutions, and they see no evidence of factor noninvariance for major groups in the data, then factor scores may be appropriate. However, if these conditions do not hold, they should consider refraining from the use of factor scores, given the potential problems that may arise.

Summary of EFA

EFA has proven to be a useful tool for researchers in a wide variety of disciplines. It has been used to advance theoretical understanding of the latent processes underlying observed behaviors, as well as to provide validity evidence for psychological and educational measures. In addition, a closely allied procedure, PCA, is often employed to reduce the dimensionality within a set of data and thereby make subsequent analyses more tractable. Given its potential for providing useful information in such a broad array of areas, and its ubiquity in the social sciences, it is important for researchers to have a good understanding regarding its strengths and limitations, and a sense for how it can best be used. It is hoped that this chapter has provided some measure of understanding to the reader.

In reality, EFA can be seen as a series of allied statistical procedures rather than as a single analysis. Each one of these procedures requires the data analyst to make decisions regarding the best course of action for their particular research problem. Quite often it is not obvious which approach is best, necessitating the use of several and subsequent comparison of the results. The first stage of analysis is the initial extraction of factors. As described above, there are a number of potential approaches that can be used at this step. Perhaps the most important decision at this stage involves the selection of PCA or one of the other extraction techniques. As noted, PCA focuses on extracting total variance in the observed variables while EFA extracts only shared variance. While results of the two approaches obtained for a set of variables may not differ dramatically in some cases, they are conceptually very different and thus are most appropriate in specific situations. One guideline for deciding on which approach to use is whether the goal of the study is understanding what common latent variables might underlie a set of observed data, or simply reducing the number of variables, perhaps for use in future analyses. In the first case, an EFA approach to extraction (e.g. PAF, MLE) would be optimal, whereas in the latter the researcher may elect to use PCA. Within the EFA methods of extraction, it is more difficult to provide an absolute recommendation for practice, although trying multiple approaches and comparing the results would be a reasonable strategy.

Once the initial factor solution is obtained, the researcher must then decide upon the type of rotation that is most appropriate. Given that rotation is designed solely to make the factor loadings conform more closely to simple structure and thus more interpretable, multiple strategies may be employed and the one providing the most theoretically reasonable answer retained. Of course, the first major decision in this regard is whether to use an orthogonal or oblique rotation. In general practice, I would recommend using an oblique approach first in order to obtain the factor correlation matrix. If the factors appear to be correlated with one another, then the Pattern matrix values can be used to determine how the variables grouped together into factors. On the other hand, if the interfactor correlations are negligible, the researcher could simply rerun the analysis using an orthogonal rotation and then refer to the factor loading matrix. It should be noted that some research has shown that quite often in practice the selection of rotation method will not drastically alter the substantive results of the study; i.e. which observed variables load on which factors (Finch, in press; Finch, 2006).

Typically, a researcher will investigate multiple factor solutions before deciding on the optimal one. This decision should be based first and foremost on the theory underlying the study itself. The best solution in some sense is the one that is most defendable based upon what is known about the area of research. Thus, a key to determining the number of factors (as well as the extraction/rotation strategy to use) can be found in the factor loading table. In conjunction with these loadings, there are a number of other statistical tools available to help identify the optimal factor solution. Several of the most popular of these were described previously. A key issue to keep in mind when using these is that no one of them can be seen as universally

optimal. Rather, the researcher should make use of many, if not most of them, in order to develop some consensus regarding the likely best number of factors. The extent to which these agree with one another, and with the substantive judgments made based on the factor loadings matrix, will dictate the level of confidence with which the researcher can draw conclusions regarding the latent variable structure.

EFA is somewhat unusual among statistical procedures in that frequently there is not a single, optimal solution that all data analysts can agree upon. When one uses multiple regression and the assumptions underlying the procedure are met, all can agree that the resulting slope and intercept estimates are, statistically speaking at least, optimal. Such is not the case with EFA. Two equally knowledgeable and technically savvy researchers can take the same set of data and come up with two very different final answers to the question of how many latent variables there are for a set of observed variables. Most importantly, there will not be a statistical way in which one can be proven "better" than the other. The primary point of comparison will be on the theoretical soundness of their conclusions, with the statistical tools for identifying the optimal number of factors playing a secondary role. Quite often this lack of finality in the results makes researchers who are used to more definite statistical answers somewhat uncomfortable. However, this degree of relativity in EFA solutions also allows the content area expert the opportunity to evaluate theories in a much more open environment. Indeed, some very interesting work at the intersection of EFA and theory generation has been done recently, showing great promise for this use of the technique (Haig, 2005). It is hoped that this chapter will help the applied researcher needing to use EFA with some confidence in the basic steps of the methodology and the issues to consider.

REFERENCES

Aluja, A., García, Ó., & García, L. F. (2004). Replicability of the three, four and five Zuckerman's personality super-factors: Exploratory and confirmatory factor analysis of the EPQ-RS, ZKPQ and NEO-PI-R. *Personality and Individual Differences*, *36*(5), 1093–1108.

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

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Browne, M. W. (2001). An overview of analytic rotations in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), Available online: http://pareonline.net/getvn.asp?v=14&n=20

Finch, W. H. (in press). A comparison of factor rotation methods for dichotomous data. *Journal of Modern Applied Statistical Methods*.

Finch, H. (2006). Comparison of the performance of Varimax and Promax rotations: Factor structure recovery for dichotomous items. *Journal of Educational Measurement*, *43*(1), 39–52.

Finch, W. H., & French, B. F. (2012). The impact of factor noninvariance on observed composite score variances. *International Journal of Research and Reviews in Applied Sciences*, *1*, 1–13.

Gorsuch, R. L. (1983). Factor analysis. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*,*6*(4), 430–450.
Guttman, L. (1958). Some necessary conditions for common factor analysis. *Psychometrika*, 19*(*2), 149–161.
Haig, B. D. (2005). Exploratory factor analysis, theory generation and the scientific method. *Multivariate Behavioral Research*, *40*(3), 303–329.
Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.
Jennrich, R. I. (2007). Rotation methods, algorithms, and standard errors. In R. Cudek & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 315–335). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
Kaiser, H. F., & Caffrey, J. (1965). Alpha factor analysis. *Psychometrika*, *30*(1), 1–14.
Lawley, D. N., & Maxwell, A. E. (1963). *Factor analysis as a statistical method*. London: Butterworth.
Manos, R. C., Kanter, J. W., & Luo, W. (2011). The behavioral activation for depression scale–short form: Development and validation. *Behavior Therapy*, *42*(4), 726–739.
Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57, 519–530.
Mashal, N., & Kasirer, A. (2012). Principal component analysis study of visual and verbal metaphoric comprehension in children with autism and learning disabilities. *Research in Developmental Disabilities*, *33*(1), 274–282.
Mulaik, S. A. (2010). *Foundations of factor analysis*. Boca Raton, FL: Chapman & Hall/CRC.
Patil, V. H., McPherson, M. Q., & Friesner, D. (2010). The use of exploratory factor analysis in public health: A note on parallel analysis as a factor retention criterion. *American Journal of Health Promotion*, *24*(3), 178–181.
Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, *45*, 73–103.
Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson.
Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.
Thurstone, L .L. (1947). *Multiple factor analysis*. Chicago: University of Chicago press.
Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327
Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. Albany: State University of New York Press.

JASON W. OSBORNE & SHEVAUN D. NEUPERT

9. A BRIEF INTRODUCTION TO HIERARCHICAL LINEAR MODELING

INTRODUCTION

Hierarchical linear modeling (HLM; also referred to as multilevel modeling or MLM) is becoming more common throughout all areas of the social sciences because of its flexibility and unique advantages not present in more traditional techniques (Osborne, 2000). Our goal in this chapter is to briefly introduce the reader to the important concepts related to HLM, particularly the advantages of HLM over more traditional techniques like regression on aggregated or disaggregated data, repeated measures ANOVA, etc. We will also give some examples of how it can be used in educational research and the broader field of social science, and will give the reader an example of a simple, but powerful type of analysis: growth curve analysis. Further, we will demonstrate the same example within two popular software packages for performing HLM: HLM (SSI; http://www.ssicentral.com/hlm/) and SAS (www.sas.com).

One of the reasons HLM is becoming so common within social sciences research is the thorny problem of hierarchical or nested data structures, and the fact that most researchers do not appropriately deal with this issue unless they are using HLM.

WHAT IS A HIERARCHICAL DATA STRUCTURE?

People (and most living creatures, for that matter) tend to exist within organizational structures, such as families, schools, business organizations, churches, towns, states, and countries. In education, students exist within a hierarchical social structure that can include family, peer group, classroom, grade level, school, school district, state, and country. Workers exist within production or skill units, businesses, and sectors of the economy, as well as geographic regions. Health care workers and patients exist within households and families, medical practices and facilities (a doctor's practice, or hospital, e.g.), counties, states, and countries. Many other communities exhibit hierarchical data structures as well.

Raudenbush and Bryk (2002) also discuss two other types of data hierarchies that are less obvious but equally important and well-served by HLM: repeated-measures data and meta-analytic data. In this case, we can think of repeated measures as data that are nested or clustered within individuals, and meta-analytic data similarly involves clusters of data or subjects nested within studies.

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 187–198.

Once one begins looking for hierarchies in data, it becomes obvious that data repeatedly gathered on an individual are hierarchical, as all the observations are nested within individuals. While there are ways of adequately dealing with nested and partially nested data in ANOVA paradigms that have existed for decades, they are often not easily or effectively used. Further, the assumptions relating to them are challenging, whereas procedures relating to hierarchical modeling require fewer assumptions that are easily met.

WHY IS A HIERARCHICAL OR NESTED DATA AN ISSUE?

Hierarchical, or nested, data present several problems for analysis. First, people or creatures that exist within hierarchies tend to be more similar to each other than people randomly sampled from the entire population. For example, students in a particular third-grade classroom are more similar to each other than to students randomly sampled from the school district as a whole, or from the national population of third-graders. This is because in many countries, students are not randomly assigned to classrooms from the population, but rather are assigned to schools based on geographic factors or other characteristics (e.g., aptitude). When assigned based on geography, students within a particular classroom tend to come from a community or community segment that is more homogeneous in terms of morals and values, family background, socio-economic status, race or ethnicity, religion, and even educational preparation than a similar-sized sample randomly sampled from the entire population as a whole. When assigned based on similarity in other characteristics, students are obviously more homogenous than a random sample of the entire population. Further, regardless of similarity or dissimilarity of background, students within a particular classroom share the experience of being in the same environment—the same teacher, physical environment, curriculum, and similar experiences, which may increase homogeneity over time.

The Problem of Independence of Observations

This discussion could be applied to any level of nesting, such as the family, the school district, county, state, or even country. Based on this discussion, we can assert that individuals who are drawn from a group, such as a classroom, school, business, town or city, or health care unit, will be more homogeneous than if individuals were randomly sampled from a larger population. This is often referred to as a *design effect*.

Because these individuals tend to share certain characteristics (environmental, background, experiential, demographic, or otherwise), observations based on these individuals are not fully independent, yet most statistical techniques require independence of observations as a primary assumption for the analysis. Because this assumption is violated in the presence of hierarchical or nested data, ordinary least squares regression (and ANOVA, and most other parametric statistical procedures)

produces standard errors that are too small (unless these so-called design effects are incorporated into the analysis). In turn, this leads to an inappropriately increased probability of rejection of a null hypothesis than if: (a) an appropriate statistical analysis was performed, or (b) the data included truly independent observations.

The Problem of How to Deal with Cross-Level Data

It is often the case in educational research that a researcher is interested in understanding how environmental variables (e.g., teaching style, teacher behaviors, class size, class composition, district policies or funding, or even state or national variables, etc.) affect individual outcomes (e.g., achievement, attitudes, retention, etc.). But given that outcomes are gathered at the individual level, and other variables at classroom, school, district, state, or nation level, the question arises as to what the unit of analysis should be, and how to deal with the cross-level nature of the data.

One strategy (called dis-aggregation) would be to assign classroom or teacher (or other group-level) characteristics to all students (i.e., to bring the higher-level variables down to the student level). The problem with this approach, is all students within a particular classroom assume identical scores on a variable, clearly violating assumptions of independence of observation.

Another way to deal with this issue (called aggregation) would be to aggregate up to the level of the classroom, school, district, etc. Thus, we could talk about the effect of teacher or classroom characteristics on average classroom achievement. However, there are several issues with this approach, including: (a) that much (up to 80–90%) of the individual variability on the outcome variable is lost, which can lead to dramatic under- or over-estimation of observed relationships between variables (Raudenbush & Bryk, 2002), and (b) the outcome variable changes significantly and substantively from individual achievement to average classroom achievement.

Neither of these strategies constitute a best practice, although they have been commonly found in educational research. Neither of these strategies allow the researcher to ask truly important questions—such as what is the effect of a particular teacher variable on student learning? A third approach, that of HLM, becomes necessary in this age of educational accountability and more sophisticated hypotheses.

HOW DO HIERARCHICAL MODELS WORK? A BRIEF PRIMER

The goal of this paper is to introduce the concept of hierarchical modeling, and explicate the need for the procedure. It cannot fully communicate the nuances and procedures needed to actually perform a hierarchical analysis. The reader is encouraged to refer to Raudenbush and Bryk (2002) and the other suggested readings for a full explanation of the conceptual and methodological details of hierarchical linear modeling.

The basic concept behind hierarchical linear modeling is similar to that of OLS regression. On the base level (usually the individual level, or the level where repeated measures are taken within a particular individual, referred to here as level 1, the lowest level of your data), the analysis is similar to that of OLS regression: an outcome variable is predicted as a function of a linear combination of one or more level 1 variables, plus an intercept, as so:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_1 + \ldots + \beta_{kj}X_k + r_{ij}$$

where β_{0j} represents the intercept of group j, β_{1j} represents the slope of variable X_1 of group j, and r_{ij} represents the residual for individual i within group j. On subsequent levels, the level 1 slope(s) and intercept become dependent variables being predicted from level 2 variables:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_1 + \ldots + \gamma_{0k}W_k + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_1 + \ldots + \gamma_{1k}W_k + u_{1j}$$

and so forth, where γ_{00} and γ_{10} are intercepts, and γ_{01} and γ_{11} represent slopes predicting β_{0j} and β_{1j} respectively from variable W_1. Through this process, we accurately model the effects of level 1 variables on the outcome, and the effects of level 2 variables on the outcome. In addition, as we are predicting slopes as well as intercepts (means), we can model cross-level interactions, whereby we can attempt to understand what explains differences in the relationship between level 1 variables and the outcome. Those of you more mathematically inclined will also note that several different error terms (i.e., *r* and *u* terms) are computed in this process, rather than just a simple residual present in OLS regression.

The advantages of HLM over aggregation and disaggregation have been explored in many places, including Osborne (2000; 2008). In brief, failing to appropriately model multilevel data can lead to under-estimation of standard errors, substantial mis-estimation of effects and variance accounted for, and errors of inference.

ADVANCED TOPICS IN HLM

As many authors have discovered in the years since HLM became available, there are many applications for these analyses. Generalizations to 3- and 4- level models are available, as are logistic regression analogues (e.g., HLM with binary or polytomous outcomes), applications for meta-analysis, powerful advantages for longitudinal analysis (as compared to other methods such as repeated measures ANOVA), and many of the fun aspects of OLS regression (such as modeling curvilinear effects) is possible in HLM as well.

There is little downside to HLM, aside from the learning curve. If one were to use HLM on data where no nesting, dependence, or other issues were present, one would get virtually identical results to OLS regression from statistical software packages such as SPSS or SAS or R.

The rest of this chapter is devoted to two simple examples that represent common questions within educational (and many areas of social science) research: (a) how do individual- and school-level variables affect student achievement, and (b) can we understand growth or change in an individual as a function of individual or environmental traits?

MODELING VARIABLES AT DIFFERENT LEVELS

Our first example is an application of HLM to use variables from different levels. In this case, we have two variables at the student level (family socio-economic status and student locus of control) and two school-level variables (percent of students who meet a particular definition of economic need in the USA (receiving free lunch in school) and percentage of students who belong to disadvantaged racial minority groups) predicting student achievement test scores.

AN EMPIRICAL COMPARISON OF THE THREE APPROACHES TO ANALYZING HIERARCHICAL DATA

In this section we illustrate the outcomes achieved by each of the three possible analytic strategies for dealing with hierarchical data:

- disaggregation (bringing school level data down to the individual level),
- aggregation (bringing individual level data in summarized fashion up to the school level), and
- hierarchical linear modeling (appropriately modeling variables at the level they were gathered).

Data for this example were drawn from the National Education Longitudinal Survey of 1988 (http://nces.ed.gov/surveys/nels88/), a nationally- representative sample of approximately 28,000 eighth graders in the United States. The analysis we performed predicted composite achievement test scores (math, reading combined) from student socioeconomic status (family SES), student locus of control (LOCUS), the percent of students in the school who are members of racial or ethnic minority groups (%MINORITY), and the percent of students in a school who receive free lunch (%LUNCH, an indicator of school poverty). We expect SES and LOCUS to be positively related to achievement, and %MINORITY and %LUNCH are expected to be negatively related to achievement. In these analyses, 995 of a possible 1004 schools had sufficient data to be included.

Disaggregated analysis. In order to perform the disaggregated analysis, the level 2 values were assigned to all individual students within a particular school. A standard multiple regression was performed via SPSS entering all predictor variables simultaneously. The resulting model was significant, with $R = .56$, $R^2 = .32$, $F\ (4,22899) = 2648.54$, $p < .0001$. The individual regression weights and significance tests are presented in Table 1.

Table 1. Comparison of three analytic strategies

	Disaggregated			*Aggregated*			*Hierarchical*		
Variable	*B*	*SE*	*t*	*B*	*SE*	*t*	*B*	*SE*	*t*
SES	4.97_a	.08	62.11***	7.28_b	.26	27.91***	4.07_c	.10	41.29***
LOCUS	2.96_a	.08	37.71***	4.97_b	.49	10.22***	2.82_a	.08	35.74***
%MINORITY	-0.45_a	.03	−15.53***	-0.40_a	.06	−8.76***	-0.59_b	.07	−8.73***
%LUNCH	-0.43_a	.03	−13.50***	0.03_b	.05	0.59	-1.32_c	.07	−19.17***

Note: B refers to an unstandardized regression coefficient, and is used for the HLM analysis to represent the unstandardized regression coefficients produced therein, even though these are commonly labeled as betas and gammas. SE refers to standard error. Bs with different subscripts were found to be significantly different from other Bs within the row at $p < .05$.
**** $p < .0001$.*

All four variables were significant predictors of student achievement. As expected, SES and LOCUS were positively related to achievement, while %MINORITY and %LUNCH were negatively related.

Aggregated Analysis

In order to perform the aggregated analysis, all level 1 variables (achievement, LOCUS, SES) were aggregated up to the school level (level 2) using school-based means. A standard multiple regression was performed via SPSS entering all predictor variables simultaneously. The resulting model was significant, with $R = .87$, $R^2 = .75$, $F\ (4,999) = 746.41$, $p < .0001$. Again as expected, both average SES and average LOCUS were positively related to achievement, and %MINORITY was negatively related. In this analysis, %LUNCH was not a significant predictor of average achievement.

HLM Analysis

Finally, a hierarchical linear analysis was performed via HLM, in which the respective level 1 and level 2 variables were modeled appropriately. Note also that all level 1 predictors were centered at the group mean, and all level 2 predictors were centered at the grand mean. The resulting model demonstrated goodness of fit (Chi-square for change in model fit = 4231.39, 5 df, $p < .0001$). As seen in Table 1, this analysis reveals expected relationships— positive relationships between achievement and the level 1 predictors (SES and LOCUS), and strong negative relationships between achievement and the level 2 predictors (%MINORITY and %LUNCH). Further, the analysis revealed significant interactions between SES and both level 2 predictors, indicating that the slope for SES gets weaker as %LUNCH and as %MINORITY

increases. Also, there was an interaction between LOCUS and %MINORITY, indicating that as %MINORITY increases, the slope for LOCUS weakens. There is no clearly equivalent analogue to R and R^2 available in HLM.

COMPARISON OF THE THREE ANALYTIC STRATEGIES AND CONCLUSIONS

We assume that the third analysis represents the best estimate of what the "true" relationships are between the predictors and the outcome. Unstandardized regression coefficients (b in OLS, β and γ in HLM) were compared statistically via procedures outlined in Cohen and Cohen (1983).

Neither of the first two analyses appropriately modeled the relationships of the variables. The disaggregated analysis significantly overestimated the effect of SES, and significantly and substantially underestimated the effects of the level 2 effects. The standard errors in this analysis are generally lower than they should be, particularly for the level 2 variables (a common issue when assumptions of independence are violated).

The second analysis overestimated the multiple correlation by more than 100%, overestimated the regression slope for SES by 79% and for LOCUS by 76%, and underestimated the slopes for %MINORITY by 32% and for %LUNCH by 98%.

These analyses reveal the need for multilevel analysis of multilevel data. Neither OLS analysis accurately modeled the true relationships between the outcome and the predictors. Additionally, HLM analyses provide other benefits, such as easy modeling of cross-level interactions, which allows for more interesting questions to be asked of the data. For example, in this final analysis we could examine how family and school poverty interact, something not possible unless the multilevel data are modeled correctly.

MODELING LONGITUDINAL CHANGE OVER TIME

Our example attempts to explain changes in individual mood (or affect) over time as a function of individual traits such as neuroticism. Neuroticism, and the constant elevated levels of negative affect that accompany the trait over years or decades, can lead to a negative emotion "hair trigger" (Kendler, Thornton, & Gardner, 2001; Wilson, Bienas, Mendes de Leon, Evans, & Bennett, 2003). This process suggests that with the passage of time, people high in neuroticism may become more susceptible to elevated negative affect. Because neuroticism is associated with more variability in behavior and experience (Eid & Diener, 1999; Eysenck & Eysenck, 1985; Moskowitz & Zuroff, 2004; Neupert, Mroczek, & Spiro, 2008), we use the current example to examine whether individual differences in neuroticism are associated with differential trajectories of negative affect over time. Before we get into the example, however, we should stop and discuss the challenges of working with longitudinal data. First, it is often the case that longitudinal studies have difficulty measuring all individuals at exactly the same time, or within identical time intervals,

yet that is an assumption of RMANOVA. Next, assumptions of RMANOVA are rarely met in practice, potentially seriously compromising the validity of the results. Finally, missing data can severely cripple a RMANOVA analysis, and missing data are rarely handled appropriately (for more on missing data, see (Osborne, 2012, Chapter 6). However, HLM has none of these drawbacks. So long as any individual has one or more data points, they can be included in a repeated measures HLM analysis. Furthermore, unequal time intervals between measurements can be explicitly modeled to remove as much potential for error variance as possible. Growth curves are easily modeled as in OLS regression (i.e., using quadratic and cubic terms to model curvilinearity), and the estimation procedures in HLM tend to produce smaller standard errors, all of which make HLM a best practice for longitudinal data analysis.

Data for the example are from the National Study of Daily Experiences (NSDE) and are publicly available (www.icpsr.umich.edu). Respondents were 1031 adults (562 women, 469 men), all of whom had previously participated in the Midlife in the United States Survey MIDUS), a nationally representative telephone-mail survey of 3032 people, aged 25–74 years, carried out in 1995–1996 under the auspices of the MacArthur Foundation Research Network on Successful Midlife Development (for descriptions of the MIDUS project, see Brim, Ryff, & Kessler, 2004; Keyes & Ryff, 1998; Lachman & Weaver, 1998; Mroczek & Kolarz, 1998). Respondents in the NSDE were randomly selected from the MIDUS sample and received $20 for their participation in the project. Over eight consecutive evenings, respondents completed short telephone interviews about their daily experiences. Data collection was planned to span an entire year (March 1996 to March 1997), so 40 separate "flights" of interviews with each flight representing the eight-day sequence of interviews from approximately 38 respondents were used. The initiation of flights was staggered across the day of the week to control for the possible confounding between day of the study and day of week. Of the 1242 MIDUS respondents we attempted to contact, 1031 agreed to participate, yielding a response rate of 83%. Respondents completed an average of 7 of the 8 interviews resulting in a total of 7229 daily interviews.

The equations below were used to examine change in negative mood over time as a function of individual differences in neuroticism. In this example, individual variability is represented by a two-level hierarchical model where level 1 reflects the daily diary information nested within the person-level information at level 2.

$$\text{Level 1: MOOD}_{ti} = \beta_{0ti} + \beta_{1ti}(\text{DAY}) + r_{ti} \tag{1}$$

$$\text{Level 2: } \beta_{0i} = \gamma_{00} + \gamma_{01}(\text{NEUROT}) + u_{0i} \tag{2}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}(\text{NEUROT}) + \text{u}_{1i} \tag{3}$$

Following the guidelines by Raudenbush and Bryk (2002), the lettered subscripts in the equations depict the nesting structure. Days/timepoints are represented by *t* (level 1) and individuals are represented by *i* (level 2).

In Equation 1, the intercept (β_{0ti}) is defined as the expected level of negative mood for person *i* on the first day of the study (i.e., DAY = 0) because the variable was uncentered. Although it would have been possible to person-mean or grand-mean center DAY, we chose to leave this variable uncentered so that the interpretation of the intercept would be associated with a particular timepoint (i.e., first day of the study). The change slope, β_{1ti}, is the expected change in negative mood associated with time. The error term (r_{ti}) represents a unique effect associated with person *i* (i.e., individual fluctuation around their own mean). The level 1 intercept and slope become the outcome variables in the level 2 equations. Equation 2 includes a main effect of neuroticism and therefore tests to see if neuroticism is related to the average level of psychological distress (γ_{01}). The intercept (γ_{00}) represents the average level of negative mood for someone with average neuroticism scores because neuroticism was centered at the grand mean (CNEUORT [centered neuroticism] = 0). We chose to grand-mean center neuroticism to maintain an interpretable value of the intercept and to reduce nonessential multicollinearity for the cross-level interaction. Equation 3 provides the estimate (γ_{10}) representing change for the sample: the average relationship between day and negative mood. A cross-level interaction is represented by γ_{11} and tests whether there were neuroticism differences (Level 2) in change in negative mood over time (Level 1 relationship). Interindividual fluctuations from the average level and slope are represented by u_{0i} and u_{1i}, respectively.

We chose to present this example using SAS PROC MIXED (1997) because many people like the ability to reduce, manage, and analyze in a single software package. Detailed descriptions of the commands are described elsewhere (e.g., Neupert, in press; Singer, 1998), so we focus on the main components here. Figure 1 represents the commands that were used to test Equations 1–3. DAY (Level 1), CNEUROT (Level 2 grand-mean centered neuroticism), and DAY*CNEUROT (cross-level interaction) are included as predictors in the MODEL statement. The /SUBJECT = command specifies the nesting structure and alerts SAS that DAY is a level 1 variable and CNEUROT is a level 2 variable. The MODEL statement provides γ_{01} (CNEUROT)

```
proc mixed data=merged noclprint covtest;
title 'neuroticism differences in change of daily negative affect over time';
class caseid;
model mood= day cneurot day*cneurot
/solution ddfm=bw;
random intercept day /subject=caseid type = un;
run;
```

Figure 1. SAS commands.

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	CASEID	12.8791	0.7394	17.42	<.0001
UN(2,1)	CASEID	-1.1138	0.09028	-12.34	<.0001
UN(2,2)	CASEID	0.1539	0.01345	11.44	<.0001
Residual		5.0161	0.09936	50.49	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	2.7587	0.1265	1017	21.81	<.0001
DAY	-0.2153	0.01695	6133	-12.70	<.0001
cneurot	2.2995	0.1903	1017	12.09	<.0001
DAY*cneurot	-0.1393	0.02551	6133	-5.46	<.0 001

Figure 2. SAS output for random effects (covariance parameter estimates) and fixed effects (solution for fixed effects).

from Equation 2 as well as the default γ_{00} (intercept) from Equation 2. Estimates from Equation 3 are γ_{10} (DAY) and γ_{11} (DAY*CNEUROT). Adding a variable name to the RANDOM statement allows the slope between the specified variable and the dependent variable to vary across level 2 units. Therefore, only level 1 variables can be added to the RANDOM statement. In this example, DAY was added to the RANDOM statement to allow the change in negative mood over time to vary across people. Note that this corresponds to u_{1i} in Equation 3. If DAY was not added to the RANDOM statement, the change (β_1) slope would be constrained to be equal across all level 2 units (people). An option has been added that specifies the structure of the variance-covariance matrix for the intercepts and slopes.

Figure 2 displays the SAS output for the fixed and random effects. The four rows for Covariance Parameter Estimates correspond to the four random effects. The first row (UN 1,1) corresponds to τ_{00}, reflecting the remaining level 2 variance in the level of MOOD after accounting for CNEUROT. The second row (UN 2,1) corresponds to τ_{10}, reflecting the covariance between the intercept and slope. The third row (UN 2,2) corresponds to τ_{11}, reflecting the variance around the slope between DAY and MOOD. The fourth row (Residual) corresponds to σ^2 and reflects the remaining level 1 variance in MOOD after accounting for DAY. Note that all four of the random effects are significant. This indicates that there is still significant variance left to explain at level 1 (σ^2) and level 2 (τ_{00}) and it also shows that there is a significant relationship between the intercept of MOOD and the relationship between DAY and MOOD (significant covariance: τ_{10}). Lastly, the significant τ_{11} indicates that there is variance across people in the relationship between DAY and MOOD; that is, not all people change the same way with respect to their mood.

The Solution for Fixed Effects provides the output for the four gamma coefficients (represented in Equations 2 and 3). The Intercept corresponds to γ_{00} and indicates that

the average level of negative mood on the first day of the study for someone with average neuroticism was 2.76. The next row corresponds to γ_{10} and indicates that there is a significant and negative relationship between day and mood. For each additional day that someone stays in the study, their negative mood decreases by 0.2153 units. Notice that the number of degrees of freedom for this relationship is 6133, reflecting the fact that DAY is a level 1 variable and is based on the number of days rather than the number of people in the sample (i.e., *df* for the Intercept and CNEUROT effects are based on the number of people). The third row corresponds to γ_{01} and indicates that there are significant neuroticism differences in the level of negative mood. Not surprisingly, people with higher levels of neuroticism report more negative mood compared to people with lower levels of neuroticism. The final row represents γ_{11} and indicates that changes in negative mood over time depend on individual differences in neuroticism. Decomposing this interaction reveals that people high in neuroticism (Mean + 1SD) decreased their negative mood at a faster rate compared to people low in neuroticism (Mean − 1SD). Given the large individual differences in negative mood as a function of neuroticism, this pattern may reflect a kind of floor effect for those with low neuroticism who started at lower levels of distress.

CONCLUSION

In this chapter we highlighted important concepts related to HLM, particularly the advantages of HLM over more traditional techniques like regression on aggregated or disaggregated data, repeated measures ANOVA, etc. We demonstrated how it can be used in educational research and the broader field of social science, and provided an example of a growth curve analysis. HLM is widely regarded as a best practice and readers are strongly urged to consider using it because it addresses interesting questions.

REFERENCES

Brim, O. G., Ryff, C. D., & Kessler, R. C. (2004). *How healthy are we? A national study of well-being at midlife*. Chicago: University of Chicago Press.

Eid, M., & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology, 76*, 662–676.

Eysenck, H. J., & Eysenck, M. W. (1985). *Personality and individual differences: A natural science approach.* New York: Plenum.

Kendler, K. S., Thornton, L. M., & Gardner, C. O. (2001). Genetic risk, number of previous depressive episodes, and stressful life events in predicting onset of major depression. *American Journal of Psychiatry, 158*, 582–586.

Keyes, C. L. M., & Ryff, C. D. (1998). Generativity in adult lives: Social structural contours and quality of life consequences. In D. P. McAdams, & E. de St. Aubin (Eds.), *Generativity and adult development: How and why we care for the next generation* (pp. 227–263). Washington, DC: American Psychological Association.

Lachman, M. E., & Weaver, S. L. (1998). Sociodemographic variations in the sense of control by domain: Findings from the MacArthur studies on midlife. *Psychology and Aging, 13*, 553–562.

Moskowitz, D. S., & Zuroff, D. C. (2004). Flux, pulse, and spin: Dynamic additions to the personality lexicon. *Journal of Personality and Social Psychology, 86*, 880–893.

Mroczek, D. K., & Kolarz, C. M. (1998). The effect of age on positive and negative affect: A developmental perspective on happiness. *Journal of Personality and Social Psychology, 75*, 1333–1349.

Neupert, S. D. (in press). Emotional reactivity to daily stressors using a random coefficients model with SAS PROC MIXED: A Repeated Measures Analysis. In G. D. Garson (Ed.), *Hierarchical linear modeling handbook*. Thousand Oaks, CA: Sage.

Neupert, S. D., Mroczek, D. K., & Spiro, A. III. (2008). Neuroticism moderates the daily relation between stressors and memory failures. *Psychology and Aging, 23*, 287–296.

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation, 7*(1).

Osborne, J. W. (2008). *Best practices in quantitative methods*. Thousand Oaks, CA: Sage Publishing.

Osborne, J. W. (2012). *Best practices in data cleaning*. Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* (Vol. 1). thousand oaks, CA: Sage Publications.

SAS Institute (1997). SAS/STAT software: Changes and enhancements through Release 6.12. Cary, NC: SAS Institute.

Singer, J. D. (1998). Using SAS Proc Mixed to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*, 323–355.

Wilson, R. S., Bienas, J. L., Mendes de Leon, C. F., Evans, D. A., & Bennett, D. A. (2003). Negative affect and mortality in older persons. *American Journal of Epidemiology, 158*, 827–835.

D. BETSY MCCOACH, JOHN P. MADURA, KAREN E. RAMBO-HERNANDEZ, ANN A. O'CONNELL & MEGAN E. WELSH

10. LONGITUDINAL DATA ANALYSIS

INTRODUCTION TO LONGITUDINAL DATA ANALYSIS

Longitudinal data analysis is a very broad, general term for the analysis of data that are collected on the same units across time. Longitudinal data are sometimes referred to as repeated measures data or panel data (Hsiao, 2003; Frees, 2004). A variety of statistical models exist for analyzing longitudinal data. These models include autoregressive or Markov chain models, latent transition models, individual growth curve models, and growth mixture models, just to name a few. To determine the correct model for the analysis of longitudinal data, first the researcher must have a substantive theory about whether and how the data should change over time and what the relationships are among the observations across time. For example, imagine that a researcher collects mood data on adults every day for three months. These data are longitudinal. Although the researcher would expect to see day to day changes in mood, he or she would probably not expect to see any "growth" in mood across time. Is mood on any given time predicted by a person's overall mean mood and some amount of random daily fluctuation or error? Is today's mood related to yesterday's mood? Is the relationship between mood on day one and mood on day 3 completely mediated by mood on day 2? If so, then the analysis of such data requires a model that allows for a correlation between adjacent time points, but does not require a model that allows for growth over time. One model common longitudinal model that allow for correlations across time are called autoregressive models or Markov chain models, and are quite common in the structural equation modeling literature (Bast & Reitsma 1997; Curran 2000; Kenny & Campbell 1989; Marsh 1993). In autoregressive models, "a variable is expressed as an additive function of its immediately preceding value plus a random disturbance" (Bollen & Curran, 2006, p. 208). For more information about models of this type, the interested reader should consult (Bollen & Curran, 2004).

It is impossible to do justice to all potential longitudinal models within one chapter. Thus, in this chapter, we will focus on one specific type of longitudinal model that has become quite popular in the research literature over the past decade: the individual growth model. We will present this model within a multilevel framework. Our choice to focus on individual growth models stems from their popularity and their applicability to a large range of research questions and problems that involve the estimation of systematic growth or decline over time. We choose the multilevel

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 199–230.

framework, given that multilevel growth models seamlessly handle unbalanced data. Data are balanced if all units are measured on the same data collection schedule (i.e., at the same time points). Data are considered unbalanced if data are collected on different schedules or at different time points (Skrondal & Rabe-Hesketh, 2008). In our experience, multilevel growth models accommodate a wide range of data structures and handle a wide range of data analytic problems. Further, the framework can be easily modified to include other types of models (e.g. random intercept models or growth models with more complex error covariance structures), adding to the flexibility of the approach. However, we caution the reader not to treat the hammer that we present in this chapter as the only tool to deal with longitudinal data.

Introduction to Models of Individual Growth within a Multilevel Framework

Anytime we ask questions about growth or decline in some area, we are implicitly asking questions that involve the measurement of systematic change over time. Such questions might include: How do students' reading skills develop between kindergarten and fifth grade? Is this growth steady or does the rate of growth change over time? What is the shape of this growth trajectory? Do different people tend to change in the same way over time? Or is there a great deal of variability between people in terms of their rate of change over time? Finally, we often want to understand what factors help to predict the rate at which change occurs, or which variables allow us to understand inter-individual differences in the rate of change. In this chapter, we briefly introduce readers to the estimation of individual growth curve models using multilevel modeling. Fuller and more technical treatments of this topic appear in Raudenbush and Bryk (2002), Singer and Willet (2003).

Why Do We Need Growth Curve Modeling?

Before we embark on our journey into individual growth curve modeling, it is important to understand the inadequacies inherent in using two wave studies to measure change. The simplest type of change is a difference score, which attempts to model the difference between post-test and pre-test achievement as a function of the presence of a treatment or some other educational variable. Although simple to calculate, there are inherent difficulties in using difference scores to examine student growth (Cronbach & Furby, 1970).

First, measurement error in pre-test or post-test scores reduces the precision of the estimate of the treatment effect (Rogosa, Brandt, & Zimowski, 1982; Raudenbush, 2001). When measurement error is intertwined with the pre-test or post-test scores (or both), then "true" change and measurement error become confounded, thus the observed change between two scores may either overestimate or underestimate the degree of "true" change. For example, a student's pre-test score could be too high and their post-test score could be too low because of measurement error, leading to an erroneous conclusion that the treatment had little or no effect when, in reality, measurement error

is masking the true effect. Multiple data points are needed to extricate the confounded nature of the measurement error and true change (Singer & Willett, 2003). In addition, with only two time points, all change must be linear (and perfectly so). Thus, there is no way to examine the shape of the change trajectory across time.

Rogosa et al. (1982) recommend that "when used wisely, multiwave data will yield far better determinations of individual change than two wave data" (p. 745). The conceptualization of growth as how individuals change across time and interest in modeling the variables that predict change between as well as change within people allows for a much fuller picture of change; however, such models require the use of longitudinal data (Singer & Willett, 2003). Therefore, analyses of growth or change require data collected across at least three time points.

Multivariate Repeated Measures

Another common way to examine change is to use multivariate repeated measures (MRM) designs, of which the most common analysis is repeated measures analysis of variance (RANOVA). Although MRM allow for several waves of data collection, there are several restrictions that traditional MRM place on the measurement of change. One problematic restriction of MRM is the requirement of a fixed time-series design. The distance between time points must be consistent across all persons, and the data collection must occur at the same time for all persons (Raudenbush & Bryk, 2002). If any student is missing data at any time point during the data collection schedule, that student is typically deleted from the analyses and all information provided by that student is lost (Raudenbush & Bryk, 2002). This has two adverse consequences. First, it decreases statistical power and lowers the precision of the estimates of growth. Second, it introduces a selection bias issue into the data analysis. Therefore, by eliminating these people from the analysis, we are likely to introduce bias into our estimates of growth. Luckily, using multilevel growth models, researchers can retain units even when observations from some time points are missing, and they can fit growth models to time unstructured data.

What Do We Need to Measure Change Using Multilevel Growth Models?

To study change, we need data collected from the same units across multiple time points. As alluded to earlier, using growth modeling techniques also requires collecting at least three waves of data. However, growth curve models with only three time points only allow for the estimation of linear growth trajectories. The estimation of curvilinear growth trajectories (as shown in Figure 1) requires data collected across 4 or more time points. With larger numbers of time points, it is possible to fit increasingly complex growth functions, which can be very informative if we want to understand how units change over time. When designing longitudinal studies, it is important to consider both the number and the spacing of data collection points to accurately capture change across time. When data points are too infrequent,

or when there are too few data points, it may not be possible to accurately model the functional form of the change.

In addition to collecting data on the same units over at least three waves, growth curve modeling requires two more conditions. First, there must be an accurate measure of time. If scores are collected across three time points, we need to know how much time elapsed between time point one and time point two and how much time elapsed between time point two and time point three. Conceptually, time represents the *x*-axis in a growth curve model (see Figure 1), and the score on the outcome variable is plotted on the *y*-axis. We need to know the distance between testing occasions so that we can plot the dependent variable or the "*y*" score, on the correct location of the *x*-axis to correctly model the functional form of the growth. Several measures of time are equally reasonable, e.g., a person's age in months at each measurement occasion or the amount of time (weeks/months/years) that has elapsed between measurement occasions (McCoach et al., 2012).

The second requirement is that the assessment score must be psychometrically sound (e.g., scores are valid and reliable) and must be comparable over time (Singer & Willett, 2003). The measurement scale must also remain consistent across administrations so that a unit that has not changed across time would receive the same score at each measurement occasion. This requirement is met when either the same assessment is used at multiple time points or when the assessments have had their scores placed onto the same metric through a process called vertical scaling (Singer & Willett, 2003).

If assessments have had their scores placed on the same scale so that we can directly compare scores over time, they are said to be vertically scaled. Because vertically scaled assessments yield comparable scores, they are useful for modeling growth across time for constructs such as achievement that cannot be measured using the same assessment across multiple time points. Think of the vertical scaling procedure as placing the results of multiple years of data on the same equal interval "ruler" so that growth may be measured in the same metric. Height in inches yields an equivalent metric across time; a height of 5 feet references the same amount of height regardless of who is measured or the age at which they are measured. In the absence of vertical scaling, the difference between the two scores does not measure growth in any meaningful way because the two scores are on two different, unlinked scales. For example, if a teacher gives a 25 word spelling test every week, and the words on the spelling test differ from week to week, there is no way to determine the amount of growth that a student has made in spelling throughout the year by plotting the spelling test scores across time. Because many academic tests are scaled within specific content area but are not designed to place scores along the same metric across time points, comparing students' scores across time cannot provide information on student growth. In addition to having a scale that provides a common metric across time, the validity of the assessment must remain consistent across multiple administrations of the assessment (Singer & Willett, 2003).

HLM Models

HLM individual growth models allow for the measurement time points to vary across units and have the ability to capture the nested nature of the data. For example, in educational contexts, observations across time are nested within students and those students are nested within schools (Kline, 2005; Raudenbush & Bryk. 2002). In HLM growth models, both individual and group trajectories are estimated (Raudenbush & Bryk, 2002). The primary advantage to using HLM to model individual growth is that HLM allows for a great degree of flexibility in the structure of time. Therefore, every person within a dataset can have their own unique data collection schedule (Stoel & Garre, 2011). When the length of time between data collection points varies from person to person, we refer to the data as "time unstructured." Conventional multilevel models handle time unstructured data seamlessly because time is represented as an explicit independent variable within the dataset.

The Basic Two-Level HLM Model for Linear Growth

In an HLM individual growth model, level 1 describes an individual (or unit)'s growth trajectory across time. A simple two-level linear growth model is illustrated below.

Level 1:

$$y_{ti} = \pi_{0i} + \pi_{1i}(time_{ti}) + e_{ti} \tag{1}$$

Level 2:

$$\pi_{0i} = \beta_{00} + \beta_{01}(gender_i) + r_{0i}$$
$$\pi_{1i} = \beta_{10} + \beta_{11}(gender_i) + r_{1i}$$

The observations across time are nested within persons. The level-1 equation models individual trajectories or within individual variability across time. The dependent variable (y_{ti}) is the score for individual *i* at time *t*. We predict that y_{ti}, person *i*'s score at time *t* is a function of three things: 1) the intercept, π_{0I}, π_{ti}, (which is the predicted value of y_{it} when time = 0); 2) the product of a constant rate of change and time, $\pi_{1I}(time_{ti})$, and 3) individual error, e_{ti}. In a simple linear model, the time slope, π_{1I}, represents the linear rate of change over time. Notice that both the slope and the intercept contain a subscript *i*. This means that a separate slope and intercept are estimated for each person in the sample. The deviation of an individual from his/her predicted trajectory (e_{ti}) can be thought of as the measurement error associated with that individual's estimate at that time point. The pooled amount of error variability within individuals' trajectories is estimated by the variance of e_{ti}[var(e_{ti}) = σ^2] (Bryk & Raudenbush, 1988; Raudenbush & Bryk, 2002).

The level-2 equation models the average growth trajectories across students and deviations from those averages. The second level of the multilevel model specifies that the randomly varying intercept (π_{0i}) for each individual (i) is predicted by an overall intercept (β_{00}), the effects of level-2 variables on the intercept, and r_{0i}, the level-2 residuals represent the difference between the model implied intercept and the individual i's observed intercept. Likewise, the randomly varying linear growth slope (π_{0i}) for each individual (i) is predicted by an overall intercept (β_{10}), the effects of level-2 variables on the linear growth slope, and r_{0i}, the level-2 residual, which represents the difference between person i's model predicted linear growth slope and his or her actual growth slope.

The level-2 model allows for person-specific variables to explain variation in individuals' growth trajectories. In other words, individual growth patterns can be explained by person level predictors such as gender, socio-economic status, treatment group, etc. Ideally, person-level covariates should help to explain some of the inter-individual variability in terms of where people start (the intercept) or how fast they grow (the slope).

In our current example, gender is coded as male = 0, female = 1. Time is coded 0, 1, and 2. Therefore, the intercept, (π_{0i}) represents the predicted initial status of person i. Thus, if the student is female, the intercept (π_{0i}) is predicted from the expected value of male students on the initial measure (β_{00}) and the expected differential between males and females in initial scores (β_{0i}). The difference between the model predicted intercept, based on the level 2 model, and the person's actual intercept is captured in the random effect, r_{0i}. Likewise, the linear growth parameter (π_{0i}) is predicted from the mean growth of all male students (β_{i0}) and the expected differential in growth between males and females (β_{1i}). The difference between the model predicted slope, based on the level 2 model, and the person's actual slope is captured in the random effect, r_{0i}. The amount of between person variability in the intercept after accounting for gender is estimated by the variance of u_{0i} [$var(u_{0i}) = \tau_{00}$], and the amount of between person variability in the time slope after accounting for gender is estimated by the variance of u_{0i} [$var(u_{1i}) = \tau_{11}$] (Bryk & Raudenbush, 1988; Raudenbush & Bryk, 2002).

The linear growth model is the simplest model. However, this model can be extended through the incorporation of time varying covariates, piecewise regression terms, or polynomial terms to model non-linearity that occurs in the growth trajectory. We briefly consider the use of time varying covariates and piecewise regression models. Then we provide a more detailed description of polynomial (quadratic) growth models and provide an example of a quadratic model of growth.

Time-Varying Covariates

Time-varying covariates are variables whose values can change over time and that can enhance the model's capacity to appropriately capture observed patterns of individual change. Adding a time varying covariate (TVC) to equation 1, and removing gender as a level-2 predictor, results in the following model:

$$
\begin{aligned}
Y_{ti} &= \pi_{0i} + \pi_{1i}(time)_{ti} + \pi_{2i}(TVC)_{ti} + e_{ti} \\
\pi_{0i} &= \beta_{00} + r_{0i} \\
\pi_{1i} &= \beta_{10} + r_{1i} \\
\pi_{2i} &= \beta_{20} + r_{2i}
\end{aligned} \tag{2}
$$

By estimating a randomly varying slope for the TVC (indicated by inclusion of its associated random effect, r_{2i}), the relationship between the time varying covariate and the dependent variable varies across people. In other words, for some people the effect of the time varying covariate on the dependent variable could be quite strong and positive, whereas for others it could be weak, or even negative. Although the value of the time-varying covariate changes across time within people, the parameter value estimating the effect of the time-varying covariate on the dependent variable is assumed to be constant across time. In other words, the effect of the time varying covariate is constant across time within person, but varies across people. For example, in a study of students' reading growth over time, the number of minutes that a student spends engaged in independent reading per week could be an important time-varying covariate. At every assessment point, the researcher measures both the dependent variable (reading comprehension), and the independent variable (the number of minutes of independent reading per week). Although the number of minutes of independent reading that a student engages in per week can change at each data collection point, the estimated relationship between independent reading and reading comprehension remains constant across time for each person.

There are ways to ease this assumption that the relationship between the time varying covariate and the response variable is constant across time within persons. For example, one can build an interaction term between time and the time-varying covariate by creating a variable that equals the product of the two variables (Singer & Willett, 2003). Adding the interaction term to the model results in the following equation:

$$
\begin{aligned}
Y_{ti} &= \pi_{0i} + \pi_{1i}(time)_{ti} + \pi_{2i}(TVC)_{ti} + \pi_{3i}(time*TVC)_{ti} + e_{ti} \\
\pi_{0i} &= \beta_{00} + r_{0i} \\
\pi_{1i} &= \beta_{10} + r_{1i} \\
\pi_{2i} &= \beta_{20} + r_{2i} \\
\pi_{3i} &= \beta_{30} + r_{3i}
\end{aligned} \tag{3}
$$

The parameter estimate for the interaction term, β_{30}, helps to capture the differential effect of the time varying covariate across time. If the time varying covariate is a continuous variable, it should be centered to aid in the interpretation of the parameter estimates. In our example above, if the researcher centers the number of minutes a student reads per week at the grand mean for all occasions and persons in

the sample, then β_{00} is the overall estimated initial reading score for students who read an average number of minutes per week. β_{10} represents the expected change in reading scores over time for students reading an average number of minutes per week; and β_{20} represents the effect of an additional minute of reading per week on reading comprehension when time is equal to 0 (i.e., at initial status or baseline). β_{30} captures the differential in the effect of the time varying covariate across time. For example, assume that the growth model yields positive values for β_{10} and β_{20}, the estimates for the linear growth slope and for the effect of time spent reading per week, respectively. In that case, a negative value for β_{30} would indicate that the effect of the time varying covariate (minutes read per week) weakens (gets progressively less positive) across time. If, on the other hand, the parameter estimate for β_{30} is positive, this would indicate that the effect of the time varying covariate on the dependent variable strengthens (get more positive) over time. Allowing the interaction term to vary across people by adding the random effect term r_{3i} implies that the interaction effect, or the change in the effect of the time varying covariate, varies across people. While the introduction of an (randomly varying) interaction between time and a time-varying covariate provides great flexibility in modeling growth, it does increase the number of estimated parameters in the model. For example, the variance/covariance matrix for the random effects now would require ten unique elements, rather than six (as estimating r_{3i} adds a variance and three covariances to the model).

Incorporating time-varying covariates can be a very effective strategy for modeling non-linearity and/or discontinuities in growth trajectories (McCoach & Kaniskan, 2010). Time-varying covariates may be continuous, as in the example above, or categorical. Correct and creative coding of time-varying variables can help to more adequately capture the nature of change in the phenomenon of interest, and thus more accurately reflect the process of change, as well as correlates of that change.

Piecewise Growth Models

Often, growth trajectories may not be modeled well by a single linear slope or rate of change, even after adjusting for time-varying covariates. There may be scenarios in which a growth pattern might be more aptly represented by dividing the trajectory into growth segments corresponding to fundamentally different patterns of change (Collins, 2006). For example, imagine that a reading researcher collects achievement data on elementary students across an entire calendar year, amassing six data points from the beginning of September through the end of August (start of the next academic year). In this case, the time points between September and June capture the span of time for the change in achievement across the school year, whereas the period between June and the end of August captures the span of time for the change in reading scores during the summer (non-instructional) months. The achievement slope is likely to be substantially steeper and constant during instructional months and flatter (or perhaps even negative) during the summer, when students receive

no academic instruction; a single linear growth parameter would not represent the data well in this situation. Piecewise linear growth models "break up the growth trajectories into separate linear components" (Raudenbush & Bryk, p. 178), and can be particularly valuable when comparison of growth rates between the separate components are of interest, or to investigate differences in substantive predictors of growth between the components. Note that a sufficient number of time-points are required to enable modeling of a separate slope for each component.

Piecewise regression techniques conveniently allow for changes in a linear growth slope across time. To achieve these representations, we include multiple time variables into the model to capture the multiple linear growth slopes. If we expect one rate of growth for time points 1–4, and another rate of growth for time points 4–6, we would introduce two time variables. The second time variable always clocks the passage of time, starting at the point at which the discontinuity or change in slope is expected. Following our above example, our two-piece linear growth model would then be expressed as follows:

$$\begin{aligned} y_{ti} &= \pi_{0i} + \pi_{1i}(time\_piece1_{ti}) + \pi_{2i}(time\_piece2_{ti}) + e_{ti} \\ \pi_{0i} &= \beta_{00} + r_{0i} \\ \pi_{1i} &= \beta_{10} + r_{1i} \\ \pi_{2i} &= \beta_{20} + r_{2i} \end{aligned} \tag{4}$$

There are two different ways to code the first piece of the piecewise model, and they will result in different interpretations for the piecewise parameters. The first option is to use the same linear time variable that we introduced earlier, which is centered at the initial time point and continues to clock the passage of time for the duration of the study. This coding system is demonstrated in Table 1. Using this coding scheme, β_{10} is the parameter estimate for first time variable (time_piece1) and captures the baseline growth rate; β_{20} is the parameter estimate for the second time variable and captures the deflection from that baseline growth rate.

Piecewise coding scheme for capturing growth rate and a deflection from baseline growth

Table 1. Coding for baseline and deflected growth pieces

WAVE	*Time_piece1*	*Time_piece2*
1	0	0
2	1	0
3	2	0
4	3	0
5	4	1
6	5	2

The second option creates two separate growth slopes, one that captures the growth rate during the first piece of the piecewise model and one that captures the growth rate during the second piece of the model. To model the piecewise growth as two different growth slopes, we need to create two time variables, each of which clocks the passage of time during only one segment or piece of the piecewise model. In other words, we "stop" the first time variable (time_piece1) and "start" the second time variable (time_piece2) simultaneously, as is demonstrated starting in wave 4 of Table 2. Under this coding scheme, β_{10}, the parameter estimate for first time variable (time_piece1), captures linear growth rate for the first time period (from waves 1–4); and β_{20}, the parameter estimate for the second time variable (time_piece2), captures the linear growth rate for the second time period (waves 4–6). Note that although the coding for time_piece2 is identical across the two coding schemes, it is actually the parameter estimate for β_{20} (the slope for time_piece2) that changes meaning across the two different coding schemes. Also, notice that the coding schemes in Tables 1 and 2 are linearly dependent. Therefore, these two models are statistically equivalent. Further, one can compute the deflection parameter (β_{20}) under coding option 1 directly from the results of coding option 2. To do this, simply subtract β_{20} from β_{10} found from the coding scheme used in Table 2. Similarly, one can compute the second linear growth slope from the coding scheme used in Table 2 by summing β_{20} and β_{10} from the coding scheme used in Table 1.

Multiple changes in linear growth rates can be captured through piecewise models as well. For example, imagine that reading growth is measured in the fall and spring across four school years. Thus, we have 8 data collection points. Theoretically, we might expect reading scores to increase during the school year and remain flat (or even decrease) over the summer. Therefore, one might want to fit two growth trajectories: one for school year growth and another for summer growth. To model these multiple trajectories, we can create two time variables: one that clocks the passage of time from the beginning of the study that occurs during the school year (time_piece1), and another that clocks the passage of time during the summer (time_piece2). If we could assume that school year growth remained constant within child across the multiple years of the study and summer growth also remained constant within child across the study, we could capture the zig-zag pattern of growth across

Table 2. Piecewise coding scheme for capturing two separate growth rates

WAVE	*Time_piece1*	*Time_piece2*
1	0	0
2	1	0
3	2	0
4	3	0
5	3	1
6	3	2

the multiple years of the study with only two different slope parameters: β_{10}, which would capture the school year slope, and β_{20} which would capture the summer slope. The coding for this piecewise model is demonstrated in Table 3.

In summary, creative use of piecewise regression models can capture a variety of patterns of non-linear change as well as discontinuities in growth.

Quadratic Growth Models

For a quadratic growth, the model at level-1 takes the form

$$y_{ti} = \pi_{0i} + \pi_{1i}(time_{ti} - L) + \pi_{2i}(time_{ti} - L)^2 + e_{ti} \quad (5)$$

It is important to note that in most cases, a specific time centering constant, L, for the level-1 predictors should be introduced. Raudenbush and Bryk (2002) note that the choice of the centering constant influences the interpretation of the first order coefficient, π_{1i}. If, for example, time is centered on the first time point, then π_{1i} is defined as the "instantaneous rate of growth at the initial time point." The authors note, however, that centering at the midpoint instead of the first time point has two distinct advantages in quadratic models. The first is that the π_{1i} parameter is then understood as the "average rate of growth." The second advantage is that centering on the midpoint minimizes the correlation between the instantaneous velocity and acceleration parameters, which then has the "effects of stabilizing the estimation procedure" (Raudenbush and Bryk, 2002, p. 182). The choice of centering also affects the interpretation of the intercept, π_{0i}, which represents the predicted value of the individual *i* at time L. Nevertheless, the choice of a centering constant for all longitudinal models, even those with higher order terms, should consider the research design, data analysis goals, and the interpretability of the results. In contrast to the first order coefficient, the π_{2i} does not depend on the choice of centering. The quadratic

Table 3. Piecewise coding scheme for capturing multiple changes in two separate growth rates

WAVE	*Time piece 1*	*Time piece 2*
1	0	0
2	1	0
3	1	1
4	2	1
5	2	2
6	3	2
7	3	3
8	4	3

coefficient, π_{2i}, provides a "curvature" or "acceleration/deceleration" parameter for each person for the entire growth trajectory (Raudenbush and Bryk, 2002).

EXAMPLE: THE EFFECTS OF SCHOOL DEMOGRAPHICS ON SCHOOL ACHIEVEMENT IN SCIENCE

We illustrate the use of a growth model with data on school achievement in 5th grade science over a four year time span. In the study, annual school-level scale scores on the science section of the Connecticut Mastery Test (CMT) were collected for each of the 578 elementary and middle schools that participated in assessment between 2008 and 2011. It is important to note that the units in this example are schools, not students. We expected that the percentage of special education students, the percentage of English language learners in the school, and the percentage of students receiving free and/or reduced lunch in a given school would affect a school's achievement score on a state science test in the 5th grade. We were less clear about how those variables might influence the school's growth on the science achievement test over the four years of the study. For all the schools in the study, science achievement (y_{ti}) was measured over four consecutive time points representing scores for spring 2008, 2009, 2010, and 2011.

A Random-Coefficient Regression Model

Most of the schools in the study displayed a monotonically increasing trend suggesting that science scores are improving over time. A close visual inspection of a sample of four individual school science achievement curves displayed in Figure 1, however, suggests that the data are not best represented by the typical "straight" or linear path. In fact, the growth patterns for the sample schools appear to follow a higher order polynomial (quadratic) growth trajectory.

The graphical picture also suggests that scores plateau to a single vertex, which eliminates the possibility that the polynomial has a degree higher than two. Familiarity with polynomial functions, suggests that, for at least the given trend, the

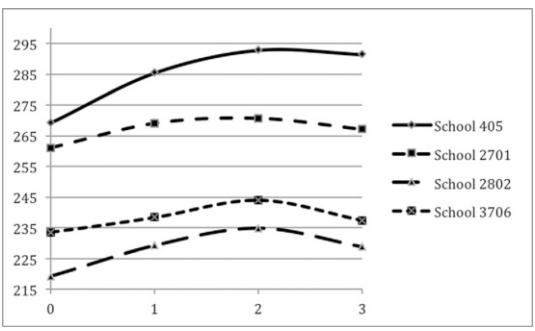


Figure 1. Science achievement for selected schools (2008 to 2011).

data are best described as quadratic function, such as the type depicted in Figure 2. In a given dataset, we can estimate one fewer random effect than we have time points. This implies that we must have at least two more time points than the order of the polynomial for model specification. Therefore, if we have collected data across 4 time points, we can estimate three random effects. Thus, if we wish to estimate random effects for each of the growth parameters at level 1 (i.e., a randomly varying intercept and a randomly varying slope for each of the growth parameters), then we can fit a linear model with three time points, a quadratic (second order polynomial) model with four time points, a cubic (third order polynomial) with five time points, etc. Also of note, it still possible to even fit a simple linear model.

In the context of the present research example, a quadratic (or second order polynomial) function contains three pieces of information. The first is a constant that represents average school science achievement at time L, the second is a coefficient for the instantaneous rate of change at time L (the centering point), and the third is a coefficient for the acceleration term. In this model, the instantaneous rate of change can be positive (indicating an upward trend in school mean science achievement when time = L) or negative (indicating a downward trend in school mean science achievement when time = L). Another implication is that the instantaneous rate of change itself is changing. Thus, the quadratic parameter, π_{2i}, describes the change in the rate of change. Along a quadratic trajectory, growth can be accelerating (indicating increasing rates of change) or decelerating (indicating decreasing rates of change). Data that demonstrate a full parabolic trajectory can have both "growth and decline" as well as "acceleration and deceleration" over different intervals of time.

Oftentimes, however, only fragments of the parabola are represented by the data. Under these conditions, there can be many combinations of "growth and decline" and "acceleration and deceleration" In Figure 3, we illustrate four parabolic fragments, each of which is defined by a positive or negative π_{1i}, and a positive or negative π_{2i}, In the top left corner, the parabolic fragment with a positive π_{1i} and a negative π_{2i} depicts a curve that begins as a positive growth trajectory; however, the rate of increase is decelerating across time. In the top right corner, the parabolic fragment with a positive

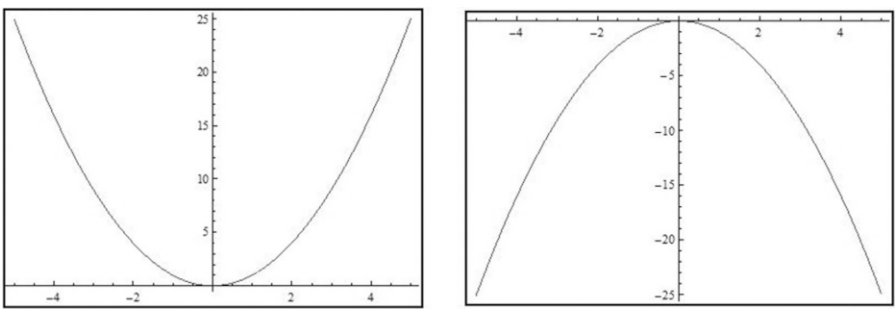


Figure 2. Plots of quadratic functions.

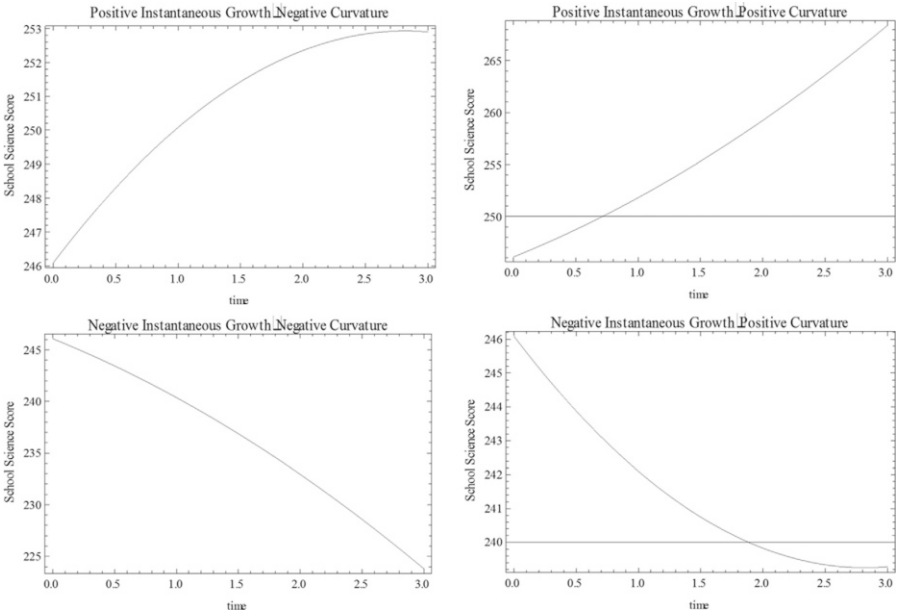


Figure 3. Growth and curvature combinations for quadratic function fragments.

π_{1i} and a positive π_{2i} depicts a curve that shows positive and accelerating growth across time. In the bottom left hand corner, the parabolic fragment with a negative π_{1i} and a negative π_{2i} illustrates negative growth (or decline) that becomes increasingly rapid over time. In the bottom right hand corner, the parabolic fragment with a negative π_{1i} and a positive π_{2i} depicts negative growth (or decline) that decelerates over time. Based on the plots of 5th grade school science achievement, scores increase, but the rate of increase decelerates across time. Thus, the curve we should anticipate most closely resembles the curve depicted in the top left corner of Figure 3.

Prior to running any statistical models, we recommend visually inspecting both individual growth trajectories and a plot of the change in the means on the outcome variable across time. No modeling technique, no matter how novel or sophisticated, can substitute for a solid understanding of the data. Our data appears to be best modeled by the quadratic function in the top left panel of Figure 3. Between the first (time = 0) and most recently scored administrations of the assessment (time = 3) the average school scores appear to grow, but their rates of change appear to slow as time passes.[1] Under this model specification, π_{0i} represents the school's science achievement in 2008, while π_{1i} is the instantaneous initial growth rate in 2008. The curvature parameter, π_{2i}, represents the curvature parameter or the acceleration of 5th grade school science achievement. The time variable was centered on the first administration of the science assessment.[2]

Table 4. Quadratic model of growth in science achievement (unconditional model)

Fixed Effect	*Coefficient (SE)*	*t Ratio*	*p-Value*
School mean achievement (π_0)			
Intercept (β_{00})	246.09 (1.07)	229.66	< 0.001
School mean growth rate (π_1)			
Intercept (β_{10})	4.85 (0.49)	9.79	< 0.001
School mean acceleration rate (π_2)			
Intercept (β_{20})	–0.86 (0.16)	–5.46	< 0.001
Estimation Method: Restricted Maximum Likelihood			
Random Effect	*Variance Component*	*x^2(df)*	*p-Value*
Variance in intercept (r_0)	609.79	8346.87 (540)	< 0.001
Variance in linear slope (r_1)	25.84	681.48 (540)	< 0.001
Variance in accel. slope (r_2)	2.49	675.91 (540)	< 0.001
Variance within (σ^2)	43.28		

If the *a priori* assumption is that the quadratic model is likely the best fitting model, then it is sensible to begin with an unconditional level-1 model with an intercept, first-order, and quadratic parameters. In hypothesizing a quadratic fit to the data, it is then necessary to test the statistical significance of this specification. If the quadratic term is not statistically significant and the quadratic model does not provide statistically significantly better fit than the linear model (using the chi-square difference test), then the quadratic term is probably not necessary, and the data can likely be fit with a simpler (i.e., linear) model.

The results of the unconditional model indicate a function with an intercept of 246.09, an instantaneous growth slope of 4.85 points per year and a curvature parameter of -.86, which indicates that the slope is becoming more negative across time.

$$y_{ti} = 246.09 + 4.85t - 0.86t^2 \tag{6}$$

These results suggest that, on average, schools begin with a science scale score of 246.09 out of a possible 400 points on the science achievement test. The model also indicates that school scores are improving (rather than declining) since the instantaneous growth rate in 2008 (represented by the β_{10} parameter) was estimated to be approximately 4.85 points per year. In addition, it appears that the growth of science in schools is, as predicted, is slowing down.

For linear and quadratic growth models, it is possible to determine whether the outcome measure is increasing or decreasing by using the first derivative test of the function.[3]

$$score = y_{ti} = 246.09 + 4.85t - 0.86t^2 \quad (7)$$
$$rate\,of\,change = \frac{dy}{dt} = 4.85 - 1.72t$$

Once the level-1 model is specified, the first derivative is calculated, and the values over the interval of interested are substituted into the function. If the values are positive, the function is increasing; if the values of the function are negative, the function is decreasing. Table 5 uses the first derivative test to evaluate 4 possible quadratic functions. Table 6 provides the first derivative test for our school science achievement data.

The overall results of the first derivative test for our level-1 function are found in the first row of Table 5 and Table 6. The results indicate that the science scores have been increasing over time since the assessment was first introduced although the growth rate has been steadily declining. By time point 3, it appears that science growth has completely leveled out, and may even be declining slightly.

Table 5. First derivative test to determine increasing or decreasing growth of hypothetical models

Parameter Coefficient Description	y_{ti}	*Interval**	y'_{ti}	*Conclusion*
Pos. Inst. Growth (π_{1i}) Neg. Curvature (π_{2i})	$246.09 + 4.85t - 0.86t^2$	[0, 3]	$4.85 - 1.72t$	Increasing between year 1 and just before year 3; decreasing at year 3
Neg. Inst. Growth (π_{1i}) Neg. Curvature (π_{2i})	$246.09 - 4.85t - 0.86t^2$	[0, 3]	$-4.85 - 1.72t$	Decreasing from year 1 through year 3
Pos. Inst. Growth (π_{1i}) Pos. Curvature (π_{2i})	$246.09 + 4.85t + 0.86t^2$	[0, 3]	$4.85 + 1.72t$	Increasing from year 1 through year 3
Neg. Inst. Growth (π_{1i}) Pos. Curvature (π_{2i})	$246.09 - 4.85t + 0.86t^2$	[0, 3]	$-4.85 + 1.72t$	Decreasing between year 1 and just before year 3; increasing after year 3

* *Note*: Since time is centered on first administration of the science test in 2008, the measurement occasions 2008, 2009, 2010, and 2011 now correspond to times 0, 1, 2, and 3.

Table 6. First derivative test results for the school science achievement

Time Point	*Rate of Change Calculation*	*Rate of Change Interpretation*
0	$4.85 - 1.72(0) = 4.85$	Growth
1	$4.85 - 1.72(1) = 3.13$	Growth
2	$4.85 - 1.72(2) = 1.41$	Growth
3	$4.85 - 1.72(3) = -0.31$	Decline

Unlike linear growth models which have only one constant growth or decline parameter, quadratic functions have a non-zero acceleration coefficient. This implies that the growth or decline can be accelerating or decelerating. To determine whether a function is accelerating or decelerating, the second derivative test is used in a manner similar to the first derivative.

$$\begin{aligned} score = y &= 246.09 + 4.85t - 0.86t^2 \\ rate\,of\,change &= \frac{dy}{dt} = 4.85 - 1.72t \\ acceleration &= \frac{d^2y}{dt^2} = -1.72 \end{aligned} \tag{8}$$

After calculating the second derivative of the level-1 model, the values from the time interval of interest are substituted into the second derivative. Positive values indicate locations along the time interval where the function is accelerating while negative values indicate ranges where the function is decelerating. In our example, the result of the second derivative is −1.72. It is a constant and indicates that the rate of change is steadily declining. In quadratic functions, the acceleration (or change in the rate of change) is always constant, making the calculation relatively trivial in this example. However, for higher order polynomial models, such as cubic models, the acceleration rate is itself variable, and the value of the second derivative should be evaluated at each time point in a manner consistent with the first derivative test.

The results for our function are found in the first row of Table 7. The model suggests that the rate of change is becoming more negative, thus, growth in school science achievement is slowing down.

Once the unconditional model is estimated, it is then appropriate to fit the level-1 model to the data using any additional time varying covariates. One of the interesting features of the data was access to variables that had the potential to be treated as either time-varying or time-invariant covariates. In this study, some variables were measured at every time point, and thus, could conceivably be

Table 7. Second derivative test determine accelerating or decelerating growth model (concavity)

Parameter Coefficient Description	y_{ti}	*Interval*	y''_{ti}	*Conclusion*
Neg. Curvature (π_{2t})	$246.09 \pm 4.85t - 0.86t^2$	[0, 3]	−1.72	Instantaneous growth (slope) is decreasing (concave down)
Pos. Curvature (π_{2t})	$246.09 \pm 4.85t + 0.86t^2$	[0, 3]	1.72	Instantaneous growth (slope) is increasing (concave up)

treated as either time-varying covariates or time-invariant group-level variables. Researchers in similar research designs often assume that because data are collected at every time point they should be treated as time varying. If, however, there is not sufficient variability in the data across time within unit, then treating the variable as time varying may not be advantageous. For example, demographic data such as socioeconomic status (SES) often appears in growth models and fits this very situation. In these types of variables, although it is possible to measure school characteristics over time, there may not be enough within cluster variance in the variable to justify its treatment as time-varying covariate. The most sensible approach to determining whether to treat a variable time-varying covariate or group-level variable is to calculate the intraclass correlation coefficient (ICC) for each variable in question. The ICC is known as the "cluster effect" and it measures the proportion of variance in the outcome that is "between groups" (i.e., the level-2 units). To use the ICC to determine whether variables should be treated as time-varying covariates, first set each variable of interest as the outcome and obtain parameter estimates for the within-group variability (σ_2) and the between-group variability (π_{00}). With these parameter estimates, it is possible to calculate the ICC, which is given by the formula

$$\rho = \frac{\tau_{00}}{(\tau_{00} + \sigma^2)} \tag{9}$$

and is therefore best described as the between-group variance divided by the total variance. ICC values close to 1 indicate very little within-cluster variability across time. ICC values near zero indicate large within-cluster variability across time.

In the school science achievement study, many of the demographic variables used have the potential to function as either time-variant or time-invariant. As a result, ICCs were calculated for all the variables considered in the model.

In our experience, variables with ICCs of 0.85 and above reflect a low degree of within cluster variance and are best treated as time-invariant group-level variables.

The high ICCs for the percentage of English language learners (ELL) and the percentage of free and/or reduced lunch students (LUNCH) in Table 8 suggest that virtually all the variance in these variables occurs between schools rather than within schools. As a result, these two variables do not display enough variability within

Table 8. ICCs for science achievement demographic variables

Variable	τ_{00}	σ^2	*ICC*
ELL	70.25	11.59	0.86
LUNCH	1100.2	43.88	0.96
SPED	8.93	20.26	0.31

schools to be treated as time-varying covariates at level-1. In response to these results, the average percentages for these three variables were calculated over the four years of testing and used at level-2 (in this case, school) to best understand the growth in scores over time. In contrast, the percentage of special education students taking the science achievement test (SPED) had an ICC of 0.31, which is fairly low. Thus, we treated the percentage of special education students in the grade as a time-varying covariate.

Once the appropriate predictor variables are chosen for the level-1 model, it is critical to define their "location" to correctly interpret the intercept, instantaneous growth rate, and curvature of the model. Raudenbush and Bryk (2002) identify four possible locations for the predictor variables: the natural metric, grand mean centering, group mean centering, and "specialized" locations. In the natural metric centering approach, the zero is defined as the absence of any of the predictor variable. This is theoretically plausible since the demographic variables are "percentages" of student population characteristics. For example, in our model, it is possible to have zero percentages of special education students. However, we chose the second approach, grand mean centering for the purpose of this study. Grand mean centering is the standard choice of location in the classical analysis of covariance model. It allows the intercept to be interpreted as the expected value for a school that is at the mean on the variable that has been grand mean centered. In our example, because we grand mean centered the free lunch and ELL variables, we can think of the intercept as the expected science score for an "average" school in our sample, or more specifically, as the expected school science score for a school that has an average percentage of free lunch and ELL students (for our sample). The third strategy, group mean centering, allows for the centering of level-1 predictor variables around the mean of their level-2 unit, which is school is the present study. While there are countless "specialized" centering strategies for multilevel models, in the growth modeling framework the most common centering approaches will likely be ones that define the metric such that the intercept corresponds to an outcome at a specific time point.[4] For our study, the level-1 SPED predictor variable was grand mean centered, our level-2 variables (percentage of free lunch students and percentage of ELL students) were grand mean centered, and we added one time varying covariate, the percentage of special education students in the fifth grade.

The results indicate that percentage of special education students (the only time-varying covariate in the study) is statistically significant. The SPED predictor variable has the effect of reducing the school science score by 0.37 points for each percentage point increase in the number of special education students taking the test. Recall that the relationship between this time-varying covariate and school achievement is consistent across time. In addition, given that we are already estimating 3 random effects for a model with 4 time points, we are unable to estimate a random effect for this variable. Thus, the effect of the special education variable on school science achievement is assumed to be constant across schools as well.

Table 9. Level-1 quadratic growth model for science achievement

Fixed Effect	*Coefficient (SE)*	*t Ratio*	*p-Value*
School mean achievement (π_0)			
Intercept (β_{00})	245.96 (1.06)	232.44	<0.001
School mean growth rate (π_1)			
Intercept (β_{10})	5.01 (0.48)	10.34	<0.001
School mean acceleration rate (π_2)			
Intercept (β_{20})	–0.89 (0.15)	–5.78	<0.001
SPED slope (π_3)			
Intercept (β_{30})	–0.37 (0.05)	–8.41	<0.001
Estimation Method: Restricted Maximum Likelihood			
Random Effect	*Variance Component*	*x^2(df)*	*p-Value*
Variance in intercept (r_0)	597.46	8661.15 (540)	<0.001
Variance in linear slope (r_1)	26.09	689.04 (540)	<0.001
Variance in accel. slope (r_2)	2.43	680.39 (540)	<0.001
Variance within (σ^2)	40.85		

An Intercepts- and Slopes-as-Outcomes Model of the Effects of the Percentages of English Language Learners and Students Receiving Free and/or Reduced Lunch

With an appropriately identified level-1 model, it then possible to test the full hypotheses of the school science achievement study. We hypothesized that the percentages of English language learners (ELL) and free and/or reduced lunch recipients (LUNCH) would affect the initial status, the instantaneous growth rate, and the curvature of school science achievement. We also hypothesized that schools would vary in terms of their initial achievement as well as the shape and trajectory of their growth; therefore, we estimated random effects for each of those growth parameter components. In the final trimmed model reported in Table 10, we did not include any parameters that were not statistically significant in the full model.

The results in Table 10 suggest that the initial school mean achievement in science depended jointly on the percentage of English language learners and free and/or reduced lunch recipients taking the science test in the school. Each percent increase in English language learners taking the test lowered the school's science achievement by 0.28 points, after controlling for the percentage of students receiving free and/or reduced lunch. In addition, each percent increase in test-takers that participated in the free and/or reduced lunch program resulted in initial mean science scores that were approximately 0.68 points lower, after controlling for the percentage of English language learners taking the test. In contrast, only the percentage of free and/or reduced lunch students taking the test had a statistically

Table 10. Full quadratic growth model for science achievement

Fixed Effect	*Coefficient (SE)*	*t Ratio*	*p–Value*
School mean achievement (π_0)			
Intercept (β_{00})	245.94 (0.49)	498.74	<0.001
ELL (β_{01})	–0.28 (0.06)	–4.49	<0.001
LUNCH (β_{02})	–0.62 (0.02)	–35.74	<0.001
School mean growth rate (π_1)			
Intercept (β_{10})	4.96 (0.49)	10.25	<0.001
LUNCH (β_{11})	–0.02 (0.01)	–3.59	0.001
School mean acceleration rate (π_2)			
Intercept (β_{20})	–0.87 (0.15)	–5.62	<0.001
SPED slope (π_3)			
Intercept (β_{30})	–0.37 (0.04)	–9.68	<0.001
Estimation Method: Restricted Maximum Likelihood			
Random Effect	*Variance Component*	*x^2(df)*	*p-Value*
Variance in intercept (v_0)	96.71	1832.55 (538)	<0.001
Variance in linear slope (v_1)	27.07	693.86 (539)	<0.001
Variance in acceleration (v_2)	2.49	681.92 (540)	<0.001
Variance within (σ^2)	40.76		

significant effect on π_{1t} a school's initial instantaneous growth rate in science. School science test scores increased 0.02 points more slowly for each percentage increase in the number of free and/or reduced lunch recipients taking the test. However, the same demographic variables that contributed to modeling the initial status and initial instantaneous growth rate did not make a statistically significant contribution to modeling the acceleration/curvature parameter. Table 11 provides predicted values for schools with three different demographic compositions. Whereas all three types of schools do increase across the 4 time points, the high special education/high ELL/ high free lunch school's predicted initial scores are substantially below those of the average and low special education/ ELL/ free lunch schools. Further, whereas, the average schools are expected to gain about 7 points between 2008 and 2011 and the low special education/ ELL/ free lunch schools are expected to gain about 9 points during that time period, the high special education/ ELL/ free lunch schools are expected to gain less than 3 points. Therefore, the high special education/ ELL/ free lunch schools, who start out with the lowest science scores, also make the slowest growth, and their growth levels off the most quickly, which is a disturbing finding, given the original performance gap.

Table 11. Predicted values for schools of three hypothetical demographic variable groups

School Characteristics	τ_0	τ_1	τ_2	τ_3
High Special Education (%)				
High English Language Learner (%)	222.72	223.51	225.19	225.12
High Free/Reduced Lunch (%)				
Average Special Education (%)				
Average English Language Learner (%)	245.94	250.03	252.38	252.99
Average Free/Reduced Lunch (%)				
Low Special Education (%)				
Low English Language Learner (%)	271.8	276.56	279.58	280.86
Low Free/Reduced Lunch (%)				

The "high" and "low" categories are defined as values that are 1 standard deviation above and below the mean for each independent variable.

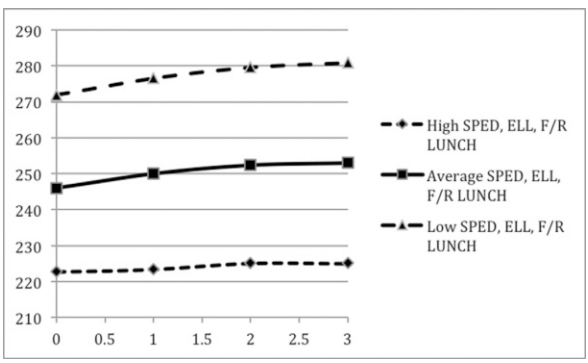


Figure 4. Predicted values for schools of three hypothetical demographic variable groups.

Although the quadratic model provided an appropriate fit of the data, it makes theoretical sense to compare this model to a simpler linear model for the sake of parsimony. As mentioned earlier, if the fit of the simpler model is not statistically significantly worse than the fit of the more complicated one, then the simpler model should be preferred.

Generally, when the number of clusters is small, the parameter estimates should be computed using restricted maximum likelihood estimation (REML). REML estimates of variance-covariance components adjust for the uncertainty about fixed effects (Raudenbush & Bryk, 2002). However, when using chi-square difference tests to compare models that differ in terms of their fixed effects, it is necessary to use deviances that were computed under full maximum likelihood (McCoach & Black, 2008). The full linear model is identical to the quadratic model in terms of

Table 12. Full linear growth model for science achievement

Fixed Effect	*Coefficient (SE)*	*t Ratio*	*p-Value*
School mean achievement (p_0)			
Intercept (β_{00})	246.76 (0.48)	519.01	<0.001
ELL (β_{02})	–0.28 (0.06)	–4.58	<0.001
LUNCH (β_{03})	–0.62 (0.02)	–35.99	<0.001
School mean growth rate (p_1)			
Intercept (β_{10})	2.35 (0.16)	14.60	<0.001
LUNCH (β_{11})	–0.02 (0.01)	–3.46	0.001
SPED slope (π_2)			
Intercept (β_{20})	–0.37 (0.04)	–8.77	<0.001
Estimation Method: Restricted Maximum Likelihood			
Random Effect	*Variance Component*	$x^2(dt)$	*p-Value*
Variance in intercept (ρ_0)	92.27	1436.31 (536)	<0.001
Variance in linear slope (ρ_1)	4.32	729.26 (537)	<0.001
Variance in SPED slope (ρ_2)	0.13	596.19 (538)	0.041
Variance within (sig sq)	44.68		

Table 13. Statistics for covariance components models

Model	*Number of Parameters*	*Deviance*
1. Level-2 Linear Growth	10	15850.99
2. Level-2 Quadratic Growth	14	15807.99

the demographic variables used to predict the intercept and growth rate. The only difference is the deletion of the acceleration/deceleration fixed and random effects in the model. Therefore, we can use deviances computed under REML to compare the fit of the two models.

Model selection should be guided by theory and informed by data. Adding parameters is likely to improve fit and cannot lead to worse model fit (Forester, 2000). The critical issue is whether the improvement in the fit of the model justifies the inclusion of additional parameters. Thus, the principle of parsimony is paramount (Burnham & Anderson, 2004). Additionally, using data to compare several plausible competing hypotheses often provides more useful information than comparing a given model to an often implausible null hypothesis (Burnham & Anderson, 2004).

When we judge the fit of a quadratic versus a linear model we are comparing nested models. If two models are nested, the deviance statistics of two models can

be compared directly. The deviance of the simpler model (D_1) minus the deviance of the more complex model (D_2) provides the change in deviance ($\Delta D - D_1 - D_2$). The simpler model always will have at least as high a deviance as the more complex model, and generally the deviance of the more complex model will be lower than that of the simpler model. In evaluating model fit using the chi-square difference test, the more parsimonious model is preferred, as long as it does not result in significantly worse fit. Therefore, when the change in deviance (ΔD) exceeds the critical value of chi-square with ($p_1 - p_2$) degrees of freedom, we favor the more complex model. However, if the more complex model does not result in a statistically significant reduction in the deviance statistic, we favor the parsimonious model (McCoach & Black, 2008).

For comparisons of models that differ in their fixed effects, it is necessary to use full information maximum likelihood (FIML) to estimate the deviances of the models. The model comparison between the level-2 linear and level-2 quadratic models, shown in Table 14, suggests that the more complicated quadratic model is, in fact, the more appropriate model for the data.

A visual inspection of the plot of the average of the actual school science achievement scores compared the score estimates provided by the linear and quadratic models confirms the conclusion that the growth is best modeled by the addition of a second degree polynomial. In fact, our quadratic predictions are almost identical to the scores from the raw data. It is always a good idea to compare the model implied level-1 growth trajectory to the actual data: a well specified growth model should be able to recover the original shape of the data fairly accurately.

Table 14. Model comparisons of science achievement growth models

Model Comparison	Δx^2	*df*	*p-Value*
Level-2 Quadratic vs. Level-2 Linear	42.85	4	< 0.001

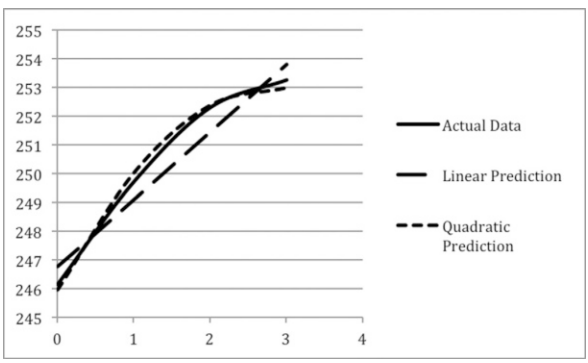


Figure 5. Predicted linear and quadratic science achievement scores compared to actual science scores.

DISCUSSION

This longitudinal modeling of school-level science achievement is one approach to the study of school effectiveness. For this simple example, we focused on three key demographic variables that the literature has identified as related to school achievement. We employed a quadratic growth model to gain insights into how school science achievement changes over time.

The addition of a curvature parameter to the linear growth function improved the model of the school science achievement data. Although school science means were increasing, the instantaneous rate of that increase was slowing down. This result is consistent with our instinct regarding school level improvement on large scale assessments. Since the first time point coincides with the first year the science assessment was administered, we can see that improvement is steepest in the beginning, as schools become more comfortable with the new test. However, the rate of growth slows across the 2008–2011 period.

Another implication of the model is that the percentage of free lunch students has an influence not only on mean school achievement, but also the instantaneous growth rate. Each percentage increase in the number of students receiving free and/or reduced lunch reduces the instantaneous growth rate by two hundredths of a point each year. This effect might seem trivial, but this means that schools where 85% of the students receive free or reduced price lunch are expected to grow 1.5 points more slowly ***per year*** than schools where 10% of the students receive free or reduced price lunch. Thus, over the time period in question, the gap between those two types of schools would widen by 4.5 points. Overall, we were able to conclude that school science achievement improved in the three years following the first administration of the science assessment, but the improvement levels off or slows down across time.

Growth modeling has the potential to provide insights that go beyond static measures of achievement differences. Educational policymakers will always be interested in the status of school achievement, but these concerns are being augmented by questions about growth and decline. The consequence of these concerns is that "effectiveness" is better understood as dynamic rather than static school characteristic. As a result, multilevel modeling of repeated measures of school achievement data can serve as one straightforward technique for understanding school dynamics and change over time.

GROWTH MODELS: ISSUES AND PITFALLS

Using individual growth models with real data presents a variety of issues and challenges. We address several key concerns: measurement of the dependent variable, regression to the mean, measurement error, floor or ceiling effects, the scale of the dependent variable, non-normal or non-interval level data, and changes in the distribution across time.

Measurement of the Dependent Variable

Growth models with continuous variables require an assumption that the dependent variable is normally distributed at each time point. However, when using the same measure across time, the distribution of the dependent variable could change. For example, scores on an assessment could be positively skewed at time point 1, normally distributed at time point 2, and negatively skewed at time point 3. Imagine giving the same assessment to students across three time points. Perhaps on the first testing occasion, the test is difficult for the most of the students, resulting in a positively skewed distribution. On the second testing occasion, students are fairly normally distributed. By the third testing occasion, many students have made so much growth that the assessment is now fairly easy for them, resulting in a negatively skewed distribution. Situations such as this are not uncommon, resulting in changes in the shape of the distribution of the dependent variable across time.

In this situation, transforming the dependent variable across all of the time points is problematic, as the transformation would alleviate the problem at one time point but exacerbate the problem at another time point. However, applying different transformations across the different time points is also not possible, as that creates different scales across time. Some have suggested standardizing the dependent variable at each time point to normalize the distribution and to try to ensure equitability of the variable across time. This is a poor idea. The fact that standardized scores have a mean of 0 and a standard deviation of 1 (and thus a variance of 1) leads to two important outcomes. First, because the mean across time points is standardized to be 0, growth models using standardized scores are not capturing growth per se, instead, they capture change in relative status. Second and more importantly, standardizing scores at each time point constrains the variance of the measure to be equal across time, which is often an unrealistic assumption. Educational and psychological research has consistently shown that the variance in achievement, skills, or ability generally increases across time (Bast & Reitsma, 1998; Gagné, 2005; Kenny, 1974). "Such standardization constitutes a completely artificial and unrealistic restructuring of interindividual heterogeneity in growth" (Willett, 1989) and constraining scores in this way is likely to produce distorted results (Thorndike, 1966; Willett, 1989). Thus Willett (1989) recommends against the standardization of the dependent variable when conducting analyses of change.

Measurement Challenges

There are several measurement issues that should be considered before embarking on growth analyses. These challenges relate to the ability to adequately capture growth using any measure. The reliability of scores used as dependent variables in growth analysis is of particular concern. Regression to the mean, or the tendency for those with extreme initial scores to score closer to the average score on subsequent assessments, can bias growth measures, overestimating the growth of low achieving

students and underestimating the growth of high achieving students. Measurement error, or the degree of imprecision in test scores, is also of concern (McCoach, Rambo, & Welsh, 2012).

Regression to the Mean

Regression to the mean is an important, but commonly misunderstood statistical phenomenon. When using an independent variable (such as a test score at Year 1) to predict scores on a dependent variable (such as a test score at Year 2), errors in prediction will occur whenever the correlation between the two variables is less perfect (+1.0 or −1.0) (Campbell & Kenny, 1999). These errors in prediction will make it appear that people with initially extreme scores have scores closer to the mean on the posttest. Therefore, the scores of high achieving individuals will grow at a smaller rate than low or average achieving individuals. People who score very low at the initial time point are more likely to demonstrate steeper growth rates than average or high achieving individuals (McCoach et al., 2012).

Measurement Error

The measurement of psycho-educational constructs is fraught with error. For example, educators use scores on achievement tests to infer a person's level of content mastery in a domain. However, a person's score on the test is **not** a perfect measure of his or her achievement level. There are a variety of factors that could cause the test score to be either an over or an underestimation of the person's actual achievement level. The content sampling of items on the test, the format of the items, the testing conditions, and many other factors can cause the observed test score to deviate from the underlying trait value. All of these factors are subsumed under the general term measurement error. Reliability is a related concept in that it describes the consistency of scores across time, test forms, or internally within the test itself. Measurement error and reliability are inversely related: the greater the measurement error, the lower the reliability of the scores. The goal, of course, is to minimize the degree of measurement error in scores; however, it is impossible to completely eliminate (McCoach et al., 2012).

Both unconditional and conditional errors of measurement influence the reliability with which we can estimate the scores of high ability students. Conditional errors of measurement are errors that depend on the location of a score on the scale (Lohman & Korb, 2006), whereas unconditional errors of measurement are evenly distributed across the entire range of scores. The reliability coefficient or the traditional standard error of measurement both assumes errors of measurement to be constant across the score distribution. However, in general, the amount of error in test scores is not uniform across the distribution of scores (Lohman & Korb, 2006). Instead, it is U-shaped: the error is lowest for people in the middle of the score distribution and highest for the people at the extremes of the score distribution (McCoach et al., 2012).

Floor or Ceiling Effects

A somewhat related issue is that of ceiling effects. A test may not contain questions to assess or distinguish among the lowest or highest scoring individuals. When an individual hits the ceiling of an assessment, there is no way to assess how much more the person knows and can do. Conversely, the floor does not accurately capture a person's true level of functioning. If a person's performance is far above or below the range of the assessment for one or more of the testing occasions, we cannot accurately estimate growth across time. Obviously, if the floor of a test is too high or the ceiling of a test is too low, then there is no way to accurately measure the achievement or the growth of the individuals whose performance falls outside the range of the assessment. If people outgrow the test, or if it is too difficult at the outset, estimates of growth may be greatly distorted. Therefore, when designing longitudinal studies, it is extremely important to consider the range of the assessment and whether the range of the assessment will be able to capture the full range of abilities of a diverse set of people across all of the time periods included in the study (McCoach et al., 2012).

Attrition and Missing Data

Individual growth modeling can easily handle missing observations at level-1, assuming "that the probability of missingness is unrelated to unobserved concurrent outcomes (conditional on all observed outcomes)" (Singer & Willett, 2003, p. 159). When level-1 data are missing completely at random or missing at random, individual growth modeling should still produce valid results (Singer & Willett, 2003). However, when attrition is systematic and is related to scores on the outcome variable of interest (after controlling for the independent variables in the model), the estimates of the growth parameters are likely biased, leading to invalid inferences about the phenomenon of interest. Thus it is very important to examine the nature of the missingness within the sample prior to conducting growth analyses. The interested reader should consult Enders (2010) for an excellent introduction to the issue of missing data.

RECOMMENDATIONS

When a researcher is interested in capturing growth or change over time, it is best to collect three of more data points. The more complex the shape of the growth trajectory is expected to be, the greater the number of time points required to estimate the model. In addition, the reliability of the growth slope is dependent not only on the reliability of the outcome measure, but also on the number of observations. Increasing the number of observations collected increases the reliability of the growth slope. Of course, there are diminishing returns to increasing the number of observations. The timing and spacing of the measurements is also important. For an excellent treatment

of these issues, see Raudenbush and Xiao-Feng (2001). Researchers should consider the expected shape of the trajectory and their goals for conducting the analysis during the design phase when they decide how frequently and how many times they plan to measure participants on the outcome measure of interest. Further, it is critical to visually inspect the shape of the individual growth trajectories prior to conducting any statistical analyses to understand and correctly model the functional form of the data. No statistical analysis can supplant the wealth of information that is provided by the visual examination of individual growth trajectories.

Measurement issues are especially salient when analyzing change. The reliability of a measure of change is bounded by the reliability of the initial measure. Therefore, to accurately capture the nature of change, it is important to use measures that exhibit strong psychometric qualities. One possible solution to this issue is to use multiple indicators of the construct of interest and to impose a measurement model onto the analysis of change. This is easily accomplished in the SEM framework. While it is possible to build crude measurement models in an HLM framework under certain limiting assumptions, the SEM framework lends itself more naturally to the modeling of latent variables. For details on using SEM models to estimate growth models, see (Ferrer-Caja & McArdle, 2003; McArdle, 2001; McArdle, 2006). Further, to increase the reliability of growth slopes, increase the number of data collection points.

We hope that this introduction to growth modeling within the multilevel framework provides useful advice for researchers who are interested in analyzing change or growth. We end where we started, by reminding the reader once again that the most important ingredient necessary to build a successful longitudinal model is a substantive theory about the nature of the change in the variable of interest over time. When conducting longitudinal models, no complex statistical models can ever substitute for a combination of substantive theory, knowledge of the data, and good old fashioned common sense.

NOTES

[1] The centering parameter, L, was set at 2008, because this represents the first year the science achievement portion of the CMT was administered to 5th graders in the state.

[2] The choice of choice of a centering constant L of 2008 placed the interpretation of the intercept as the average school science achievement at the first administration of the test. This makes substantive sense since the study is most interested in the initial achievement of schools as well as their growth rates. It should also be noted that a midpoint centering of 2.5 years is unhelpful from both a substantive and interpretive standpoint largely because it represents a test administration time point that does not exist. The assessment is given yearly in the spring making interpolation of data points extremely problematic.

[3] Since the function we chose to model the data is monotonic it is also differentiable. The derivative of a polynomial can be calculated by applying the Power Rule: $\frac{d}{dx}x^n = nx^{n-1},\ n \neq 0.$

[4] The centering of level-2 variables is not as critical as the choices of centering for level-1 predictors. As a result, Raudenbush and Bryk (2002) suggest that it is "often convenient" to center all of the level-2 predictors on their corresponding grand means.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

Bast, J., & Reitsma, P. (1997). Matthew effects in reading: A comparison of latent growth curve models and simplex models with structured means. *Multivariate Behavioral Research, 32,* 135–167.

Bast, J., & Reitsma, P. (1998). Analyzing the development of individual differences in terms of Matthew effects in reading: Results from a Dutch longitudinal study. *Developmental Psychology, 34,* 1373–1399.

Bollen, K. A., & Curran, P. J. (2004). Autoregressive latent trajectory (ALT) models: A synthesis of two traditions. *Sociological Methods Research*, *32*, 336–383.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley Interscience.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*, 65–108.

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

Campbell, D. T., & Kenny, D. A. (1999). *A primer of regression artifacts*. New York: Guilford Press.

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, *57*, 505–28.

Cronbach, L. J., & Furby, L. (1970). How we should measure 'change': Or should we? *Psychological Bulletin, 74*, 68–80.

Curran, P. J. (2000). A latent curve framework for studying developmental trajectories of adolescent substance use. In J. Rose, L. Chassin, C.Presson, & J. Sherman (Eds.), *Multivariate applications in substance use research.* Hillsdale, NJ: Erlbaum.

Duncan, T. E., Duncan, S. C., & Stryker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum and Associates.

Enders, C. K. (2010). Applied missing data analysis*.* New York: Guilford Press.

Ferrer-Caja, E., & McArdle, J. J. (2003). Alternative structural equation models for multivariate longitudinal data analysis. *Structural Equation Modeling. 10*, 493–524.

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, 44, 205–231.

Frees, E. (2004). *Longitudinal and panel data.* Cambridge University Press.

Gagné, F. (2005). From noncompetence to exceptional talent: Exploring the range of academic achievement within and between grade levels. *Gifted Child Quarterly, 49*, 139–153.

Hsiao, C. (2003). *Analysis of panel data.* Cambridge University Press.

Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.) Belmont, CA: Thompson-Wadsworth.

Kenny, D. (1974). A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin, 82*, 342–362.

Kenny, D. A., & Campbell, D. T. (1989). On the measurement of stability in over-time data. *Journal of Personality, 57,* 445–481.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY US: Guilford Press.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*, 767–778.

Lohman, D. F., & Korb, K. A. (2006). Gifted today but not tomorrow? Longitudinal changes in ability and achievement during elementary school. *Journal for the Education of the Gifted*, *84*.

Marsh, H. W. (1993). Stability of individual differences in multiwave panel studies: Comparison of simplex models and one-factor models. *Journal of Educational Measurement, 30,*157–183.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics, 31,* 35–62.

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analyses. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.). *Structural equation modeling: Present and future.* Lincolnwood, IL: SSI.

McArdle, J. J. (2006). Dynamic structural equation modeling in longitudinal experimental studies. In K. van Montfort, H. Oud, & A. Satorra (Eds.). *Longitudinal Models in the Behavioural and Related Sciences*. Mahwah, NJ: Erlbaum.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica: The RAND Corporation.

McCoach, D. B., & Black, A. C. (2008). Assessing model adequacy. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 245–272). Charlotte, NC: Information Age Publishing.

McCoach, D. B., & Kaniskan, B. (2010). Using time varying covariates in multilevel growth models. *Frontiers in Quantitative Psychology and Measurement*, *1*(17). DOI: 10.3389/fpsyg.2010.00017

McCoach, D. B., Rambo, K., & Welsh, M. (2012). Issues in the analysis of change. *Handbook of measurement, assessment, and evaluation in higher education.*

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, 36, 318–324.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: John Wiley.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*(3), 259–284.

Muthén, B., Brown, C., Masyn, K., Jo, B., Khoo, Yang C., Liao, J. (2002). General growth mixture modeling for randomized preventative interventions. *Biostatistics, 3*, 459–475.

O'Connell, A. A., Logan, J., Pentimonti, J, & McCoach, D. B. (in press). *Linear and quadratic growth models for continuous and dichotomous outcomes*.

Popham, W. J. (1999). *Classroom assessment: What teachers need to know* (2nd Ed.) Boston: Allyn & Bacon.

Ram, N., & Grimm, K. J. (2009). Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development, 33,* 565–576.

Raudenbush, S. W. (2001). Toward a coherent framework for comparing trajectories of individual change. In A. G. Sayer (Ed.), *New methods for the analysis of change.* (pp. 35–64). Washington, DC US: American Psychological Association.

Raudenbush S., & Bryk, A (2002). *Hierarchical linear models,* 2nd Ed. London: Sage Publications.

Raudenbush, S. W., & Xiao-Feng, L. (2001). Effect of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6*, 387–401.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*, 726–748. Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Sclove, L. S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* 52, 333–343.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY US: Oxford University Press.

Skrondal, A., & Rabe-Hesketh, S. (2008). Multilevel and related models for longitudinal data. In J. deLeeuw & E. Meijer (Eds.), *Handbook of multilevel analysis*, (pp. 275–300). New York: Springer Science+Business Media.

Stoel & Garre, 2011 Growth curve analysis using multilevel regression and structural equation modeling. (pp. 97–111). In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis*. New York, NY: Routledge.

Thorndike, R. L. (1966). Intellectual status and intellectual growth. *Journal of Educational Psychology, 57*(3), 121–127.

Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in a growth mixture models. In G. R. Hancock & K. M. Samuelson (Eds.), *Advances in latent variable mixture models.* Charlotte, NC: Information Age.

Wang, M., & Bodner, T. E. (2007). Growth mixture modeling: Identifying and predicting unobserved subpopulations with longitudinal data. *Organizational Research Methods*, *10*, 635–656.

Willett, J. B. (1989). Questions and answers in the measurement of change. *Review of Research in Education, 15,* 345–422.

SPYROS KONSTANTOPOULOS

11. META-ANALYSIS

INTRODUCTION

The objective of scientific investigations is to gain knowledge and understanding about phenomena through careful and systematic observations and analyses. Arguably, scientific research has an inherent cumulative nature since dependable information of prior scientific inquiries guides future studies as well as facilitates knowledge building. Scientific investigations seek to describe and explain phenomena that relate to a wide range of individuals and settings. Indeed, knowledge building is especially valuable when it can be generalized or transferred to different population groups and settings. However, it is rare that a comprehensive and generalizable body of knowledge will result from one single study (see Cook et al., 1992). By and large, a single effort of data collection will describe a restricted population of individuals in a specific geographic area. Thus, acquiring research evidence produced from multiple studies is essential to draw more general conclusions.

The last 50 years the overwhelming growth of scientific endeavors has led to an abundance of research studies. In particular, the last three decades a large body of quantitatively oriented empirical studies that focus on specific topics of research (e.g., teacher effectiveness) and discuss similar associations of interest has been produced. The amount of research related to various topics of scientific interest poses the question of how to group, organize, and summarize findings in order to identify and utilize what is known as well as guide research on promising areas (Garvey & Griffith, 1971). Hence, the development of systematic methods for organizing information across related research studies that focus on specific topics to produce generalizable knowledge has become particularly important. This need for accumulating research evidence in a specific research area has led to the development of systematic methods for synthesizing research quantitatively the last 30 years (Cooper, Hedges, & Valentine, 2009). The main purpose of integrating empirical evidence is to make generalizations about a specific topic of interest.

There are multiple ways of summarizing results from a sample of related studies that discuss similar relationships of interest (e.g., teacher characteristics and student achievement). For example, narrative reviews of related literature have been common practice for a long time. In such review studies, expert reviewers in a specific field summarize findings from a sample of studies that they have selected. Another example, which has gained ample attention the last 20 years, is systematic reviews or research syntheses (see Cooper et al., 2009). In research syntheses there are clear

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 231–246.

sets of rules about searching for related studies, selecting the final sample of studies, extracting quantitative information from each study, and analyzing the quantitative indexes to produce summary statistics (Borenstein, Hedges, Higgins, Rothstein, 2009). One would argue that the mechanisms involved in research synthesis reviews are more transparent than those involved in narrative reviews mainly because the criteria and processes that are followed in the review is specified very clearly in research synthesis (Borenstein et al., 2009).

Currently, the use of research syntheses that include statistical methods to summarize results from various empirical studies that test the same hypothesis is widespread in education, psychology, medicine, and the social science and health science research in general. A crucial part of research synthesis is the statistical analysis involved in combining quantitative information among related studies. Although a few quantitative methods have been described for accumulating research evidence, meta-analysis (e.g., Borenstein et al., 2009; Glass, 1976; Hedges & Olkin, 1985; Lipsey & Wilson, 2001) is widely considered to be the most popular and the most appropriate. The term meta-analysis of primary research evidence was first introduced by Glass (1976), who defined it as the "analysis of analyses" (p. 3).

Meta-analysis refers to the statistical methods that are used to combine quantitative evidence from different primary research studies that test comparable hypotheses for the purposes of summarizing evidence and drawing general conclusions (Cooper et al., 2009). In meta-analysis first the results of individual studies are described via numerical indexes, also called effect size estimates (e.g., correlation coefficient, standardized mean difference, odds ratio). Second, these numerical estimates are combined across studies to obtain summary statistics such as a weighted mean (e.g., a standardized mean difference or an association). The importance of each study estimate is demonstrated via a weight that is used in the computation of the summary statistics across the studies in the sample. That is, essentially, meta-analysis is a statistical procedure that uses study specific weights to compute an average estimate in a sample of studies. Once the weighted average estimate and its standard error have been computed, a formal test can be used (e.g., *a z* test) to determine the statistical significance of the mean.

The present chapter focuses on the meta-analysis part of research synthesis. The structure of the chapter is as follows. First, we define research synthesis and we delineate its advantages. Second, we discuss the types of effect sizes used in research synthesis. Third, we present fixed and random effects models in meta-analysis. Univariate meta-analysis is assumed, that is, only one effect size per study. Finally, we show how fixed and random effects models can be applied to real data. The examples we discuss are from educational research.

RESEARCH SYNTHESIS AND ITS ADVANTAGES

Research synthesis refers to very clearly defined steps or activities that are followed in the process of combining quantitative evidence from a sample of related studies.

The ultimate objective is to make a general statement about relationships or effects in a research area of interest. Now, meta-analysis is the part of research synthesis that is related to the statistical methods used to combine quantitative study specific indexes (Hedges & Olkin, 1985). That is, meta-analysis is the statistically related component of research synthesis. Of course, research synthesis involves other important components that are non-statistical. For example, first a topic should be identified and primary or secondary research questions need to be formulated. For example, what is the association between class size and achievement? Then, a careful and exhaustive literature search needs to be conducted. This step typically involves electronic searches of large databases such as ERIC, of the internet (e.g., google scholar), relevant journals and books, as well as investigating citations that appear in retrieved studies and contacting researchers who are experts in the specific area, etc. This step also involves identifying specific criteria for including or excluding studies (e.g., year of study, type of design, age of individuals, etc). The third component involves indentifying information from each study that can be used to construct quantitative indexes (e.g., a standardized mean difference) as well as study specific characteristics (e.g., year of publication, type of research design used, type of setting, geographic location, etc). The last component refers to the interpretation of the results produced from the meta-analysis step. This is an important step because it produces general statements about an effect or a relationship of interest. Excellent sources of the non-statistical aspects of research synthesis are available in Cooper (1989), Cooper et al. (2009), and Lipsey and Wilson (2001).

An important advantage of research synthesis is that it produces robust results and knowledge that can be generalizable across different samples and settings (Cooper et al., 2009). This constitutes a unique aspect of research synthesis that is crucial for the external validation of the estimates (see Shadish, Cook, & Campbell, 2002). Generally, the estimates that are produced from research syntheses have higher external validity than estimates reported in single studies. As a result, one can make more valid inferences about associations or effects of interest. That is, the summary statistics that are generated by meta-analyses can verify or refute theories, identify promising areas of research, advance substantive theory, and guide future research. In addition, the results of research synthesis can inform policy (Borenstein et al., 2009). For example, the results of a research synthesis can attest that a school intervention improves student achievement, or that treatments or drugs improve human health, etc. From a statistical point of view, the tests used in meta-analysis have higher statistical power than those from individual studies, which increases the probability of detecting the associations or effects of interest (Cohn & Becker, 2003).

EFFECT SIZES

Effect sizes are quantitative indicators that summarize the results of a study. Effect sizes reflect the magnitude of an association between variables of interest or of a treatment effect in each study. There are different types of effect sizes, and the effect

size used in a research synthesis should be chosen carefully to represent the results of each study in a way that is easily interpretable and is comparable across studies. The objective is to use effect sizes to put results of all studies "on a common scale" so that they can be readily interpreted, compared, and combined. It is important to distinguish the effect size estimate reported in a study from the effect size parameter (i.e., the true effect size) in that study. The effect size estimate will likely vary somewhat from sample to sample whereas the effect size parameter is fixed. An effect size estimate obtained from the study with a very large (essentially infinite) sample should be very close to the effect size parameter.

The type of effect size that will be used depends on the designs of the studies, the way in which the outcome is measured, and the statistical analysis used in each study. Typically, the effect size indexes used in the social sciences fall into one of three categories: the standardized mean difference, the correlation coefficient, and the odds ratio.

THE STANDARDIZED MEAN DIFFERENCE

In many studies in education the dependent variable is student achievement and the independent variable is a school related intervention. When the outcome is on a continuous scale and the main independent variable is dichotomous a natural effect size is the standardized mean difference. The standardized mean difference is computed by first subtracting the mean outcome in the control group from the mean outcome in the treatment group. Then in order to standardize this mean difference we divide the difference by the within group standard deviation (or pooled standard deviation). Namely, the standardized mean difference is

$$d = \frac{\overline{Y}_T - \overline{Y}_C}{S}, \tag{1}$$

where $\overline{Y}_T$ is the sample mean of the outcome in the treatment group, $\overline{Y}_C$ is the sample mean of the outcome in the control group, and S is the within-group standard deviation of the outcome

$$S = \sqrt{\frac{(n_T - 1)S_T^2 + (n_C - 1)S_C^2}{(n_T + n_C - 2)}}, \tag{2}$$

where n indicates sample size, T indicates the treatment group, C indicates the control group, and S indicates the standard deviation.

The corresponding standardized mean difference parameter is

$$\delta = \frac{\mu_T - \mu_C}{\sigma}, \tag{3}$$

where μ_C is the population mean in the treatment group, μ_C is the population mean outcome in the control group, and σ is the population within-group standard deviation of the outcome. This effect size is expressed in standard deviation units. The variance of the standardized mean difference is

$$v_d = \frac{n_T + n_C}{n_T n_C} + \frac{d^2}{2(n_T + n_C)}. \tag{4}$$

This variance can always be computed so long as the sample sizes of the two groups within a study are known. Because the standardized mean difference is approximately normally distributed, the square root of the variance can be used to compute confidence intervals for the true effect size or effect size parameter δ. Specifically, a 95% confidence interval for the effect size is given by

$$d - 1.96\sqrt{v_d} \le \delta \le d + 1.96\sqrt{v_d}. \tag{5}$$

THE CORRELATION COEFFICIENT

In studies where we are interested in examining the relation between two continuous variables (e.g., motivation and achievement), the correlation coefficient is a natural measure of effect size. In order to conduct analyses, first, the correlation coefficient r is transformed into a Fisher z-transform

$$z = \tfrac{1}{2}\log\left(\frac{1+r}{1-r}\right). \tag{6}$$

The corresponding correlation parameter is ρ and the population parameter that corresponds to the estimate z is ς, the z-transform of ρ. The variance of the z-transform is now stabilized and is only a function of the sample size n of the study

$$v_z = \frac{1}{n-3}. \tag{7}$$

As in the case of the standardized mean difference, the z-transform is approximately normally distributed, and the square root of the variance can be used to compute confidence intervals for the effect size parameter ς. Specifically, a 95% confidence interval for the effect size is given by

$$z - 1.96\sqrt{v_z} \le \varsigma \le z + 1.96\sqrt{v_z}. \tag{8}$$

Once the average z transform or the upper and lower bounds are computed the z-transforms can be inverted back to correlations using the formula

$$r = (e^{2z} - 1)/(e^{2z} + 1). \tag{9}$$

THE LOG ODDS RATIO

In some studies the dependent variable is dichotomous (e.g., going to college or not) and we are interested in examining the effects of an intervention that is also dichotomous (e.g., school intervention *Vs* no intervention). In such cases a natural effect size is the log odds ratio (OR). The log odds ratio is just the natural log of the ratio of the odds of a particular outcome in the treatment group to the odds of that particular outcome in the control group (e.g., odds or attending college). That is, the log odds ratio is defined as

$$\log(OR) = \log\left(\frac{p_T/(1-p_T)}{p_C/(1-p_C)}\right) = \log\left(\frac{p_T(1-p_C)}{p_C(1-p_T)}\right), \tag{10}$$

where p_T and p_C are the proportions in the treatment and control groups respectively that have the target outcome. The corresponding odds ratio parameter is

$$\omega = \log\left(\frac{\pi_T/(1-\pi_T)}{\pi_C/(1-\pi_C)}\right) = \log\left(\frac{\pi_T(1-\pi_C)}{\pi_C(1-\pi_T)}\right), \tag{11}$$

where π_T and π_C are the population proportions in the treatment and control groups, respectively that have the target outcome.

The large sample variance of the log odds ratio is

$$v_{\log(OR)} = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}, \tag{12}$$

where n indicates the counts and i, j indicate the row and column in the 2×2 table. As in the case of the standardized mean difference, the log odds ratio is approximately normally distributed, and the square root of the variance can be used to compute confidence intervals for the effect size parameter ω. Specifically, a 95% confidence interval for the effect size is given by

$$\log(OR) - 1.96\sqrt{v_{\log(OR)}} \le \omega \le \log(OR) + 1.96\sqrt{v_{\log(OR)}}. \tag{13}$$

There are several other indexes in the odds ratio family, including the risk ratio (the ratio of proportion having the target outcome in the treatment group to that in the control group or p_T/p_C) and the risk difference (the difference between the proportion having a particular one of the two outcomes in the treatment group and that in the control group or $p_T - p_C$). For a discussion of effect size measures in studies with dichotomous outcomes, including the odds ratio family of effect sizes, see Fleiss (1994).

FIXED EFFECTS MODELS

Statistical inference involves making projections from samples to populations. In meta-analysis there are two kinds of models that have been used to facilitate inferences from samples to populations: the fixed and the random effects models (Cooper et al., 2009). When the inference pertains only to the particular sample of studies used in a research synthesis (or other sets of identical studies) then the fixed effects model seems appropriate (Hedges, 2009). The underlying assumption in this meta-analytic model is that the effect size parameter is unknown, but fixed at a certain value. That is, the collection of the specific studies in the sample at hand has a common true effect size (Borenstein et al., 2009). Another way of thinking about fixed effects models in meta-analysis is via the homogeneity of effect sizes. For example, when a treatment yields comparable effects across primary studies, it is reasonable to combine the effect size estimates of all the studies in the sample and summarize the treatment effect by a single common estimate. The fixed effects models assume that the between study heterogeneity of the effects is zero.

In the simplest case, the fixed effects model involves the computation of one average effect size. Specifically, the meta-analyst combines the effect size estimates across all studies in the sample using weights to compute an overall weighted average. Now, let θ_i be the unobserved effect size parameter in the ith study, let T_i be the corresponding observed effect size estimate from the ith study, and let v_i be its variance. The data from a set of k studies are the effect size estimates T_1, …, T_k and their corresponding variances v_1, …, v_k.

The effect size estimate T_i is modeled as the effect size parameter plus a sampling error ε_i. That is

$$T_i = \theta_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, v_i).$$

The parameter θ is the mean effect size parameter for all studies and has the interpretation that θ is the mean of the distribution from which the study-specific effect size parameters (θ_1, θ_2, ..., θ_k) were sampled. This is not conceptually the same as the mean of θ_1, θ_2, ..., θ_k, the effect size parameters of the k studies that were observed. The effect size parameters are in turn determined by a mean effect size β_0, that is

$$\theta_i = \beta_0,$$

which indicates that the θ_i's are fixed and thus in a single equation the effect size estimate

$$T_i = \beta_0 + \varepsilon_i. \tag{14}$$

Note that in meta-analysis, the variances (i.e., the v_i's) vary from study to study. That is, each study has a *different* sampling error variance. In meta-analysis these

variances are known and are a function of the sample size of the study. The amount of sampling uncertainty is not identical in every study, and thus, the precision of the estimates varies from study to study. If an average effect size is to be computed across studies, it seems reasonable to utilize a weighting scheme and assign more weight to estimates with more precision (i.e., smaller variance) than those with less precision.

The weighted least squares (and maximum likelihood) estimate of the overall effect, β_0 under the model is

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} \tag{15}$$

where $w_i = 1/v_i$. This estimator corresponds to a weighted mean of the T_i's and gives more weight to the studies whose estimates have smaller variances. This weighted average can be computed using a weighted regression model that includes only the constant term.

The variance $v_\bullet$ of the estimate $\hat{\beta}_0$ is simply the reciprocal of the sum of the weights across studies,

$$v_\bullet = \left(\sum_{i=1}^{k} w_i \right)^{-1}. \tag{16}$$

and the standard error SE($\hat{\beta}_0$) of $\hat{\beta}_0$ is just the square root of $v_\bullet$. Under this model $\hat{\beta}_0$ is normally distributed, and a $100(1-\alpha)$ percent confidence interval for the parameter β_0 is given by

$$\hat{\beta}_0 - t_{\alpha/2}\sqrt{v_\bullet} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2}\sqrt{v_\bullet}, \tag{17}$$

where t_α is the 100α percent point of the t-distribution with $(k-1)$ degrees of freedom. Alternatively, a two-sided test of the hypothesis that $\beta_0 = 0$ at significance level α uses the test statistic $Z = \hat{\beta}_0 / \sqrt{v_\bullet}$ and rejects if $|Z|$ exceeds $t_{\alpha/2}$.

RANDOM EFFECTS MODELS

When the statistical inference pertains to generalizations beyond the observed sample of studies used in a research synthesis, then the random effects model seems appropriate (Hedges, 2009). In this case the specific samples of studies may not be the main interest since it is simply a sample of studies drawn from a population. Thus, different sets of studies drawn from the population may differ in characteristics and

effect size parameters. In this model the effect size parameters are treated as if they were a random sample from a population of effect size parameters (DerSimonian & Laird, 1986; Hedges, 1983; Raudenbush & Bryk, 2002). Another way of thinking about random effects models is via the heterogeneity of effect sizes. For example, it is possible that the study specific effects might be inconsistent. That is, the size of the treatment effect may vary considerably across studies. In these cases, the meta-analytic methods need to take into account this variability in the effects. As a result, meta-analytic models become more complicated since they must be designed to take into account two sources of variation. The first source of variation of study estimates is the sampling error. Even under the assumption that there is a common effect size parameter fixed at a specific value, the effect size estimates will vary from study to study due to sampling error, which is the within study variability. The second source of variation in study estimates arises when the effect size parameter is itself random and has its own distribution (i.e., it varies across sets of studies). This component of variation of the effect size parameters across studies represents the inconsistency or heterogeneity in effects across studies (see Raudenbush & Bryk, 2002).

When the effect size parameter is a random effect, the applicable meta-analytic model is called a random effects model, because it captures random variation among studies (or between-study variation) (see Schmidt and Hunter, 1977; Hedges, 1983; DerSimonian and Laird, 1986). This model introduces heterogeneity among the effect size parameters that is captured by the between-study variance (Hedges & Vevea, 1998). However, the between-study variance has to be non-negligible or statistically significant in order for a random effects model to be appropriate. The random effects model can be thought of as a generalization of the fixed effects model that incorporates random variation across studies. In other words, the fixed effects model is a special case of the random effects model where the between-study variance is zero.

The simplest random effects model involves the estimation of an average effect size across studies. Again, the analyst combines the effect size estimates across all studies in the sample to compute a weighted average. However, in this case a natural way to describe the data is via a two-level model with one model for the data at the study level and another model for the between-study variation. The within-study level is as defined for the fixed effects models earlier. In the between-study level, the effect size parameters are modeled around a mean effect size β_0 plus a study-specific random effect η_i. That is

$$\theta_i = \beta_0 + \eta_i, \qquad \eta_i \sim N(0, \tau^2).$$

The above equation suggests that θ_i is random and has distribution with a specific variance τ^2. Now, the η_i represent differences between the effect size parameters from study to study. The variance parameter τ^2, often called the between-study variance component, describes the amount of variation across studies in the random effects (the η_i's), and therefore the effect size parameters (the θ_i's).

The two-level model described above can be written as a one-level model as follows

$$T_i = \beta_0 + \eta_i + \varepsilon_i = \beta_0 + \xi_i, \tag{18}$$

where ξ_i is a composite error defined by $\xi_i = \eta_i + \varepsilon_i$. Writing this as a one-level model, we see that each effect size is an estimate of the population parameter β_0 with a variance that depends on both v_i and τ^2. Hence, it is necessary to distinguish between the variance of T_i assuming a fixed θ_i and the variance of T_i incorporating the variance of the θ_i as well. Since the sampling error ε_i and the random effect η_i are assumed to be independent and the variance of η_i is $\hat{\tau}^2$, it follows that the variance of T_i is the sum of the two variances, $v_i^* = v_i + \hat{\tau}^2$.

The least squares (and maximum likelihood) estimate of the mean a_0 under the model is

$$\hat{\beta}_0^* = \frac{\sum_{i=1}^{k} w_i^* T_i}{\sum_{i=1}^{k} w_i^*}, \tag{19}$$

where $w_i^* = 1/(v_i + \hat{\tau}^2) = 1/v_i^*$ and $\hat{\tau}^2$ is the between-study variance component estimate. This estimator corresponds to a weighted mean of the T_i, giving more weight to studies with estimates that have smaller variances.

The sampling variance $v_\bullet^*$ of weighted average $\hat{\beta}_0^*$ is simply the reciprocal of the sum of the weights,

$$v_\bullet^* = \left(\sum_{i=1}^{k} w_i^* \right)^{-1}, \tag{20}$$

and the standard error SE $(\hat{\beta}_0^*)$ of $\hat{\beta}_0^*$ is just the square root of $v_\bullet^*$. Under this model $\hat{\beta}_0^*$ is normally distributed so a $100(1 - \alpha)$ percent confidence interval for β_0 is given by

$$\hat{\beta}_0^* - t_{\alpha/2}\sqrt{v_\bullet^*} \le \beta_0 \le \hat{\beta}_0^* + t_{\alpha/2}\sqrt{v_\bullet^*}, \tag{21}$$

where t_α is the 100α percent point of the t-distribution with (k – 1) degrees of freedom. Similarly, a two-sided test of the hypothesis that $\beta_0 = 0$ at significance level α uses the test statistic $z = \hat{\beta}_0^* / \sqrt{v_\bullet^*}$ and rejects if $|Z|$ exceeds $t_{\alpha/2}$.

Estimation of the Between-Study Variance τ^2

The estimation of τ^2 can be accomplished without making assumptions about the distribution of the random effects or under various assumptions about the distribution

of the random effects using other methods such as maximum likelihood estimation. Maximum likelihood estimation is more efficient if the distributional assumptions about the study-specific random effects are correct, but these assumptions are often difficult to justify theoretically and difficult to verify empirically. Thus distribution free estimates of the between-studies variance component are often preferred.

A simple, distribution free estimate of τ^2 is given by

$$\hat{\tau}^2 = \begin{cases} \dfrac{Q-(k-1)}{a} & if\ Q \geq (k-1) \\ 0 & if\ Q < (k-1) \end{cases} \tag{22}$$

Where a is given by

$$a = \sum_{j=1}^{k} w_i - \frac{\sum_{j=1}^{k} w_i^2}{\sum_{j=1}^{k} w_i}, \tag{23}$$

$w_i = 1/v$ and Q is defined as

$$Q = \sum_{i=1}^{k} ((T_i - \hat{\beta}_0)^2 / v_i), \tag{24}$$

where $\hat{\beta}_0$ is the estimate of β_0 that would be obtained from equation (15) under the hypothesis that $\tau^2 = 0$. The statistic Q has the chi-squared distribution with $(k - 1)$ degrees of freedom if $\tau^2 = 0$. Therefore, a test of the null hypothesis that $\tau^2 = 0$ at significance level α rejects the hypothesis if Q is greater than the $100(1 - \alpha)$ critical value of the chi-square distribution with $(k - 1)$ degrees of freedom. Estimates of τ^2 are set to 0 when $Q - (k - 1)$ yields a negative value, since τ^2, by definition, cannot be negative.

EXAMPLE

To illustrate the usefulness of fixed and random effects models let's consider an example about modified school calendars. The data include studies on schools that modified their calendars without extending the length of the school year (see Cooper, Valentine, Charlton, & Melson, 2003). All studies assessed students from grade one through grade nine and reported achievement differences between students attending year round calendar schools and traditional calendar schools. The achievement differences were expressed in standard deviation units (i.e., standardized mean differences) to ensure all estimates were on the same scale. Our

data included information on mathematics achievement. The sample of studies used here is somewhat different than that used in the Cooper et al. study, but it suffices for the purposes of the example. Overall, 46 studies were included in the sample. The data are reported in Table 1. The studies were retrieved from databases including ERIC, PsychINFO, and Dissertation Abstracts.

RESULTS

All standardized mean differences or effect sizes in the data do not reflect adjustments for covariates, and thus, they are unadjusted differences between year round calendar and traditional calendar schools. Data from 46 studies were included in the computations. The effect size estimates ranged from −1.19 to 1.12 with a mean of −0.0368 and a standard deviation of 0.4288. Negative effect sizes indicate that students attending traditional (nine-month) calendar schools outperformed their counterparts in year-round calendar schools. In contrast, positive effect sizes point to higher student achievement in year-round calendar schools compared to traditional calendar schools.

First, let's consider the fixed effects model. We computed the fixed effects weighted average using SAS proc reg (see Appendix). Essentially, this is a weighted least squares model. One can also use SAS proc means to compute the weighted mean. The weighted average effect size using the data in Table 1 is −0.0282 and the standard error of the estimate is 0.0096. The standard error of the weighted average is computed as the ratio of the standard error (0.0300) of the estimate produced by proc reg (where only the constant is included in the equation) to the square root of the mean square error in the ANOVA table (3.1128). In this case, the square root of the mean square error in the ANOVA table is the same as the standard deviation of the weighted average estimate obtained from in proc means. The z test in this case is significant at the 0.05 level indicating that overall mathematics achievement was significantly higher in traditional calendar schools than year-round calendar schools. The 95% CI (−0.0476, −0.0088) does not include zero as expected.

Now let's consider a random effects model. One can use SAS proc mixed to conduct the meta-analysis (see Appendix). SAS proc mixed uses maximum likelihood estimation to compute the estimates (Littell et al., 1996; Singer, 2008). Conceptually the random effects model is a two-level model, where the first level involves a within-study model and the second level a between-study model (Konstantopoulos, 2011). The estimate of the between-study variance is 0.1327 and it is significant at the 0.05 level, indicating systematic differences in effect sizes across studies. To indicate the magnitude of the variance estimate notice that the between-study variance is nearly five times larger than the average of the 46 effect size variances. The overall weighted average estimate across studies is −0.0103 and its standard error is 0.0576. The weighted average in this case is nearly one-half as large as that in the fixed effects model, whilst the standard error of the random effects mean estimate is more than three times larger than the standard error of the

mean estimate in the fixed effects model. One would expect a larger standard error of the mean estimate in the random effects model whenever the between-study variance is considerable. In addition, because the weights used in the computation of the mean estimate incorporate the between-study variance one would expect that the mean estimates in the fixed and the random effects models would differ. The *z* test used indicated that the overall effect was not different from zero (i.e., non-significant effect). The 95% CI (–0.126, 0.106) includes zero as expected.

CONCLUSION

This chapter described research synthesis and meta-analysis and discussed fixed effects and random effects models used in meta-analysis. To demonstrate the applicability of the models we used data from studies that compared mathematics achievement between schools that followed a traditional calendar and schools that modified their calendar (year-round schools).

The weighted averages both in fixed and random effects models were negative and in the fixed effects case the estimate was significantly different than zero. This indicates that the traditional calendar schools performed higher than the year-round calendar schools in mathematics. The fixed effects estimate was twice as large as that produced from the random effects analysis. In contrast, but as expected, the standard error of the random effects model mean estimate was a few times larger than that of the fixed effects mean estimate. This is expected whenever the between-study variance is nontrivial. In addition, because the weights used in the random effects analysis are different than those in the fixed effects analysis the overall average estimates are also different in the two models. Again, the difference in the mean estimates will be more pronounced when the between-study variance is substantial.

In sum, analysts use research synthesis to combine evidence across studies and make generalizable statements. The statistical analysis part of the research synthesis, known as meta-analysis, is essential in this process. Analysts can use fixed or random effects models to analyze the data. Their decision depends on the inferences they want to make as well as on the consistency of the effects in the sample of studies.

APPENDIX

Fixed effects model: Using proc reg in SAS

```
proc reg data=temp;
     model effectsize = / ;
     weight wt;
```

where wt = 1/variance. The standard error of the estimate needs to be devided by the square root of the Mean Square Error.

Random effects model: Two-level unconditional meta-analysis using proc mixed in SAS

```
proc mixed data=temp covtest;
      class studyid;
      model effectsize = / solution notest ;
      random int / sub = studyid;
      repeated / group = studyid;
      parms (0.1)
(0.145) (0.120) (0.148) (0.138) (0.023) (0.043) (0.012)
(0.016) (0.016) (0.019) (0.020) (0.015) (0.015) (0.017)
(0.017) (0.019) (0.007) (0.005) (0.004) (0.020) (0.018)
(0.019) (0.023) (0.020) (0.022) (0.006) (0.007) (0.007)
(0.007) (0.007) (0.015) (0.012) (0.009) (0.001) (0.001)
(0.001) (0.001) (0.001) (0.001) (0.001) (0.001) (0.030)
(0.034) (0.031) (0.030) (0.033)
/ eqcons=2 to 47;
run;
```

REFERENCES

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester West Sussex, U.K.: Wiley.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods*, *8*, 243–253.

Cook, T. D., Cooper, H., Cordray, D., Hartmann, H., Hedges, L., Light, R., Louis, T., Mosteller, F. (Eds.). (1992). *Meta-Analysis for explanation: A casebook*. New York: Russell Sage Foundation.

Cooper, H. (1989). *Integrating research* (2nd Ed.). Newbury Park, CA: Sage Publications.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd Ed.). New York: Russell Sage.

Cooper, H., Valentine, J. C., Charlton, K., & Melson, A. (2003). The effects of modified school calendars on student achievement and on school and community attitudes: A research synthesis. *Review of Educational Research*, *73*, 1–52.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

Garvey, W., & Griffith, B. (1971). Scientific communication: Its role in the conduct of research and creation of knowledge. *American Psychologist*, *26*, 349–361.

Fleiss, J. L. (1994). Measures of effect size for categorical data. Pages 245–260 in H. Cooper and L. V. Hedges, *The handbook of research synthesis*. New York: The Russell Sage Foundation.

Glass, G. V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher*, *5*, 3–8.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*, 388–395.

Hedges, L. V. (2009). Statistical considerations. In H. Cooper, L. V. Hedges, & J. C. Valentine, *The handbook of research synthesis and meta-analysis* (pp. 357–376). New York: Russell Sage.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random effects models in meta analysis. *Psychological Methods*, *3*, 486–504.

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis? *Research Synthesis Methods*, *2*, 61–76.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute INC.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*. Newbury Park, CA: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimentaldesigns for generalized causal inference*. Boston, MA: HoughtonMifflin.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel growth models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *24*, 323–355.

Table 1. Effect sizes for school calendar studies: mathematics

Study	*Effect Size*	*Variance*
1	–0.330	0.145
2	–0.410	0.120
3	–0.560	0.148
4	–1.190	0.138
5	0.470	0.023
6	0.310	0.043
7	0.220	0.012
8	0.080	0.016
9	–0.260	0.016
10	0.460	0.019
11	–0.630	0.020
12	–0.050	0.015
13	0.110	0.015
14	0.000	0.017
15	0.050	0.017
16	0.081	0.019
17	0.324	0.007
18	0.251	0.005
19	–0.063	0.004
20	0.000	0.020
21	0.190	0.018
22	–0.280	0.019
23	–0.610	0.023
24	–0.150	0.020
25	–0.620	0.022
26	0.130	0.006
27	0.000	0.007
28	–0.030	0.007

(*Continued*)

Table 1. Effect sizes for school calendar studies: mathematics - Continued

Study	*Effect Size*	*Variance*
29	–0.100	0.007
30	0.110	0.007
31	0.230	0.015
32	1.120	0.012
33	0.690	0.009
34	–0.130	0.001
35	–0.090	0.001
36	0.002	0.001
37	0.040	0.001
38	–0.030	0.001
39	–0.050	0.001
40	–0.070	0.001
41	–0.190	0.001
42	–0.320	0.030
43	1.030	0.034
44	–0.520	0.031
45	0.020	0.030
46	–0.930	0.033

MAURICIO SALGADO & NIGEL GILBERT

12. AGENT BASED MODELLING

INTRODUCTION

In this chapter, we describe the main characteristics of agent-based modelling. Agent-based modelling is a computational method that enables researchers to create, analyse, and experiment with models composed of autonomous and heterogeneous agents that interact within an environment, in order to identify the mechanisms that bring about some macroscopic phenomenon of interest. Here, we explain what agent-based modelling is all about, addressing some theoretical issues and defining the main elements of an agent-based model. We also discuss the relation between this method and the quest for causal explanations in the social sciences and the important issue about the identification of social mechanisms. Then, we suggest a standardized process consisting of a sequence of steps to develop agent-based models for social science research. Finally, we present a useful example of an agent-based model that tackles an important phenomenon in educational research: differential school effectiveness.

WHAT IS AGENT-BASED MODELLING?

Theoretical Background

Over the past forty years, a new kind of research method has increasingly been used in the social sciences: that of the *agent-based modelling* (from here on *ABM*). ABM is an outstanding modelling technique to build explanations of social processes, based on ideas about the emergence of complex behaviour from simple activities (Axelrod, 1997; Epstein & Axtell, 1995; Gilbert & Troitzsch, 2005). With this technique we can study properties of emergent orders that arise from local interactions among a multitude of independent components. And we can understand the ways in which such emergent orders can influence or constrain the individual actions of those components. This process is known as 'self-organisation' and is characterised by the concepts of 'bottom-up' and 'downward causation'.

There is an increasing interest in ABM as a modelling approach in the social sciences, since it enables researchers to build computational models where individual entities and their cognition and interactions are directly represented. In comparison to alternative modelling techniques, such as variable-based approaches using structural equations (or statistical modelling) or system-based

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 247–265.

approaches using differential equations (or mathematical modelling), ABM allows modellers to simulate the emergence of macroscopic regularities over time, such as ants colonies, flock of birds, norms of cooperation, traffic jams, or languages, from interactions of autonomous and heterogeneous agents (Gilbert, 2007). The emergent properties of an agent-based model are then the result of 'bottom-up' processes, the outcome of agent interactions, rather than 'top-down' direction. In fact, the absence of any form of top-down control is the hallmark of ABM, since the cognitive processes, behaviours, and interactions at the agent-level bring about the observed regularities in the system- or macro-level. For this reason, ABM is most appropriate for studying processes that lack central coordination, including the emergence of macroscopic patterns that, once established, impose order from the top down.

Agent-based models involve two main components. Firstly, these models entail the definition of a population of *agents*. Secondly, they involve the definition of some relevant *environment*. In the following, we discuss the core concepts of 'agents' and their 'environment' in more detail.

The Agents

The agents are the computational representation of some specific social actors – individual people or animals, organisations such as firms or bodies such nation-states – capable of interacting, that is, they can pass informational messages to each other and act on the basis of what they learn from these messages. Thus, each agent in the model is an autonomous entity. The artificial population can include heterogeneous agents, which is useful when the researcher wants to build a model of a certain phenomenon with different agents' capabilities, roles, perspectives or stocks of knowledge. In ABM, agents are conventionally described as having four important characteristics (Abdou, Hamill, & Gilbert, 2012):

- *Perception*: Agents can perceive their environment, including other agents in their vicinity.
- *Performance*: They have a set of behaviours that they are capable of performing such as moving, communicating with other agents, and interacting with the environment.
- *Memory*: Agents have a memory in which they record their previous states and actions.
- *Policy*: They have a set of rules, heuristics or strategies that determine, given their present situation and their history, what they should do next.

Agents with these features can be implemented in many different ways. Different architectures (i.e. designs) have merits depending on the purpose of the simulation. Nevertheless, every agent design has to include mechanisms for receiving input from the environment, for storing a history of previous inputs and actions, for devising what to do next, for carrying out actions and for distributing outputs.

The Environment

ABM also involves the definition of some relevant environment. The environment is the virtual world in which the agents act. It may be an entirely neutral medium with little or no effect on the agents, as in some agent-based models based on game theory, where the environment has no meaning. In other models, the environment may be as carefully designed as the agents themselves, as in some ecological or anthropological agent-based models where the environment represents complex geographical space that affects the agents' behaviour.

ABM and Complexity Theory

Finally, it can be said that the interest in ABM reflects a growing attention in complex adaptive systems by social scientists, that is to say, the possibility that human societies may be described as highly complex, path-dependent, non-linear, and self-organising systems (Castellani & Hafferty, 2009; Macy & Willer, 2002; Miller & Page, 2007). Complexity theory and the accompanying trappings of complex systems provide the theoretical basis for agent-based models. Thus, for instance, a complex system is a set of entities connected to each other and the external environment in a way that gives it an overall identity and behaviour. An agent-based model in its most basic form represents a system of such discrete entities (Manson, Sun, & Bonsal, 2012). For this reason, while modellers are usually interested in addressing specific theoretical questions and working in particular substantive areas, they almost invariably draw on complexity concepts when using an agent-based approach. The emphasis on processes and on the relations between entities that bring about macroscopic regularities, both of which can be examined by ABM, accounts for the developing link between complexity theory and ABM research.

MECHANISMS, EXPLANATIONS AND ABM

One of the main objectives of ABM is to test, by experimental means, the hypothesised mechanisms that bring about the macroscopic phenomenon the researcher is interested in explaining. Following the definition provided by Hedström (2005), a *mechanism* describes a constellation of entities (i.e., agents) and activities (i.e., actions) that are organised such that they regularly bring about a particular type of outcome. Therefore, social researchers explain an observed macroscopic phenomenon by referring to the mechanisms by which the phenomenon is regularly brought about.

In ABM these mechanisms are translated as the model *microspecifications*, that is to say, the set of behavioural and simple rules that specify how the agents behave and react to their local environment (which includes, of course, other agents). Once the population of agents and the environment are defined, the researcher can implement the microspecifications and run the computer simulation in order to evaluate whether these rules bring about or 'generate' the macro phenomenon of interest, over the

simulated time. The motto of ABM is then: 'if you did not grow it, you did not explain it'. Equation (1) represents the same motto in the notation of first order logic:

$$(\forall x)(\neg Gx \supset \neg Ex) \quad (1)$$

Note that, in this perspective, there is a sharp distinction between generative explanations from the mere description or mere discovery of regularities. Clearly, it is not sufficient to identify, for instance, the statistical association between two or more variables. In ABM, what defines an explanation is the explicit representation, in a computer code, of the underlying generative mechanism, which is a deeper reconstruction of the social regularity. As Hedström and Swedberg (1996, p. 287) claimed, "[t]he search for generative mechanisms […] helps us distinguish between genuine causality and coincidental association, and it increases the understanding of why we observe what we observe." Therefore, understanding is obtained or enhanced by making explicit the underlying generative mechanisms that link one state or event to another – and in the social sciences, individual actions constitute this link.

Agent-based models can be used to perform highly abstract thought experiments that explore plausible mechanisms that may underlie observed patterns. That is precisely one of the promises of ABM: given the limitations of experimental methods and the complexity of social phenomena, agent-based models are important for this kind of endeavour (Hedström & Ylikoski, 2010). ABM allows systematic exploration of consequences of modelling assumptions and make it possible to model much more complex phenomena than was possible earlier.

DEVELOPING AGENT-BASED MODELS IN SOCIAL SCIENCE RESEARCH

Research in ABM has developed a more or less standardized research process, consisting of a sequence of steps. In practice, several of these steps occur in parallel and the whole process is often performed iteratively as ideas are refined and developed. Figure 1 shows the main steps that researchers should follow in order to build an agent-based model. The sequence begins with the social theory and finishes with 'the target' or the social process the researcher is interested in modelling.

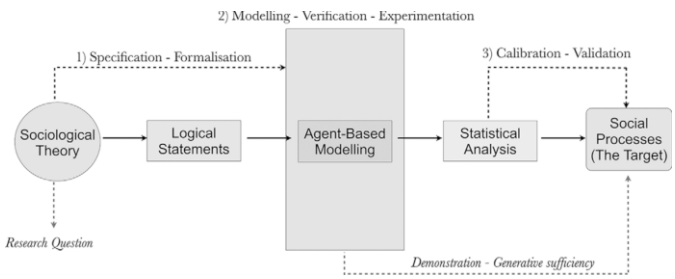


Figure 1. Main steps and stages to build an agent-based model.

Between these two points, a series of steps must be achieved in order to develop a sound agent-based model. In order to simplify the presentation, we have identified three major stages: 1) Specification and formalisation; 2) Modelling, verification and experimentation; and 3) Calibration and validation. The first stage involves translating the theoretical hypothesis that explains the social process of interest, usually expressed in natural languages, into *formal languages*, using logics or mathematics. The second stage includes the modelling itself, in which the researcher builds and verifies the model by experimental means. The third step includes the calibration of the model with empirical data and the consequent validation of it using appropriate statistical tests. In the following subsections we describe these major stages in detail.

Specification and Formalisation

It is essential to define precisely the research question (or questions) that the model is going to address at an early stage. The typical research questions that researchers try to answer when using ABM are those that explain how regularities observed at the societal or macro level can emerge from the interactions of agents at the micro level. According to Epstein (1999, 2007)features distinguishing the approach from both ldquoinductiverdquo and ldquodeductiverdquo science are given. Then, the following specific contributions to social science are discussed: The agent-based computational model is a new tool for empirical research. It offers a natural environment for the study of connectionist phenomena in social science. Agent-based modeling provides a powerful way to address certain enduring – and especially interdisciplinary – questions. It allows one to subject certain core theories – such as neoclassical microeconomics – to important types of stress (e.g., the effect of evolving preferences, researchers using ABM try to solve the following question: '*How could the decentralised local interactions of heterogeneous agents generate or bring about the given macro phenomenon?*' This is the typical research question that modellers working with ABM aim to answer for any macro-property they want to explain.

To define precisely the research question, it is also needed that the model is embedded in existing *social theories*. Reviewing existing theories relating to the model's research question is important to identify the causal mechanisms that are likely to be significant in the model. Thus, it is important to choose the theory that provides the most plausible and empirically testable causal mechanism. In this sense, one important feature of ABM is that it does not impose any *a priori* constraints on the mechanisms assumed to be operating. ABM is not based on any specific theory of action or interaction (Hedström & Ylikoski, 2010). It is a methodology for deriving the social outcomes that groups of interacting actors are likely to bring about, whatever the action logics or interaction structures may be.

Evidently, a model is always a simplification of the 'real world'. In fact, this is the reason why scientists build models: they want to reduce the complexity of the world and isolate the main elements that bring about the phenomenon to be explained. For

this reason, researchers using ABM aim to *specify* the causal mechanisms underlying some phenomenon. According to Miller and Page (2007), to specify a theory is to reduce the world to a fundamental set of elements (equivalence classes) and laws (transition functions), and on this basis to better understand and possibly predict key aspects of the world. As we discussed before, in ABM researchers have to specify the agents that are to be involved in the model and the environment in which they will act. For each type of agent in the model, the attributes and behavioural rules need to be specified – that is, the set of simple rules that specify how the agents behave and react to their local environment. An attribute is a characteristic or feature of the agent, and it is either something that helps to distinguish the agent from others in the model and does not change, or something that changes as the simulation runs.

Once the appropriate theory – the one that provides plausible causal explanations about the target social process – has been identified and the behavioural rules have been specified, the researcher is equipped with *theoretical hypothesis* to be tested, in this case, *in silico*, that is, using computers as laboratories to run experiments. In the social sciences, hypotheses are expressed usually in textual form; in *natural languages*. However, natural languages are usually ambiguous and concepts are not always rigorously defined. For this reason, when we have a theory of how individuals behave in the situation we are analysing, it is useful to express it in the form of a procedure or *formal or artificial language*, using logics or mathematics. An advantage of using logic is that it gives conditions – these results hold when and if the following are true. Thus, researchers can use conditional rules such as '*If C1, C2, and C3, then EP*', where *C* represents some condition and *EP* is the emergent or macro property they want to explain. This kind of formalisation facilitates the ultimate aim in ABM: to formalise theoretical hypothesis in the form of a computer program.

As Gilbert and Troitzsch (2005, p. 5) argue, "[t]he process of formalization, which involves being precise about what the theory means and making sure that it is complete and coherent, is a very valuable discipline in its own right". In this respect, ABM has a similar role in the social sciences to that of mathematics in the physical sciences.

Modelling, Verification and Experimentation

Once the theory has been specified and formalised into logics or mathematics, modellers can translate this formalisation into computer programs. Thus, the formal model can be programmed and run on the computer, and the behaviour of the simulation can be observed and tested. To build a model is similar to design an experiment. Thus, given the macroscopic *explanandum* – a regularity to be explained – the canonical agent-based experiment is as follows (Epstein, 1999, 2007)features distinguishing the approach from both ldquoinductiverdquo and ldquodeductiverdquo science are given. Then, the following specific contributions to social science are discussed: The agent-based computational model is a new tool for empirical research. It offers a natural environment for the study of connectionist phenomena in social science. Agent-based modeling provides a powerful way to

address certain enduring - and especially interdisciplinary - questions. It allows one to subject certain core theories - such as neoclassical microeconomics - to important types of stress (e.g., the effect of evolving preferences: '*situate an initial population of autonomous heterogeneous agents in a relevant environment; allow them to interact according to local rules, and thereby generate – or grow – the macroscopic regularity from the bottom up*'. The ultimate aim of researchers using ABM is to establish an account of the configuration's attainment by a decentralised system made of heterogeneous and autonomous agents.

However, when writing computer programs, especially complicated ones, it is very common to make errors. The process of checking that a program does what it was planned to do is known as *verification*. In the case of agent-based models, the difficulties of verification are compounded by the fact that many simulations include random number generators, which means that every run is different and that it is only the distribution of results which can be anticipated by the theory. It is therefore essential to 'debug' the simulation carefully, preferably using a set of test cases, perhaps of extreme situations where the outcomes are easily predictable and run multiple experiments in order to explore and measure the *parameter space* – the set of values of parameters encountered in a particular model.

When the agent-based model can generate the type of outcome to be explained, then the researcher has provided a *computational demonstration* that a given microspecification (or mechanism) is in fact sufficient to generate the macrostructure of interest. This demonstration, called *generative sufficiency* (Epstein, 1999)features distinguishing the approach from both ldquoinductiverdquo and ldquodeductiverdquo science are given. Then, the following specific contributions to social science are discussed: The agent-based computational model is a new tool for empirical research. It offers a natural environment for the study of connectionist phenomena in social science. Agent-based modeling provides a powerful way to address certain enduring - and especially interdisciplinary - questions. It allows one to subject certain core theories - such as neoclassical microeconomics - to important types of stress (e.g., the effect of evolving preferences, provides a candidate mechanism-based explanation of the macro-phenomenon. The agent-based modeller can then use relevant data and statistics to estimate the generative sufficiency of a given microspecification by testing the agreement between 'real-world' and the generated macrostructures in the computer simulation (we will discuss more about this in the following subsection). On the other hand, when the agent-based model cannot generate the outcome to be explained, the microspecification is not a candidate explanation of the phenomenon and the researcher has demonstrated the hypothesised mechanism to be false.

Calibration and Validation

Once researchers have specified some a substantively plausible agent-based model, and this model generates the emergent macro pattern of interest, they can use empirical data to estimate the size of various unknown parameters of this model and

the agreement between the predicted and the real data. This is reached by calibrating and validating the model.

While verification concerns whether the program is working as the researcher expects (as discussed in the previous subsection), *validation* concerns whether the simulation is a good model of the real system, the 'target'. A model that can be relied on to reflect the behaviour of the target is 'valid'. A common way of validating a model is to compare the output of the simulation with real data collected about the target. However, there are several caveats which must be borne in mind when making this comparison. For example, exact correspondence between the real and simulated data should not be expected. So, the researcher has to decide what difference between the two kinds of data is acceptable for the model to be considered valid. This is usually done using some *statistical analysis* to test the significance of the difference. While goodness-of-fit can always be improved by adding more explanatory factors, there is a trade-off between goodness-of-fit and simplicity. Too much fine-tuning can result in reduction of explanatory power because the model becomes difficult to interpret. At the extreme, if a model becomes as complicated as the real world, it will be just as difficult to interpret and offer no explanatory power. There is, therefore, a paradox here to which there is no obvious solution. Despite its apparently scientific nature, modelling is a matter of judgement.

Finally, it is important to distinguish the different ways in which an agent-based model can be validated and calibrated. According to Bianchi, Cirillo, Gallegati, & Vagliasindi, (2007), there are three ways of validating an agent-based model, namely:

- *Descriptive output validation*, or matching computationally generated output against already available data. This kind of validation procedure is probably the most intuitive one and it represents a fundamental step towards a good model's calibration;
- *Predictive output validation*, or matching computationally generated data against yet-to-be-acquired system data. Obviously, the main problem concerning with this procedure is essentially due to the delay between the simulation results and the final comparison with actual data;
- *Input validation*, or ensuring that the fundamental structural, behavioural and institutional *initial* conditions incorporated in the model reproduce the main aspects of the actual system.

Since the empirical validation of ABM is still a brand new topic, at the moment there are only a limited number of contributions in the literature dealing with it (see for instance, Axtell, Axelrod, Epstein, & Cohen, 1996; Bianchi et al., 2007; Fagiolo, Moneta, & Windrum, 2007; Kleindorfer, O'Neill, & Ganeshan, 1998).

Softwares and Modelling Environments

Once the agent-based model has been formalised, an important decision is whether to write a special computer program (using a programming language such as Java,

C++, C#, or Visual Basic) or use one of the packages or toolkits that have been specially created to help in the development of simulations. It is almost always easier to use a package than to start afresh writing one's own program. This is because many of the issues that take time when writing a program have already been dealt with in developing the package. For example, writing code to show plots and charts from scratch is a skilled and very time-consuming task, but most packages provide some kind of graphics facility for the display of output variables. At least some of the bugs in the code of packages will have been found by the developer or subsequent users (although you should never assume that all bugs have been eliminated). The disadvantage of packages is that they are, inevitably, limited in what they can offer. It is difficult to find one that is easy to debug, has a good graphics library, can be compiled efficiently and is portable across different computers.

There is a choice of several packages for ABM, most of them are available free of charge and they are well suited for social scientists. Table 1 provides a comparison between four popular modelling environments on a number of criteria. The choice of the implementation tool depends on several factors, especially one's own expertise

Table 1. A comparison of Swarm, RePast, Mason and NetLogo

	Swarm	*RePast*	*Mason*	*NetLogo*
Licence*	GPL	GPL	GPL	Free, but not open source
Documentation	Patchy	Limited	Improving, but limited	Good
User Base	Diminishing	Large	Increasing	Large
Modelling Language(s)	Objective-C, Java	Java, Python	Java	NetLogo
Speed of Execution	Moderate	Fast	Fastest	Moderate
Support for graphical user interface development	Limited	Good	Good	Very easy to create using 'point and click'
Built-in ability to create movies and animations	No	Yes	Yes	Yes
Support for systematic experimentation	Some	Yes	Yes	Yes
Easy of learning and programming	Poor	Moderate	Moderate	Good
Easy of Installation	Poor	Moderate	Moderate	Good
Link to geographical information system	No	Yes	Yes	Yes

Source: Gilbert (2007)

*GPL General Public Licence, http://www.gnu.org/copyleft/gpl.html

in programming and the complexity and the scale of the model. NetLogo is the quickest to learn and the easiest to use, but may not be the most suitable for large and complex models. Mason is faster than RePast, but has a significantly smaller user base, meaning that there is less of a community that can provide advice and support.

NetLogo (Wilensky, 2011) is currently the best of the agent-based simulation environments for beginners and even for many serious scientific models. NetLogo is a distant descendant of the Logo programming language that was created in the 1960s as a tool for schoolchildren. It still retains some aspects of its heritage; for example, there are 'turtles' and they move around on 'patches'. Its programming language uses a very simple syntax that is supposed to resemble English, and provides a simplified programming language and graphical interface that lets users build, observe and use agent-based models without needing to learn the complex details of a standard programming language. Just as importantly, it has a growing user community and the NetLogo team at Northwestern University provides a complete, useful and professional set of documentation and tutorial materials.

AN EXAMPLE: AGENT-BASED MODEL OF DIFFERENTIAL SCHOOL EFFECTIVENES

In this section we explain a simple agent-based model that addresses the educational phenomenon known as 'differential school effectiveness', that is to say, we present a model that aim to explain why some schools differ significantly in terms of their effectiveness for particular pupil groups (Coleman, 1966; Nuttall, Goldstein, Prosser, & Rasbash, 1989; Sammons, Nuttall, & Cuttance, 1993). This model is based on some theories about social behaviours and interactions in the classroom among pupils and teachers. Since the primary goal of this chapter is to explain ABM, we are *not* interested in precisely identifying the underlying process that causally explains differences in school effectiveness. Hence, the mechanism we implement as the model microspecification must be understood as a *candidate* mechanism-based explanation, a provisional hypothesis that, although sufficient to generate the observed differences, is subject to further empirical falsification.

The agent-based model we present in this Chapter was calibrated using a well-known dataset, the *London Educational Authority's Junior Project* (Nuttall et al., 1989). This was a longitudinal study of around 2000 children. For this chapter, we used a subsample of this data, for pupils' mathematics progress over 3 years from entry to junior school to the end of the third year in junior school. The subsample consists of 887 pupils from 48 schools, with five relevant variables, namely:

- *School ID*, an identification number assigned to each school, from 1 to 48,
- *Occupational Class*, a variable representing father's occupation, where 'Non Manual Occupation' = 1 and 'Other Occupation' = 0,
- *Gender*, a variable representing pupils' gender, where 'Boy' = 1 and 'Girl' = 0, and
- *Math 3* and *Math 5*, pupils' scores in maths tests in year 3 and in year 5 respectively, with a range from 0 to 40.

In the following sections we explaining the theoretical framework, from which we define the causal mechanisms that (likely) bring about differences in school performance. Then, we describe de model and their main components and dynamics. Finally, we present some simulation results.

A Mechanism-Based Explanation of Differential School Effectiveness

Although the literature in differential school effectiveness does not provide any canonical mechanism to explain the observed differences, a starting point is to recognise the importance of the social ties in which a pupil is embedded and its effect on her or his educational attainment. It can be assumed that the school structures the opportunities for closer friendship ties, which in turn affect peer outcomes. Since it has been demonstrated that pupils' characteristics within the classroom (such as socio-economic status and educational achievement) have an impact on the achievement of their peers (Beckerman & Good, 1981; Hanushek, Kain, Markman, & Rivkin, 2003), it is likely that the pupils' friendship networks are playing an important role as a *mediating factor*. Individuals' actions are often influenced by the people they interact with and, especially, by the actions of significant others, such as friends (Bearman, Moody, & Stovel, 2004; Kandel, 1978). There is no reason to rule out, *a priori*, the hypothesis that a similar mechanism might produce a *peer effect* in educational achievement (Jæger, 2007).

There is a growing interest in the literature in using network topologies and friendship ties instead of school-grade cohorts as the relevant peer group (e.g., Calvó-Armengol, Patacchini, & Zenou, 2009; Weinberg, 2007). Halliday and Kwak (2012) recently estimated peer effects using a mix of empirical data on friendship ties and school-grade cohorts. Their results suggest that the behaviour observed at the school-grade level is essentially a reduced-form approximation of a two-step process in which students first sort themselves into peer groups and then behave in a way that determines educational achievement. Although the exact sorting mechanism is not addressed in their research, it can be assumed that pupils choose their friends by a combination of contextual opportunities, geographical proximity (within the school or classroom) and *homophily* (McPherson, Smith-Lovin, & Cook, 2001), that is, the tendency of students to group themselves with other similar to them in gender, occupational class and educational performance.

Ever since the observational study carried out by Rist (2000) in the 1970s, educational researchers have been aware of the impact student-teacher relationships have on pupils' learning. Schools where teachers have higher expectations regarding the future of their students perform better than others where teachers have lower expectations (de Vos, 1995). These expectations, identified as 'teacher expectations bias' or 'false-positive teacher expectations' (Jussim & Harber, 2005), determine which pupils are defined as 'fast learners' and which ones as 'slow learners'. In this way, teachers' behaviour contributes to a 'self-fulfilling prophecy' (Merton, 1968), that is, pupils who are considered 'slow learners' in advance receive less

attention and educational feedback, and consequently, they perform worse compared to pupils who are considered 'fast learners'. Empirical research has corroborated this teacher expectation bias effect (de Boer, Bosker, & van der Werf, 2010; Rosenthal & Jacobson, 1968). Although the evidence is still inconclusive, particularly about the real magnitude of this effect and whether it accumulates or dissipates over time, it seems that powerful self-fulfilling prophecies do occur in the classroom, and they may selectively be directed towards students from stigmatised social groups and low-achieving students.

The proposed ABM presented here takes into account the two mechanisms described above as the model *microspecification*. That is, the model assumes that a combination of friendship dynamics based on homophily and self-fulfilling prophecy based on teacher expectations bias can produce differential achievement among students and schools. This explanatory mechanism can be established as follows. Firstly, there is a peer effect among pupils, which is brought about by the pupils' tendency to sort themselves in groups with similar others. This lateral group formation mechanism affects their individual learning and progress, producing groupings of pupils with different academic performances. Secondly, the differences among groups determine the way in which teachers interact with their pupils, since groups of high-performance pupils capture more attention and receive more feedback from teachers compared to groups of low-performance pupils. This vertical behavioural mechanism also affects the pupils' academic performance.

Model Description

The mechanisms described in the previous Subsection were implemented in an agent-based model. In order to build this model, we refined Wilensky's model (1997) to replicate the group formation mechanism. Students form groups with others similar to them following group formation rules present at the school level. For simplicity, we assume that these rules are stable and similar for all the individuals within the school (Akerlof & Kranton, 2002). We are not interested in giving an account of the emergence of these rules; we take for granted they exist. We also assume for simplicity that teachers always discriminate between 'slow learner' and 'fast learner' groups, and give the corresponding feedback to the students in those groups. The agent-based model was build in NetLogo (Wilensky, 2011), and the statistical analysis was performed in the statistical software R (2011).

The ABM was designed following two basic assumptions. The first concerns the way in which pupils' group themselves. To simulate how students sort themselves into peer groups, an initial number of spots where students can 'hang out' is defined. Every school has three tolerance criteria that are adopted by the students to decide whether to stay in a specific group or to move to another one. Taking into account the available data, we defined three tolerance levels: 1) *Educational tolerance*, which reflects the students' tolerance of accepting others with different attainments in *Math*

3; 2) *Gender tolerance*, which indicates the students' tolerance for people of the opposite sex; and 3) *Occupational class tolerance*, the pupils' tolerance for peers of a different occupational class. Each pupil is endowed with these three tolerance levels, which are identical for all the pupils within a school, but differ between schools. Tolerance levels range between 0% and 100%.

Pupils who belong to the same spot establish a group. If they are in a group that has, for example, a higher percentage of people of the opposite sex than the school's tolerance, then they are considered 'uncomfortable', and they leave that group for the next spot. Movement continues until everyone at the school is 'comfortable' with their group. This may result in some spots becoming empty.

Figure 2 shows an example of the result of this process: the student network at the end of a simulation for school 32. Male and female pupils are coloured dark grey and light grey respectively; round and square shaped nodes represent low and high occupational class respectively; and the numbers indicate their previous attainment in *Math 3*. In this scenario, education, gender and class tolerances are 0.9, 0.3 and 0.9 respectively. There are 39 students in school 32 and these sort themselves into 13 groups of 'friends'.

The second assumption concerns the way in which pupils' learning of one specific subject evolves over time. It seems reasonable to assume that this learning can be modelled as a logarithmic function of the educational feedback received on the subject. Thus, there is an initial period of rapid increase, followed by a period where the growth in learning slows (evidence supporting this pattern of learning may be found in Baloff, 1971).

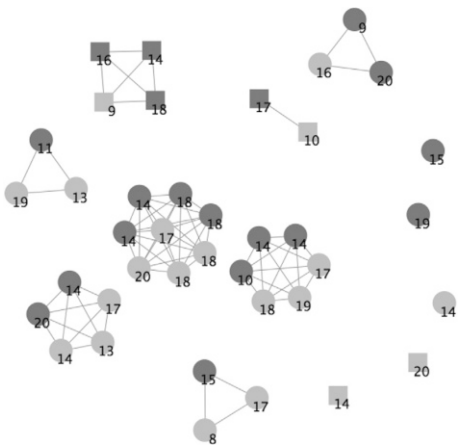


Figure 2. Simulated students social network in school # 32. Boys and girls are coloured dark grey and light grey respectively; round and square shaped nodes represent low and high occupational class respectively; and the numbers show the students' previous attainment in math 3.

In order to model pupils' learning in maths from year 3 to year 5, we define a *student learning curve*. Firstly, we assume that learning maths is a continuous process in which the student receives feedback on the subject from the teacher. This learning process starts at the first maths lesson, *lesson* 0, and finishes when the knowledge of maths is measured in year 5 (or *Math 5*). Because we do not have any measure of the educational feedback involved in this process, we arbitrarily define 1,000 as the amount of feedback that the entire learning process involves. Figure 3 shows this learning curve. Students' marks are assumed to be a function of the amount of teachers' feedback that they have received. We also assume that when the test *Math 3* is applied, students have learned half of the topics they were supposed to learn. Further, since both *Math 3* and *Math 5* range between 0 and 40, we transform *Math 3* by dividing it by 2.

Secondly, we assume that the feedback that students receive from their teachers depends on the sorting process that pupils perform within their schools. That is, in this simulation teachers' feedback is a function of the pupils' group average achievement. Teachers use this average group achievement as a signal about the future performance of all the pupils in the group. Teachers then adjust the amount of effort they invest in educational feedback accordingly.

Let g_k be a group in school j, s_{ik} is a student in the group k and $Math3_k$ the average of *Math 3* marks of group g_k. The amount of feedback that the students of group k receive is:

$$tk = (e^{2\times Math3_k})^{\frac{1}{\vartheta}}. \tag{2}$$

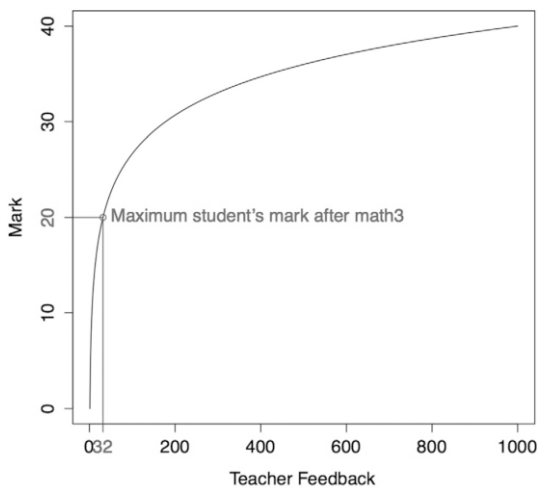


Figure 3. Simulated pupils' learning curve, which relates the teacher's feedback the students receive and the mark or score they obtain in their tests.

where ϑ in the exponent allows us to fit a logarithmic function that maps 'Teacher Feedback' into 'Mark' (see Figure 3). Since we want to fit a logarithmic function, we define $\vartheta \approx 5.790593$; for we know that $\log(1{,}000^{\vartheta}) \approx 40$ which is the scale of the test scores. Then, the simulated student's score $simMath5_{ik}$ is shown in Equation (3), where $t_{ik} = t_k + t_{math3,ik}$ and $t_{math3,ik}$ is the amount of feedback the pupils of group k have had when their attainment is measured as *Math 3*.

$$simMath5_{ik} = \log(t_{ik}^{\vartheta}) \tag{3}$$

Agent-Based Model Calibration

The ABM was initialised with the pupils' performance in *Math 3* and the parameter space given by the three tolerance levels was explored. The objective was to find a set of tolerance levels for each school that minimises the differences between the data and the simulation results. Let *dj* be such a difference for school *j*. Then, Equation (4) expresses this difference

$$d_j = \sum_{i=1}^{n_j} |Math5_{ij} - simMath5_{ij}|/2 \tag{4}$$

where $Math5_i$ and $simMath5_i$ are the score in *Math 5* of student *i* obtained from the data we described above and from the simulations, respectively. In the example shown in Figure 2, $d_{32} = 2.231$, which means that the simulated score in *Math 5* differs, on average, from the data by ±2.231 units. In order to explore the parameter space of the model, we ran 126,720 simulations. This represents all the possible combinations of the three tolerance levels (varying among 0.3, 0.5, 0.7 and 0.9) and the number of spots (varying among 15, 20 and 25) across the 22 schools. In order to have more robust results, we ran each setting 30 times and then took the average of d_j over all 22 schools as the aggregate outcome.

Model Results

Table 2 shows the results for the parameter setting that minimises d_j. We present the average distance (in the same units as the data) between the predicted scores and the real scores in *Math 5* for the simulation ('ABM (d_j)'). The table also shows the number of groups ('Final Groups') in which all the pupils were happy with their group membership, given the values in the 'Tolerance Levels' for education, gender and occupational class (the last three columns of Table 2). Recall that these three last variables were set as simulation parameters, and the specific values presented in the table correspond to those combinations at the school level that minimise the distance between the simulated and the data scores in *Math 5*.

Comparing the averages between the predicted and the observed scores, we see that the predictions errors of the ABM are not high; in fact, the distance averaged

Table 2. Calibration results. Comparison between the predicted pupils' scores by the agent-based model and the observed scores. Results are presented by schools

	Number		*Final*	*Tolerance Levels*		
School Id	*of pupils*	*ABM (d_j)*	*groups*	*Education*	*Gender*	*Occupational Class*
1	25	3.36	13	90%	50%	30%
4	24	3.12	12	90%	90%	50%
5	25	2.26	12	90%	70%	90%
8	26	2.82	12	90%	70%	30%
9	21	2.91	12	90%	70%	30%
11	22	3.10	12	90%	30%	70%
12	19	3.55	12	90%	50%	30%
20	28	2.62	12	90%	30%	70%
22	18	3.63	10	90%	30%	70%
23	21	3.19	12	90%	90%	50%
25	20	3.50	11	90%	30%	50%
26	19	2.79	12	90%	70%	50%
29	20	3.36	12	90%	70%	30%
30	35	2.56	14	70%	90%	70%
31	22	3.60	12	90%	70%	50%
32	39	2.71	15	90%	30%	90%
33	25	3.04	12	90%	30%	90%
35	27	2.44	13	90%	70%	30%
41	38	3.25	16	90%	30%	70%
45	30	2.62	12	90%	30%	70%
46	62	2.96	15	90%	90%	70%
47	22	3.61	12	90%	50%	90%

over all schools equals 3.04 on a scale of 40 points. Thus, the ABM, despite its simplicity, offers a reasonable fit to the data.

The simulation results suggest a high educational tolerance, since most of the values equal 90% (except from school 30, in which the tolerance level equals 70%). On the other hand, the tolerance levels of occupational class and gender vary across the schools. Therefore, the group formation mechanism in our simulation seems to be ruled by the variables occupational class and gender, while previous attainment in maths does not discriminate much between groups.

The hypothesised mechanism that bring about the differences in school effectiveness seems to be justified. The simulation results indicate that the mechanism

of group formation helps to minimise the distance between the predicted and the real scores, allowing a better fit with the data. For instance, when we compare the number of groups with the number of pupils, we can see that in general we have fewer groups than students in each school (for a graphical example, see Figure 2). If the numbers of groups made no difference in the simulation, then the number of groups and the number of pupils would tend to be similar (at least in those schools with 25 or fewer pupils, which is the maximum number of groups the ABM calibration allowed). This is clearly not the case. Thus, the sorting mechanism that has been implemented in this simulation and the groups reflecting that mechanism seem to be important in explaining the differences in effectiveness among schools.

CONCLUSIONS

In this chapter, we have explained the main concepts of ABM and the relation between this relatively new research method and scientific explanations in the social sciences. We have discussed the process of designing and building an agent-based model and we have recommended a set of standard steps to be used when building agent-based models for social science research. The first, and the most important, step in the modelling process is to identify the purpose of the model and the question(s) to be addressed. The importance of using existing theories to justify a model's assumptions and to validate its results has also been stressed. To explain these concepts, we have exemplified describing an agent-based model that addresses differential school effectiveness.

RECOMMENDED READINGS

There is an increasing introductory and more advanced literature on ABM, some of it well suited for social scientists. For those interested in going into this research method in greater depth, we recommend the following readings:

Epstein, J. M. (2007). *Generative social science: Studies in agent-based computational modeling*. New Jersey: Princeton University Press.

Gilbert, N. (2007). *Agent-based models*. California: Sage Publications Ltd.

Gilbert, N., & Troitzsch, K. G. (2005). *Simulation for the social scientist* (2nd ed.). Glasgow: Open University Press.

Railsback, S. F., & Grimm, V. (2011). *Agent-based and individual-based modeling: A practical introduction*. New Jersey: Princeton University Press.

REFERENCES

Abdou, M., Hamill, L., & Gilbert, N. (2012). Designing and building an agent-based model. In A. J. Heppenstall, A. T. Crooks, L. M. See, & M. Batty (Eds.), *Agent-based models of geographical systems* (pp. 141–165). Dordrecht: Springer Netherlands.

Akerlof, G. A., & Kranton, R. E. (2002). Identity and schooling: Some Lessons for the economics of education. *Journal of Economic Literature*, *40*(4), 1167–1201.

Axelrod, R. (1997). Advancing the art of simulation in the social sciences. *Complexity*, *3*(2), 16–22.

Axtell, R., Axelrod, R., Epstein, J. M., & Cohen, M. D. (1996). Aligning simulation models: A case study and results. *Computational & Mathematical Organization Theory*, *1*(2), 123–141.

Baloff, N. (1971). Extension of the learning curve – some empirical results. *Operational Research Quarterly (1970–1977)*, *22*(4), 329–340.

Bearman, P. S., Moody, J., & Stovel, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *The American Journal of Sociology*, *110*(1), 44–91.

Beckerman, T. M., & Good, T. L. (1981). The classroom ratio of high- and low-aptitude students and its effect on achievement. *American Educational Research Journal*, *18*(3), 317–327.

Bianchi, C., Cirillo, P., Gallegati, M., & Vagliasindi, P. (2007). Validating and calibrating agent-based models: A case study. *Computational Economics*, *30*(3), 245–264.

Calvó-Armengol, A., Patacchini, E., & Zenou, Y. (2009). Peer effects and social networks in education. *Review of Economic Studies*, *76*(4), 1239–1267.

Castellani, B., & Hafferty, F. W. (2009). *Sociology and complexity science*. Berlin, Heidelberg: Springer.

Coleman, J. S. (1966). *Equality of educational opportunity study (EEOS)*. Washington: National Center for Educational Statistics.

de Boer, H., Bosker, R. J., & van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology*, *102*(1), 168–179.

de Vos, H. (1995). Using simulation to study school effectiveness. Presented at the *The Annual Meeting of the European Council on Educational Research*, Bath, England.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, *4*(5), 41–60.

Epstein, J. M. (2007). *Generative social science: Studies in agent-based computational modeling*. New Jersey: Princeton University Press.

Epstein, J. M., & Axtell, R. L. (1995). *Growing artificial societies: Social science from the bottom up*. Washington, D.C.: Brookings Institution, U.S.

Fagiolo, G., Moneta, A., & Windrum, P. (2007). A critical guide to empirical validation of agent-based models in economics: Methodologies, procedures, and open problems. *Computational Economics*, *30*(3), 195–226.

Gilbert, N. (2007). *Agent-based models*. California: Sage Publications Ltd.

Gilbert, N., & Troitzsch, K. G. (2005). *Simulation for the social scientist* (2nd ed.). Glasgow: Open University Press.

Halliday, T. J., & Kwak, S. (2012). What is a peer? The role of network definitions in estimation of endogenous peer effects. *Applied Economics*, *44*, 289–302.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, *18*(5), 527–544.

Hedström, P. (2005). *Dissecting the social: On the principles of analytical sociology*. Cambridge: Cambridge University Press.

Hedström, P., & Swedberg, R. (1996). Social mechanisms. *Acta Sociologica*, *39*(3), 281–308.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology*, *36*(1), 49–67.

Jæger, M. M. (2007). Economic and social returns to educational choices. *Rationality and Society*, *19*(4), 451–483.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, *9*(2), 131–155.

Kandel, D. B. (1978). Homophily, selection, and socialization in adolescent friendships. *The American Journal of Sociology*, *84*(2), 427–436.

Kleindorfer, G. B., O'Neill, L., & Ganeshan, R. (1998). Validation in simulation: Various positions in the philosophy of science. *Management Science*, *44*(8), 1087–1099.

Macy, M. W., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, *28*(1), 143–166.

Manson, S. M., Sun, S., & Bonsal, D. (2012). Agent-based modeling and complexity. In A. J. Heppenstall, A. T. Crooks, L. M. See, & M. Batty (Eds.), *Agent-based models of geographical systems* (pp. 125–139). Dordrecht: Springer Netherlands.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444.

Merton, R. K. (1968). *Social theory and social structure*. New York: Free Press.

Miller, J. H., & Page, S. E. (2007). *Complex adaptive systems: An Introduction to computational models of social life*. Princeton University Press.

Nuttall, D. L., Goldstein, H., Prosser, R., & Rasbash, J. (1989). Differential school effectiveness. *International Journal of Educational Research*, *13*(7), 769–776.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from http://www.R-project.org.

Rist, R. (2000). Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. *Harvard Educational Review*, *70*(3), 257–302.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development* (First Printing.). New York: Holt, Rinehart & Winston.

Sammons, P., Nuttall, D., & Cuttance, P. (1993). Differential school effectiveness: Results from a reanalysis of the inner London education authority's junior school project data. *British Educational Research Journal*, *19*(4), 381–405.

Weinberg, B. A. (2007). Social interactions with endogenous associations. *National Bureau of Economic Research Working Paper Series*, *No. 13038*.

Wilensky, U. (1997). *NetLogo party model*. Evanston, IL.: Center for connected learning and computer-based modeling, Northwestern University. Retrieved from http://ccl.northwestern.edu/netlogo/models/Party.

Wilensky, U. (2011). NetLogo. *Center for connected learning and computer-based modeling. Northwestern University, Evanston, IL.* Retrieved from http://ccl.northwestern.edu/netlogo.

JAMES HALL & PAMELA SAMMONS

13. MEDIATION, MODERATION & INTERACTION

Definitions, Discrimination & (Some) Means of Testing

INTRODUCTION

In 1986 Baron and Kenny set out to clarify the terms "Mediation" and "Moderation" as used in the social sciences (with the origins of each described by Roe, 2012). Twenty six years later, the seminal paper that this collaboration resulted in (Baron & Kenny, 1986) has been cited over 35,000 times (35,672 via Google Scholar as of 09/01/2013). However, despite this extensive record of citation, uncertainty continues to surround the use of these terms in social science research and they have received relatively little attention in specifically educational research (cf. Kraemer, Stice, Kazdin, Offord, & Kupfer, 2001). Partly in response to this uncertainty, and partly in response to advances made in the application of more complex statistical analyses in educational research (e.g. Creemers, Kyriakides, & Sammons, 2010; Goldstein, 2003; Luyten & Sammons, 2012; Tatsuoka, 1973), this chapter is made-up of four sections which together provide the quantitative educational researcher with an up to date understanding of these terms as well as examples of their current implementation to test theoretical models and address notions of causality. These four sections are titled:

1. Unambiguous Definitions
2. Discriminating Mediation, Moderation, and Interaction
3. Some means of testing Mediation and Moderation
4. Testing Moderation: An example through three equivalent statistical
5. analyses

Together, the first two sections of this chapter present simple, clear definitions that distinguish "Mediation", "Moderation", and "Interaction" both from each other as well as from a number of other commonly-used terms. Section 3 then presents a number of statistical methods by which these terms can be statistically operationalised. This third section pays particular attention to Moderation as the statistical methods associated with it (in comparison to Mediation) are particularly varied and numerous. The final section of this paper (Section 4) then builds upon the focus on Moderation within Section 3 by presenting an example Moderation from educational research conducted within the early years (for children under age

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 267–286.

5 years) which is then statistically operationalised and tested by three equivalent parallel analyses.

UNAMBIGUOUS DEFINITIONS

Mediation

This is a trivariate one-tailed hypothesis concerning mechanisms of effect. A pre-established causal relationship between two variables is theorised to exist due to an intermediate third variable (see Figure 1). While the additional (third) variable that is hypothesised to have this effect is known as a "*mediator*" it is also sometimes referred to as an "*intermediate variable*" (Kraemer *et al*., 2001), or "*explanatory link*" (Rose, Holmbeck, Coakley, & Franks, 2004). Further, mediators have "*mediating effects*" which are otherwise labelled "*indirect effects*", "*surrogate effects*", "*intermediate effects*" and/or "*intervening effects*" (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; Wu & Zumbo, 2007).

Moderation

This is also a trivariate one-tailed hypothesis – but one that is quite different from mediation and with a completely separate historical origin (Roe, 2012). Hypotheses of moderation ask, "Under what conditions/for whom/when is a pre-established causal relationship observable?" (cf. mechanisms in mediation). The presence of a third "*moderator*" variable is also termed an *"effect-modifier"* (Hinshaw, 2002) and/or a "*causal interaction effect*" (Wu & Zumbo, 2007).

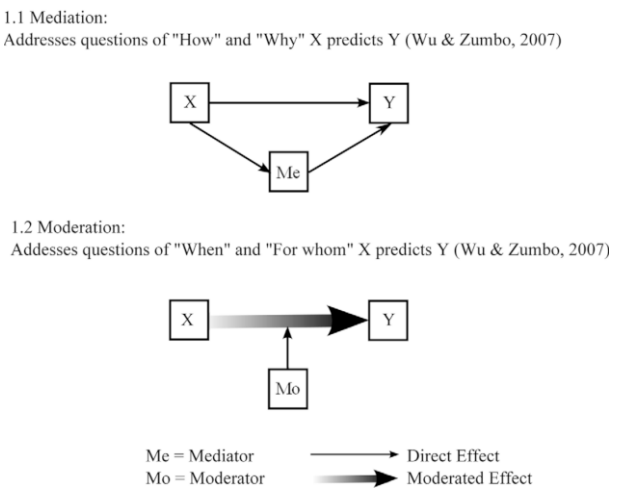


Figure 1. Graphical illustration of the hypotheses of mediation and moderation.

This last alternative name is also important as it evidences the close association (and therefore also sometimes confusion) between the terms "Moderation" and "Statistical Interaction" (an explanation for the alternative name of "casual interaction effect" is given in Section 3). The origins of this association go back to the first reported use of "Moderation" – commonly cited as Saunders (1955) – in which this term was adopted as a synonym for what quantitative researchers now refer to as a "*(Statistical) Interaction Effect*" (again, see Section 3). This change in meaning over the past 58 years and the close relationship that today's definition of "Moderation" has to "(Statistical) Interaction" (see below) is just one reason why confusion continues with the use of these terms.

(Statistical) Interaction

This is a two-tailed hypothesis implying that two or more concepts, "*work together*" or, "*have a combined effect*" in eliciting a third (for example in: Kraemer *et al*., 2001; Talamini *et al*., 2002) which should in no way be mistakenly confused with the concept of behavioural or psychological or gene-environment interactions (e.g. Rutter & Silberg, 2002). One of the common points of difficulty (explored further in Section 3) is that Moderation is a hypothesis that is often answered by the specification of a 'Statistical Interaction' which is then, in-turn, commonly tested with statistical artefacts known as *"Statistical Interaction Effects/Terms"*.

DISCRIMINATING MEDIATION, MODERATION, AND INTERACTION

Mediation ≠ Moderation

Although Mediation and Moderation are distinct trivariate research hypotheses, confusion continues not only over their distinction, but also over which is the more appropriate for any given research project as well as how these hypotheses can be combined. The first of these difficulties (distinguishing Mediation and Moderation) continues partly due to the simple similarity of the two words, partly due to their changing definitions over time, and partly due to the similar purposes for which both are used in research. Considering this third point in more detail: Mediation and Moderation are both, *"theories for refining and understanding a causal relationship"* (Wu & Zumbo, 2007) and both are unidirectional (i.e. "A and B affect C" rather than "there is an association between A, B, and C") trivariate hypotheses. The problems that arise over the application of these distinct hypotheses is also evident in the continuing confusion concerning the term "*Indirect Effect*" which although having a specific meaning encompassing Mediation (see Preacher & Hayes, 2004) also has an additional and more intuitive meaning: "*any and all effects other than those direct*". This additional understanding of the term "Indirect effect" has led to its use in reference to Statistical Interaction and thereby also Moderation. The paper by Goodnight, Bates, Staples, Petit, and Dodge (2007) provides an example of this

more intuitive usage of the term, although for clarity we recommend such usage should be avoided,

> …However, in addition to direct main-effects-type links between temperament and behavior problems, there are also more indirect, interaction-effect-type links involving temperament...

With this background of confusion over the meaning and usage, the terms "Mediation" and "Moderation" have continued to be discussed long after the paper of Baron and Kenny (1986). Table 1 provides an overview of a selection of five journal articles in different fields since the turn of the millennium that have all aimed to provide clarifying guidelines. The problems researchers continue to encounter with these terms is evidenced in the inconsistent guidelines across these papers.

Table 1. A selection of past guidelines (since 2000) for distinguishing mediation from moderation

Authors:				
(Kraemer, *et al*., 2001)	(Hinshaw, 2002)	(Nicholson, Hursey, & Nash, 2005)	(Essex, *et al*., 2006)	Wu & Zumbo (2007)
Journal:				
American Journal of Psychiatry	*Development and Psychopathology*		*Archives of General Psychiatry*	*Social Indicators Research*
Mediation:				
		That being mediated has temporal precedence		
		Mediator and that mediated are correlated		
Either co-domination of mediated and mediator (partial) OR			Either co-domination of mediated and mediator (partial) OR	Answers, "how" and "why"
Mediator dominates that mediated (total)			*Mediator dominates that mediated (total)*	*Mediator is a state*
				Mediator is observed or manipulated
Moderation:				
		Moderator has temporal precedence		
		Moderator and that moderated are uncorrelated		
Co-domination of moderated and moderator				Answers, "for whom" and "when"
				Moderator is a trait
				Moderator is observed

A casual examination of Table 1 also reveals that none of the articles originate from the field of Educational Research and, to the best of our knowledge prior to this chapter; Educational Researchers have never had tailor-made guidance written for them on the issues that surround Mediation and Moderation.

Of the guidelines distinguishing Mediation from Moderation that are presented in Table 1, only two are consistent across all the articles:

1. The varying importance of temporal precedence (e.g. Cole & Maxwell, 2003)
2. The varying importance of which measures should/should not be significantly correlated with one another.

The first of these is particularly paramount given that both Mediation and Moderation are viewed as causal unidirectional hypotheses of effect. As a result, they require appropriate quantitative data to be tested: that which is appropriate for testing *any* unidirectional hypotheses. This is a condition of gathered quantitative data that is most commonly resolved by collecting data with a temporal element (i.e. data that is *longitudinal* in the case of correlational/survey research or *repeated-measures* in the case of experimental designs). Having data with the correct clear temporal precedence (establishing 'causal priority'; Preacher & Hayes, 2004) is perhaps the most important precondition that researchers can and should establish for both Mediation and Moderation.

Real-world ambiguities. Unfortunately, even when educational researchers hold clear unambiguous definitions of Mediation and Moderation there still remain real-world occasions in which the appropriateness of one over the other is ambiguous. Within developmental science (a catch-all label that includes much quantitative educational research), this can often be attributed to the time-frame under consideration. For example, it is often possible for the same set of measures to be related first as a mechanism (mediation) but then later as a conditional effect (moderation). The paper by Masten (2007) provides an example of this. The relationship between background adversity, an individual's stress-regulators, and their subsequent stress-response begins with stress regulators being shaped by adversity as they develop. However, once stress-regulators are developed, their relationship with adversity changes: stress-regulators are now deemed to operate by altering the stress-response to adversity. Thus in the first period, a *mechanism* (mediation) is at work while in the second, a *condition*al effect (moderation) comes into evidence (see Figure 2).

Another common difficulty that researchers can face when determining whether it is more appropriate to specify a hypothesis of Mediation or Moderation is that these hypotheses can also be combined. Again Masten (2001) provides an example - this time of, *"...a risk-activated moderator analogous to an automobile airbag or immune system response"*. Figure 3 illustrates this effect with generalised labels. For the educational researcher in particular, this is also an excellent description of the ideal functioning of social interventions such as Head Start in the USA (see Currie

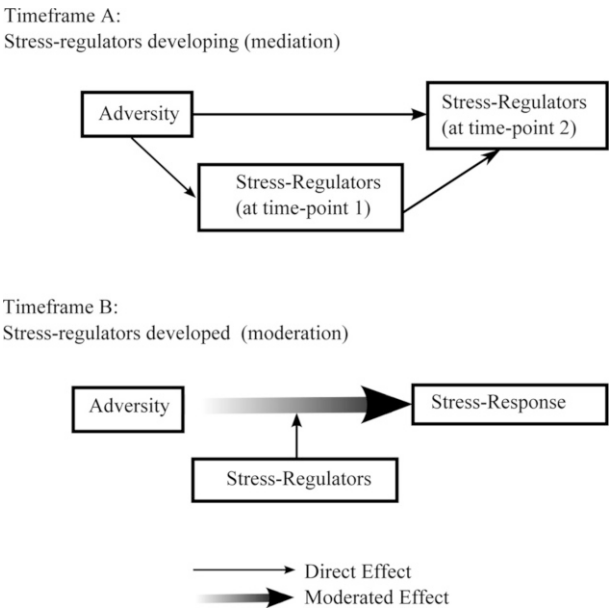


Figure 2. The plausibly of mediation and moderation as appropriate hypotheses varying by the time-frame under investigation (adapted from masten, 2007).

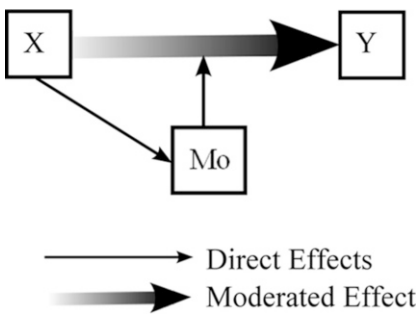


Figure 3. An example moderation that contravenes common guidelines but which has real-world application (adapted from the "risk-activated moderator" of Masten, 2001).

and Thomas, 1995) and Sure Start in the UK (see Glass, 1999). At the same time, although this example has obvious application in the real-world it also contravenes the only two consistent guidelines about when to hypothesise Moderation that are shown in Table 1.

The "*risk-activated moderator*" of Masten (2001; Figure 3) is just one example of how Mediation and Moderation may be integrated as hypotheses. Two more

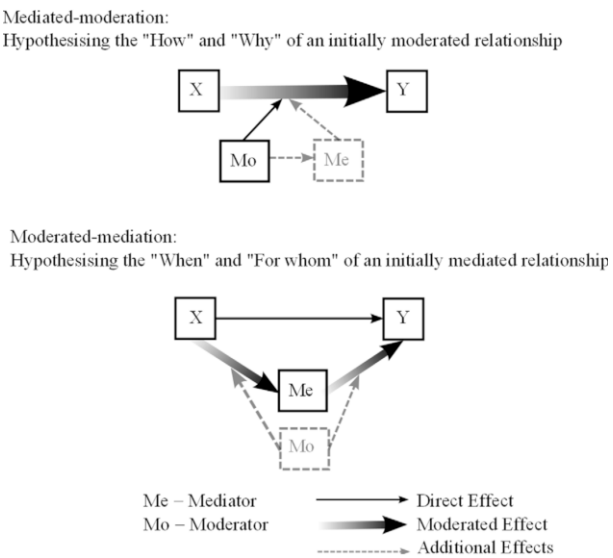


Figure 4. An illustration of the hypotheses of mediated-moderation and moderated-mediation (adapted from Wu & Zumbo, 2007).

examples of these hypotheses in combination are *"Moderated-mediation"* and *"Mediated-moderation*" (see Figure 4). Both hypotheses concern causal relationships hypothesised between at least four measures [W, X, Y, Z] and both postulate *conditional* (moderated) *mechanisms* of effect (mediation) whereby X affects Y. Further in-depth description and discussion of these combinations of Mediation and Moderation can be found in Muller, Judd, and Yzerbyt (2005), Wu and Zumbo (2007), Preacher, Rucker, and Hayes (2007), and Edwards and Lambert (2007). These papers also outline the various approaches for the statistical testing of these hypotheses.

Moderation ≠ (Statistical) Interaction

Just as hypotheses of Moderation are often confused with Mediation so too is Moderation often confused (or viewed as synonymous with) Statistical Interaction. This is a problem that is at-least partly due to the overlap between the two concepts, one where Moderation can be viewed as a more restricted version of Statistical Interaction as evidenced by the alternative name for Moderation given by Wu and Zumbo (2007): "*Causal Interaction*". The relationship between the concepts of Statistical Interaction and Moderation can be understood as the difference between a two-tailed hypothesis and a more restrictive one-tailed hypothesis. Thus, although the statistical methods that are used to test hypotheses of (Statistical) Interaction can also be applied to hypotheses of Moderation, to conclude Moderation from

these methods necessitates relying heavily upon pre-existing knowledge, be this from past research findings, broader substantive theories, or from other top-down sources of knowledge (e.g. Nicholson, Hursey, & Nash, 2005). When there is a lack of sufficient pre-existing knowledge to justify putting forward a hypothesis of (one-tailed) Moderation at the beginning of a research project, subsequent evidence of a (two-tailed) Statistical Interaction should not over-interpreted as inferring Moderation (as in Rutter & Silberg, 2002; Kraemer *et al*., 2001) . For example, finding evidence that educational outcomes are significantly related to the interaction of parental background and various educational factors that a child experiences (thus: attainment=background + ed.factor + background *x* ed.factor) should not be over-interpreted to conclude that education can alter the effects of parental background unless there is additional top-down information to warrant this (e.g. Burchinal, Peisner-Feinberg, Bryant, & Clifford, 2000; Hall, *et al*., 2009; NICHD, 2000).

(Statistical) Interaction ≠ Statistical Interaction Terms

While Statistical Interaction is a two-tailed hypothesis that two or more concepts "*work together*"/"*have a combined effect*" upon a third, Statistical Interaction Terms are two-tailed statistical artefacts (defined as the product of two variables) and are often specified to test these hypotheses – commonly within regression- (of bivariate form: $Y = b_0 + b_1X + b_2Z + B_3X.Z + e$) and ANOVA-based statistical analyses.

The relationship between Moderation, Statistical Interaction, and Statistical Interaction Terms takes the following form: *Moderation* is a more restricted one-tailed alternative to the two-tailed hypotheses of *Statistical Interaction* although both are often tested through the specification of *Statistical Interaction Terms.*

SOME MEANS OF TESTING MEDIATION AND MODERATION (THEORY IN PRACTICE)

Mediation

Although MacKinnon and colleagues (2002) discuss fourteen statistical methods to test hypotheses of Mediation, here we note only the four main methods and direct readers to Hayes (2009) for a fuller while also contemporary discussion of Mediation as well as the methods that are available for its testing.

1. The 'C*ausal Steps Approach'* of Baron and Kenny (1986). This is a technique that has been strongly criticized as having the least statistical power to accurately detect Mediation Effects (Fritz & MacKinnon, 2007; Hayes, 2009)
2. The *Sobel Test* (Sobel, 1982) is a more formal test of mediation compared to the Causal Steps Approach. Multiple regression analyses are conducted and the results of each are combined (see Preacher & Hayes, 2004). Various macros and online calculators are available for this additional step (for example from: http://www.danielsoper.com)

3. *Bootstrapping*. One of the problems with the Sobel Test is that it assumes normality in the distribution of variables which limits its appropriate application. One alternative that does not make this assumption is to conduct statistical boostrapping to estimate Mediation Effects. Not only is this non-parametric technique applicable with non-normally distributed variables, it is also retains its reliability with lower sample sizes compared to the Sobel Test (for more detail see Preacher & Hayes, 2004).
4. Statistical *Path Analysis* (often within "Structural Equation Modelling", SEM) commonly incorporates the above *Bootstrapping* approach within a broader statistical modelling framework that Reynolds and Ou (2003) note as an especially suitable technique for*, "theory driven tests of hypotheses of causal mediation"* (p.451, Reynolds & Ou, 2003). A good overview is provided by Tatsuoka (1973) who documents both the historical origins of path analysis and provides an account of its initial take-up by educational researchers.

Moderation

Compared to the methods available for testing a hypothesis of Mediation, the options available to researchers interested in Moderation are both more numerous and more complex (with this at least partly attributable to the changing definition of Moderation over time and partly due to its conceptual relationship with Statistical Interaction). Back in 1986 Baron and Kenny discussed the statistical methods suitable for testing hypotheses of moderation (statistical interaction terms within either regressions or ANOVAs) and presented detailed guidelines for deciding the appropriateness of one over the other. For this, Baron and Kenny emphasised that the suitability of a method depended upon the level of measurement by which each of the three concepts featured in the Moderation were measured, be this continuous or categorical. Fifteen years later however and Kraemer and colleagues (2001) noted a "*struggle*" existed between two statistical approaches used for testing hypotheses of moderation: 1) Sub-group comparisons (that commonly dichotomise continuous moderator variables), 2) Statistical Interaction Terms. For the educational researcher in particular however, there is at least one additional statistical technique not covered by either Baron and Kenny (1986) or Kraemer and colleagues (2001) and which also directly tests hypotheses of Moderation, has nothing to do with Statistical Interaction Terms (in ANOVAs or regressions), and for which quantitative educational data is frequently suitable: *Random Slope Effects*. These are typically examined in multilevel models that explore the hierarchical structure of nested data in social or educational contexts where students are clustered in classes, themselves clustered in schools, in turn clustered in neighborhoods etc (see Goldstein, 2003; Luyten & Sammons, 2010). Here examples of hypotheses that may be tested include that the shape of relationships between prior attainment and later attainment (the slope) may differ between higher level units (for example classes or schools) and also for different groups of students within different schools (for example by SES or gender).

An in-depth discussion of the three methods (sub-group comparisons, statistical interaction terms, random slope effects) that are particularly suitable for educational researchers interested in testing hypotheses of moderation follows below. First however, it is worth mentioning that these options may be grouped in two different ways: 1) Explicit versus Implicit Tests, and 2) Variable-based versus Personbased. The Explicit/Implicit distinction refers to whether a technique is, or by contrast is not, a literal and direct test of the trivariable causal hypothesis of Moderation that is illustrated in Figure 1 (random slope effects) or whether a conclusion of Moderation is instead only inferred from a Statistical Interaction (sub-group comparisons, statistical interaction terms). The Variable/Person-based distinction refers to whether a method emphasises a pattern of statistical relationship between *variables* (as in random slope effects, statistical interaction terms) or statistical differences between *units of analysis* (commonly people; as in sub-group comparisons).

1. *Sub-group comparisons* (indirect person-based test of moderation). Here, evidence of moderation is obtained by establishing that a bivariate relationship is significantly different between two or more groupings of the unit of analysis (be these people, schools etc). If the moderator variable was originally measured with a continuous variable, then an intermediate, though strongly criticised (Frazier, Barron, & Tix, 2004; MacCallum, Zhang, Preacher, & Rucker, 2002; McClelland & Judd, 1993), step is necessary: sub-group creation through dichotomisation/ categorisation.
2. *Statistical Interaction Terms* (indirect variable-based test of moderation). This is a multi-stage procedure that only actually tests the existence of a combined working-together of two or more variables as they jointly impact another. It is then up to the researcher to interpret whether this also constitutes evidence of moderation (see Wu and Zumbo, 2007). As an act of inductive reasoning, this additional step should be informed by broader conceptual understanding such as the findings from previous research. This means of testing.
3. explains the alternative name for Moderation as a, "*Causal Interaction Effect*": A non-causal bi-directional relationship is established (the Interaction Effect) before post-hoc reasoning is undertaken to establish a causal relationship from this. The procedure for testing a Statistical Interaction Term in an OLS regression is as follows:
   a. Mean-centre your predictor [X] and moderator [Z] variables.
   b. Construct a new 'interaction' variable of the form: predictor x moderator [XZ]
   c. Use this variable as a predictor of your outcome [Y] along with the original variables [X, Z] in a regression equation of the form: $Y = X + Z + XZ$ [$+e$]
   d. Interpret only the *un*standardised regression co-efficient from the [XZ] statistical interaction term.
   e. Plot any significant Statistical Interaction Term to aide interpretation.
4. *Random Slope Effects* (direct variable-based test of moderation). Unlike Sub-group Comparisons and the use of Statistical Interaction Terms, Random Slope

Effects test a hypothesis of Moderation directly, not via the intermediate step of first establishing a Statistical Interaction. Random Slope Effects refer to when a statistical regression relationship (the Slope, [*s*]) between two variables [*X*, *Y*] is allowed to vary as a function of a third [*Z*]. Unfortunately, this most direct means of testing a hypothesis of Moderation is also the most restricted in terms of the requirements it imposes on quantitative data. Random Slope Effects require clustered or nested data (as noted above) and therefore multilevel (hierarchical linear) statistical modelling techniques. On top of this, for a Random Slope Effect to test a hypothesis of Moderation a very specific set of relationships must to be specified between variables: The Moderating variable [*Z*] must be at the *between* level (level 2) while the Moderated relationship [*Y* on *X*] must be at the *within* level (level 1). Fortunately, the requirement for nested quantitative data is one that educational research frequently meets due to the nested nature of educational systems (e.g. children within classes within schools within districts/ neighbourhoods).

TESTING MODERATION: AN EXAMPLE THROUGH THREE EQUIVALENT STATISTICAL ANALYSES

The final section of this paper presents a worked example of some of the main issues so far discussed. An educational research question is expressed as a hypothesis of Moderation and this is then tested with the three statistical approaches discussed above in Section 3: A Sub-group Comparison, specification and testing of a Statistical Interaction Term, and the testing of a Random Slope Effect.

Theoretical Background

A mother's age at the birth of her child is known to significantly impact this child's cognitive development: *Children of younger mothers are likely to demonstrate poorer cognitive development* (Borkowski, *et al*., 1992; Fergusson & Lynskey, 1993). However, higher 'quality' (Currie, 2001) preschool has been found to partial 'protect' (Rose, *et al*., 2004) children from such adverse outcomes (Burchinal, Peisner-Feinberg, Bryant, & Clifford, 2000; Hall, *et al*., 2009; NICHD, 2000). This set of relationships can be expressed as a hypothesis of Moderation: Attendance at a preschool of higher quality may *moderate* the relationship between a mother's age (at child-birth) and her child's subsequent cognitive development.

Method

To test the hypothesis of Moderation suggested above, a secondary analysis of the Effective Preschool, Primary, and Secondary Education (EPPSE; Sylva, Melhuish, Sammons, Siraj-Blatchford, & Taggart, 2010, 2012) dataset was undertaken. Adopting a longitudinal research design, EPPSE was the first large scale British

research project to examine the quality and effectiveness of various programmes of pre-, primary, and secondary schools as predictors of the development and educational attainment of over 3,000 British children from 3 years of age to adulthood.

Participants. 2857 participating families with children in attendance at n=141 preschools (for at least 10 weeks already) were recruited when these children were of mean age 36 months. This recruitment of families from preschools ensured that a sufficient level of nesting of data was achieved (families within preschools) such that preschool effects on child outcomes could be reliably estimated (e.g. Goldstein, 1987, 2003).

Measures. For this example, the outcome measure [Y] was each child's General Cognitive Ability (GCA) as measured by the British Ability Scales (Elliot, NFER-NELSON, Smith, & McCulloch, 1996) and as assessed at mean child age 58 months (n=2574; mean=96.73; Standard Deviation, SD=14.51). The predictors of GCA at mean child age 58 months were:

- GCA at 36 months (n = 2764; mean = 91.36; SD = 13.90)
- Mother-age at child-birth [X] assessed at parental interview at enrolment (n = 2779) with a six category ordinal scale (1 = *16–20*, 2 = *21–25*, 3 = *26–35*, 4 = *36–45*, 5 = *46–55*, 6 = *56–65*) that is here treated as continuous for solely pedagogical purposes (thus: mean = 3.16; SD = 0.66)
- The hypothesised moderator [*Z*]: The overall 'quality' of the preschool that each child attended as measured by the Early Childhood Environmental Rating Scale (ECERS-R; Harms, Clifford, & Cryer, 1998; preschool n = 141; child n = 2857; child mean = 4.47; SD = 1.00)

With reference to the guidelines of Table 1, it should be noted that the hypothesised moderator, preschool quality, was uncorrelated with the variable whose effect quality was hypothesised to moderate (mother's age). For more details on these measurements see Sylva and colleagues (2010).

Analytic techniques. Each of the three statistical techniques for testing a hypothesis of Moderation that were discussed in Section 3 (Sub-group Comparisons, Statistical Interaction Terms, Random Slope Effects) were conducted within the statistical framework of Multilevel Structural Equation Modelling (SEM) using Version 6 of the Mplus Software (Muthén & Muthén, 2010). Version 6 of the Mplus Software estimated missing data using maximum likelihood procedures as an integral part of all three analyses (Muthén & Muthén, *ibid*). As the following results serve only as an example of the methods discussed above, we do *not* report the results in full as we would in a purely substantive piece of work as a detailed substantive interpretation is not the aim (full results are of course available from the authors).

Results

Sub-group Comparisons. Comparisons between sub-groups based on the quality of preschools were made possible through the specification of a "multi-level mixture model". Given that preschool quality was originally measured on a continuous scale, the specification of sub-groups necessitated an initial step of dichotomisation. A mean ± 1 standard deviation dichotomisation strategy was used to form two groups of $n = 623$ and $n = 461$ children who had attended $n = 55$ 'low' and 'high' quality preschools respectively (the remaining and excluded $n = 1773$ children attended $n = 86$ preschools where quality was within the mean ± 1SD range). The following effects of mother-age on GCA at 58 months were found:

- In the 'Low' preschool quality group: Older mothers had children who demonstrated significantly higher GCA (standardised regression coefficient, $\beta = 0.10$, $p = 0.001$)
- In the 'High' preschool quality group: There was no significant relationship between mother age and child GCA ($\beta = -0.01$, $p = 0.85$)
- Further, the relationship between mother age and child GCA was significantly higher in the 'Low' quality preschool group than it was in the 'High' quality group ($\beta = 0.10$ vs. $\beta = -0.01$; $t_{1080} = 2.66$, $p < 0.01$)

In conclusion, a differential impact of mother's age upon GCA at 58 months was found in low versus high quality preschools. For children attending 'high' quality preschools, the children of younger mothers had (on average) indistinguishable levels of GCA compared to children of older mothers: This was not so for children attending 'low' quality preschools.

```
MPLUS SUBGROUP COMPARISON (VIA DICHOTOMISATION) SYNTAX:
MISSING ARE ALL (-999999);
idvariable = childid;
CLUSTER = centreid;
WITHIN = bgcam q53am;
CENTERING = GRANDMEAN (bgcam, q53am);
CLASSES = group (2);
KNOWNCLASS = group (group1=0 group1=1);
DEFINE:
IF (ecers_r LE 3.4690) THEN group1=0;
IF (ecers_r GE 5.4691) THEN group1=1;
ANALYSIS:
TYPE = MIXTURE TWOLEVEL;
ALGORITHM = INTEGRATION;
```

```
MODEL:
%WITHIN%
%OVERALL%
rgcam on bgcam;
rgcam on q53am;
%group#1%
rgcam on bgcam;
rgcam on q53am;
%group#2%
rgcam on bgcam;
rgcam on q53am;
```

Statistical interactionterms. A "multi-level path model" was specified in which the Statistical Interaction Term introduced in Section 3 ($Y = X + Z + XZ$) was specified at the preschool (between) level. The following effects of mother-age on GCA at 58 months were found (bearing in mind that standardised results were unavailable for this model):

- There was a positive and statistically significant effect of mother-age on GCA at 58 months (unstandardised regression coefficient, $b = 4.92$, $p < 0.001$)
- There was also a positive effect of preschool quality on GCA at 58 months although this did not quite reach the 95% confidence level ($b = 1.780$, $p = 0.064$)
- There was also a statistically significant negative effect from the Statistical Interaction Term *mother-age x preschool quality* ($b = –0.685$, $p = 0.020$)

In conclusion: mother's age had a decreasing effect on children's GCA at 58 months as the preschools that these children attended increased in quality. The children of younger mothers had (on average) lower GCA at 58 months but this was less apparent when these children had attended higher quality preschools. In other words, this analysis and its results lead to the same substantive conclusion as that returned from the Sub-group Comparisons.

```
MPLUS INTERACTION TERM SYNTAX:
MISSING ARE ALL (-999999);
idvariable = childid;
```

```
CLUSTER = centreid;
BETWEEN = ecers_r;
WITHIN = q53am bgcam qualage;
CENTERING = GRANDMEAN (bgcam, q53am, qualage);
DEFINE: qualage = ecers_r*q53am;
ANALYSIS:
TYPE = RANDOM TWOLEVEL;
ALGORITHM = INTEGRATION;
MODEL:
%WITHIN%
rgcam on bgcam q53am qualage;
%BETWEEN%
rgcam on ecers_r;
```

Random Slope Effects. Once again, a "multilevel path model" was specified, but this time also featuring "random effects". In this analysis only one random effect was estimated: the statistical regression slope (s) between mother's age and child GCA at 58 months was allowed to vary between children and this variation was specified to depend upon the quality of the preschools that children attended. As with the estimation of the Statistical Interaction Term, the specification of a Random Slope Effect meant standardised results were again unavailable. The following effects of mother-age on GCA at 58 months were found:

- There was a positive and statistically significant effect of mother-age on GCA at 58 months (unstandardised regression coefficient, $b = 4.54$, $p = 0.001$)
- There was no effect of preschool quality on GCA at 58 months ($b = -0.44$, $p = 0.208$)
- The significant relationship between mother-age on child GCA at 58 months was significantly diminished by increasing preschool quality ($b = -0.63$, $p = 0.031$)

In conclusion, mother's age had a smaller effect on child GCA at 58 months when these children were in attendance at higher quality preschools. Once again, this conclusion is essentially the same as that drawn from the Sub-group Comparisons and the specification/testing of the Statistical Interaction Term discussed above.

```
MPLUS RANDOM SLOPES SYNTAX:
MISSING ARE ALL (-999999);
idvariable = childid;
CLUSTER = centreid;
BETWEEN = ecers_r;
WITHIN = q53am bgcam;
CENTERING = GRANDMEAN (bgcam, q53am);
ANALYSIS:
TYPE = RANDOM TWOLEVEL;
ALGORITHM = INTEGRATION;
MODEL:
%WITHIN%
rgcam on bgcam;
s | rgcam on q53am;
%BETWEEN%
s on ecers_r;
rgcam on ecers_r;
```

Discussion

Although all three methods led to the same substantive conclusion, the robustness of the relationship between this conclusion and the various statistical results/evidence varied. For example, the use of Sub-group Comparisons meant that no estimation was possible of any direct effect from preschool quality on child GCA at 58 months. Furthermore, although the specification of a Statistical Interaction Terms did include this estimate, the operationalisation of the hypothesised moderation was weaker than with the Sub-group Comparisons. This was because the specification of the Statistical Interaction Term was equivalent to using a two-tailed statistical technique to test a one-tailed hypothesis. It was only through the use of a Random Slope Effect that an appropriate one-tailed statistical test was carried out for a one-tailed hypothesis while the conducted analysis also fully estimated the effect of preschool quality upon child GCA at 58 months (leaving aside the potential problems that come through dichotomising moderators as here in the Sub-group Comparisons). The three worked examples show that it is important for educational researchers to

specify their causal hypotheses carefully and to be aware that the robustness of the results and the conclusions drawn may be affected by their conceptualisation and choice of statistical methodology. It is helpful to consider whether hypotheses can be tested in more than one way and to establish if the conclusions remain broadly similar across different approaches used.

CONCLUSIONS

This chapter sought to provide a critical up-to-date review of the terms Mediation, Moderation, and Interaction as they are being commonly defined, discriminated, used, and tested as of 2013 and to consider some of their implications for quantitative educational research. That said, with a greater consideration on the historical origins of these terms as well as 'real-life' difficulties and ambiguities, we have also tried to equip educational researchers with the working knowledge to use these ideas with greater precision and clarity in their own research and to raise awareness of the broader substantive and methodological literatures which often vary in their chosen terminology.

From our review of both historic and current guidelines and practice, a number of recommendations emerge. First, it is essential that educational researchers have sufficient evidence to put forward and clearly distinguish one-tailed hypotheses such as Mediation and Moderation. Second, educational researchers must then gather (or have access to) data that is suitable to address these hypotheses with particular emphasis on the correct causal ordering of measures. Whether the quantitative research is correlational/survey or experimental in nature, for a one-tailed hypothesis to be adequately tested, there must be clear evidence of the appropriate temporal precedence between measures (as there was in the example provided in Section 4 above). Third, with the increasing availability, uptake, and sophistication of statistical software packages, it is our recommendation that quantitative researchers seriously consider the merits of Structural Equation Modelling (SEM) programmes such as EQS, LISREL, AMOS, and MPLUS. Not only have many of the historic statistical methods for testing Mediation, Moderation, and Statistical Interaction been incorporated into these packages, but they also facilitate the testing of these hypotheses when they are chained-together (e.g. multiple mediations as "indirect effects") and combined (e.g. mediated-moderation) in ways that allow the researcher to address and model interesting and complex topics in educational contexts.

Finally and with the aims of this chapter aside, the critical reader might ask themselves, "*given the difficulties with these terms, are they really worth all the trouble?*" and this is an understandable question. Perhaps unsurprisingly, it is our opinion that the greater adoption and use of all of these terms by educational researchers is of direct benefit to the substantive knowledge of the field. In particular, the hypotheses of Mediation and Moderation are tools that give the educational researcher the ability to specify increasingly complex *and yet still testable* research hypotheses and enable them to explore causality in more plausible ways in complex and messy educational and social research contexts. Thus, Mediation and Moderation

are tools that may empower the researcher to specify and test a greater number of clear hypotheses and, done with awareness and defensibly (if not the largely unobtainable "correctly"), this fosters the development of pyramid(s) of knowledge that can help to advance the knowledge base and possibilities of studying important educational research questions.

REFERENCES

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Borkowski, J. G., Whitman, T. L., Wurtz-Passino, A., Rellinger, E. A., Sommer, K., Keogh, D., et al. (1992). Unraveling the "new morbidity": Adolescent parenting and developmental delays. In N. Bray (Ed.), *International review of research in mental retardation* (Vol. 18, pp. 159–196). New York: Academic Press.

Burchinal, M., Peisner-Feinberg, E., Bryant, D., & Clifford, R. (2000). Children's social and cognitive development and child-care quality: Testing for differential associations related to poverty, gender, or ethnicity. *Applied Developmental Science, 4*(3), 149–165.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modelling. *Journal of Abnormal Psychology, 112,* 558–577.

Creemers, B. P. M., Kyriakides, L., & Sammons, P. (2010). Background to educational effectiveness research. *Methodological advances in educational effectiveness research*. New York: Routledge.

Currie, J. (2001). Early childhood education programs. *The Journal of Economic Perspectives, 15*(2), 213–238.

Currie, J., & Thomas, D. (1995). Does head start make a difference? *American Economic Review, 85*, 341–364.

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods, 12*(1), 1–22.

Elliot, C. D., NFER-NELSON, Smith, P., & McCulloch, K. (1996). *British ability scales second edition (BAS II)*. Windsor: NFER-NELSON.

Essex, M. J., Kraemer, H. C., Armstrong, J. M., Boyce, W. T., Goldsmith, H. H., Klein, M. H., et al. (2006). Exploring risk factors for the emergence of children's mental health problems. *Archives of General Psychiatry, 63*, 1246–1256.

Fergusson, D. M., & Lynskey, M. T. (1993). Maternal age and cognitive and behavioural outcomes in middle childhood. *Paediatric and Perinatal Epidemiology, 7*(1), 77–91.

Frazier, P. A., Barron, K. E., & Tix, A. P. (2004). Testing moderator and mediator effects in counselling psychology. *Journal of Counselling Psychology, 51*(1), 115–134.

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science, 18*, 233–239.

Glass, N. (1999). Sure start: The development of an early intervention programme for young children in the United Kingdom. *Children & Society, 13*(4), 257–264.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Charles Griffin and Co.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London, Edward Arnold.

Goodnight, J. A., Bates, J. E., Staples, A. D., Petit, G. S., & Dodge, K. A. (2007). Temperamental resistance to control increases the association between sleep problems and externalizing behavior development. *Journal of Family Psychology, 21*(1), 39–48.

Hall, J., Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2009). The role of pre-school quality in promoting resilience in the development of young children. *Oxford Review of Education, 35*(3).

Harms, T., Clifford, R., & Cryer, D. (1998). *Early childhood environment rating scale, Revised Edition*. New York: Teachers' College Press.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*(4), 408–420.

Hinshaw, S. P. (2002). Intervention research, theoretical mechanisms, and causal processes related to externalizing behavior patterns. *Development and Psychopathology, 14*, 798–818.

Kraemer, H. C., Stice, E., Kazdin, A., Offord, D., & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry, 158*, 848–856.

Luyten, H., & Sammons, P. (2010). Multilevel modelling. In B. P. M. Creemers, L. Kyriakides & P. Sammons (Eds.), *Methodological advances in educational effectiveness research*. New York: Routledge.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83–104.

Masten, A. S. (2001). Ordinary magic: Resilience processes in development. *American Psychologist, 56,* 227–238.

Masten, A. S. (2007). Resilience in developing systems: Progress and promise as the fourth wave rises. *Development and Psychopathology, 19*(3), 921–930.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Quantitative Methods in Psychology, 114*(2), 376–390.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89*(6), 852–863.

Muthén, L. K., & Muthén, B. O. (2010). Mplus software (Version 6). Los Angeles, CA: Authors.

NICHD. (2000). The interaction of child care and family risk in relation to child development at 24 and 36 months. *Applied Developmental Science, 6*(3), 144–156.

Nicholson, R. A., Hursey, K. G., & Nash, J. M. (2005). Moderators and mediators of behavioral treatment for headache. *Headache, 45*, 513–519.

Preacher, K. J., & Hayes, A. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers, 36*(4), 717–731.

Preacher, K. J., Rucker, D. D., Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42,* 185–227.

Reynolds, A. J., & Ou, S. (2003). Promoting resilience through early childhood intervention. In S. S. Luthar (Ed.), *Resilience and vulnerability: Adaptation in the context of childhood adversity*. Cambridge: Cambridge University Press.

Roe, R. A. (2012). What is wrong with mediators and moderators? *The European Health Psychologist, 14*(1), 4–10.

Rose, B., Holmbeck, G., Coakley, R., & Franks, E. (2004). Mediator and moderator effects in developmental and behavioral pediatric research. *Journal of Developmental and Behavioral Pediatrics, 25*, 1–10.

Rutter, M., & Silberg, J. (2002). Gene-environment interplay in relation to emotional and behavioral disturbance. *Annual Review of Psychology, 53*, 463–490.

Saunders, D. R. (1955). The 'moderator variable' as a useful tool in prediction. *Proceedings of the conference on testing problems. Educational Testing Service*, 54–58.

Sobel, M. E. (1982). Aysmptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 290–212). San Francisco: Jossey-Bass.

Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2010). *Early childhood matters: Evidence from the effective pre-school and primary education project*. Abingdon: Routledge.

Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2012). *Effective pre-school, primary and secondary education 3–14 Project (EPPSE 3–14) - Final report from the key stage 3 phase: Influences on students' development form age 11–14*: Department for Education.

Talamini, R., Bosetti, C., La Vecchia, C., Dal Maso, L., Levi, F., Bidoli, E., et al. (2002). Combined effect of tobacco and alcohol on laryngeal cancer risk: A case-control study. *Cancer Causes and Control, 13*, 957–964.

Tatsuoka, M. M. (1973). Multivariate analysis in educational research. *Review of Research in Education, 1*, 273–319.

Wu, A. D., & Zumbo, B. D. (2007). Understanding and using mediators and moderators. *Social Indicators Research: An International Interdisciplinary Journal for Quality of Life Measurement.*

SECTION 3

STRUCTURAL EQUATION MODELS

MATTHEW W. GALLAGHER & TIMOTHY A. BROWN

14. INTRODUCTION TO CONFIRMATORY FACTOR ANALYSIS AND STRUCTURAL EQUATION MODELING

Confirmatory factor analysis (CFA) is a powerful and flexible statistical technique that has become an increasingly popular tool in all areas of psychology including educational research. CFA focuses on modeling the relationship between manifest (i.e., observed) indicators and underlying latent variables (factors). CFA is a special case of structural equation modeling (SEM) in which relationships among latent variables are modeled as covariances/correlations rather than as structural relationships (i.e., regressions). CFA can also be distinguished from exploratory factor analysis (EFA) in that CFA requires researchers to explicitly specify all characteristics of the hypothesized measurement model (e.g., the number of factors, pattern of indicator-factor relationships) to be examined whereas EFA is more data-driven. In this chapter we will provide a general introduction to how CFA and SEM can be used within educational research and other areas of psychology. We will begin with a nontechnical overview of the purpose of and methods underlying CFA and SEM before describing the various potential uses of CFA and SEM in educational research. We will then discuss the advantages of CFA and SEM over traditional methods of data analysis, provide an overview of the core steps in conducting CFA and SEM analyses, and discuss some practical issues in conducting these analyses such as software options. We then provide a brief summary of some of the more advanced methods in which CFA and SEM can be extended to conduct sophisticated analyses. We conclude with an illustrative series of example models in which the relationship between academic self-efficacy and academic performance is examined using CFA and SEM.

OVERVIEW AND GOALS OF CFA AND SEM

The goals of both CFA and SEM are to identify latent variables using a set of manifest indicators and to then evaluate hypotheses regarding the relationships among the latent variables. The conceptual background for conducting these analyses is the common factor model (Thurstone, 1947), which states that each manifest indicator is a linear function of one or more common factors and a unique factor. Factor analytic techniques therefore attempt to partition the variance of an indicator into (1) common variance, or the proportion of variance that is due to the latent variable, and (2) unique variance, which is a combination of random error variance

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 289–314.

(e.g., measurement error) and reliable variance that is specific to a particular item. Both EFA and CFA and SEM attempt to reproduce the observed intercorrelations/covariances between items with a more parsimonious set of latent variables. The primary difference, as mentioned above, is that CFA and SEM require researchers to explicitly specify every aspect of the models to be evaluated. CFA and SEM therefore require that researchers have a strong conceptual or empirical foundation to guide the specification and evaluation of models.

Common Uses of CFA and SEM

Some of the most common uses of CFA in educational and other areas of research include scale validation, construct validation, and evaluating measurement invariance. It is now considered standard practice to conduct a series of factor analyses when developing a new measure in psychological research. The standard progression is for researchers to begin by specifying an EFA model to evaluate an initial pool of items, and to then move to a CFA framework to provide a more rigorous evaluation of how a theoretical model represents the observed data. Through this process, researchers are able to determine the number of latent variables that best represents the constructs of interest and the pattern of relationships (i.e. factor loadings) between the observed items and latent variables. Thus, for instance, CFA can help researchers determine whether they should focus on the total score of a measure or subscales comprised of particular items from that scale. CFA also provides superior methods of evaluating other psychometric properties (e.g., reliability) of a scale than traditional methods such as Cronbach's alpha. For these reasons, educational researchers are strongly encouraged to use CFA when developing and validating new scales.

Another common application of CFA is to evaluate whether the measurement properties of an assessment are invariant. This is often an important second step in scale development. Measurement invariance can be tested cross-sectionally between groups or longitudinally between assessments of the same individuals. The use of CFA to evaluate measurement invariance across groups is discussed in detail in Chapter XX; in brief, these methods allow researchers to evaluate whether the relationship between indicators and latent variables is consistent between groups. For example, researchers could use CFA to evaluate measurement invariance between sexes on a test of mathematical proficiency. This analysis could help researchers determine whether any observed differences between sexes represent true differences between males and females or merely indicate that the items on a particular assessment function differently between sexes. The evaluation of measurement invariance is also a very important but underappreciated issue in longitudinal research, as the demonstration of measurement invariance across assessments provides the foundation for researchers to conclude that change in a latent variable across time truly represents growth or decline rather than inconsistent measurement. For additional information about how to test measurement invariance, readers are encouraged to consult Brown (2006), and Cheung and Rensvold (2002).

A third area in which CFA is commonly used is construct validation. CFA and SEM provide a useful framework for demonstrating both convergent and discriminant validity of theoretical constructs. Convergent validity is indicated by evidence that multiple indicators of theoretically linked constructs are strongly interrelated; for example, results on a series of tests that all purport to measure mathematical aptitude load on a single factor. Discriminant validity is indicated by evidence that indicators of theoretically distinct constructs do not correlate strongly with one another; for example, indicators of verbal and mathematical aspects of intelligence load on separate factors and correlate more so with indicators within the same domain of intelligence than with indicators within a different domain of intelligence. One of the most robust ways in which CFA can be used in construct validation is with the use of multitrait-multimethod techniques (Campbell & Fiske, 1959; Kenny & Kashy, 1992), a powerful yet infrequently used technique in which several constructs are measured using multiple methods and then modeled such that common variance due to method effects is separated from common variance due to latent traits.

Advantages of CFA and SEM

CFA and SEM have numerous advantages over traditional statistical techniques such as correlation and regression. One of the primary advantages of CFA and SEM is that they allow researchers to estimate the relationships between variables while accounting for measurement error. Traditional statistical techniques impose the generally unrealistic assumption that variables have been measured perfectly with no error. This assumption of error-free measurement is rarely appropriate in educational research or other areas of psychological research and results in parameter estimates that are biased to an unknown degree due to the failure to account for measurement error. By specifying latent variables that allow for the estimation of measurement error, researchers are able to obtain more accurate, reliable, and valid estimates of the relationships among latent constructs. This can also result in increased statistical power as the relationships between variables can be more precisely estimated after properly accounting for the role of measurement error. An important strength of CFA is the ability to model complex error structures among indicators to account for method effects (e.g., two self-report indicators of intelligence may correlate more strongly with one another than peer and teacher evaluations of intelligence would). Another important advantage of latent variable techniques such as CFA and SEM is that they permit the specification of complex longitudinal models that can help researchers to evaluate sophisticated theoretical models regarding change (e.g., latent growth curve models). A few of the more advanced methods are discussed later in this chapter. It is worth noting, however, that there are many circumstances in which CFA and SEM may not be the ideal method of data analysis. Most notably, if researchers are focusing on manifest variables that do not include measurement error (e.g., gender, grade point average), latent variable modeling techniques such as CFA and SEM may not be necessary.

CORE STEPS IN CFA AND SEM ANALYSES

CFA and SEM are complex statistical techniques that are performed in an iterative process and that present researchers with a number of important decisions during the process. The steps identified subsequently will provide readers with a general outline of the most common steps that researchers will follow when conducting CFA and SEM analyses. The subsequent steps presume that a researcher has already collected an appropriate dataset and has screened the data for outliers, univariate normality, and multivariate normality (Kline, 2011).

Specify Theoretical Model

The first step in conducting CFA and SEM analyses is for the researcher to clearly specify the theoretical model they are interested in testing. As mentioned previously, CFA and SEM differ from more data-driven procedures such as EFA and it is therefore crucial that researchers have a very clear idea of the specific models they want to test in advance. It is often helpful to diagram the planned models using common SEM notation and symbols. A fully notated example for a two-factor, six indicator CFA model can be seen in Figure 1. When diagramming SEM models, circles are used to denote latent variables, squares or rectangles are used to denote manifest or observed variables, correlations (standardized solutions) or covariances

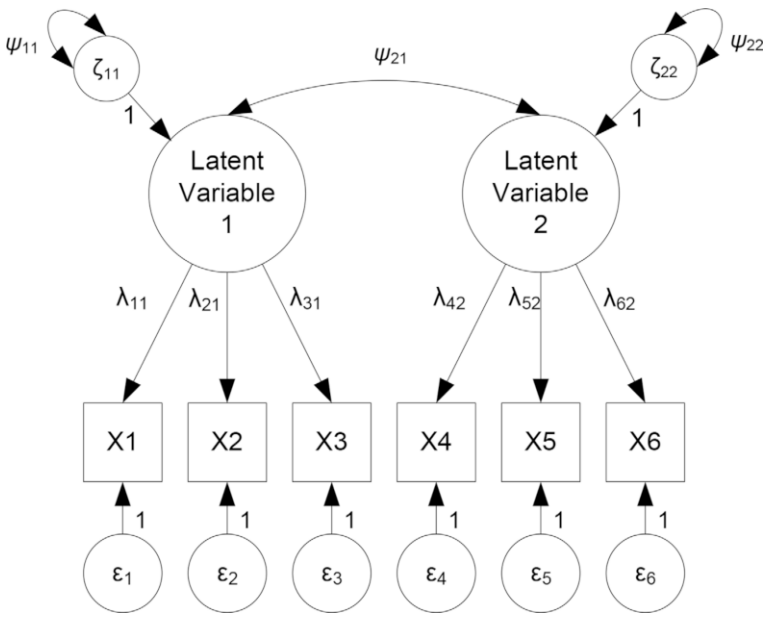


Figure 1. Example two factor CFA measurement model with six manifest indicators.

(unstandardized solutions) are denoted using double headed arrows, and single headed arrows are used to denote direct effects such as factor loadings or effects of one latent variable on another. Other common notations in SEM include the use of lambda (λ) to denote factor loadings, psi (ψ) to denote variances and covariances/correlations, and theta (θ) to indicate residuals and residual covariances.

Specify Measurement Model

After clearly specifying a theoretical model to be tested, researchers should next evaluate the measurement model for the latent variables of interest. The test of the measurement model should always be conducted prior to evaluating structural equation models. There are two important issues that researchers must consider when specifying the measurement model. The first issue is that researchers must ensure that CFA and SEM models are statistically identified. Adequate identification occurs when the number of parameters to be estimated in a model does not exceed the number of pieces of information in the variance-covariance input matrix. If a model is not adequately identified, then a solution cannot be solved as there are an infinite number of potential solutions. In CFA, the number of known pieces of information is determined by the size of the input variance/covariance matrix and can be calculated using the formula $b = [p * (p + 1)]/2$, where b is the number of elements in the input matrix and p is the number of variables included in the input matrix. For example, an input matrix of three variables would provide six pieces of information (three variances and three covariances) while an input matrix of two variables would only provide 3 pieces of information (two variances and 1 covariances). It is therefore only possible to freely estimate six parameters (i.e. three factor loadings and three residuals or two factor loadings, three residuals and the variance of the latent variable) in a model that uses an input matrix with three variables. When the number of freely estimated parameters equals the number of elements in the input matrix, then a model is just-identified and will fit the data perfectly. When the number of elements in the input matrix is greater than the number of freely estimated parameters, then a model is over-identified and the degrees of freedom (df) for the model can be determined by subtracting the number of freely estimated parameters from the number of known elements. When a model is overidentified, researchers are able to obtain goodness of fit statistics (discussed in more detail subsequently) that provide information about how well the specified CFA model reproduced the observed relationships in the sample data. It is also important to consider whether a model is locally identified in addition to being globally identified. Large models that include many variables will usually be over-identified but researchers should take care that each latent variable within a model is adequately identified. Situations in which the overall model is over-identified but certain components of the model are not locally identified are referred to as empirically under-identified solutions (e.g., selection of a marker variable that is unrelated to the other indicators that are

specified to load on the same factor; see next paragraph). In these cases, either the model cannot be estimated or the model will converge but contain out of bounds parameter estimates (i.e., negative residual variances).

The second important issue in specifying CFA measurement models is setting the scale of latent variables. Latent variables do not have an inherent metric so the scale of these variables must always be set using one of the three methods. The most widely used method is the marker variable method, which involves fixing the factor loading of one indicator for each latent variable to be 1.0. This method results in setting the scale of the latent variable to the metric of the marker variable. Another common method is to standardize the factor variance, which involves fixing the variance of the latent variable to 1.0. The fixed factor method results in a standardized solution for the factor loadings and residuals. A third, but less common, method for setting the scale of latent variables is the effects coding approach (Little, Slegers, & Card, 2006). This method involves constraining the loadings of a latent variable to average 1.0. This is done by freely estimating all but one of the factor loadings and then fixing the remaining factor loading to equal the number of indicators minus each of the freely estimated factor loadings. The advantage of the effects coding method is that the parameters in a model (i.e., variances, means) reflect the observed scale of the indicator variables. The disadvantage of the effects coding method is that it requires slightly more complicated syntax. Each method is valid and will produce identical results in terms of model fit.

Estimate and Evaluate Measurement Model

The next step is to estimate the model using one of the many software packages designed for latent variable analysis (discussed later). The estimation process in CFA and SEM involves a fitting function (most commonly maximum likelihood; ML) that iteratively produces parameter estimates in an attempt to minimize the differences between the model-implied variance-covariance matrix and the sample variance-covariance matrix. For a more thorough description of the procedures involved in ML estimation and the circumstances in which other estimation methods are preferred, the reader is referred to Brown (2006), and Eliason (1993).

If a model converges successfully (i.e., a solution is obtained through ML estimation), researchers can then evaluate how acceptable the model fit the data. There are three primary components of the results that researchers should focus on when evaluating model fit. The first is overall goodness of fit, which reflects the degree to which the estimates of the CFA model reproduce the relationships between variables in the observed sample. A variety of fit statistics have been developed and it is generally recommended that researchers report multiple fit indices as they provide a more conservative and comprehensive evaluation of model fit. The classic goodness of fit index is model chi-square (χ^2). If the χ^2 value of a model exceeds the critical value from the χ^2 distribution (determined by the model's degrees of freedom), then the null hypothesis of adequate model fit is rejected. Although

χ^2 provides a very straightforward test of model fit, it has significant limitations including that it is overly sensitive to sample size and is therefore likely to reject very good models if the sample is large. For this reason, it is generally recommended that researchers report χ^2, but focus more on other fit indices when evaluating model fit. The most widely accepted global fit indices are the root mean square error of approximation (RMSEA; Browne & Cudeck, 1992; Steiger & Lind, 1980), the comparative fit index (CFI; Bentler, 1990), the Tucker-Lewis index (TLI; Tucker & Lewis, 1973) which is sometimes also referred to as the non-normed fit index, and the standardized root mean square residual (SRMR; Bentler, 1995). For each of these fit statistics values generally range from 0 to 1. For the SRMR and RMSEA, values closer to 0 indicate better model fit, while values closer to 1 indicate better model fit for CFI and TLI. Recommendations vary in terms of what values of these fit statistics should be considered acceptable. Early guidelines for model fit suggested that CFI and TLI values greater than .9, and RMSEA values less than .1 should be considered acceptable (Bentler, 1990; MacCallum et al., 1996). More recently, the results of one of the most comprehensive simulation studies examining model fit (Hu & Bentler, 1999) suggested the following guidelines for considering a model to have good fit: (1) SRMR values close to or below .08, (2) RMSEA values close to or below .06, and (3) CFI and TLI values close to or above .95. It is important to recognize that these guidelines should used as general recommendations rather than rigid guidelines and that model fit should always be evaluated in terms of multiple fit indices rather than just a single fit statistic.

The second aspect of the results that researchers should examine when evaluating model fit is localized areas of poor fit. The global fit indices (e.g., RMSEA, CFI) provide a useful evaluation of the overall fit of a model but it is possible for a model to have good overall fit while poorly reproducing specific aspects of the model. The most common method for identifying localized misfit is by examining modification indices. Modification indices reflect the approximate change in the overall model χ^2 if a fixed or constrained parameter were to be freely estimated. Modification indices can be conceptualized as a χ^2 with 1 degree of freedom so modification indices of 3.84 or greater (i.e., the critical value of χ^2 with 1 df, $\alpha = .05$) suggest that the model fit could be significantly improved by freely estimating the parameter in question. Large modification indices may therefore provide researchers with information about how a particular model may be misspecified (e.g., the need for specifying a residual covariance between two indicators to account for a method artifact). However, it is important that researchers not make revisions to a model solely based on modification indices without a theoretical or empirical basis as this can lead to model overfitting and inappropriate capitalization on chance associations in the sample data (MacCallum, Roznowski, & Necowitz, 1992).

The third aspect of model evaluation is the interpretability, strength, and statistical significance of parameter estimates. It is important to confirm that model results do not contain any out of range values such as negative variances (often referred to as Heywood cases or offending estimates). This outcome can indicate significant problems in how

the model was specified or problems with the sample data. Nonsignificant parameter estimates may indicate unnecessary parameters or items that are poor indicators of a latent construct. It is also useful to examine the completely standardized parameter estimates at this stage as these can be interpreted as correlations in the case of associations between latent variables, standardized regression coefficients in the case of factor loadings, and the proportion of variance unexplained in indicators in the case of residual variances. For example, a correlation approaching 1.0 between two latent variables may indicate that the two constructs are not truly distinct and that it may be more appropriate and parsimonious to collapse the variables into a single latent construct.

Consider Model Revisions

After estimating the measurement model and evaluating goodness of fit, the next step for researchers is to decide whether any revisions to the model are warranted. As mentioned previously, potential model revisions can be indicated based on modification indices or evaluation of the significance and strength of the parameter estimates. Any revisions should be made in an iterative fashion as modification indices are not independent of one another and minor changes in how a model is specified can produce large changes in both model fit and the parameter estimates. Researchers should err on the side of not making *post hoc* model revisions unless there is a strong theoretical or empirical foundation so as not to artificially inflate model fit by incorporating revisions that merely reflect sample-specific characteristics.

Specify Structural Models (If Applicable)

After establishing a satisfactory measurement model using CFA, researchers can then begin to specify structural equation models. Structural models allow researchers to explicitly model hypothesized relationships beyond the basic associations that are specified in CFA measurement models (i.e., factor covariances). More specifically, it is at this point that researchers can test hypothesizes regarding the presence or absence of regression effects among the latent variables, test models that involve the estimation of indirect effects to evaluate mediation hypotheses, test models that involve the estimation of interaction effects to evaluate moderation hypotheses, and test models that involve the specification of complex patterns of longitudinal growth such as latent growth curve models (Preacher et al., 2008). These are just a few of the many types of structural equation models that can be specified and researchers need to take care to use models that are appropriate for testing their specific theoretical hypotheses and models.

Reporting Results

The final step in conducting CFA and SEM analyses is to report the analyses in a clear and understandable manner so that it is possible for others to understand

exactly how the models were specified and to replicate the models in independent samples (cf. McDonald & Ho, 2002). Given the complexity of many CFA and SEM models it is not always feasible to report every single parameter estimate from a model, but for transparency there are certain aspects of models that should always be presented. Researchers should clearly state how a model was specified, including information about the method of scale-setting used and justification of any *post hoc* model modifications. Multiple indices of model fit should be reported, preferably all five of the fit statistics described previously (i.e., χ^2, RMSEA, SRMR, CFI, and TLI). Researchers should also report information regarding the specific parameters of interest in a model, in both unstandardized and standardized form. Presentation of model parameters can often be accomplished most easily by presenting the results in a figure. It is also preferable to include the descriptive statistics (means and standard deviations) and the correlation/covariance matrix used to estimate the models so that readers can see the input matrix that was used to estimate the model (e.g., for data re-analysis).

PRACTICAL ISSUES IN USING CFA AND SEM

Software

There are now numerous software packages that are capable of estimating CFA and SEM models. Some of the most widely used programs include Mplus (Muthén & Muthén, 2008–2012), LISREL (Jöreskog & Sörbom, 1996), AMOS (Arbuckle, 2010), EQS (Bentler, 2006), CALIS (SAS Institute, 2005), Mx (Neale, Boker, Xie, & Maes, 2003), and multiple packages within the R statistical framework including SEM (Fox, 2006) and LAVAAN (Rosseel, 2011). All of these programs allow for the specification of CFA and SEM models either through the creation of syntax files or graphical interfaces. Each program is capable of estimating the most common CFA and SEM models but certain programs have unique characteristics or advantages. Mplus is in some ways the most flexible software program as it allows users to specify advanced models such as exploratory structural equation modeling (Asparaouhov & Muthén, 2009), multilevel structural equation modeling (Muthén & Asparouhov, 2008), and to use Bayesian estimation procedures that are not available or not easily specified in other programs. LISREL is a good program for didactic purposes as it allows researchers to specify models in terms of the matrices that comprise SEM models (e.g., lambda matrix for factor loadings). Mx has some advanced capabilities for estimating twin models and is the most common latent variable program in genetics research. All of the SEM packages within the R framework (e.g., LAVAAN) are open-source and free. AMOS and EQS both provide users with the option to specify models using a graphical interface. Although this capability may seem appealing, researchers should take caution as it is very easy to misspecify models when using graphical interfaces and ultimately, it is often easier to specify complex models using a syntax file.

Sample Size Requirements

As with any other area of research, the issue of statistical power is an important one when conducting CFA and SEM. There are multiple methods of determining power in CFA and SEM. One approach is based on statistical power for evaluating a model using RMSEA (MacCallum et al., 1996; Preacher & Coffman, 2006). This method requires researchers to specify the degrees of freedom for a model, alpha (typically .05), desired power (typically .80), and null and alternative values for RMSEA, and provides the sample size necessary to achieve the desired level of power in terms of the RMSEA model evaluation. An alternative and more flexible method of estimating power for latent variable models is the Monte Carlo method. Monte Carlo simulation studies allow researchers to evaluate the bias in specific parameter estimates and to determine power for detecting significant parameters based on population parameter estimates and varying sample sizes specified by the researcher. For a more detailed overview of methods for calculating power in CFA and SEM models and the use of Monte Carlo methods, readers are referred to Brown (2006), and Muthén and Muthén (2002).

Handling Missing Data

A common issue that applied researchers face when conducting CFA, SEM or any other form of statistical analysis is determining the most appropriate method for handling missing data. It is rare that researchers will collect a dataset in which no data are missing, and there are many reasons that data may be missing. Current typologies of missing data distinguish between three forms of missing data. In some situations data can be considered to be missing completely at random (MCAR) if, for example, a particular questionnaire was accidentally omitted in assessment packets for a few individuals. Data can also be considered missing at random (MAR) if, for example, attrition in a longitudinal study of academic outcomes is related to other variables in the data set such as academic engagement or motivation. Finally, data can be missing not at random or nonignorable if the pattern of missingness is related to some unobserved variable(s). A more complete description of the nature and implications of these patterns of missingness can be found in Enders (2010), but for the purposes of this chapter we will focus on what researchers can do to manage MAR and MCAR situations.

Many of the traditional methods of handling missing data (e.g., pairwise or listwise deletion) are inappropriate, as it has been repeatedly demonstrated that these approaches result in reduced statistical power and often produce biased parameter estimates (Allison, 2003; Enders, 2010; Schafer & Graham, 2002). The two methods that are the most appropriate strategies for handling missing data are full information maximum likelihood (FIML) estimation (also commonly called direct maximum likelihood), and multiple imputation. Both approaches are appropriate when data can be considered to be MAR or MCAR. We will focus our discussion on the FIML approach as this approach is easily implemented in many latent variable modeling software packages

(e.g., Mplus, LISREL), and is generally regarded by methodologists as the most straightforward method of handling missing data (Allison, 2003). FIML methods use all of the available data to provide appropriate estimates of parameters and standard errors for a model while accounting for missing data. FIML is now the default estimator in Mplus and can be easily used in LISREL by including the MI keyword in the data line and indicating the missing data code in the dataset (e.g., MI = 9).

ADDITIONAL APPLICATIONS OF CFA AND SEM

Examining Mediation Using Structural Equation Modeling

Mediation can be defined as a process in which the effect of one variable (X) on another variable (Y) occurs through an intervening variable (M) (Baron & Kenny, 1986; MacKinnon, 2008). Mediation is an increasingly popular focus of research in educational research. SEM provides a very useful framework for evaluating mediational hypotheses. The use of latent variables allows researchers to obtain more accurate estimates of the overall indirect effect as well as the constituent parts of the indirect effect (i.e., M on X, Y on M). SEM also allows researchers to simultaneously evaluate multiple mediators and to extend mediation models to a longitudinal framework to evaluate how mediational processes unfold over time. Furthermore, it is possible to directly obtain bias-corrected and accelerated bootstrapped confidence intervals of the indirect effect, the current best-practice recommend method (Preacher & Hayes, 2008; MacKinnon, 2008) within SEM software packages such as Mplus. An example of how mediation can be examined within an SEM framework is presented later in this chapter.

Examining Moderation Using Structural Equation Modeling

Moderation (i.e., interactions) is also an increasingly popular area of research within education and other social sciences domains. Moderation can be tested in SEM for both categorical and continuous moderators. Categorical moderators can be evaluated using multiple groups models in which the parameters of interest are specified for each category of the moderator, with differences in the relationships between the groups considered evidence of moderation that can be tested for statistical significance using equality constraints. There are also multiple methods for evaluating continuous moderators within SEM. Little, Bovaird, and Widaman (2006) describe an approach in which a latent interaction term is specified by orthogonalizing the respective indicators of the independent variable and the moderator. Example syntax for how this approach be applied can be found in Schoemann (2010).

Longitudinal Extensions of Structural Equation Modeling

One of the most useful ways in which CFA and SEM can be extended is to examine longitudinal data. There are many ways in these methods can be extended to

model complex patterns of change. Cross-lagged panel models allow researchers to evaluate how interindividual standing in latent constructs changes over time (Burkholder & Harlow, 2003). Latent growth curve models allow researchers to examine intraindividual trajectories of change and can be used to evaluate non-linear and other complex patterns of change (Bollen & Curran, 2006; Preacher et al., 2008). Latent difference score models are a third approach for modeling longitudinal change and allow researchers to examine intraindividual change in latent constructs between two assessments (McArdle, 2009). Each of these methods is well-suited to studying a variety of research topics and researchers interested in learning more about these topics are encouraged to consult Collins (2006), Selig and Preacher (2009), or Little, Bovaird, and Card (2007).

Multilevel Structural Equation Modeling

The final extension of CFA and SEM that we will mention is multilevel structural equation modeling (MSEM; Muthén & Asparouhov, 2008). MSEM is a relatively recent development and combines all of the advantages of hierarchical linear modeling (e.g., accounting for nested dependencies in data) and SEM (e.g., accounting for measurement error). MSEM is therefore an extremely robust statistical framework as it allows researchers to specify models that are not possible when using either hierarchical linear modeling or SEM. MSEM remains an infrequently used method given its complexity. However, descriptions of how these methods can be used in applied research are increasingly common (e.g., Preacher, Zyphur, & Zhang, 2010) and MSEM is likely to be a major area of growth in the next decade.

Illustrative Study

An example study will now be presented to demonstrate the sequence of steps that researchers will typically follow when conducting a study involving CFA and SEM. The data for these example models come from a study in which undergraduates completed a series of self-report questionnaires during their first semester of college to identify the psychological variables (e.g., self-efficacy, hope, engagement) that best predict academic performance during the first four years of college (Gallagher & Lopez, 2008). Participants were 229 students (129 males, 100 females) at a large Midwestern university who participated in exchange for psychology course credit. Prior to completing their first semester, participants completed the academic self-efficacy scale (Chemers, Hu, & Garcia, 2001), identified their goal for their GPA after four years of college, and provided consent to have their academic performance (semester GPA) tracked by the investigators through the University Registrar's office.

For illustration purposes, we will focus on just the relationships between academic self-efficacy, self-reported goals for GPA during the first semester of college, and cumulative GPA after four years of college. The descriptive statistics and correlation matrix used for these analyses are presented in Table 1. There were no missing data

Table 1. Sample correlations, standard deviations (SD) and means (M) for self-efficacy (SE) items, college grade point average goal (GPAGOAL), and four year college grade point average (GPA)

	SE1	*SE2*	*SE3*	*SE4*	*SE5*	*SE6*	*SE7*	*SE8*	*GPAGOAL*	*GPA*
SE1	1									
SE2	.408	1								
SE3	.365	.533	1							
SE4	.256	.247	.363	1						
SE5	.540	.432	.497	.325	1					
SE6	.432	.422	.507	.374	.756	1				
SE7	.246	.354	.441	.254	.476	.464	1			
SE8	.385	.385	.375	.316	.576	.579	.421	1		
GPAGOAL	.032	.051	.110	.148	.234	.309	.122	.142	1	
GPA	.263	.166	.278	.203	.302	.371	.102	.215	.350	1
N	229	229	229	229	229	229	229	229	228	147
M	5.14	5.45	4.92	4.83	4.95	5.07	4.67	5.92	3.39	2.96
SD	1.50	1.39	1.39	1.60	1.14	1.29	1.31	1.15	.34	.49

for the self-efficacy variables but data were missing for one person's GPA goals and 82 people were missing data on four year college GPA. These missing data were considered MAR and were accommodated using FIML. A series of four models will be presented to demonstrate the common steps researchers may take when using CFA and SEM. The first model uses CFA to evaluate the measurement model of the academic self-efficacy scale. The second model is an extension of the one-factor measurement model to include a correlated residual between two items. The third model examines the effect of academic self-efficacy on cumulative college GPA using SEM. The fourth model tests a mediation model in which participants' goals for GPA reported during their first semester of college partially mediates the effects of academic self-efficacy beliefs on cumulative GPA after four years of college. Mplus syntax for each of these examples will be presented, but each model could be conducted in the other latent variable software programs mentioned previously.

Evaluating the Measurement Model

The first step in evaluating the effects of academic self-efficacy on academic performance is to determine how well the latent construct of academic self-efficacy was measured. This can be accomplished using a basic one-factor CFA model. Annotated Mplus syntax and selected output from this model are presented in Table 2. As can be seen in Table 2, the syntax required for specifying a one-factor

Table 2. Mplus syntax and selected output of confirmatory factor analysis of the academic self-efficacy scale

SYNTAX:

```
TITLE: Academic Self-Efficacy Confirmatory Factor Analysis
DATA: FILE IS acaselfeff.dat;
VARIABLE:
    NAMES ARE id gpagoal gpa4year aselfe1-aselfe8;          !Identify all
    variables in data set
    USEVARIABLES ARE aselfe1-aselfe8;   !Specify variables to be used in model
    MISSING are all (-9);
ANALYSIS:
    TYPE IS GENERAL;
    ESTIMATOR IS ML;
MODEL:
    acaeffic by aselfe1-aselfe8;                !Specify 8 items as indicators
                                !Mplus defaults to marker variable identification
OUTPUT: MODINDICES(4) STANDARDIZED;          !Request completely
                                standardized results and !modification indices
```

SELECTED OUTPUT:

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

Value	58.290
Degrees of Freedom	20
P-Value	0.0000

CFI/TLI

CFI	0.944
TLI	0.921

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.091
90 Percent C.I.	0.065 0.119

SRMR (Standardized Root Mean Square Residual)

Value	0.044

MODEL RESULTS

	Estimate	S.E	Est./S.E.	Two-Tailed P-Value
ACAEFFIC BY				
ASELFE1	1.000	0.000	999.000	999.000
ASELFE2	0.901	0.130	6.946	0.000
ASELFE3	1.008	0.135	7.491	0.000
ASELFE4	0.804	0.142	5.643	0.000
ASELFE5	1.132	0.123	9.189	0.000

ASELFE6	1.260	0.142	8.887	0.000
ASELFE7	0.864	0.125	6.934	0.000
ASELFE8	0.899	0.114	7.886	0.000
Intercepts				
ASELFE1	5.140	0.099	51.838	0.000
ASELFE2	5.445	0.092	59.276	0.000
ASELFE3	4.921	0.091	53.814	0.000
ASELFE4	4.825	0.106	45.686	0.000
ASELFE5	4.948	0.075	65.923	0.000
ASELFE6	5.074	0.085	59.473	0.000
ASELFE7	4.672	0.087	54.013	0.000
ASELFE8	5.917	0.076	78.046	0.000
Variances				
ACAEFFIC	0.745	0.165	4.518	0.000

COMPLETELY STANDARDIZED MODEL RESULTS

	Estimate	S.E	Est./S.E.	Two-Tailed P-Value
ACAEFFIC BY				
ASELFE1	0.575	0.049	11.835	0.000
ASELFE2	0.559	0.050	11.146	0.000
ASELFE3	0.628	0.045	14.009	0.000
ASELFE4	0.434	0.058	7.511	0.000
ASELFE5	0.861	0.024	36.073	0.000
ASELFE6	0.842	0.025	33.350	0.000
ASELFE7	0.570	0.049	11.644	0.000
ASELFE8	0.677	0.040	16.787	0.000
Variances				
ACAEFFIC	1.000	0.000	999.000	999.000

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
ASELFE1	0.331	0.056	5.917	0.000
ASELFE2	0.313	0.056	5.573	0.000
ASELFE3	0.395	0.056	7.005	0.000
ASELFE4	0.188	0.050	3.755	0.000
ASELFE5	0.741	0.041	18.037	0.000
ASELFE6	0.709	0.043	16.675	0.000
ASELFE7	0.325	0.056	5.822	0.000
ASELFE8	0.458	0.055	8.393	0.000

MODEL MODIFICATION INDICES

Minimum M.I. value for printing the modification index 4.000

	M.I.	E.P.C.	Std E.P.C.	StdYX E.P.C.
WITH Statements				
ASELFE2 WITH ASELFE1	4.201	0.204	0.204	0.145
ASELFE3 WITH ASELFE2	21.306	0.409	0.409	0.330
ASELFE4 WITH ASELFE3	4.267	0.224	0.224	0.145
ASELFE5 WITH ASELFE1	4.856	0.140	0.140	0.197
ASELFE5 WITH ASELFE2	5.477	–0.138	–0.138	–0.207
ASELFE5 WITH ASELFE3	5.311	–0.132	–0.132	–0.213
ASELFE5 WITH ASELFE4	4.298	–0.146	–0.146	–0.176
ASELFE6 WITH ASELFE1	5.362	–0.169	–0.169	–0.197
ASELFE6 WITH ASELFE2	4.483	–0.144	–0.144	–0.179
ASELFE6 WITH ASELFE5	13.271	0.203	0.203	0.505
ASELFE7 WITH ASELFE3	4.567	0.177	0.177	0.153

CFA model in Mplus is straightforward. The first few lines of syntax involve providing a title for the analysis, identifying the location of the data file (for Mplus the data file can be either tab-delimited, comma-delimited, or a fixed width ASCII file), providing variable names for all variables included in the dataset, selecting the specific variables that are included in the model to be analyzed, and identifying what numeric value used to indicate missing data (blanks can also be used as a missing data code if the data are in a fixed width ASCII file). The syntax for specifying the one-factor CFA model requires only two lines in Mplus. The first line signifies that the latent construct of Academic Self-Efficacy is identified by the eight items of the academic self-efficacy scale (Chemers et al., 2001). For this model, the latent construct of academic self-efficacy is identified using the marker variable method. This method of model identification is the default method in Mplus and simply requires that the corresponding indicators for the latent variable are specified (e.g., acaeffic by aselfe1-aselfe8;); by default, Mplus uses the first indicator after the "by" keyword (aselfe1) as the marker variable by fixing its unstandardized factor loading to 1.0. The final line of syntax instructs Mplus to provide additional output in the form of modification indices that equal 4.0 or above, and the standardized/completely standardized estimates.

Because the CFA model of the academic self-efficacy scale converged successfully with no error messages, the first step is to examine the model fit statistics. The χ^2 test

of model fit indicated significant model misfit ($p < .001$). However, as previously mentioned, the χ^2test is an overly conservative test and it is therefore more important to focus on the remaining model fit statistics. Although the SRMR is consistent with good model fit (.044), the CFI, TLI, and RMSEA indicate marginal fit (values of .944,.922, and .091, respectively). Taken together, these results suggest the specified model fit does not provide a good representation of the data, so the next step is to examine the modification indices to determine whether it may be possible to improve fit by respecifying the model. As noted earlier, this should only be done if substantively justified, as high modification indices do not necessarily indicate a relationship that is theoretically meaningful.

In the results presented in Table 2, the largest modification index is for the residual covariance between items 2 and 3 of the scale. The value for this modification index (21.31) is well above 3.84 and indicates there is a relationship between these two items that is not sufficiently accounted for by the latent variable of academic self-efficacy. An examination of the content of these two items reveals that this may be explained by a method effect arising from similar wording. Given the common stem of these items, it was deemed appropriate to specify a residual covariance between these two items to account for the method effect.

Revising the Measurement Model

The syntax and selected output from a second measurement model of the academic self-efficacy scale in which a residual covariance between items two and three is specified is presented in Table 3. As seen in Table 3, including the residual covariance requires an additional line of syntax (aselfe2 with aselfe3;). An examination of the model fit for this second model reveals that the inclusion of the residual covariance between the two items significantly improved model fit. The CFI and TLI values are both above .95, SRMR is below .08, and RMSEA equals .06. Together, these model fit statistics indicate good model fit for the one-factor measurement model of the academic self-efficacy scale that includes the residual covariance between items two and three. Inspection of modification indices indicates there are no remaining salient focal areas of ill fit.

An examination of the completely standardized factor loadings in this revised measurement model indicates that all eight of the items of the academic self-efficacy scale have moderate to large factor loadings (range = .43 to .87). The square of these loadings represents the proportion of the variance in the indicators explained by the latent constructs. Thus, the magnitude of these loadings indicates that a moderate proportion of the variance in the indicators could be explained by the latent variable of academic self-efficacy. It appears that the eight items are all adequate indicators of academic self-efficacy. Furthermore, an examination of the residual covariance parameter estimate indicates that this relationship was

Table 3. Mplus syntax and selected output of confirmatory factor analysis of the academic self-efficacy scale with residual covariance specified between items 2 and 3

SYNTAX:

```
TITLE: Academic Self-Efficacy Confirmatory Factor Analysis with Residual
   Covariance
DATA: acaselfeff.dat;
VARIABLE:
   NAMES ARE id gpagoal gpa4year aselfe1-aselfe8;
   USEVARIABLES ARE aselfe1-aselfe8;
   MISSING are all (-9);
ANALYSIS: ESTIMATOR IS ML;
MODEL:
   acaeffic by aselfe1-aselfe8;
   aselfe2 with aselfe3;                    ! specify residual covariance
OUTPUT: STANDARDIZED;
```

SELECTED OUTPUT:

TESTS OF MODEL FIT

Chi-Square Test of Model Fit	
Value	36.638
Degrees of Freedom	19
P-Value	0.0088
CFI/TLI	
CFI	0.974
TLI	0.962
RMSEA (Root Mean Square Error Of Approximation)	
Estimate	0.064
90 Percent C.I.	0.031 0.094
SRMR (Standardized Root Mean Square Residual)	
Value	0.036

MODEL RESULTS

	Estimate	S.E	Est./S.E.	Two-Tailed P-Value
ACAEFFIC BY				
ASELFE1	1.000	0.000	999.000	999.000
ASELFE2	0.857	0.129	6.644	0.000
ASELFE3	0.973	0.134	7.256	0.000
ASELFE4	0.799	0.144	5.566	0.000
ASELFE5	1.153	0.126	9.179	0.000
ASELFE6	1.278	0.145	8.835	0.000
ASELFE7	0.861	0.126	6.852	0.000
ASELFE8	0.906	0.116	7.838	0.000

ASELFE2 WITH				
ASELFE3	0.413	0.098	4.222	0.000
COMPLETELY STANDARDIZED MODEL RESULTS				
				Two-Tailed
	Estimate	S.E	Est./S.E.	P-Value
ACAEFFIC BY				
ASELFE1	0.572	0.049	11.683	0.000
ASELFE2	0.529	0.052	10.140	0.000
ASELFE3	0.603	0.047	12.942	0.000
ASELFE4	0.429	0.058	7.377	0.000
ASELFE5	0.870	0.023	37.511	0.000
ASELFE6	0.849	0.025	34.235	0.000
ASELFE7	0.564	0.049	11.422	0.000
ASELFE8	0.677	0.040	16.778	0.000
ASELFE2 WITH				
ASELFE3	0.317	0.063	5.059	0.000

statistically significant. All subsequent models therefore include the residual covariance between items two and three.

Extending to Structural Equation Modeling with an Outcome

After establishing an appropriate measurement model, the next step would be to begin examining the relationship between academic self-efficacy and GPA. In this situation, cumulative GPA is a manifest variable outcome and we therefore do not include an intermediate model in which the measurement model for the outcome is also evaluated. The syntax and selected output from the structural equation model in which we examine the effect of academic self-efficacy on cumulative college GPA four years later is presented in Table 4. As seen in Table 4, the only modifications to the syntax required to specify this model is to add the GPA variable to the usevariables list and to add an additional line of syntax (gpa4year on acaeffic;). The inclusion of GPA as an outcome and the specification of the effect of academic self-efficacy on GPA did not worsen fit: as with the previous model, CFI and TLI values are both above .95, SRMR is below .08, and RMSEA equals .06. The results of this model indicate that academic self-efficacy is a significant predictor of cumulative college GPA four years later. The unstandardized effect was B = .239, SE = .055, $p < .001$. The completely standardized effect was $\beta = .415$ and academic self-efficacy predicted 17.2% of the variance in cumulative college GPA. The completely standardized results of this model are presented in Figure 2. These results support

Table 4. Mplus syntax and selected output of structural equation model examining the effect of academic self-efficacy scale on four-year college grade point average

SYNTAX:

```
TITLE: Academic Self-Efficacy SEM with 4year GPA as outcome
DATA: FILE IS acaselfeff.dat;
VARIABLE:
   NAMES ARE id gpagoal gpa4year aselfe1-aselfe8;
   USEVARIABLES ARE gpa4year aselfe1-aselfe8;
   MISSING are all (-9);
ANALYSIS: ESTIMATOR IS ML;
MODEL:
   acaeffic by aselfe1-aselfe8;
   aselfe2 with aselfe3;
   gpa4year on acaeffic;    !Estimate effect of Academic Self-Efficacy on GPA
OUTPUT: STANDARDIZED;
```

SELECTED OUTPUT:

TESTS OF MODEL FIT

Chi-Square Test of Model Fit	
Value	44.345
Degrees of Freedom	26
P-Value	0.0139
CFI/TLI	
CFI	0.974
TLI	0.964
RMSEA (Root Mean Square Error Of Approximation)	
Estimate	0.056
90 Percent C.I.	0.025 0.083
SRMR (Standardized Root Mean Square Residual)	
Value	0.039

UNSTANDARDIZED MODEL RESULTS

	Estimate	S.E	Est./S.E.	Two-Tailed P-Value
GPA4YEAR ON				
ACAEFFIC	0.239	0.055	4.327	0.000

COMPLETELY STANDARDIZED MODEL RESULTS

	Estimate	S.E	Est./S.E.	Two-Tailed P-Value
GPA4YEAR ON				
ACAEFFIC	0.415	0.078	5.337	0.000

R-SQUARE				
Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
GPA4YEAR	0.172	0.065	2.669	0.008

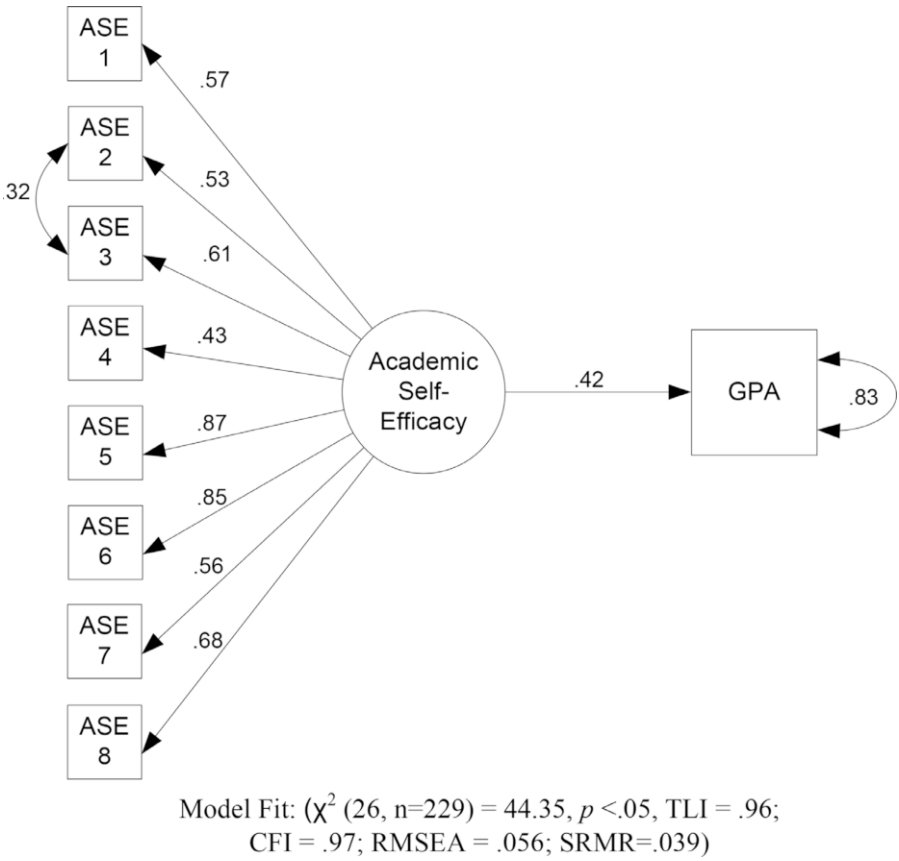


Figure 2. Example figure for presenting SEM results. Results correspond to the completely standardized results in Table 4.

the hypothesis that academic self-efficacy is a predictor of academic outcomes and provides the basis for examining potential mechanisms of the effects of academic self-efficacy on cumulative GPA.

Evaluating a Mediation Model

The final example model is a mediation model in which we examine whether the effects of academic self-efficacy on cumulative college GPA four years later are partially mediated by the GPA goals students set during their first semester of college. The syntax and selected output from the SEM in which we examine the indirect effect of academic self-efficacy on cumulative college GPA four years later via GPA goals are presented in Table 5. As seen in Table 5, the specification of this mediation model requires just a few minor additions to the syntax of the previous SEM model. The usevariables line is modified to include the additional variable of

Table 5. Mplus syntax and selected output of structural equation model examining the indirect effect of academic self-efficacy on four-year college grade point average via gpa goals in 1st semester

SYNTAX:

```
TITLE: Mediation Model: Aca Self-Efficacy → GPA Goal → 4yearGPA
DATA: FILE IS acaselfeff.dat;
VARIABLE:
   NAMES ARE id gpagoal gpa4year aselfe1-aselfe8;
   USEVARIABLES ARE gpagoal gpa4year aselfe1-aselfe8;
   MISSING are all (-9);
ANALYSIS: ESTIMATOR IS ML;
MODEL:
   acaeffic by aselfe1-aselfe8;
   gpa4year on acaeffic;
   gpa4year on gpagoal;
   gpagoal on acaeffic;
   aselfe2 with aselfe3;
Model Indirect:      !specify estimation of indirect effect
   gpa4year ind gpagoal acaeffic;
OUTPUT: CINTERVAL STANDARDIZED;
```

SELECTED OUTPUT:

```
TESTS OF MODEL FIT
Chi-Square Test of Model Fit
   Value                    58.633
   Degrees of Freedom       33
   P-Value                  0.0039
CFI/TLI
   CFI                      0.965
   TLI                      0.952
```

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.058	
90 Percent C.I.	0.033	0.082

SRMR (Standardized Root Mean Square Residual)

Value	0.043

UNSTANDARDIZED MODEL RESULTS

				Two-Tailed
	Estimate	S.E.	Est./S.E.	P-Value
GPA4YEAR ON ACAEFFIC	0.172	0.043	3.996	0.000
GPAGOAL ON ACAEFFIC	0.097	0.024	3.995	0.000
GPA4YEAR ON GPAGOAL	0.413	0.115	3.595	0.000

COMPLETELY STANDARDIZED MODEL RESULTS

				Two-Tailed
	Estimate	S.E.	Est./S.E.	P-Value
GPA4YEAR ON ACAEFFIC	0.345	0.079	4.349	0.000
GPAGOAL ON ACAEFFIC	0.284	0.067	4.232	0.000
GPA4YEAR ON GPAGOAL	0.282	0.075	3.759	0.000

R-SQUARE

				Two-Tailed
Observed Variable	Estimate	S.E.	Est./S.E.	P-Value
GPAGOAL	0.081	0.038	2.116	0.034
GPA4YEAR	0.254	0.068	3.737	0.000

TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

				Two-Tailed
	Estimate	S.E.	Est./S.E.	P-Value
GPA4YEAR GPAGOAL ACAEFFIC	0.040	0.015	2.723	0.006

CONFIDENCE INTERVALS OF INDIRECT EFFECTS

	Lower .5%	Lower 2.5%	Estimate	Upper 2.5%	Upper. 5%
GPA4YEAR GPAGOAL ACAEFFIC	0.002	0.011	0.040	0.069	0.078

GPA goals, the effect of academic self-efficacy on GPA goals is specified (gpagoal on acaeffic;), the effect of GPA goals on cumulative GPA is specified (gpa4year on gpagoal;), the estimation of the indirect effect is requested by including "Model Indirect: gpa4year ind gpagoal acaeffic;", and CINTERVAL is added to the output line so that confidence intervals of the indirect effect can be evaluated to determine whether there is evidence of mediation. The model fit for this mediation model was good: CFI and TLI values are above .95, SRMR is below .08, and RMSEA equals .06. The results indicated that there was a significant indirect effect of academic self-efficacy on cumulative college GPA four years later via GPA goals. The estimate of the

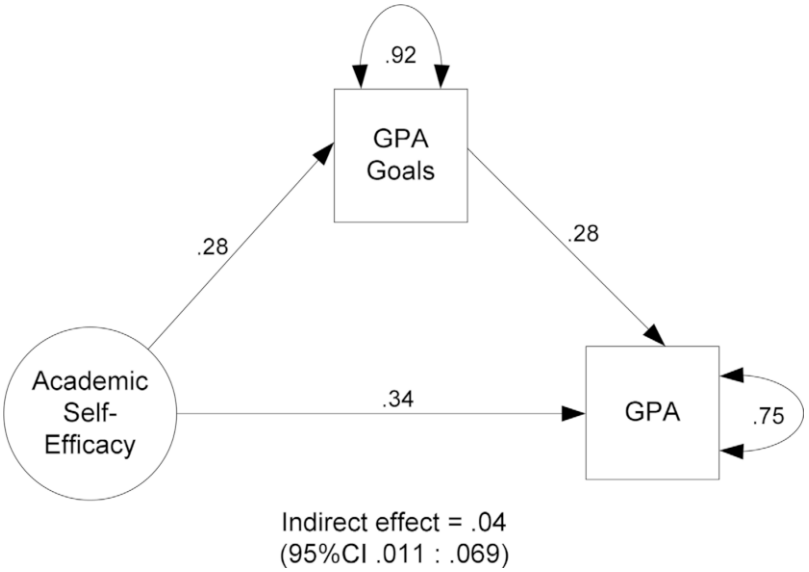


Figure 3. Example figure for presenting SEM mediation model results. results correspond to the completely standardized results in Table 5.

indirect effect was significant (B = .040, SE = .015, $p < .01$) and the 95% confidence interval of the indirect effect (.011 : .069) did not include 0. A path diagram with the completely standardized results of this mediation model can be seen in Figure 3. These results suggest that the academic self-efficacy beliefs may promote superior academic performance in college by causing students to set higher GPA goals for themselves. The theoretical implications of these results are not important for the purposes of this chapter, but the models described here and presented in Tables 2–5 provide an introduction to how CFA and SEM can be used in educational research.

SUMMARY

CFA and SEM are powerful statistical tools that have become increasingly popular in education research. The topics discussed within this chapter are just some of the many ways that these techniques can be used to evaluate measurement models and test complex theoretical models. The growth of these techniques has coincided with the development of more user-friendly statistical software for conducting these analyses and an increasing amount of publications providing didactic information about how these techniques can be applied to various research topics. Below we provide a few recommendations for resources that educational researchers may find helpful for additional information about how to apply these techniques in their own research programs.

SUGGESTIONS FOR FURTHER READING

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Byrne, B. M. (2011). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge Press.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201–226.

Teo, T., & Khine, M. S. (Eds.) (2009). *Structural equation modeling in educational research: Concepts and applications*. Rotterdam: Sense Publishers.

REFERENCES

Allison, P. D. (2003). Missing data techniques for structural equation models. *Journal of Abnormal Psychology, 112,* 545–557.

Arbuckle, J. L. (2010). *IBM SPSS Amos 19 User's Guide.* Crawfordville, FL: Amos Development Corporation.

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

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality & Social Psychology, 51,* 1173–1182.

Bentler, P. M. (1990). Comparative fit indices in structural equation models. *Psychological Bulletin, 28*, 97–104.

Bentler, P. M. (2006). *EQS for Windows (Version 6.0)* [Computer software]. Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606.

Bollen, K. A., & Curran, P. J. (2006). Latent curve models: A structural equation perspective. Hoboken, NJ: Wiley-Interscience.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford Press.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*, 230–258.

Burkholder, G. J., & Harlow, L. L. (2003). An illustration of a longitudinal cross-lagged design for larger structural equation models. *Structural Equation Modeling*, *10*, 465–486.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Chemers, M. M., Hu, L., & Garcia, B. (2001). Academic self-efficacy and first-year college student performance and adjustment. *Journal of Educational Psychology, 93*, 55–65.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Myths and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112,* 558–577.

Collins, L. M. (2006): Analysis of longitudinal data; The integration of theoretical model, temporal design, and statistical model, *Annual Review of Psychology 57,* 505–528.

Eliason, S. R. (1993), *Maximum likelihood estimation*. Newbury Park, CA: Sage.

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: Guilford Press.

Fox, J. (2006). Structural equation modeling with the sem package In R. *Structural Equation Modeling, 13*, 465–486.

Gallagher, M. W., & Lopez, S. J. (2008, August). *The unique effects of hope and self-efficacy on academic performance*. Poster presented at the American Psychological Association 116th Annual Convention, Boston, MA.

Hu, L., & Benter, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.

Kenny, D. A., & Kashy, D. A. (1992). The analysis of the multitrait-multimethod matrix using confirmatory factor analysis. *Psychological Bulletin, 112,* 165–172.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Little, T. D., Bovaird, J. A., & Card, N. A. (Eds.). (2007). *Modeling contextual effects in longitudinal studies*. Mahwah, NJ: Erlbaum.

Little, T. D., Bovaird, J. A., & Widaman, K. F. (2006). On the merits of orthogonalizing powered and interaction terms: Implications for modeling interactions among latent variables. *Structural Equation Modeling, 13*, 497–519.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling, 13,* 59–72.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201–226.

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

MacCallum, R. C., Roznowski, M., & Necowitz, L. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490–504.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis.* Mahwah, NJ: Erlbaum.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577–605.

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82.

Muthén, B. O., & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143–165). Boca Raton, FL: Chapman & Hall/CRC.

Muthén, B. O., & Muthén, L. K. (2008–2012). *Mplus user's guide.* Muthén & Muthén: Los Angeles, CA.

Muthén, L. K. and Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4,* 599–620.

Neale, M. C., Boker, S. M., Xie, G., Maes, H. H. (2003). *Mx: Statistical modeling. 6.* Department of Psychiatry, Virginia Commonwealth University/.

Preacher, K. J., & Coffman, D. L. (2006, May). Computing power and minimum sample size for RMSEA [Computer software]. Available from http://www.quantpsy.org/.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods, 40, 879–891.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling.* Thousand Oaks, CA: Sage Publications.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15,* 209–233.

SAS Institute (2005). *SAS/STAT User's Guide,* Version 9. Cary, NC: SAS Institute Inc.

Schafer. J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.

Schoemann, A. D. (2010). *Latent variable moderation.* Retrieved January 22nd, 2012, from http://www.quant.ku.edu/pdf/16.1%20latent%20variable%20moderation.pdf.

Selig, J. P., & Preacher, K. J. (2009). Mediation models for longitudinal data in developmental research. *Research in Human Development, 6,* 144–164

Steiger, J. H., & Lind, J. C. (1980). *Statistically-based tests for the number of common factors.* Paper presented at the annual Spring Meeting of the Psychometric Society in Iowa City. May 30, 1980.

Teo, T., & Khine, M. S. (Eds.) (2009). *Structural equation modeling in educational research: Concepts and applications*. Rotterdam: Sense Publishers.

Thurstone, L. L. (1947). *Multiple-Factor Analysis*. Chicago: University of Chicago Press.

DANIEL A. SASS & THOMAS A. SCHMITT

15. TESTING MEASUREMENT AND STRUCTURAL INVARIANCE

Implications for Practice

BACKGROUND ON INVARIANCE TESTING

Measurement validation in the behavioral sciences is generally carried out in a psychometric modeling framework that assumes unobservable traits/constructs (i.e., latent factors) created from the observed variables (often items measuring that construct) are the variables of interest. Unfortunately, many researchers compare these constructs across populations/groups (e.g., males & females) assuming that they have the same psychometric properties and association between unobserved and observed across groups of interest. In other words, researchers will often make the premature and untested assumption that the theoretical constructs they are interested in are *invariant* (or equivalent) from one group to another. When researchers assume a measure is invariant, they are failing to investigate whether the construct has *factorial invariance* (Byrne, Shavelson, & Muthén, 1989; Millsap, 1998). In particular, they fail to test whether the latent factor scores were generated in a similar fashion across groups, thus producing the same *metric* (unstandardized factor loadings) and *scalar* (intercept or threshold) parameters.

With the advent of computers and accessible latent modeling software, invariance testing within multi-group confirmatory factor analysis (MCFA) and multi-group structural equation modeling (MSEM) literature has increased considerably over the past 20 years (Meade, Johnson, & Braddy, 2008; Vandenberg & Lance, 2000). This has enabled researchers to more easily explore whether latent factors, along with the relationship between latent factors, are invariant across populations. Despite methodological guidelines, statistical procedures, and widely available software, researchers continue to struggle with the numerous decisions that need to be made when testing for invariance. As delineated below, several authors have assessed the utility and applicability of invariance procedures to provide guidelines including: 1) setting the measurement scale, 2) evaluating model fit and statistical power, and 3) estimating the appropriate model depending on the data characteristics. The present chapter aims to provide an up-to-date review of important considerations when conducting MCFA and MSEM analyses; especially as they relate to assessing whether or not the latent factors (i.e., metric and scalar parameters), latent factor mean scores, and structural coefficients are equal across populations. A demonstration

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 315–345.

will also be carry-out to illustrate how different statistical models can affect the estimation of parameters within invariance testing.

REVIEW OF TERMINOLOGY

One of the primary difficulties in conducting invariance analyses is deciphering the statistical jargon, especially because researchers sometimes use these terms differently. Although seemingly simple, the word "invariance" or the idea that something is "invariant" across groups is often a source of confusion. Invariance essentially implies that the parameters tested (whether it is factor loadings, means, structural coefficients, etc.) are equal across groups of interest. Therefore, when structural coefficients are invariant (or equal) it suggests that the same association between latent factors exists across groups (e.g., subjects in the treatment and control group). Invariance can also be conceptualized in the framework of moderation analysis, because if structural coefficients differ across groups (i.e., not invariant), group membership is thought to moderate this relationship. A source of confusion may arise from the fact that variability (the opposite of invariance) is what many researchers seek when comparing group means. However, in the context of MCFA, it is generally invariance of parameters across groups that researchers seek to meet the statistical assumptions of the test. This is similar to the homogeneity of variance test, where researchers want the variances to be equal across groups and the means to be different.

A critical point to remember is that invariance analyses compare the unstandardized coefficients, which are not frequently reported by researchers. For this reason, researchers and readers need to be aware that large standardized coefficient differences across groups are not always indicative of non-invariance, as the unstandardized coefficients are influenced by the observed variable and latent factor variances. Using more traditional statistical analyses as an example, a statistically significant difference between two bivariate correlations (e.g., the correlation between X and Y for males and females) may draw a different conclusion when comparing two unstandardized slopes across these groups using the same data. For this reason, it is critical that researchers and readers understand exactly what statistics (standardized vs. unstandardized) are being compared.

There exists several common invariance classification terms, with the first related to the measurement model. The term *measurement invariance* is frequently used when the configural invariance, weak factorial or metric invariance, and strong factorial or scalar invariance models are all deemed invariant. *Configural invariance* exists when the same model is estimated for each group simultaneously with the estimated parameters free to vary across groups (i.e., not constrained to be equal). In other words, the same model is estimated, but the parameter estimates are allowed to differ across groups. This model is important to establish a baseline for which more restrictive models can be compared, as all subsequent models are tested with increasingly more

restrictions. *Metric invariance* tests whether the unstandardized factor loadings are equal across groups, meaning that the association between observed (often items) and unobserved (latent factors) is relatively equal across groups.

Following tests of metric invariance, *scalar invariance* evaluates whether the observed variables metric (either intercepts or thresholds) are relatively equal across groups. Intercepts are examined when the researcher assumes the observed variables are continuous, whereas models assuming ordered categorical observed variables estimate the thresholds. Intercept invariance tests if the observed variable means are proportionally equal across groups, whereas threshold invariance tests whether the thresholds (or distribution cut points) are equal across the groups. Stated differently, threshold invariance occurs when the cut points on the unobserved normal distribution are equal across groups for each observed variable. It is worth mentioning that when researchers test for threshold invariance they assume the underlying observed variable's distribution is normal and has a continuous scale. For analysis purposes, it is also important to ensure that the number of observations per cell is sufficiently large to estimate the thresholds adequately. If not, researchers might need to recode their data to obtained adequate representation. Finally, invariance analyses require the same number of thresholds per observed variable across the groups. Therefore, if item 1 has response data in five categories for Group 1, then Group 2 must also have responses in all five categories. If not, the data will need to be recoded to create the same number of categories across the groups being compared.

After metric and scalar invariance has been obtained, researchers may test for *strict factorial invariance*, which tests whether the observed variable's residuals (a.k.a., uniquenesses or scale factors) are equal across the groups. These tests are possible when using a maximum likelihood estimator or weighted least squares mean and variance (WLSMV) estimator using the theta estimation method. However, WLSMV using the delta estimation method (Mplus default) fixes these scale factors at one for both groups and, therefore, does not allow for such analyses. In any case, many researchers argue that these comparisons are of less interest (Bentler, 2005; Widaman & Reise, 1997) and may not even be worth testing (Selig, Card, & Little, 2008). Researchers should also be aware of the abundance of comparisons that could be made within the measurement model (see Marsh et al., 2009; Vandenberg & Lance, 2000); however, we only provide the two most commonly evaluated aspects of the measurement model.

Although metric and scalar invariance (defined hereafter as only measure invariance) are required for valid latent factor mean score comparisons, they are also required to test for equality of covariance between latent factors and structural invariance. Past literature defines structural invariance in numerous ways. For clarity purposes, and to mirror the terminology used with single-group analyses, we use the following definitions. *Covariance invariance* analyses test whether the unstandardized relationship between latent factors is equal across groups, whereas *structural invariance* analyses test whether the unstandardized relationship between

latent variables (either correlational or predictive) is equal across groups. Thus, covariance invariance analyses are often tested within the confirmatory factor analysis (CFA) framework, whereas structural invariance analyses focus on the "causal relationships" within the structural equation modeling (SEM) framework. These analyses are critical when researchers seek to test whether relationships between latent factors, both predictive and non-predictive, are moderated by group membership within a theoretical model.

To provide a larger context, measurement invariance is also required when researchers desire to compare observed mean scores [e.g., general linear models (GLM)] or associations between observed variables [e.g., ordinary least squares (OLS) correlation and regression analyses] across groups. One shortcoming of these analytic approaches (e.g., GLM or OLS) is that they assume the variables are observed (i.e., measured without error), which automatically implies these variables are invariant. Unfortunately, simply assuming measurement invariance can result in inaccurate conclusions when this assumption is violated (Hancock, Lawrence, & Nevitt, 2000; McDonald, Seifert, Lorenzet, Givens, & Jaccard, 2002).

Comparing Latent Factor Means and Structural Coefficients

Although there has been an increase in studies comparing latent factor mean scores within a MCFA framework over the last decade (Millsap & Meredith, 2007), MSEM studies remain scarce. Testing measurement and structural invariance is essential because they (a) estimate and adjust for measurement error within each factor, (b) assess factorial validity, and (c) test whether the measurement and structural models are invariant across groups. Overall, this approach has the benefit of testing an assortment of statistical assumptions and research questions within a single modeling framework. Research (Hancock et al., 2000; McDonald et al., 2002) indicates that the statistical conclusions drawn from mean comparisons may differ, or be invalid, depending on the type of analysis conducted (e.g., GLM vs. MCFA) and whether measurement invariance is achieved. Thus, invariance testing is essential for making valid inferences across populations (see Meredith, 1993; Vandenberg & Lance, 2000).

While a number of methodological studies demonstrate the importance of having measurement invariance prior to testing the equality of latent factor means, the literature base is less abundant for MSEM studies. While researchers frequently test for covariance invariance, they rarely test for structural invariance. This is somewhat concerning because while the interfactor covariance might be invariant, the predictive relationship between variables might not be invariant after adjusting for other variables in the model. Thus, researchers should also test the invariance of the structural coefficients in cross-validation studies, to ensure that their theoretical (i.e., structural or path) model generalizes across either independent samples or different groups. Similar to mean comparisons, it is important to remember that measurement invariance is a prerequisite to testing structural coefficients across groups.

Potential Causes of Measurement Non-invariance

When a factor model varies across groups (whether due to bias or non-invariant), the metric (factor loadings) and/or scalar (intercepts or thresholds) parameters make differential contributions to the means, which prevents valid mean comparisons or relationship differences between groups (Meredith, 1993). Factor loadings can be thought of as the unstandardized weights resulting from regressing the observed variable on the latent factor, thus it represents the strength of the relationship between the factor and the observed variable. Metric non-variance results from observed variables making unequal slope contributions to the latent factor across groups. In other words, the slopes in the equation used to compute the factor scores differ across groups. This can result from a number of sources. The first is the conceptual interpretation (perhaps for cultural reasons) of the construct differing across groups; the second is the meaning of items changes when the scales are translated across languages and/or cultures; and third the response scale range or meaning differs across groups (Chen, 2008).

Intercept invariance tests whether or not the observed variable has the same intercept or origin across groups. Therefore, if a factor is invariant subjects with the same latent factor score should have similar responses on average for an observed variable. Note that when testing for intercept invariance, researchers assume that the observed variables (e.g., items) are continuous. Again, threshold invariance is similar to intercept invariance, except the observed variables are treated as ordered categorical variables (Millsap & Yun-Tein, 2004). Thresholds are the cutoff points on the unobserved normal distribution where, on average, respondents vary between two different response options. Thresholds divide the distribution into the number of categories minus one, thus a 5-point response scale contains four thresholds. Threshold invariance holds if the distribution cut points (i.e., cut-offs between response option categories) on the observed variable distributions are equal across groups. In general, metric non-invariance can result from (a) social desirability or social norm perceptions, (b) particular groups displaying a propensity to respond more strongly to an observed variable despite having the same latent factor mean, and/or (c) certain groups having different reference points when making statements about themselves (Chen, 2008).

Because numerous factors can contribute to factorial invariance, observed variable content should be inspected carefully when a measure is suspected to be non-invariant. These differences must be distinctive to a particular item or set of items, because when all items are equally influenced (e.g., biased) by the aforementioned factors, measurement invariance would likely hold, even though biases are still likely.

Historically, testing for measurement invariance has been encouraged as a prerequisite to compare latent factor means or structural coefficients (Millsap & Meredith, 2007). From a measurement perspective, the usefulness of invariance testing far exceeds simply allowing for valid comparisons across groups. Invariance analyses are extremely beneficial to understand when and how groups differ at the

observed variable level. Consequently, these analyses are useful to purify tests of measurement error and bias, while facilitating the understanding of observed variable group disparities. Along with examining factor loading magnitudes, measurement invariance should play an important role in measurement development. For example, an observed variable with a smaller factor loading is arguably better than an observed variable that functions differently (i.e., non-invariant) across groups. Therefore, researchers need to create and select scales that not only have a strong factor structure, but are also invariant across variables of interest.

CONSIDERATIONS WHEN TESTING INVARIANCE

To supplement the increase in MCFA and MSEM use, considerable empirical research has been conducted. These developments have focused on three major issues: (a) setting the common factor scale, (b) assessing model fit of invariant models, (c) determining the appropriate estimator and invariance approach, and (d) considerations for non-invariant measures. Each of these issues will be discussed below.

Setting the Factor Scale

When conducting an MCFA or MSEM model, researchers must choose how to identify the model or set the factor scale equal across groups for model standardization or identification reasons (see Cheung & Rensvold, 1999, 2000). For identifying the model and setting the scale, there are three main methods available: the reference-group method, the marker-variable method, and the effects-coding method (Little, Slegers, & Card, 2006). The reference-group method fixes the latent factor means (often fixed at zero) and latent factor variances (often fixed at one) across the groups. This approach is useful in that it allows for invariance testing on every observed variable's metric and scalar parameter. The limitation of this approach is it assumes homogeneity of latent factor variances, which may not be met. The marker-variable method is similar to the analysis of variance dummy-coding model and involves fixing one intercept of each latent factor to zero and its unstandardized factor loading to one. Thus, the variances of all latent factors are estimated with scales equivalent to the chosen marker variable. The limitation with this approach is it assumes that observed variable is invariant across the groups, but, again, this may be a false assumption. The effects-coding method constrains a set of unstandardized factor loadings and intercepts to sum to 0.0 and 1.0, respectively. This method estimates the latent factor means and variance based on the observed variables and weighted based on how well each observed variable represents the latent factor.

Although each identification method has its strengths and weakness, any of the three methods are appropriate when the statistical assumptions are met. Moreover, each procedure provides model fit, tests for various multiple group invariance analyses, yields comparable estimates of latent effect sizes, and permits the

computation of differences in latent factor means. The marker-variable method is more commonly employed to identify the model (Vandenberg & Lance, 2000), which may explain why it is the default in most SEM software packages. Regardless, this decision should not be made without considerable attention and analysis as to whether the appropriate assumptions are viable.

Empirical research using the marker-variable method suggests that reference indicators should not be selected arbitrarily, as a non-invariant indicator can influence the invariance conclusions (Cheung & Rensvold, 1999; French & Finch, 2008; Millsap, 2001; Steiger, 2002). Specifically, if the assumption of factor loading invariance is not met, invariance conclusions for other loadings may be incorrect resulting in biased parameter estimates and model fit statistics. This creates a paradoxical situation because the "most invariant unstandardized factor loading" cannot be determined without specifying the model, but model specification requires an invariant unstandardized referent factor loading (French & Finch, 2008). To circumvent this dilemma, Cheung and Rensvold (1999, 2001) developed a search procedure called the factor-ratio test, which French and Finch (2008) found to perform fairly well. The main shortcoming of this procedure is that its very time consuming to perform all the individual invariance tests. Recently, Cheung and Lau (in press) demonstrated a bias-corrected bootstrap procedure that simplified the search for a reference invariant observed variable to a single model using the factor-ratio test. An alternative procedure to identify the "most invariant" item is to set the latent factor variances equal across groups and then identify the "most invariant" observed variable based on the smallest change in chi-square or the modification indices. However, this procedure assumes homogeneity of latent factor variances, which may not be a valid assumption and may lead to the incorrect reference observed variable. Reiterating, constraining different observed variables, or the latent factor variances, can lead to contrasting invariance conclusions, thus the selection of this indicator should be scientific.

Assessing Model Fit

Sample and configural invariance model. Generally, the first step in invariance testing is to confirm factorial validity by testing the model fit separately for each group, while also evaluating the model parameters (e.g., size of the standardized factor loadings). After individual group model fit is obtained, a test of configural invariance is conducted to obtain a baseline model that can be compared to more restrictive invariance models. As with any model, an assortment of model fit statistics should be evaluated that consider the various model components (model complexity, sample size, etc.) that may influence the results. When using maximum likelihood (ML) estimation, the configural invariance model fit is equal to the sum of the c^2 statistics for the individual group analyses. Likewise, the sum of the *df* for the individual group analyses should equal the *df* of the configural invariance model. These sums are useful to compute as they can helpensure the configural

model was properly specified. It is important to note that for WLSMV or robust ML estimation only the *df* is summative and the c^2 is not summative. Regardless, researchers should carefully evaluate the all model statistics to ensure their models are correctly specified. In summary, the configural invariance model tests the extent to which the underlying structure fits the data with no between group constraints; thus, prior to determining whether parameter estimates are equal across groups, the configural invariance model must fit reasonably well.

To evaluate model fit, the individual groups and configural invariance model often consider the c^2 statistic, Comparative Fit Index (CFI), Tucker-Lewis index (TLI), Root Mean Square Error Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR) with ML or robust ML (MLR) estimation. WLSMV uses the sample model fit statistics, but replaces the SRMR with the Weighted Root Mean Square Residual (WRMR). The c^2 statistic is valuable because it allows researchers to make a statistical inference related to model fit in the population, while being the only true test statistic of model fit. However, researchers need to be aware that the c^2 is often overly sensitive to model rejection with large sample sizes and/or complex models (Saris, Satorra, & van der Veld, 2009). Due to this limitation, it is common to place less emphasis on the c^2. That being said, it is still important that researchers evaluate models with a significant c^2 to ensure that their models are not severely misspecified (Barrett, 2007; McDonald & Marsh, 1990; McIntosh, 2007; Saris et al., 2009).

Approximate fit indices (AFI), such as the CFI, TLI, RMSEA, & SRMR/WRMR, are appealing because they adjust for sample size and model complexity, but are limited because it is difficult to establish guidelines as to what constitutes "good model fit." These guidelines are also very subjective and have been shown to perform poorly and/or vary across different models (Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004). In any case, researchers commonly use criteria by Hu and Bentler (1999) to deem a model as having a good fit: CFI and TLI > 0.95, RMSEA < 0.06, and SRMR < .08.

Invariance models. The difficulties in evaluating fit for invariance models are well-known (see Chen, Sousa, & West, 2005). While the Δc^2 allows a statistical comparison between nested models, this test statistic presents the same concerns (i.e., sensitivity to sample size and model complexity) as the c^2 statistic (Chen, 2007; Marsh & Hocevar, 1985). However based on personal experience and research (Sass, Schmitt, & Marsh, in press), this test statistic often works well to detect evidence of non-invariance, which can later be evaluated for practical significance (i.e., the amount of differences in estimated parameters across groups and whether inferences related to latent factor mean scores and structural coefficients are influenced).

One important distinction between ML, MLR, and WLSMV estimation methods is that the traditional $\Delta\chi^2$ (e.g., measurement invariance model χ^2 minus configural invariance model χ^2) cannot be employed for nested models with WLSMV and MLR. Instead, researchers need to employ the DIFFTEST procedure in Mplus for

WLSMV and the Satorra-Bentler scaled difference test for MLR (Satorra & Bentler, 2001). The difference testing procedure using the MLR χ^2 can be computed using the output results from Mplus (see Brown, 2006, pp. 379–387; Mplus web site for more details); although, the strictly positive Satorra-Bentler $\Delta\chi^2$ should be used when the values are negative (see Satorra & Bentler, 2010). As this tends to be an area of confusion, it is worth mentioning that the $\Delta\chi^2$ between two nested models will not equal the traditional Δc^2 when using MLR and WLSMV estimation.

To supplement the Δc^2, the change in approximate fit indices (ΔAFI, i.e., ΔCFI, ΔTLI, ΔRMSEA, ΔSRMR, & ΔWRMR) are commonly considered when doing invariance testing. To date, several studies (Chen, 2007; Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008; Sass et al., in press) have been conducted to evaluate appropriate ΔAFI guidelines. Although provided in more detail within each article, the following cutoff criteria were proposed by Chen (2007): reject ΔCFI < -0.01, ΔRMSEA > 0.01, and ΔSRMR > 0.015. Meade et al.'s (2008) cutoff criteria were similar to Chen's (2007), but had the following differences: reject ΔCFI < -0.002 and ΔRMSEA > 0.007. The ΔTLI and ΔWRMR standards have not been established to date; however, Marsh et al. (2010) argued that a ΔTLI closer to zero may be more appropriate. Our simulation study (Sass et al., in press) using ordered categorical data suggested that the ΔWRMR should not be used due to its sensitivity to sample size and model complexity. In fact, all the ΔAFI with WLSMV should be used with extreme caution, as the WLSMV estimator does not allow for a direct comparison between models (Sass et al., in press).

Given the limitations of the c^2 (or Δc^2) and approximate fit indices, researchers often need to make a subjective decision associated with what constitutes "good model fit." As a result, researchers must provide an impartial and evidence based assessment of whether invariance exists. Although there is no universal approach for evaluating model fit in invariance testing, the approach taken here was to consider the: (a) statistical significance of the Δc^2, (b) change in approximate fit statistics, and (c) magnitude of difference between the parameter estimates.

Steps for Testing Measurement Invariance

Although there are numerous types of invariance models that can be discussed (Chen et al., 2005; Cheung & Rensvold, 2000; Marsh, 1985; Vandenberg & Lance, 2000), we only consider the steps for testing parameters (i.e., metrics & scalars) required to be invariant for valid mean and correlation/regression comparisons. Many methodologists and applied researchers employ the following set of steps for testing invariance: individual group models, CI model, metric invariance model, and scalar invariance model. Muthén and Muthén (1998–2010, pp. 433–435) have argued for testing the factor loadings and intercepts (or thresholds) in tandem given that they both influence the item characteristic curve (ICC) simultaneously. Furthermore, any indication of non-invariance, whether they are caused by the factor loadings or intercepts (or thresholds), is concerning for item quality and the source of this

non-invariance can be detected with follow-up analyses. For these reasons, we tend to constrain the metric and scalar parameters to be invariant in tandem when testing for invariance.

A counterargument to testing factor loadings and intercepts simultaneously is that factor loadings and intercepts (or thresholds) influence different facets of the ICC. For this reason, researchers may elect to test them sequentially (factor loading before the intercepts/thresholds) or follow one of the many other procedures (e.g., see Marsh et al., 2009; Vandenberg & Lance, 2000). Regardless, researchers have the option of testing an array of other equality constraints (e.g., residual variances, interfactor covariance, etc.) that may prove fruitful in understanding measurement or structural differences across groups. The most common approach is the *forward approach* (sequentially adding more model constraints), whereas the *backward approach* (sequentially removing model constraints) appears less common (Dimitrov, 2010).

Considerations for Non-invariant Measures

Researchers who test for measurement invariance are frequently presented with the arduous task of deciding how to utilize non-invariant measures. After a measure is defined as non-invariant from a statistical ($\Delta\chi^2$) and/or practical (ΔCFI, ΔTLI, ΔRMSEA, & ΔSRMR) perspective, there are several options for treating these measures. These options include 1) deleting the non-invariant observed variables and only use invariant observed variables for statistical analyses, 2) applying a partial measurement invariance model, 3) using all the observed variables and assume any differences are small and do not influence the results, 4) interpreting the scores independently and preclude group comparisons, and/or 4) avoid using the scale (Cheung & Rensvold, 1999; Millsap & Kwok, 2004).

For practical applications, the first three options seem most appropriate for many circumstances. Option 1 works well with longer measures when the removal of observed variables (or items) does not adversely affect the measure's psychometric properties and the researcher does not desire to compare scores to the normative sample. Caution should be employed with this approach for widely used measures, as the results might not generalize and previous psychometric conclusions might be altered. Option 2, the partial measurement invariance model (only invariant metric and scalar parameters are equal across groups, with non-invariant parameters free to vary across groups) (see Byrne, Shavelson, & Muthén, 1989; Millsap & Kwok, 2004 for more details), only constrains invariant items to be equal across groups, while relaxing the constraints for non-invariant items. This approach is problematic because when certain items are non-invariant factor scores are not perfectly comparable. However, if the number of non-invariant items is small compared to total number of items, or the overall amounts of non-invariance is small, the latent factor means used for group comparisons should not be drastically impacted. Option 3 may also be feasible for longer measures when the degree of non-invariance is minimal

and the majority of observed variables are invariant. Regardless of the approach, researchers should use caution when interpreting the findings with non-invariant observed variables. Researchers are also encouraged to compare the results under different assumptions or options to assess the impact of non-invariant variables. For example, the researcher could consider comparing the latent factor means for the measurement invariance and partial measurement invariance models to uncover whether it significantly influences practical or statistical significance.

ILLUSTRATION

This illustration demonstrates the five most fundamental and commonly investigated research questions by applied researchers, while also providing some direction and recommendations when conducting invariance analyses. Stated more generally, the research questions are as follows: 1) does the collection of measures/scales of interest in the SEM possess factorial validity and measurement invariance across groups of interest (in our case, teacher education level); 2) do the latent factor mean scores differ across these groups; 3) does the structural model fit the data; and 4) are the structural coefficients equal across these groups?

As demonstrated below, these questions allow for a relatively thorough understanding of one's measurement and structural model. This includes common statistical analyses conducted within the SEM literature (i.e., tests of whether the measurement and structural model fit the data with large and statistically significant model parameters) and whether these model parameters are equal (i.e., possess a similar measurement and structural model) across groups. As demonstrated elsewhere (Byrne, 2012), numerous other model restrictions can be tested to explore other interesting and important research questions; however, these analyses were not demonstrated here because we wanted to provide a more focused approach of what we consider the central aspects of invariance testing.

Description of Sample and Instruments

For this illustration, data were collected from 617 certified teachers working in three public school districts in the southwestern U.S. Most teachers were females (78.0%), employed by a suburban (61.3%) or urban (30.3%) school districts, White (57.5%) or Hispanic (36.4%) descent, and earned either Bachelors (61.4%) or Masters (38.4%) degrees. Teachers (n = 9) who earned their doctorate were not included in these analyses. Teachers' age ranged from 22 to 78 (M = 41.64, SD = 10.78) and taught at the following grade levels: elementary (43.3%), middle (35.7%), and high (21.1%) school. Years of teaching experience ranged from one to 44 years (M = 13.14, SD = 13.14) and most teachers (86.5%) indicating having classroom management training and/or instruction during the past five years.

Martin, Sass, and Schmitt (2012) provide more details on the research design, data collection, measures selected, and justification for the model tested here (note,

that only five of the eight variables were used for presentation purposes). Although Martin et al. (2012) supply additional information associated with the measures used in the model, a brief description of each scale employed is provided.

Taken from the Behavior and Instructional Management Scale (BIMS, Martin & Sass, 2010), the six item Instructional Management (IM) scale uses a 6-point response scale from "strongly agree" to "strongly disagree" and measures "teachers' instructional aims and methodologies, and includes aspects of monitoring seatwork, structuring daily routines, and the use of lecture and student practice." The internal consistency reliability coefficient was 0.78. It should be noted that due to sparse cell counts with responses of 5 and 6, these data were recoded to 4. Justification for this is provided below.

From the Teacher Stressor Scale (TSS, Hui & Chan, 1996), the 9-item Student Behavior Stressors (SBS) scale that measures teachers' perceived student-related concerns (i.e., lack of student motivation, working with students of mixed ability) was used. Using a 5-point response scale from 1 (no stress) to 5 (extreme stress), the internal consistency reliability was 0.92.

The Maslach Burnout Inventory – Educator Survey (MBI-ES; Maslach et al., 1996) provided measures of Emotional Exhaustion (EE) and Personal Accomplishment (PA). Using a response scale from 0 (never) to 6 (every day), Emotional Exhaustion and Personal Accomplishment assess teachers level of stress related to teaching and their ability to have a positive impact on their students, respectively. The internal consistency reliability coefficients were 0.92 and 0.79 for the Emotional Exhaustion and Personal Accomplishment scale, respectively.

The *Job Satisfaction* (McLaney & Hurrell, 1988) measure evaluated teachers overall level of job satisfaction and whether they would prefer a different job or recommend their job to others. Using a 6-point response scale from 1 (strongly disagree) to 6 (strongly agree), this four item measure had an internal consistency reliability coefficient of 0.84.

STATISTICAL ANALYSES

Below we outline several statistical considerations when conducting invariance analyses, along with useful research questions that researchers might be interested in exploring. While several statistical models could be used for invariance testing (see the special issue in the *Journal of Psychoeducational Assessment*, 2011, vol. 9), this illustration focused on invariance testing within a CFA framework using MLR and WLSMV estimation.

Model Estimation and Identification

It is known among experienced data analysts that the model estimator can significantly alter the research findings, not to mention the interpretation of the results. For this reason, data were analyzed using MLR and WLSMV estimation methods to

illustrate how the interpretation differs and perhaps conclusions. Although ML is perhaps the most popular estimation method, MLR and WLSMV are preferable in this case given that the data are slightly skewed (Lubke & Muthén, 2004; Yuan & Bentler, 2000). It is also worth noting that simulation research comparing ML and MLR within a measurement invariance framework have shown these estimators tend to produce similar change in approximate fit indices (Sass et al., in press).

For illustration purposes, statistical analyses were conducted within Mplus 6.11 (Muthén & Muthén, 1998–2012) using default estimation procedures for each model estimator. Therefore, data were treated as continuous (i.e., a covariance matrix was analyzed) for MLR estimators and categorical (i.e., a polychoric correlation matrix was analyzed with variances on the diagonals) for WLSMV using delta parameterization. The small percent of missing data (0.03%) was also treated using the default procedures within Mplus, which is a full-information maximum likelihood method for MLR and a pair-wise procedure for WLSMV (see Asparouhov & Muthén 2010).

When testing invariance with the WLSMV estimator, the number of thresholds (or response scale) per item must be equal across comparison groups. Thus, researchers should consider recoding their data into fewer categories if: 1) the number of observations per cell is too small and might results in poor parameter estimation or 2) the data do not cover the range of response options equally across groups (e.g., group 1 has data for 5 response options, whereas group 2 only has data for 4 response options). Due to sparse cell counts in categories 5 and 6 on the Instructional Management subscale for the current data, these items were recoded to a 4-point rather than 6-point response scale. The same recoded dataset was used for both estimators.

When conducting invariance tests, the scale must be set for each factor across groups for model identification purposes (Cheung & Rensvold, 1999, 2000; Little, Slegers, & Card, 2006). In the present study, we tested the homogeneity of variance assumption across education levels using the mean scale. Our analyses revealed no statistically, or practically, significant difference between group variances after a Bonferroni adjustment for Type I error. The largest difference in variances between the bachelors (BA) and Masters (MA) degree groups was on the Job Satisfaction scale, $F(1,613) = 4.167$, $p = .042$, with the difference in variances being rather small ($SD_{BA} = 1.22$ vs. $SD_{MA} = 1.13$) from a practical perspective. For this reason, the present study fixed all the factor variances at one to identify the model, thus allowing for a test of invariance for each observed variable. Note, the mean scale variances (e.g., 1.22 & 1.13) were not fixed at one, but instead the latent factor variances.

Overall Model Fit Criteria

Determining whether the overall and subgroup (e.g., BA & MA) models fit the data is the first step in invariance testing. It is an essential step because a misspecified initial model may have a substantial impact on subsequent tests of invariance and

final model fit. A misspecified model would make it very difficult, if not impossible, for a researcher to justify theoretically or statistically that it is appropriate to test for invariance because the model is incorrect. For this study, we evaluated the model fit for overall, BA, and MA sample using the following fit statistics: χ^2, CFI, TLI, and RMSEA. Hu and Bentler (1999) tentatively indicated that CFI and TLI statistics greater than 0.90 as an adequate fit, with values greater than 0.95 as a good fit. RMSEA values less than 0.08 and 0.06 were tentatively defined as mediocre and good, respectively.

Invariance Model Fit Criteria

The complications associated with assessing the change in model fit from a less restrictive (e.g., configural invariance model) to a more restrictive model (e.g., measurement invariance model) in a multi-group analysis possess the same concerns as with single-group analyses (Saris, Satorra, & van der Veld, 2009). For this study, we interpreted the change in model fit from both a statistical ($\Delta\chi^2$) and practical/approximate (ΔCFI, ΔRMSEA, & ΔSRMR) perspective. When using the $\Delta\chi^2$ with WLSMV, the DIFFTEST procedure within Mplus must be used to obtain valid test statistics, whereas the Satorra-Bentler scaled χ^2 difference test for MLR procedure should be used when employing MLR (Satorra & Bentler, 2001, 2010). Again, only ML estimation allows for a direct comparison in χ^2 statistics, thus one cannot sum or subtract the χ^2 with MLR and WLSMV.

When considering the practical/approximate perspective in determining model fit, several considerations should be taken depending on the model estimator. With MLR, researchers can use the criteria established by past research with ML (e.g., Chen, 2007) to evaluate the change in practical/approximate model fit (Sass et al., in press). Of course, these criteria should not replace sound judgment. For this study, we used the follows criteria to evaluation a satisfactory change in practical/approximate model fit statistics: ΔCFI/ΔTLI < 0.01 and ΔRMSEA < 0.015 for tests of factor loading invariance and ΔCFI/ΔTLI < 0.01 and ΔRMSEA < 0.015 for tests of intercept invariance.

The WLSMV estimator adjustment does not allow for a direct comparison between models and, therefore, the ΔCFI, ΔTLI, and ΔRMSEA can be very biased for misspecified model (Sass et al., in press). For this reason, the change in practical/approximate model fit statistics should be interpreted cautiously and greater emphasis should be placed on the $\Delta\chi^2$ under these circumstances.

Research Questions

The analyses conducted below provide an illustration of five research questions that many social science researchers would find useful to understand: 1) factorial validity, 2) measurement invariance, 3) equality of means, 4) quality of the SEM model with the full sample, and 5) equality of structural relationships between groups. While

comparing means is extremely prevalent within research, fewer studies test whether the relationships between variables are equal across groups, and even fewer studies ensuring these comparisons (i.e., mean or relationship differences between groups) are valid by first testing for measurement invariance. These important questions are not limited to any specific social science discipline, as all studies using latent variables should first test whether construct or factor scores are measuring the same thing across comparison groups of interest. For this study, we present the following five research questions for illustration purposes and test whether the results differ based on the estimator selected.

Research question 1. To answer the first research question, a five-factor CFA model was estimated using the entire sample of teachers (i.e., both Bachelors and Master's degree teachers). These model results provide the first indication of factorial validity (i.e., the factor structure matchers the proposed theoretical model) and whether the overall CFA model fits the data well. The next models test the same CFA model, but for each group separately. Without good fitting initial models that accurately represent the factor structure, all subsequent analyses would likely be biased and provide incorrect results. For this illustration, model modifications were made to improve the overall CFA model fit, and, if needed, prior to proceeding to SEM. Although there are some concerns with this approach (see Asparouhov & Muthén, 2009; Browne, 2001; MacCallum, Roznowski, & Necowitz, 1992), it is generally acceptable if the number of changes is small and the modifications can be theoretically justified. After obtaining a good fitting CFA model for the entire sample and the bachelors and masters samples, the configural invariance model is evaluated to provide the baseline model by which the other invariance models are compared. Collectively, these analyses provide tests of factorial validity prior to testing the invariance models and test the following research questions: Does the full teacher sample, Bachelor's degree teacher sample, Master's degree teacher sample, and configural invariance model possess a good model fit with large standardized factor loadings?

Research question 2 Assuming acceptable factorial validity can be obtained, the next step is testing the measurement invariance (metric and scalar invariance in tandem) of the factor model. Again, many researchers test for metric (unstandardized factor loadings) and scalar (thresholds or intercepts) invariance sequentially (unstandardized factor loadings followed by either threshold or intercept invariance); however, this illustration elected to test the omnibus effect (i.e., both metric and scalar in tandem) and only test for partial measurement invariance if the change in model fit was statistically and practically significant. Although item residual (or uniqueness) invariance could also be assessed when using MLR estimation, the scaled latent factors are fixed at one for WLSMV using delta parameterization across the groups. Due to the contrasting approaches, these analyses were not conducted within the current illustration. However, Byrne (2012) provided the Mplus code for these

analyses for interested researchers. Regardless of whether or not the residuals are tested, obtaining evidence of measurement invariance is critical to compare other statistics (means, correlations, etc.) across variables of interest. In the case of this study, we tested the following research question: Are the metric (i.e., unstandardized factor loadings) and scalar (intercept and threshold) parameters invariant across teachers with a bachelors and masters degrees.

Research question 3. The next invariance research question focuses on the invariance (or equality) of means across groups. When testing the equality of latent factor mean scores with a structural modeling framework, researchers can either 1) constrain both latent factor means to zero and assess the $\Delta\chi^2$ (this is analogous to an overall *F*-test) or 2) constrain one group's means to zero and estimate the other groups. The latter approach, which we used here, compares the groups on each latent factor using a *t*-statistic and is analogous to dummy coding. If the measurement model is non-invariant, we recommend conducting these analyses using the measurement invariance and partial measurement invariance models to identify the degree to which the results change. Regardless, this study will test the following research question: Is there a significant difference in latent factor means between teachers with bachelors and masters degrees on the five latent factors tested in Figure 1? Another way of stating this research question is as follows: Are the five latent factor means invariant across teachers with bachelors and masters degrees?

Research question 4. Assuming that the CFA model fits the data well using the entire sample, the next step is testing whether the structural model fits the data for the entire sample. To assess the SEM model, it is useful to compare the CFA and SEM model fit statistics to determine the degree of model misfit that resulted from estimating this more constrained model. Similar to the CFA using the entire sample, it may be useful to test whether the entire sample fits the data well before testing the model using the subsamples. It is worth noting that researchers could elect to test research question 5 first; however, we find it useful to diagnosis any potential concerns associated with the structural model before testing for structural invariance. Essentially, research question 4 answers the following questions: Does the proposed theoretical model adequately fit the data and are the structural coefficients large and statistically significant?

Research question 5. Once the measurement model is tested for invariance and the structural model fits the data for the entire sample, the next model tested estimates a non-invariant structural model with measurement invariance (or partial measurement invariance if needed). For this model, the structural coefficients are not constrained to be equal across the groups, whereas the metric and scalar parameters are set as invariant across groups. This model provides a baseline model when comparing the structural invariance. Assuming this baseline model fits the data well, researchers can constrain all the structural parameters to be equal (or invariant) across the

groups to assess change in model fit. If the $\Delta\chi^2$ is non-significant researchers can conclude the structural coefficients are equal (i.e., invariant or not moderated) across groups. When the $\Delta\chi^2$ is statistically significant, researchers can use theory and the modification indices to identify those structural coefficients that differ across groups. Regardless, the research question tested here is as follows: Is the structural model invariant across groups?

DATA EXAMPLE RESULTS

The steps and logic behind measurement invariance testing with MLR (or ML) estimation is nearly identical to that of WLSMV. However, a few important differences are worth reiterating. First, MLR assumes continuous observed variables. Therefore, intercept (rather than thresholds with WLSMV) equality is tested, along with unstandardized factor loading equality. In this case, the Δdf is equal to two times the number of observed variables (one Δdf for both the intercept and factor loading) that were constrained to be invariant. The Δdf differs considerably from WLSMV, which in addition to the one Δdf for each unstandardized factor loadings it also has to c-1 Δdf (c is the number of categories for that observed variable). For example, if an item has a 5 point scale the change in *df* will be five (one Δdf for the factor loading and four Δdf for the thresholds). A second important difference is the assumptions made about the data and how these models are estimated (for more details see Rhemtulla, Brosseau-Liard & Savalei, 2010). For this reason, researchers need to carefully consider what assumptions are being made about the data and which model is most appropriate. Third, MLR requires researchers to use either the robust $\Delta\chi^2$ (Satorra & Bentler, 2001) or the strictly positive robust $\Delta\chi^2$ (Satorra & Bentler, 2010), whereas WLSMV requires the use of the DIFFTEST within Mplus.

WLSMV Estimation

To provide a forecast of the individual model results, we first estimated the CFA and SEM model for the full sample (n = 617) with WLSMV to examine whether the model fits the data well and possesses statistically and practically significant parameter estimates (e.g., factor loading, structural coefficients). Results indicated that both models fit the data well (see Table 1), while also displaying large significant standardized factor loadings (all larger than .50) and structural coefficients (all larger than |.35|). However, the $\Delta\chi^2$ using the DIFFTEST procedure suggests that the SEM model fits significantly worse than the CFA model (i.e., $\Delta\chi^2(5) = 78.54$, $p <. 0001$) even though the ΔCFI, ΔTLI, and ΔRMSEA is relatively small. Thus, it is possible that a partial mediation model would provide a better representation of the data than a full mediation model for certain paths in the model. To improve understanding of this model, researchers should test the direct and indirect effects and evaluate the modification indices. While outside the scope of this illustration, the modification

indices suggest that the path between personal accomplishment and emotional exhaustion should be estimated.

After testing the full sample for CFA and SEM, we then tested these models for teachers with BA ($n = 379$) and MA ($n = 238$) degrees separately and found that both models fit the data with relatively small differences in model fit between the groups. Therefore, it was statistically appropriate to estimate the CI model. To ensure the correct CI model was estimated, researchers may use these two helpful hints. First, the sum of the *df* for each sample (see BA_{CFA} & MA_{CFA}) should equal the CI_{CFA} model (see Table 1), thus the total degrees-of-freedom in our study is 1034 (i.e., 517+517 = 1034). Second, the *df* should align the unstandardized and thresholds (we do this in EXCEL) to be certain that the correct parameters are free to vary across the groups. This step is critical as it is very easy to accidentally constrain these parameters to be equal across groups in Mplus and other software packages. Unlike ML and MLR estimation, the WLSMV χ^2 statistics for each subsample does not sum to the CI model.

Using the approximate model fit statistics from the CI_{CFA} model (see Table 1), these results suggest it is appropriate to constrain the unstandardized factor loadings and thresholds to be equal across the two groups. Results from this analysis (see MI_{CFA} in Table 1) indicated that the invariance measurement model did not fit significantly worse than the configural invariance model based on the $\Delta\chi^2$ ($p = 0.120$). While the ΔAFIs present some concern, as the change in ΔAFIs were often greater than 0.01, it is critical to remember that the ΔAFIs is often inappropriate for model comparisons using WLSMV. Based on this information, greater focus should be placed on the $\Delta\chi^2$, which again suggests the measurement model is invariant across groups. From a practical perspective, researchers can also compare the factor loadings and thresholds (see Table 2) to identify whether any of the parameter estimates differ noticeably from each other. While other measurement components could certainly be tested (see Marsh et al., 2009; Vandenberg & Lance, 2000), again this study focused solely on those parameters (i.e., unstandardized factor loadings and thresholds) that are required to be equal across groups to ensure the latent factor scores are created using the same metric and scalar weights. In summary, the current overall model, along with the change in model fit statistics, suggests the measurement invariance model fits the data well, which implies it is statistically appropriate to compare latent factor mean scores. The final steps are to test for equality of covariance matrices and structural coefficients across groups.

Starting with the equality of covariance matrices (given that the variances are fixed at one, this is equivalent to testing the equality of correlations between latent factors), the results in Table 1 indicate that the $CovIN_{CFA}$ model does not fit significantly worse than the MI_{CFA} model based on the $\Delta\chi^2$ statistic ($p = 0.832$). Again the ΔAFIs was positive, indicating that the more restrictive model ($CovIN_{CFA}$) fit significantly better than the less restrictive model (MI_{CFA}). Based on statistical theory this is not possible, as a more restrictive model can never fit better than a less restrictive model. Thus, researchers need to remember to calculate the *p*-value from the $\Delta\chi^2$ to ensure

Table 1. Model fit statistics for each model estimated using WLSMV and MLR estimation

	χ^2	*df*	$\Delta\chi^2$	Δdf	*P*	*CFI*	ΔCFI	*TLI*	ΔTLI	*RMSEA*	$\Delta RMSEA$
					WLSMV						
CFA (n = 617)	2462.67	517				0.951		0.947		0.078	
SEM (n = 617)	2523.83	522	78.54	5	< 0.01	0.949	−0.002	0.946	−0.001	0.079	0.001
BA_{CFA} (n = 379)	1659.15	517				0.954		0.950		0.076	
MA_{CFA} (n = 238)	1231.57	517				0.952		0.947		0.076	
CI_{CFA} (n = 617)	2844.23	1034				0.954		0.950		0.075	
MI_{CFA} (n = 617)	2473.68	1231	220.57	197	0.12	0.968	0.014	0.971	0.021	0.057	−0.018
$CovIN_{CFA}$ (n = 617)	2057.15	1241	5.80	10	0.83	0.979	0.011	0.981	0.010	0.046	−0.011
SEM with MI (n = 617)	2596.97	1241				0.965		0.969		0.060	
SEM with MI & SI (n = 617)	2234.56	1246	3.89	5	0.57	0.975	0.010	0.977	0.008	0.051	−0.009
					MLR						
CFA (n = 617)	1228.72	513				0.930		0.923		0.046	
SEM (n = 617)	1258.37	518	29.65	5	< 0.01	0.927	−0.003	0.921	−0.002	0.048	0.002
BA_{CFA} (n = 379)	990.28	513				0.924		0.917		0.050	
MA_{CFA} (n = 238)	986.43	513				0.892		0.882		0.062	
CI_{CFA} (n = 617)	1976.76	1026				0.911		0.903		0.055	
MI_{CFA} (n = 617)	2053.53	1094	76.77	68	0.22	0.910	−0.001	0.908	0.005	0.053	−0.002
$CovIN_{CFA}$ (n = 617)	2055.12	1104	1.59	10	≈1.00	0.911	0.001	0.910	0.002	0.053	0.000
SEM with MI (n = 617)	2082.74	1104				0.909		0.907		0.054	
SEM with MI & SI (n = 617)	2086.608	1109	3.87	5	0.57	0.909	0.000	0.908	0.001	0.053	−0.001

Note. The following models were estimate in the following order: BA (BAchelor's degree), MA (MAster's degree), CI (Configural Invariance), MI (Measurement Invariance), CovIN (Covariance INvariance), and SI (Structural Invariance). Recall that MLR estimation correlated four residual variances for four times, whereas these parameters were not estimated with WLSMV.

that the difference is statistically significant. In any case, these results indicate that the relationship between latent factors does not differ across groups, which suggests that it is unlikely that the unstandardized structural coefficients will also differ across groups. In fact, this inference was confirmed when comparing an SEM model with MI (SEM with SI) to an SEM model with structural invariance (SEM with MI and SI), as the $\Delta\chi^2$ was not statistically significant ($p = 0.566$). Structural invariance

Table 2. WLSMV estimates of the unstandardized factor loadings (UFL) and thresholds for teachers with BA and MA degrees

	UFL		*Thresholds*	
Item	*BA*	*MA*	*BA*	*MA*
IM2	0.71	0.73	(–0.93, 0.43, 1.37)	(–0.74, 0.34, 1.27)
IM3	0.60	0.63	(–0.23, 1.02, 1.82)	(–0.18, 0.99, 1.72)
IM5	0.59	0.72	(–1.05, –0.02, 0.81)	(–0.90, 0.03, 0.90)
IM6	0.59	0.64	(–1.27, –0.17, 0.74)	(–1.03, 0.07, 0.86)
IM9	0.63	0.71	(–0.28, 1.03, 2.03)	(–0.24, 1.10, 2.24)
IM12	0.81	0.86	(–0.57, 0.74, 1.72)	(–0.35, 0.70, 1.68)
SB1	0.80	0.81	(–1.89, –1.02, –0.60, –0.10, 0.27, 1.17)	(–2.13, –0.93, –0.52, –0.15, 0.30, 1.13)
SB2	0.79	0.77	(–1.78, –1.12, –0.76, –0.34, 0.07, 0.98)	(–1.78, –1.01, –0.61, –0.26, 0.07, 0.94)
SB3	0.80	0.76	(–1.25, –0.53, –0.27, 0.13, 0.46, 1.26)	(–1.19, –0.45, –0.02, 0.27, 0.48, 1.23)
SB4	0.75	0.77	(0.05, 0.71, 0.95, 1.27, 1.75, 2.21)	(–0.25, 0.48, 0.76, 1.23, 1.57, 2.03)
SB5	0.88	0.87	(–0.90, –0.14, 0.15, 0.48, 0.81, 1.39)	(–0.85, –0.09, 0.17, 0.54, 0.80, 1.30)
SB6	0.84	0.81	(–1.06, –0.42, –0.04, 0.30, 0.68, 1.39)	(–1.10, –0.42, –0.12, 0.34, 0.66, 1.37)
SB7	0.78	0.87	(–1.13, –0.52, –0.26, 0.12, 0.42, 1.03)	(–1.20, –0.60, –0.35, 0.13, 0.42, 1.03)
SB8	0.96	0.94	(0.12, 0.83, 1.15, 1.57, 1.98)	(0.06, 0.74, 1.00, 1.35, 1.78)
SB9	0.94	0.95	(–0.10, 0.50, 0.74, 1.13, 1.45, 1.89)	(–0.30, 0.45, 0.67, 0.99, 1.20, 1.95)
EE1	0.87	0.87	(–0.98, –0.04, 0.65, 1.29)	(–0.93, –0.11, 0.71, 1.30)
EE2	0.90	0.88	(–1.39, –0.44, 0.29, 0.96)	(–1.43, –0.37, 0.32, 1.04)
EE3	0.84	0.88	(–0.91, 0.07, 0.69, 1.39)	(–0.94, 0.11, 0.89, 1.46)
EE4	0.68	0.72	(–0.87, –0.08, 0.58, 1.18)	(–0.90, –0.12, 0.52, 1.27)
EE5	0.86	0.88	(–1.31, –0.31, 0.34, 1.04)	(–1.34, –0.42, 0.37, 1.04)
EE6	0.86	0.84	(–1.10, –0.09, 0.49, 1.19)	(–0.97, –0.29, 0.56, 1.34)
EE7	0.74	0.77	(–0.92, 0.11, 0.90, 1.64)	(–0.89, 0.14, 0.79, 1.56)
EE8	0.72	0.76	(–1.21, –0.22, 0.25, 0.93)	(–1.27, –0.26, 0.26, 0.87)
EE9	0.85	0.84	(–1.23, –0.26, 0.20, 0.94)	(–1.27, –0.32, 0.25, 0.92)
PA2	0.55	0.57	(–1.98, –1.64, –1.44, –1.00, –0.65, 0.18)	(–2.03, –1.72, –1.43, –0.99, –0.65, 0.12)

	UFL		*Thresholds*	
Item	*BA*	*MA*	*BA*	*MA*
PA3	0.75	0.73	(–1.98, –1.69, –1.33, –0.98, –0.66, 0.08)	(–2.03, –1.82, –1.64, –1.16, –0.81, –0.12)
PA5	0.76	0.84	(–2.22, –2.08, –1.82, –1.22, –0.82, 0.08)	(–2.39, –2.13, –1.73, –1.33, –0.85, 0.05)
PA6	0.60	0.80	(–1.85, –1.67, –1.37, –0.86, –0.45, 0.62)	(–2.24, –1.78, –1.50, –1.12, –0.56, 0.48)
PA7	0.87	0.81	(–2.55, –1.93, –1.64, –1.05, –0.74, 0.21)	(–2.39, –1.96, –1.78, –1.32, –0.89, 0.11)
PA8	0.45	0.59	(–2.30, –1.64, –1.26, –0.84, –0.44, 0.48)	(–2.03, –1.72, –1.43, –0.89, –0.57, 0.34)
JS1	0.80	0.68	(–1.59, –1.06, –0.41, –0.10, 0.51)	(–1.56, –1.03, –0.43, –0.20, 0.59)
JS2	0.89	0.93	(–2.03, –1.59, –1.41, –0.67, 0.24)	(–2.12, –1.64, –1.38, –0.73, 0.41)
JS3	0.88	0.86	(–1.46, –1.05, –0.64, –0.13, 0.64)	(–1.60, –1.25, –0.92, –0.08, 0.75)
JS4	0.79	0.72	(–1.46, –0.81, –0.20, 0.04, 0.66)	(–1.35, –0.83, –0.07, 0.30, 0.91)

Note. Instructional management (IM), Student behavior stressors (SB), Emotional exhaustion (EE), Personal accomplishment (PA), and Job Satisfaction (JS)

implies that the predictive relationships between latent variables are not moderated by teacher education level.

MLR Estimation

The initial test of the CFA model using the full sample revealed that the model fit the data relatively poorly, χ^2 (584) = 2431.92, $p < .0001$, CFI = 0.825, TLI = 0.812, RMSEA = 0.068. For this reason, the modification indices were evaluated to identify the cause of misfit. These results suggested that the following residual covariance matrices needed to be estimated to improve the model fit: SB8 with SB9, SB1 with SB3, EE1 with EE2, and EE4 with EE8. While these changes could be justified theoretically and statistically, this discussion was omitted to conserve space. Instead, interested readers should read the work of Barry and Finney (2009) and Byrne (2012) who discusses how data are collected and the shape of the observed variables influence the residual covariances. The model fit significantly improved, $\Delta\chi^2$ (4) = 341.82, $p < .0001$, ΔCFI = 0.090, ΔTLI = 0.098, ΔRMSEA = -0.024, after estimating these residual variances; however, the model fit for the entire sample [see CFA (n = 617) in Table 2] still suggested some degree of misfit. Despite this, further modifications were not made given that the$\Delta\chi^2$ and ΔAFI did not considerably change with additional modifications. Similar to WLSMV estimation, the $\Delta\chi^2$ was statistically significant when comparing the CFA and SEM model (see Table 1), which again implies that an alternative model may be more appropriate. To be consistent with theory, these changes were not made to the model.

With an adequate fitting CFA and SEM model, subsample CFA analyses were conducted using the teachers with BA and MA degrees. Analyses revealed that teachers with a BA degree (see BA_{CFA} in Table 1) fit the data better than the model using MA teachers (MA_{CFA} in Table 1). It is worth noting that the modification indices did not suggest a single large source of model misfit, thus considerable model changes would be required to improve the model fit. Despite having some concern that subsample models do not adequately fit the data, the CI_{CFA} model was estimated and displayed a marginally acceptable model fit. However, when compared to the MI_{CFA} model, the model fit did not significantly change based on the $\Delta\chi^2$ and ΔAFIs. Comparable to the WLSMV results, these analyses imply that the measurement model (i.e., unstandardized and intercepts) was invariant across groups. The unstandardized factor loadings and intercepts (see Table 3) also provide evidence that these parameters are similar across the groups. Tests of residual

Table 3. MLR estimates of the unstandardized factor loadings (UFL) and intercepts for teachers with BA and MA degrees

	UFL		*Intercepts*			*UFL*		*Intercepts*	
Item	*BA*	*MA*	*BA*	*MA*	*Item*	*BA*	*MA*	*BA*	*MA*
IM2	0.54	0.65	2.24	2.24	EE3	1.57	1.69	6.08	6.12
IM3	0.37	0.45	1.78	1.78	EE4	0.77	0.84	5.58	5.78
IM5	0.62	0.69	2.57	2.49	EE5	1.68	1.72	5.96	6.11
IM6	0.52	0.58	2.69	2.52	EE6	1.61	1.58	5.62	5.76
IM9	0.38	0.40	1.78	1.74	EE7	1.48	1.49	2.70	2.70
IM12	0.61	0.70	1.98	1.92	EE8	0.64	0.74	3.14	3.09
SB1	0.81	0.78	4.60	4.58	EE9	1.27	1.35	2.62	2.55
SB2	0.87	0.85	4.87	4.78	PA2	0.65	0.66	2.73	2.77
SB3	0.79	0.70	4.08	3.88	PA3	0.99	0.82	3.04	3.09
SB4	0.89	0.89	2.06	2.32	PA5	0.75	0.90	2.83	2.82
SB5	1.07	1.03	3.43	3.37	PA6	0.73	0.94	2.51	2.53
SB6	1.03	0.94	3.76	3.79	PA7	0.97	0.87	3.06	3.09
SB7	0.82	0.98	4.12	4.20	PA8	0.57	0.69	3.10	3.10
SB8	0.96	1.02	1.86	1.99	JS1	1.11	1.05	4.30	4.31
SB9	0.94	0.98	2.31	2.50	JS2	0.92	0.91	5.00	4.95
EE1	1.33	1.36	5.86	5.90	JS3	1.25	1.07	4.34	4.41
EE2	1.34	1.40	5.89	6.11	JS4	1.19	1.03	4.04	3.79

Note. Instructional management (IM), Student behavior stressors (SB), Emotional exhaustion (EE), Personal accomplishment (PA), and Job Satisfaction (JS)

variance invariance indicated that the residual variances were also invariant across groups, $\Delta\chi^2$ (4) = 1.842, p = 0.7648. However, the residuals were not constrained to be equal in Table 1 in an effort to align the models more closely with those resulting from the WLSMV analyses.

Shifting focus to the covariance and structural coefficients, both these models appeared invariant across groups based on the $\Delta\chi^2$ and ΔAFIs (see Table 1, $CovIN_{CFA}$ and SEM with MI and SI). Collectively, these results provide a useful demonstration of how the overall model fit can vary considerably based on the model estimator, whereas the overall conclusions related to measurement and structural invariance are generally unaltered. Similar to WLSMV results, analyses suggest the relationships (both correlation and predictive) did not differ across the groups.

Latent Factor Mean Score Comparison

As indicated above, the latent factor mean scores cannot be justifiably compared if the latent factors are not invariant across comparison groups. Fortunately, both WLSMV and MLR provided statistical evidence of MI. For both model estimation methods, the latent factor means in the BA teachers group was fixed at zero, while the MA teachers mean scores were estimated (for more detail on this procedure see Byrne, 2012, pp. 248–254). Results revealed nearly identical conclusions across model estimators (see Table 4), in that teachers with BA and MA degrees did not differ across any of the five latent factors. In fact, the Cohen's *d* [computed using

Table 4. Latent factor mean difference results between teachers with BA and MS degree using WLSMV and MLR estimation

	IM	*SB*	*EE*	*PA*	*JS*
			WLSMV		
M_{Diff}	–0.11	0.00	0.03	0.16	–0.03
t-statistic	–1.15	0.00	0.29	1.63	–0.38
p-value	0.25	1.00	0.77	0.10	0.71
Cohen's d	–0.09	0.00	0.02	0.13	–0.03
			MLR		
M_{Diff}	–0.10	0.01	0.00	0.16	–0.03
t-statistic	–1.08	0.12	–0.02	1.67	–0.36
p-value	0.28	0.90	0.99	0.10	0.72
Cohen's d	–0.09	0.01	0.00	0.13	–0.03

Note. Instructional management (IM), Student behavior stressors (SB), Emotional exhaustion (EE), Personal accomplishment (PA), and Job Satisfaction (JS).

the equation $d = 2(t)/\sqrt{df}$] indicated small effect sizes based on his (Cohen, 1988) tentative effect size standards: small $d \approx 0.20$; medium $d \approx 0.50$, and large $d \approx 0.80$.

A Closer Look at Covariance and Structural Model Results

While Tables 2 and 3 provide the descriptive statistics for the metric and scalar parameter estimates (numbers taken from the CI model), Table 5 provides the covariance matrices for both samples across the two estimators (numbers taken from the MI model). Figure 1 displays the structural coefficients across these two teacher groups using both estimators (numbers taken from the SEM with MI model). For sake of clarity, the covariance results here were obtained from the MI_{CFA} model, as these model parameters were not constrained to be equal across the groups. Recall, the $CovMI_{CFA}$ would have equal covariance matrices across groups. The numbers in Figure 1 came from the SEM with MI model, as the same structural model coefficients would be extracted from the SEM with MI and SI model.

Starting with the covariance invariance (i.e., equal covariance matrices across the groups) results (see Table 5), analyses provided evidence of equal covariance matrices across the two teacher groups based on the change in model fit statistics. In

Table 5. Provides the covariance matrix for the five latent factors for teachers with BA and MA degrees using WLSMV and MLR estimation

	1	*2*	*3*	*4*	*5*
			WLSMV		
1. Job Satisfaction	1.00	–0.31	–0.48	–0.72	0.48
2. Instructional management	–0.38	1.00	0.35	0.23	–0.60
3. Student behavior stressors	–0.51	0.33	1.00	0.60	–0.38
4. Emotional exhaustion	–0.69	0.22	0.58	1.00	–0.36
5. Personal accomplishment	0.54	–0.47	–0.40	–0.37	1.00
			MLR		
1. Job Satisfaction	1.00	–0.28	–0.51	–0.76	0.42
2. Instructional management	–0.38	1.00	0.34	0.22	–0.47
3. Student behavior stressors	–0.52	0.34	1.00	0.62	–0.41
4. Emotional exhaustion	–0.69	0.22	0.62	1.00	–0.35
5. Personal accomplishment	0.53	–0.47	–0.41	–0.35	1.00

Note. The lower left matrix represents the inter-factor covariance coefficients for teachers with BA degrees, whereas the upper right matrix provides the covariance coefficients for teachers with MA degrees.

fact, the $\Delta\chi^2$ was statistically non-significant when using both MLR and WLSMV estimation methods, thus implying that the relationship between variables does not change based on group membership. The largest difference emerged between Instructional management and Personal accomplishment (difference of 0.13) with WLSMV; however, these differences are often difficult to interpret given that the covariance depends on the variance of the both latent factors. In the case of this study, the factor variances are fixed at one and, therefore, this variation can be interpreted as the difference in correlation coefficients. Turning our attention next to MLR estimation, the largest change (difference of 0.11) was the association between Job satisfaction and Personal accomplishment, which again provides evidence of a rather small difference in the covariance matrix. It is worth recognizing that these differences would be more difficult to interpret had an unstandardized factor loading been fixed at one across the groups rather than the variances.

Figure 1 provides the unstandardized structural coefficients, which were tested for structural invariance. It is important to recognize that only unstandardized coefficients are tested for equality, thus it can be more difficult to assess whether a difference is large (differences are scale dependent). The results in Figure 1 suggest

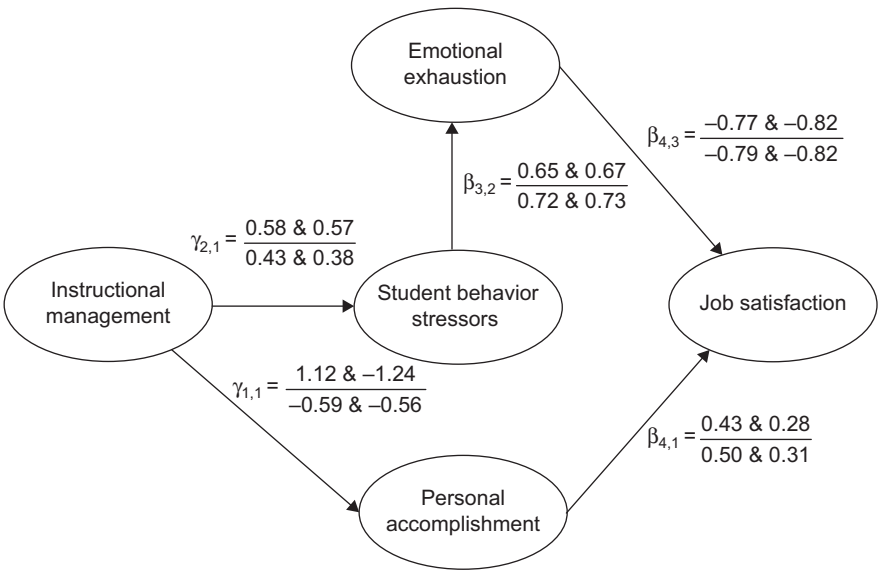


Figure 1. Provides the estimated unstandardized structural coefficients using WLSMV and MLR estimation with BA and MA teachers. The numbers above the line represent the bachelors and masters degree teachers using WLSMV, respectively, whereas the numbers below the line represent the bachelors and masters degree teachers using MLR, respectively.

few differences between the BA and MA teacher samples, whether we compare the WLSMV or MLR estimation results. The largest difference emerged for $\beta_{4,1}$, but again the $\Delta\chi^2$ provided no evidence that any of these structural coefficients differed across groups. While the unstandardized coefficients are tested for invariance (or equality of coefficients across groups), the standardized coefficients are often easier to interpret.

Although they were not statistically tested for differences, the standardized coefficients (see Figure 2) and R^2 statistics are provided here for comparison purposes across the groups. As seen here, the structural coefficients and R^2 statistics are similar across BA and MA teachers when compared within the same estimator. Similar to the unstandardized results, the largest difference was on $\beta_{4,1}$, with the prediction for MA teachers being smaller than BA teachers. Interestingly, larger differences did materialize based on the estimator selected. In most cases (the exception being $\beta_{3,2}$ and $\beta_{4,3}$), the structural coefficients were slightly larger when using WLSMV estimation compared to MLR. The largest differences was for $\gamma_{2,1}$ and

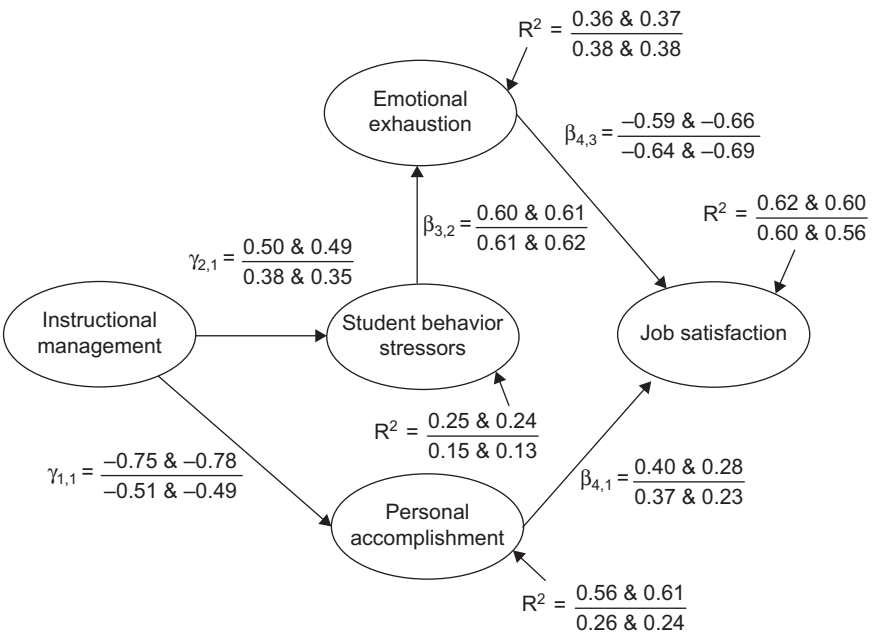


Figure 2. Provides the estimated standardized structural coefficients and R^2 statistics using WLSMV and MLR estimation with bachelors and masters degree teachers. The numbers above the line represent the bachelors and masters degree teachers using WLSMV, respectively, whereas the numbers below the line represent the bachelors and masters degree teachers using MLR, respectively.

$\gamma_{1,1}$, which also produced a larger R^2 for the student behavior stressors and personal accomplishment latent factors when using WLSMV. This simple demonstration suggests that WLSMV provided a better overall fit and generated larger predictions of the latent factors.

Research Question Summary

Turning our attention back to the initial research questions, the following conclusions were reached. Evidence of factorial validity (research question 1) was only obtained when using WLSMV given that the model fit the data well, the standardized factor loadings were all sufficiently large, and the interfactor correlations did not display any discriminate validity concerns. However, there was less confirmation when using MLR, as the model fit was less than desired and there were several correlated residual variances. Regardless of the estimator, there was evidence of measurement invariance (research question 2), equality of latent factor means (research question 3), and an invariant structural model (research question 5) across teacher education groups. While the overall SEM provided an adequate model fit with large structural coefficients and large R^2 statistics (research question 4), concern did arise given that the SEM model fit significantly worse than the CFA model. Therefore, researchers might consider revising the model slightly to provide greater support for research question 4.

DISCUSSION

The difficulty with conducting invariance tests results from the immense number of decisions placed upon the researcher. Foremost, researchers must select the "best" type of model to estimate (e.g., factor analysis vs. item response theory), the most appropriate estimation method (e.g., a least squared, maximum likelihood, Bayesian estimation just to name a few), what assumptions should be made about the data (i.e., treating the observed variables as ordered categorical or continuous), what invariance models to evaluate (measurement invariance, covariance invariance, residual invariance, structural invariance, etc.), how to set the latent factor scale, and what constitutes an "invariant model," just to name a few. Researchers also need to ensure models are nested to correctly estimate the $\Delta\chi^2$ and ΔAFI, while also being aware that direct comparisons of the $\Delta\chi^2$ (e.g., WLSMV uses the DIFFTEST procedure in Mplus and MLR using the Satorra-Bentler scaled difference test) or ΔAFI (e.g., results may not be accurate with WLSMV) may be inappropriate. As shown here, ensuring the Δdf matches the number of parameters to be fixed (or set as invariant) across the groups and comparing the modeling parameters between different models tested (e.g., CI to MI) helps the researcher understand how each model was estimated and any differences one might expect. While documentation

(e.g., Byrne, 2012) is available to help researchers analyze data, they must be cognizant of those factors that influence the results and the assumptions made about the data.

In the current illustration, the results were fairly similar across estimation methods. However, researchers cannot simply assume the estimator is irrelevant, as conclusions can vary based on the model estimator (see Sass, 2011). In fact, this study indicated that the CFA and SEM models did not fit the data that well using MLR and the structural coefficients were often smaller for MLR than WLSMV. In closing, we hope this material helps researchers appreciate the benefits and interesting research questions that can be evaluated with invariance testing, while at the same time understanding the complications and limitations with these statistical procedures.

REFERENCES

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

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*, 815–824.

Barry, C. L., & Finney, S. J. (2009). Does it matter how data are collected? A comparison of testing conditions and the implications for validity. *Research & Practice in Assessment, 3*, 1–15.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care, 44(11),* 176–181.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, N.J.: Guiford Press.

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

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14,* 464–504.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, 95, 1005–1018.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12,* 471–492.

Cheung, G. W., & Lau, R. S. (in press). A direct comparison approach for testing measurement invariance. *Organizational Research Methods*.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25,* 1–27.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling, *Journal of Cross-cultural Psychology, 31,* 187–212.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*, 121–149.

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling*, *15,* 96–113.

Hambleton, R. K., Merenda, P., & Spielberger C. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and manova in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling*, 7, 534–556.

Horn J. L., McArdle J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117–144.

Hu & Bentler (1999). Cutoff criteria for fit indexes in covariance structure analysis: Coventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Hui, E. K. P., & Chan, D. W. (1996). Teacher stress and guidance work in Hong Kong secondary school teachers. *British Journal of Guidance and Counseling, 24*, 199–211.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15*, 136–153.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling, 13,* 59–72.

Lubke, G. H., & Muthén, B. O. (2004). Applying multi-group confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11,* 514–534.

Marsh, H. W. (1985). The structure of masculintiy/feminity: An application of confirmatory factor analysis to higher-order factor structures and factorial invariance. *Multivariate Behavioral Research*, *20,* 427–449.

Marsh, H. W., Hau, K. T., & Wen, Z. L. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu & Bentler (1999) findings. *Structural Equation Modeling, 11,* 320–341.

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. *Psychological Bulletin, 97,* 562–582.

Marsh, H. W., Liem, G. A., Martin, A. J., Nagengast, B., & Morin, A. J. S. (2011). Methodological-measurement fruitfulness of Exploratory Structural Equation Modeling (ESEM): New approaches to key substantive issues in motivation and engagement. *Journal of Psychoeducational Assessment, 29*, 322–346.

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

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Applications to students' evaluations of university teaching. *Structural Equation Modeling, 16*, 439–476.

Martin, N. K., & Sass, D. A. (2010). Construct validation of the Behavior and Instructional Management Scale. *Teaching and Teacher Education, 26*, 1124–1135.

Martin, N. K., Sass, D. A., & Schmitt, T. A. (2012). Teacher efficacy in student engagement, instructional management, student stressors, and burnout: A theoretical model using in-class variables to predict teachers' intent-to-leave. *Teaching and Teacher Education. 28*, 546–559.

Maslach, C., Jackson, S. E. Leiter, M. P. (1996). *Maslach Burnout Inventory Manual*, 3rd Ed. Mountain View, CA.: CPP, Inc.

MacCallum, R. C. , Roznowski, M. & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111, 490–504.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness-of-fit. *Psychological Bulletin, 107*, 247–255.

McDonald, R. P., Seifert, C. F., Lorenzet, S. J., Givens, S., & Jaccard, J. (2002). The effectiveness of methods for analyzing multivariate factorial data. *Organizational Research Methods*, *5,* 255–274.

McIntosh, C. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences, 42*, 859–857.

McLaney, M. A., & Hurrell, J. J. (1988). Control, stress, and job satisfaction in Canadian nurses. *Work and Stress*, *2*, 217–224.

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling, 14*, 611–635.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93,* 568–592.

Meredith, W. (1993). Measurment invariance, factor analysis and factorial invariance. *Psychometrika*, *58,* 525–543.

Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research, 33,* 403–424.

Millsap, R. E. (2001). When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling, 8,* 1–17.

Millsap, R. E., & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9,* 93–115.

Millsap, R. E., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R. MacCallum (Eds.), *Factor analysis at 100* (pp. 131–152). Mahwah, NJ: Lawrence Erlbaum.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39,* 479–515.

Muthén, L. K., & Muthén, B. O. (1998–2010). Mplus User's Guide. 6th Ed. Los Angeles, CA: Muthén & Muthén.

Rhemtulla, M., Brosseau-Liard, P., & Savalei, V. (2010). How many categories is enough to treat data as continuous? A comparison of robust continuous and categorical SEM estimation methods under a range of non-ideal situations. Retrieved from http://www2.psych.ubc.ca/~mijke/files/HowManyCategories.pdf

Saris, W. E., Satorra, A., & van der Veld, W. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*, 561–582.

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (in press). *Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators*. *Structural Equation Modeling*

Selig, J. P., Card, N. A., & Little, T. D. (2008). Latent variable structural equation modeling in cross-cultural research: Multigroup and multilevel approaches. In F. J. R. van de Vijver, D. A. van Hemert & Y. Poortinga (Eds.) *Individuals and Cultures in Multi-level Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Saris, W. E., Satorra, A., & van der Veld, W. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*, 561–582.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66,* 507–514.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75,* 243–248.

Steiger, J. H. (2002). When constraints interact: A caution about reference variables, identification constraints, and scale dependencies in structural equation modeling. *Psychological Methods, 7*, 210–227.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–69.

Van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31*(1), 33–35.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Book Chapter Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Yu, C. Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. Doctoral dissertation, University of California, Los Angeles.

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In Sobel, M. E., & Becker, M. P. (Eds.), *Sociological methodology 2000* (pp. 165–200). Washington, D.C.: ASA.

GEORGE A. MARCOULIDES & RONALD H. HECK

16. MIXTURE MODELS IN EDUCATION

Mixture models are a special type of quantitative model in which latent variables can be used to represent mixtures of subpopulations or classes where population membership is not known but inferred from the data. Mixture modeling is used to assign individuals to the most likely latent class and to obtain parameter estimates of a proposed model for the classes identified. Another use of mixture modeling is to represent latent change or growth trajectory classes comprised of individuals with similar trajectories. Applications of such combined models extend our ability to examine a variety of complex relationships in educational research. In this chapter, we provide an illustration of mixture modeling using longitudinal data from high school students to describe a model of academic achievement level and growth in mathematics.

INTRODUCTION

Latent variable mixture modeling is a data analysis technique which assumes that the observations in a given cross-sectional or longitudinal dataset are sampled from a heterogeneous population containing a mixture of unobserved or latent subpopulations, each having its own unique response variable distribution. The past few decades have witnessed a tremendous growth in the accumulation of cross-sectional and longitudinal data sources within the behavioral, educational, and social sciences. Accompanying this buildup of data are the occasional inconsistencies in the results obtained from analyses that incorrectly assume population homogeneity rather than heterogeneity that may be present. As a consequence, much interest has emerged in the use of various types of latent mixture modeling approaches that can tackle cross-sectional, longitudinal, and even multilevel data situations. In general, mixture modeling can be used to analyze various subpopulations of observations (e.g., individuals, schools, etc.) where the population membership is not known ahead of time but, rather, must be inferred from the data. In this type of formulation, the subpopulations are referred to as latent classes, which can be readily defined through categorical latent variables.

Recent advances in computer software (e.g., Lanza, Collins, Lemmon, & Shafer, 2007; Lanza, Dziak, Huang, Xu, & Collins, 2011; Muthén & Muthén, 1998–2012) have not only made the use of mixture modeling and its various extensions quite straightforward, but have also increased their use by applied researchers across

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 347–366.

many disciplines. Extensions of latent mixture modeling include such models as latent transition analysis (Collins & Lanza, 2010; Marcoulides, Gottfried, Gottfried, & Oliver, 2010), associative latent transition analysis (Bray, Lanza, & Collins, 2010), and growth mixture modeling (Muthén & Shedden, 1999), to name just a few. All of the previously-mentioned models share the characteristic that population heterogeneity is accounted for by a latent grouping variable, but that each individual's (or other unit of analysis) group membership cannot be known with any degree of certainty; rather, the probability of group membership must be inferred from the data (Muthén & Muthén, 1998–2012).

In this chapter, we first develop a framework for the consideration of various types of mixture models which introduces ways that latent variables can be used to identify subsets of individuals or groups that are similar. We then provide a couple of extended examples utilizing latent variable mixture modeling with set-ups, output, and interpretation. The first is a simple latent class analysis (identifying subpopulations with individuals who are similar) which uses cross-sectional data, and the second is a growth mixture model which investigates latent classes of individuals with similar growth trajectories. There can be no doubt that mixture modeling techniques open up many new types of modeling capabilities, which are often generically referred to as mixture or latent class modeling. We hope these pragmatic examples taken from the field of education will help illustrate some of the varied uses of this type of modeling.

The chapter is organized in three sections as follows. In the first section, we provide an overview of mixture modeling in general including a taxonomy that identifies various types of cross-sectional and longitudinal mixture models. In the second section, we introduce model assumptions and estimation methods related to conducting a latent class or mixture analysis of a simple growth model. For purposes of illustration, we next develop a simplified confirmatory factor analysis measurement model, where we identify subsets of individuals who are in similar latent classes. In the last section, we combine latent mixture modeling with latent growth modeling to illustrate how we might identify mixtures (or latent classes) of individuals based on their continuous latent curve trajectories. Throughout the chapter we use a notational system generally considered to be consistent with the so-called multilevel framework, although this choice is somewhat arbitrary.

A FRAMEWORK FOR DEFINING MIXTURE MODELS

A commonly accepted taxonomy of different mixture models is provided in Table 1 (see also Muthén, 2002). As we suggest in this taxonomy, there are several common ways to differentiate existing mixture models. It is important to note that since this is an emerging general modeling framework, new models fitting into this framework are being developed constantly. First, we observe that mixture models involve combinations of continuous and categorical latent and observed outcome variables. Second, mixture modeling can be applied to both cross-sectional and longitudinal

Table 1. Taxonomy of models with latent classes

Model Name	*Variable Type*	*Cross-sectional/ Longitudinal*	*Within-Class Variation*
Latent Class Analysis (LCA)	Categorical	Cross-sectional	No
Latent Profile Analysis (LPA)	Continuous	Cross-sectional	No
Latent Transition Analysis (LTA)	Categorical	Longitudinal	No
Mixture Factor Analysis	Categorical Continuous	Cross-sectional	Yes
Mixture Structural Equation Modeling	Categorical Continuous	Cross-sectional	Yes
Growth Mixture Modeling (GMM)	Categorical Continuous	Longitudinal	Yes

data. Third, the taxonomy suggests that some types of mixture modeling support examining variation among individuals within each identified latent class (e.g., mixture factor analysis, growth mixture modeling), while others treat possible within-class variation as fixed (e.g., latent class analysis, latent transition analysis). Fourth, mixture modeling can be combined with other commonly-used modeling techniques including factor models, structural equation models, multiple group models, growth models, and multilevel models. Readers should keep in mind that the common thread running through the latent class modeling techniques summarized in Table 1 is that the underlying classes explain variability in the observed dependent variables (y) and that the overall approach provides a means of classifying individuals according to their latent class membership.

The general latent variable mixture modeling framework has two basic parts (Muthén & Muthén, 1998–2012). First is the so-called measurement part, which corresponds to the relationship between the set of observed dependent variables and one or more categorical latent variables, commonly labeled simply as (c). Second is the structural set of relationships between the latent variables, the relationships between the observed variables, and the relationships between the latent categorical (c) variables and observed covariates (x). Although the general mixture model can be extended to include continuous latent variables used to classify individuals, in this chapter we will focus only on analyses involving categorical latent variables (for additional details, see Muthén, 2002).

Types of Mixture Models

In the next section, we briefly describe several different type of possible mixture models included in the taxonomy provided in Table 1. We also provide some illustrations concerning how they may be used to study various educational phenomena.

Latent Class Analysis (LCA)

Latent class analysis at its most basic level operates much in the same way as any structural equation model (e.g., a confirmatory factor analysis model) whereby a latent variable is measured via a number of manifest or indicator variables. Often times, it is assumed that the manifest variables correspond to the distributional characteristics of a single sample or that group membership can be readily defined based on an observable variable. An example is where the sample consists of explicitly identifiable or defined groups such as male and female or experimental and control groups. In many cases, however, group membership may not be known beforehand. For the situation in which the variables that determine group heterogeneity are not known and group membership has to be inferred from the data, the groups are commonly called latent classes (i.e., membership is not observable but latent – for further details see Lubke & Muthén, 2005). As previously indicated, studies which involve analyses of potentially heterogeneous groups are also sometimes generically referred to in the literature as mixture models or simply mixture analyses (Muthén, 2002).

The latent class analysis consists of cross-sectional data with multiple items measuring a construct which is represented as a latent class variable (Muthén, 2001). The goal is to identify items that indicate the classes well, estimate the class probabilities, identify covariates that explain class membership, and classify individuals properly within each latent class. As a simple example of a latent class or mixture analysis, consider Figure 1, which represents the data distribution for a single continuous outcome variable *x*. The figure suggests that the data actually consist of different groups of individuals but that the group membership is not directly observed. In fact, the observed distribution of *x* corresponds to those of the two latent classes $c = 1$ and $c = 2$, each with different means (μ_1 and μ_2). The figure implies that the two distributions are not directly observed (the dotted lines); rather, it is only the mixture of the two distributions that is observed (the solid line). As such, a latent class or mixture analysis can be used to determine the presence and nature of the mixture and its associated parameter estimates. Assessment of the latent classes or mixture distribution can subsequently be judged by comparison to the homogeneous distribution via a measure of data-to-model fit (Muthén, 2002).

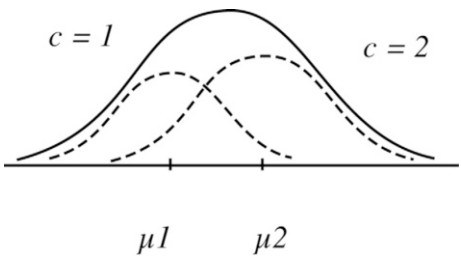


Figure 1. Distribution of a continuous outcome variable X with two mixtures.

In order to conduct a mixture analysis of the above illustrated data, an explicit model that divides the sample into various mutually exclusive latent classes must be posited (for in depth technical details on conducting latent class analyses, see Muthén & Muthén, 1998–2012). The basic assumption of such a latent class model that explains the relationships between the observed variables measured is that the population from which the sample was taken consists of *k* latent classes of unknown size (postulated to be mutually exclusive and collectively exhaustive). For example, a proposed CFA model for $k = 1, \ldots, K$ latent classes can be specified as

$$y_{ik} = \nu_k + \Lambda_k \eta_{ik} + \varepsilon_{ik}, \quad (1)$$

where for any class *k,* an individual's responses y_{ik} is a vector of observed scores, ν_k is a mean vector, Λ_k is a matrix of factor loadings, η_k is a vector of factor scores, and ε_{ik} is a vector of residual errors. In the above figure, the number of latent classes would essentially be equal to two. In our first example analysis provided in a later section, we will illustrate how one would specifically arrive at such a conclusion about the number of latent classes present in the data distribution.

LATENT PROFILE ANALYSIS

Latent profile analysis is similar in many ways to latent class analysis. The primary difference is that latent profile analysis assumes there are two or more continuous dependent variables which are indicators of the latent variable (*c*), whereas the analysis is referred to as latent class analysis when the outcomes are categorical (as we noted in Table 1). The continuous indicators might be considerably correlated (such as mathematics and reading scores), but the assumption is that the relationship is due to the mixing of several classes of individuals, each having unrelated outcomes (Muthén, 2001). Latent class and profile analysis have features similar to factor analysis, in that it is assumed the underlying latent variable (*c*) is responsible for the association observed between the observed outcomes. As we noted in Table 1, in latent class or profile analysis, the model for each class can be tested to see if it is the same or not across classes. For latent profile analysis, the mean for each outcome variable may be expected to change across classes. For latent class analysis with a dichotomous outcome, the probability of each outcome variable changes across classes (Muthén, 2001).

Mixture Factor Analysis

Readers will see that the more simplified latent class or latent profile analysis can also be extended to situations where there are several underlying factors that comprise a measurement (or factor) model. Factor analysis is an approach for describing associations among observed indicators in terms of a smaller number of latent continuous factors. The observed indicators may be continuous, ordinal,

or dichotomous. The general factor analytic approach can be extended to consider situations where individuals are clustered in groups (such as students within classrooms or schools). Using a mixture model, we could also consider situations where individuals are members of subpopulations (or latent classes) which differ in the parameters of the individual factor model. More specifically, we propose a set of subpopulations (or latent classes) which differ in the parameter values of the measurement model defined for individuals. In this situation, the mixture factor model is something like a multiple-group factor analysis, where individuals belong to latent classes differing with respect to the measurement model. An example might be a clustering of individuals who perform similarly on a mathematics achievement factor and a reading achievement factor where the grouping is unknown, rather than a situation where one might examine a possible different achievement factor model for males and females (i.e., where the two gender subpopulations are known ahead of time).

Similarly, if we had individual students nested within schools, using a mixture model we could investigate situations where the level-2 units (schools) might belong to a smaller set of latent classes which differ in systematic ways. For example, there might be classes of schools with high, average, and low student achievement in mathematics and reading. With this information, educators could then target particular schools for particular types of interventions to improve their achievement results.

Mixture SEM

Measurement models that define the relationships between observed indicators and underlying factors can be extended to include proposed predictive relationships between the underlying factors. This second type of model is sometimes referred to in the structural equation modeling (SEM) literature as the structural model. This more complex type of model facilitates the investigation of situations where the primary focus of model testing is on one or more structural parameters (e.g., slopes) in the model that might differ across classes. An example might be where an underlying factor defining student motivation has a different impact on a latent achievement factor across different latent classes of individuals.

AN EXAMPLE LATENT CLASS ANALYSIS

We next illustrate a simple latent class analysis. Consider a simple confirmatory factor analysis model based on four observed variables with one proposed common factor measuring academic intrinsic motivation of 111 high-school students at age 17 participating in the Fullerton Longitudinal Study (for details on the complete longitudinal study, see Marcoulides et al., 2007). To evaluate model fit, the Bayesian Information Criterion (BIC) index (Schwartz, 1978) is generally used because it provides an ideal way to examine the relative fit of any proposed latent class model

against the model for just one class (i.e., the case for which the considered sample is homogeneous with respect to the model considered). The BIC values for the various alternative or competing models are compared, and the model with the smaller value is considered the preferred model. Although some researchers also suggest the use of the likelihood ratio goodness-of-fit test to evaluate model fit, recent research has suggested that such an approach only works well in cases where there are not large numbers of sparse cells (Nylund, Asparouhov, Muthén & Muthén, 2007).

Mplus Model Statements

Latent class analysis is similar to factor analysis, but in contrast to factor analysis, provides a classification of individuals. Below we provide illustrative *Mplus* (Muthén & Muthén, 1998–2012) statements for latent class analysis of a model based on four observed variables with one proposed common factor measuring academic intrinsic motivation. It is important to note that to simplify matters, many of the default options are invoked. We also note that such a latent class analysis that involves continuous latent class indicators is also commonly referred to as a latent profile analysis. The CLASSES statement is used to specify the number of latent classes in the model and is the one option that must be manipulated by the user, each time specifying a different number of selected classes.

```
TITLE:      Illustrative Example of a LCA;
DATA:       FILE IS filename.dat;
VARIABLE:   NAMES ARE y1-y4;
            CLASSES = c(3);
ANALYSIS:   TYPE = MIXTURE;
OUTPUT:     TECH1 TECH8;
```

RESULTS

The following illustrative results were obtained when fitting the proposed CFA model to data using the above *Mplus* (Muthén & Muthén, 1998–2012) statements. To ensure that the examined model converged on global, rather than local solutions, random start values were used (see Muthén & Muthén, 1998–2012). A user can also elect to provide their own start values and iterations by declaring the STARTS and STITERATIONS options of the ANALYSIS command in the command file above (e.g., specifying STARTS=0, the random starts are turned off; specifying STITERATIONS=50, requests that fifty iterations be performed). At present, there is no certain automated way of determining the number of latent classes that may exist in the considered data set. Similar to exploratory factor analysis, a series of models are sometimes fit using sequentially different specified numbers of latent classes (e.g., 1, 2, 3, etc.). Model fitting criteria (i.e., BIC) can then be used to determine the appropriate number of classes to retain based upon good fit values (Muthén, 2002).

Table 2. Average posterior probabilities from the 3-class model

Class	*1*	*2*	*3*
1	.925	.023	.052
2	.048	.952	.000
3	.053	.000	.947

In this example, BIC values were examined for one-, two-, three, and four-class models and indicated that fitting a three-class model consistently results in the best BIC fit values. The proposed three class measurement model fit criterion was BIC = 3378.049, compared to model fit criteria of BIC = 3382.303, BIC = 3428.176, and BIC = 3437.049 for the one-, two-, and four-class models, respectively. Posterior probabilities for each individual can also be used to determine how well the categories define groups of individuals. Similar to classification in discriminant analysis or logistic regression, high diagonal and low off-diagonal values are indicators of good classification.

Table 2 displays information related to the quality of the classification using average posterior probabilities for the three-class model considered. The fit of the model to the data suggests that three latent classes seem to classify individuals into exclusive categories optimally based on the CFA model proposed. If desired, membership in these latent classes could be further investigated according to various individual characteristics (e.g., gender, academic background) or perhaps other latent variables.

AN ILLUSTRATION OF GROWTH MIXTURE MODELING (GMM)

In this section, we provide an introduction to growth mixture modeling (GMM). Central amongst mixture modeling methods have been models for the study of developmental change over time. One such set of models that can be used to study developmental change over time is growth mixture modeling, which is referred to interchangeably in the literature as latent change analysis, latent curve modeling, or just growth modeling analysis (see Bollen & Curran, 2006; Raykov & Marcoulides, 2006). In simple terms, a GMM analysis model enables a researcher to study the rate at which the developmental process under study is changing over time. Because in some cases the shape of the developmental process may be linear, whereas in others it might even be nonlinear (e.g., quadratic, cubic, etc.), specific hypotheses regarding the actual shape of the development can even be tested. Most commonly considered growth models usually assume that all individuals in a given sample come from a single population with one mean growth (change) pattern, and the variability around that mean growth is captured by the variance of the growth factors. In essence, these models use continuous latent variables to describe random effects (i.e., the intercept and slope, level and shape, or simply initial status and growth). The continuous latent variables describe unobserved heterogeneity in individual differences in change over time.

GMM represents an extension of the traditional growth model that permits the consideration of latent classes with differing developmental trajectories. GMM thereby permits more than one mean growth rate pattern, often assuming unique mean growth patterns for each apparently existing unobserved subpopulation. In GMM, heterogeneity in subpopulation outcome levels and developmental trajectories over time are captured by both continuous and categorical latent variables. The classes of a categorical latent variable can be proposed to represent latent trajectories which cluster individuals into sets of exclusive categories. Consequently, each latent class considered may have a different random effect growth model (Muthén, 2001). The random intercept and slopes are continuous latent variables, while the trajectory classes are considered to be categorical latent variables.

In the case of a LGMM, the classes of a categorical latent variable (c) represent latent trajectory classes which classify individuals into sets of exclusive categories (Clogg, 1995). The multiple indicators of the latent classes correspond to the repeated measures obtained over time and within-class variation is permitted for the latent trajectory classes (Muthén & Muthén, 1998–2012). This variation is also represented by random effects that influence the outcomes at all time points. For example, a variety of possible covariates can be added to the model, both to describe the formation of the latent classes and how they may be differentially measured by the repeated measures. The prediction of latent class membership is determined by the multinomial logistic regression of c on x.

For example, we may observe that four latent classes of developmental trajectories define students' growth in reading over four years, each having a different initial level of reading attainment and change trajectory over time. One set of trajectories might initially begin relatively high and descend in a linear fashion over successive measurements. Another might start below the first set but remain relatively flat over successive intervals. A third may start lower than the others, but ascend in a linear manner over time successive measurement intervals. Finally, a fourth might start similar to the second set, but ascend over the first two measurements and then slow as it rises over the final two time intervals. Additionally, covariates can be incorporated into the model (e.g., student gender, socioeconomic status, and mobility) both to describe the formation of the latent classes and to determine how they may be differentially measured by the repeated measures.

Figure 2 provides an example of a LGMM. As the figure suggests, covariates (e.g., SES) can be used to explain membership in the classes. Regression slopes between the covariates and outcome (math) can also be defined to vary across categories of c. The prediction of class membership is based on the logistic regression of c on x. The influence of x can vary across the latent classes and, in the multilevel case, the regression of c on x can vary across organizations. Including covariates can improve classification of individuals into latent classes. Thresholds (τ), similar to intercepts, are considered measurement parameters and can vary across the groups comprising latent categories (similar to nonequivalence across groups in a multiple group analysis).

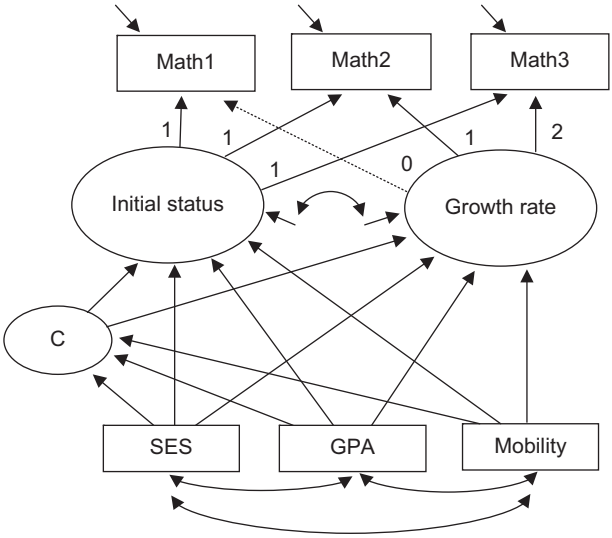


Figure 2. Adding a categorical latent variable and demographic predictors.

Specifying the Model

A growth mixture model can also be thought of as having two basic parts. The first part of the model is the general growth model for continuous and normally distributed *y* variables. This part is based on the analysis of the data based on a growth modeling approach. Two commonly used strategies to specify such a model are the Level and Shape approach and the Intercept and Slope approach (see Raykov & Marcoulides, 2006, for further discussion of basic approaches to latent growth modeling). These strategies are based on two specific kinds of parameterizations of the latent variables, the so-called Level-and-Shape (LS) model and the Intercept-and-Slope (IS) model.

The LS model was first described by McArdle (1988) and is considered to have a number of advantages over the currently popular IS model (Raykov & Marcoulides, 2006). A fundamental assumption of the IS model is that the change trajectory being studied occurs in a specific fashion and is either a linear, quadratic, cubic, or some higher order. Unfortunately, the actual change process may be quite difficult to model precisely utilizing any specific trajectory. For this reason, the less restrictive (in terms of the change trajectory) LS model is often preferable because it would generally be expected to fit the data better. Of course, the IS model can be obtained as a special case of the LS model when the coding of time is simply fixed according to the time of the repeated measurements utilized.

Different approaches to the coding of time can be utilized within the LS and the IS modeling strategy. In the IS model, the component of time is coded in increments of years (after placing the origin of time at the first age – which entails coding the first

age as 0). For example, a coding scheme for 3 consecutive assessment occasions that assumes the trajectory is constant over time (i.e., the slope is linear) would result in the following factor loading matrix for all individuals:

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \tag{2}$$

Fixing the loadings of each assessment occasion on the first factor (F1 = the Intercept factor) to a value of 1 ensures that it is interpreted as an initial true status (i.e., as a baseline point of the underlying developmental process under investigation). Specifying the change trajectory in increments of years on the second factor (F2 = the Slope factor) also ensures that the correlation between the Intercept and Slope factors reflects the relationship between the initial point and the slope of the proposed linear trajectory. In cases where a quadratic trajectory is assumed, the following factor loading matrix would be used and includes an additional factor (a quadratic factor) with squared loadings of the second factor (the linear factor):

$$\Lambda = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \tag{3}$$

For any assumed higher-order polynomial trajectories, appropriately patterned factor loading matrices would be needed.

In the LS modeling strategy the loadings on the first factor (F1 = called the Level factor) are also set to a value of 1, but the component of time is coded by fixing the loadings on the second factor (F2 = called the Shape factor) as follows (i.e., where * corresponds to a freely estimated loading):

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 1 & * \\ 1 & 1 \end{bmatrix} \tag{4}$$

Fixing the loading of the first and last assessment occasion on the second factor to a value of 0 and 1, respectively, ensures that this factor is interpreted as a change factor (regardless of the trajectory encountered - linear, quadratic, cubic, etc.). Freeing the loadings of the remaining assessment occasions on the same factor captures the change that occurs between the first and each of these later measurement occasions. In other words, specifying the change trajectory in this manner ensures that the freed loadings reflect the cumulative proportion of total change between two time points relative to the total change occurring from the first to the last time

point (regardless of the trajectory shape), and the correlation between the Level and Shape factors simply reflects their degree of overlap (Raykov & Marcoulides, 2006). In contrast to the IS model (Equation 2), this particular manner of specifying Level and Shape tends to focus on the change as a whole (i.e., over the length of the longitudinal process measured), as opposed to the incremental changes from one interval to the next or the possible acceleration or deceleration of the change process (e.g., as might be captured with a quadratic change). As an alternative to the loadings used in Equation 4, we could also specify non-linear trajectories by freeing the last repeated measurements and fixing the first to a linear growth rate (e.g., 0, 1, *).

In this case, for ease of demonstration we will adopt the more restricted Intercept and Slope strategy, where we specify that the growth follows a common polynomial (i.e., in the case of three repeated measures, this would be linear or quadratic). The time points at which the repeated y items are measured (e.g., with three time points it would be y_1, y_2, and y_3,) can be captured by fixed factor loadings and zero y intercepts. In other words, we define the first measurement as an initial status growth factor (η_0), setting the first time score measurement at 0. Subsequently, and assuming linear growth, we can define the second time score as 1 and with equidistant observations, this continues as $x_t = 0, 1, 2, \ldots, T - 1$. The residuals are normally distributed and uncorrelated with other variables, have means of zero, and a covariance matrix, which can be designated as Θ with different variances. Although some off-diagonal elements can be freed to represent covariances between residuals over time, it is usually assumed that there is no covariance structure between the residuals of the longitudinally observed variables. Of course, other types of growth trajectories (e.g., quadratic, nonlinear) can also be proposed and tested.

In our example, assume that a series of 3 repeated ordered waves of measurements on student math achievement is represented as Y_{it} (where the index i corresponds to each observed individual in the study and t corresponds to the time-ordered measurements). The following equation can be used to describe an individual's development over the repeated measurements (also sometimes called a level-1 or within-person model):

$$Y_{it} = \alpha_{yi} + \beta_{yi}\lambda_t + \varepsilon_{it} \tag{5}$$

where α_{yi} is the initial status measured at time 1 (i.e., the intercept or level) of an individual's change trajectory, β_{yi} is the slope or the shape (the change in Y_i between the consecutive measurements) of the change trajectory, λ_t corresponds to the measured time points, and ε_{it} to the model residual for each individual. Because α_{yi} and β_{yi} are random variables, these model parameters can be represented by a group mean intercept ($\mu_{\alpha y}$) and mean slope ($\mu_{\beta y}$) plus the component of individual intercept variation ($\zeta_{\alpha yi}$) and slope variation ($\zeta_{\beta yi}$), as indicated by the following so-called level-2 or between-person model equations for which, as with the above mentioned parameters, sample based estimates are generally obtained:

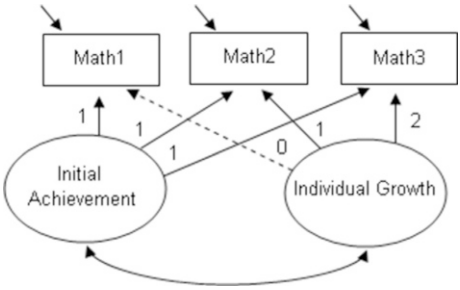


Figure 3. Example of a latent curve model assuming linear growth.

$$\alpha_{yi} = \mu_{\alpha y} + \varsigma_{\alpha yi} \tag{6}$$

$$\beta_{yi} = \mu_{\beta y} + \varsigma_{\beta yi} \tag{7}$$

A usual assumption is that there is no covariance structure between the residuals of the longitudinally observed variables, implying that the covariance matrix is diagonal. Such a model is called an unconditional model, an example of which is depicted in Figure 3.

Specifying the Latent Classes

As we noted, the second part is the specification of the latent classes as well as any covariates that might define the latent classes, as well as the intercept and slope factors. The basic latent curve model can be extended to the K latent trajectory classes of the categorical latent variable. This essentially represents the second part of the latent mixture model. In other words, the second part of the growth mixture model (i.e., the latent class part) describes individuals with similar development over time (i.e., similar growth or change trajectories) and relates these latent classes to covariates x_i (Muthén & Muthén, 1998–2012). Of interest are the class-varying means for the $K - 1$ classes. The growth factor means will likely change over the latent classes, which can result in different trajectory shapes (Muthén, 2002, 2007).

The latent growth mixture model can therefore be considered to be a combination of a continuous latent growth variable for math (represented as η_i, which consists of an intercept and slope factor) and a categorical latent variable (c) with K classes, $C_i = (c_1, c_2, \ldots c_k)'$ where $x_i = 1$ if individual i belongs to class k and 0 otherwise. For a linear growth model, the measurement model can be written as

$$y_{it} = \eta_{0i} + \eta_{1i} a_{kt} + \varepsilon_{it}, \tag{8}$$

where y_{it} ($i = 1, 2, ..., n$; $t = 1, 2, ..., T$) are outcomes influenced by the latent intercept and slope factors (η_{0i} and η_{1i}, respectively). As in traditional latent curve models, the

time scores of a are contained in a factor loading matrix specified as the Λ_k matrix. The residuals (ε_{it}) are contained in a covariance matrix Θ_k of order $T \times T$ and may vary across the latent trajectory classes (Muthén, 2002).

The intercept and slope means can then be related to one or more covariates through the following equations:

$$\eta_{0i} = \mu_{0k} + \gamma'_{0k} x_i + \varsigma_{0i}, \tag{9}$$

$$\eta_{1i} = \mu_{1k} + \gamma'_{1k} x_i + \varsigma_{1i}, \tag{10}$$

where the μ mean parameters vary across the K classes,γ are the structural parameters relating covariates to the intercept and slope coefficients such that the influence of x can vary across latent classes, and ς_{0i} and ς_{1i} are vectors of residuals assumed to be normally distributed, uncorrelated with other variables, and with mean of zero.

Let us consider the proposed growth mixture model graphically summarized in Figure 2. In this model, it is proposed that the three background variables of socioeconomic status (SES), grade point average (GPA), and mobility (i.e., whether the student changed schools) will affect students' initial status (i = Intercept) and growth rate (s = Slope) in math. In this case, the growth is assumed to be linear (coded 0, 1, 2), whereupon an Intercept-and-Slope (IS) model was utilized. The model also specifies a covariance between the initial status and slope factors. Because c is a categorical latent variable, the arrows from c to the latent growth factors indicate the intercepts of the regressions of the growth factors on the covariates vary across the classes of c. This corresponds to the regressions of the i and s factors on a set of dummy variables representing the categories of c (Muthén & Muthén, 1998–2012). In the *Mplus* model specification, the intercepts of the factors are not held equal across classes by default; however, the variances and covariances of the factors are held equal across classes as the default. Further, arrows from student SES, GPA, and mobility to c represent the multinomial logistic regressions of c on the set of background covariates (x).

As we have suggested, the latent growth mixture model provides considerable flexibility for defining across-class parameter differences. Each of the intercept, slope, and factor loading parameters can be either considered as fixed or randomly varying across classes (Asparouhov & Muthén, 2007). For example, the different shapes of the latent trajectory classes can be characterized by class-varying intercept parameters, holding Λ_k invariant across latent classes. Some classes may require class-specific variances in $\mathbf{\Theta}_k$ and Ψ_k (Muthén, 2002). Moreover, different classes may have different relations to the covariate, corresponding to class-varying γ_k coefficients (Muthén, 2002).

For purposes of illustration, we will use the default model specifications by defining the factor loadings and factor variances and covariances to be the same across the latent classes. In the *Mplus* program, this can be done by using the %OVERALL% statement without specifying differences for particular classes, as specified in

Appendix B. We could, however, subsequently relax particular restrictions of theoretical interest (e.g., factor variances and covariances, factor loadings) across classes. For interested readers, more technical descriptions and extensions of basic growth mixture models (e.g., with categorical repeated measures) can be found in Muthén (2002), Duncan et al. (2006), and the appendices of the *Mplus* User's Guide (Muthén & Muthén, 1998–2012).

Mplus Model Statements

The *Mplus* modeling statements for the proposed model presented in Figure 2 are provided in the command file provided below. In mixture modeling in Mplus, start values can be used to facilitate model estimation. For example, initially several random sets may be selected (i.e., 10 is the default). Optimization is carried out for 10 iterations for each of the 10 sets of starting values. The ending values of this initial stage of estimation are used as starting values in the final optimization stage. There is considerable flexibility in providing more thorough investigation (e.g., varying the random starts, the number of iterations used, number of optimizations carried out, and estimation algorithms). In mixture modeling, however, some starting values can generate a likelihood function with several local maxima, which suggest the importance of exploring a given model with different optimizations that are carried out with various sets of starting values (Muthén, 2002).

```
TITLE:  Growth mixture model;
DATA:   FILE IS filename.dat;

VARIABLE: Names are schcode ses gpa math1 math2 math3 moved lowses
                    acadsch;
Usevariables math1 math2 math3 gpa ses moved;
          CLASSES = c(4);

ANALYSIS: TYPE = MIXTURE;
Estimator is MLR;
starts = 100 4;
stiterations = 20;

Model:
%Overall%
i s |math1@0 math2@1 math3@2;
i on gpa ses moved;
s on gpa ses moved;
c#1 on gpa ses moved;
math1@0;
```

Plot: TYPE IS PLOT3;
SERIES IS math1-math3(*);

OUTPUT: SAMPSTAT STANDARDIZED TECH11 TECH14;

RESULTS

In this example, we examine a single-level growth mixture model for math achievement of students (N = 6,623) using three within-school covariates. The data are from a study of students' math achievement during high school. Students were observed on three occasions (with math achievement score means of 46.94, 51.64, and 55.76, respectively) with mean grade point average (GPA) = 0.06 (standardized) and SES = –0.07 (standardized), and observed mobility percentages indicating 78% remained in the same school while 22% changed schools after year 1.

To handle the presence of missing data, we used full information maximum likelihood (FIML) parameter estimation (Arbuckle, 1996). To evaluate model fit, we use the overall chi-square goodness-of-fit test, the comparative fit index (CFI), Akaike's information criterion (AIC), and the root mean square error of approximation (RMSEA) along with its associated confidence intervals. We note that because the χ^2 measure is well known to be sensitive to sample size issues, a tendency exists to reject models that are even only marginally inconsistent with the data; therefore, much more emphasis is generally placed on the other fit criteria.

A number of alternative preliminary models with different numbers of latent classes (using the strategies suggested in the latent class analysis section) were investigated using the BIC index before settling on four classes (e.g., for 3 classes, BIC = 126,134.04; for 4 classes, BIC = 125,657.85). Final classification of individuals in the identified four classes was 10.4% in class 1, 5.8% in class 2, 65.5% in class 3, and 18.3% in class 4 (with respective correctly-classified average posterior probabilities of 0.90, 0.87, 0.95, and 0.84). Just as one can use the BIC index , we note in passing that another way a researcher can examine whether she or he has a plausible number of classes is using the so-called Vuong-Lo-Mendell-Rubin test (requested with TECH11 in the *Mplus* program) and the bootstrapped parametric likelihood ratio test (requested with TECH 14). These tests compare in terms of a statistical significance *p*-value the model with *K* classes (i.e., in this case 4) to a model with (*K* – 1) classes (i.e., 3 classes). A low *p* value provides evidence that the model with *K* – 1 class is rejected in favor of the *K* class model. In this example, the Vuong-Lo-Mendell-Rubin test for the four-class model has a *p*-value of .00001. Similarly, the bootstrapped parametric likelihood ratio test has a *p* value of 0.00001. All three tests suggest that four classes fit better than three classes.

The model estimates are summarized in Table 3. First, the results suggest that students with higher GPAs and higher SES have significantly higher initial status

Table 3. Latent growth mixture model selected results

	Unstandardized Estimate	SE	Standardized Estimate
Class Invariant Estimates			
Model to Explain Intercepts			
GPA	0.506**	0.064	0.158
SES	0.798**	0.083	0.192
MOVED	0.073	0.135	0.010
Model to Explain Slope			
GPA	0.832**	0.046	0.245
SES	0.960**	0.058	0.217
MOVED	-0.208*	0.095	-0.026
Covariance			
Slope/Intercept	-1.437**	0.369	-0.163
Logistic Regression Model to Explain C#1 (Category 2 = reference group)			
GPA	0.790**	0.069	
SES	0.567**	0.070	
MOVED	0.014	0.131	
Class Specific Intercepts and Slopes			
Class #1			
Intercept	57.835**	0.231	
Slope	2.234**	0.195	
Class#2			
Intercept	32.642**	0.314	
Slope	7.060**	0.307	
Class#3			
Intercept	48.405**	0.088	
Slope	4.849**	0.058	
Class#4			
Intercept	40.227**	0.250	
Slope	3.700**	0.159	

*p < .05; **p < .01

than their peers with lower GPAs and SES, and students with higher GPAs ($\gamma = 0.832$, $p < .05$) and SES ($\gamma = 0.960$, $p < .05$) make significantly greater growth over time than their peers with lower GPAs and lower SES. Student mobility (labeled as MOVED in Table 3) is not significantly related to initial math achievement; however, there is evidence that students who move make lower growth per year than students who stay in the same school ($\gamma = -0.208$, $p < .05$). The results indicate that the means for the latent classes range from a low of 32.642 (Class 2) to a high of 57.835 (Class 1).

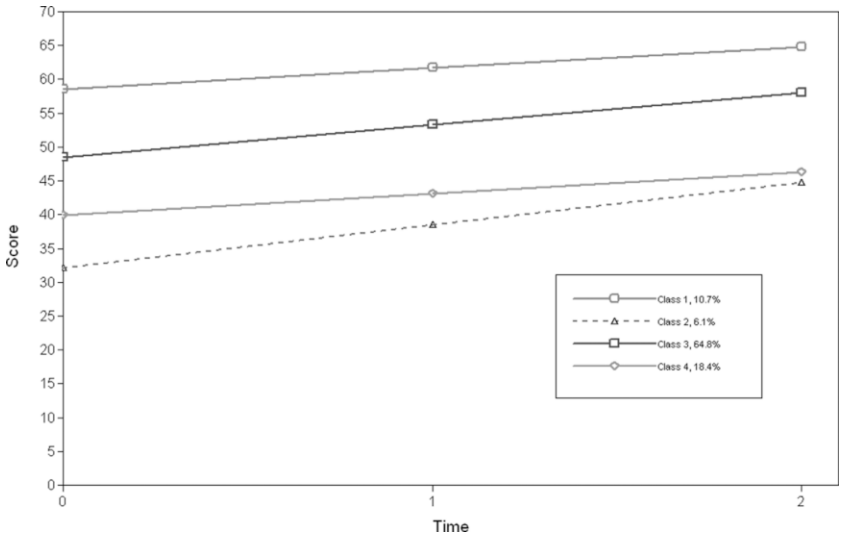


Figure 4. Estimated latent class means and growth trajectories for LGMM.

Growth means also vary considerably across the four latent classes (from a low of about 2.2 in Class 1 to a high of about 7.1 in the reference group (Class 2).

Second, regarding explaining class membership (referred to as *C*#1 in Table 3), the log odds for GPA is 0.790 ($p > .05$), the log odds for SES is 0.567 ($p > .05$), and the log odds for moved is 0.014 ($p > .10$). With respect to the reference group (i.e., Class 2, or the class with the lowest initial status and ending status), the results imply that students with higher GPAs and higher family SES are more likely to be members of other latent classes. Therefore, whether or not a student moved does not appear to affect class membership. The estimated means for each of the four latent classes over the three measurement occasions are plotted in Figure 4.

EXTENSIONS OF LGMM

The previous model provides a preliminary indication of different classes of latent growth trajectories. Such a model can also be extended to examine potential hierarchical or multilevel data structures present in the data. In such a case, the multilevel specification would use measurement models within and between groups to define the latent growth factors and within- and between-groups structural models to relate covariates at each level to the growth factors. In a multilevel mixture model, intercepts (and slopes) can be allowed to vary across both latent classes and schools. We could easily extend the previous single-level model to a multilevel model that included the possible effects of a school-level model. Using this school-level component, we could then examine the possible effects of organizational-

level variables on the level of outcomes and growth rates. For example, we might examine the relationship between two school-level variables—the SES context of the school and the quality of the school's curricular program on both the latent classes of schools with similar growth trajectories, as well as on the average level of school initial outcomes and their growth rates. The multilevel level mixture model, therefore, allows us to examine student membership in the latent trajectory classes within schools. Moreover, the latent trajectories classes themselves can be proposed to have a between-schools component as well as a within-schools component.

We note in passing that it is often more difficult to fit a multilevel mixture model to the data than a single-level model. The multilevel case can require some changes in the numerical integration process used to estimate the model (see the *Mplus User's Guide* for further discussion of these options). We believe it is important for readers to keep in mind that the choices made in estimating multilevel models with categorical outcomes (e.g., method of integration, choice of the number of random starts) can have a considerable effect on the final estimates and the time it takes to generate a solution. Models can be estimated preliminarily without using random starts, but they may not produce optimal solutions. The time it takes to estimate such models can vary considerably depending on the specification of the random start values, which was necessary to avoid local maxima in generating initial estimates. Hence, it is probably best at present to consider such multilevel results as preliminary.

CONCLUDING REMARKS

Models with categorical observed and latent variables greatly expand the range of developmental processes that can be examined in the behavioral, educational, and social sciences. Moreover, there exist a wide variety of new types of mixture models that can be considered and conducted within the SEM framework. Models such as latent growth mixture models, multilevel growth mixture models, latent transition analysis models, Markov chain models, and latent variable hybrid models are but a few of these that can be considered. As evidenced in the few example models introduced and illustrated in this chapter, these new types of models expand the ways in which we can think about how individuals' shared environments affect their individual outcomes (Asparouhov & Muthén, 2007).

It should be evident that the plethora of examples involving latent growth mixture models can vary considerably in their complexity and demands on identification and estimation time. For example, we have found that examining multilevel growth mixture models can often be more difficult to fit than simpler models. There are various options that can be considered when fitting such complex models (e.g., making changes in numerical integration process used to estimate the model, changing the number of random starts, changing the number of latent classes defined, considering whether there should be one overall model defining the latent classes or unique Level and Shape models for each latent class). All of these possibilities can affect whether a particular model converges or not and, ultimately, whether it

provides a satisfactory test of a proposed theoretical model. At the extreme, these possibilities can also lead to evidence of spurious classes. Nevertheless, despite the challenges, these kinds of models most certainly offer exciting possibilities for applied researchers to conduct studies that address questions related to particularly demanding substantive theories.

REFERENCES

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum Associates.

Asparouhov, T., & Muthén, B. (2007). Multilevel mixture models. G. R. Hancock & K. M. Samuelsen (Eds.). *Advances in latent variable mixture models*. Charlotte, NC: Information Age Publishing, Inc.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach.* New York: Wiley.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel, *Handbook of statistical modeling for the social and behavioral science* (pp. 311–359). New York: Plenum.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent growth curve modeling* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Gagné, P. (2006). Mean and covariance structure mixture models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 197–224). Greenwich, CT: Information Age Publishing.

Gottfried, A. E., Marcoulides, G. A., Gottfried, A. W., Oliver, P., & Guerin, D. (2007). Multivariate latent change modeling of developmental decline in academic intrinsic math motivation and achievement: Childhood through adolescence, *International Journal of Behavioral Development, 31,* 317–327.

Lubke, G. H*., &* Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10,* 21–39.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 561–614). New York, NY: Plenum.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, 28, 338–354.

Muthén, B. O. (2002). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.). *New developments and techniques in structural equation modeling* (pp. 1–33). Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nylund, K. L. (2007, August). Extensions of LTA with covariates and distal outcomes to study peer victimization. Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in mixture modeling: A Monte Carlo simulation. *Structural Equation Modeling, 14*(4), 535–569.

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

BARBARA M. BYRNE

17. SELECTING SEM COMPUTER PROGRAMS

Considerations and Comparisons

The rate at which structural equation modeling (SEM) has grown over the past 30 years or so has been truly quite remarkable! At least one interesting offshoot of this escalation, however, has been the somewhat parallel growth of computer software capable of handling the statistical rigors demanded by the SEM methodology. In combination, each component of this synergetic pair serves to energize the other in advancing the practice of SEM methodology. More specifically, as substantive researchers increasingly seek more comprehensive answers to an ever-widening array of research questions and SEM applications, statistical researchers are challenged to further advance the capabilities of SEM methodology, which in turn, necessitates further development of existing SEM software programs.

Since development of the first SEM program in 1974 (LISREL), the ensuing years have witnessed a steady increase in the development and revision of alternative SEM computer software such that there are now several programs from which to choose; these include: AMOS (Analysis of Moment Structures; Arbuckle, 2009), CALIS (Covariance Analysis of Linear Structural Equations; SAS), EQS (Equations; Bentler, 2005), LISREL (Linear Structural Relationships; Jöreskog & Sörbom, 2004), Mplus (Muthén & Muthén, 2007–2010), Mx (Matrix; Neale, 2002), RAMONA (Reticular Action Model or Near Approximation; Systat Software Inc., 2002), and SEPATH (Structural Equation Modeling and Path Analysis; StatSoft Inc., 2003). Although these software programs share many of the same core analytic features, they are also necessarily unique in a variety of ways. Given that a review of each is clearly beyond the scope of the current chapter, the present content focuses on only the four programs considered to be the most widely used as this edited volume goes to press. Listed alphabetically, these programs include AMOS, EQS, LISREL, and M*plus*.

Using these four programs as a backdrop to the topic of SEM application in general, and as a reference point for comparison across the varying software approaches to diverse SEM applications in particular, the intent of this chapter is to provide readers with an overview of important factors to consider in their selection of a SEM computer program. More specifically, the purposes of this chapter are fivefold: (a) to suggest factors one may wish to consider in assessing the extent to which a particular SEM program is most appropriate to one's own professional and personal needs, (b) to describe the major features of the AMOS, EQS, LISREL,

T. Teo (Ed.), Handbook of Quantitative Methods for Educational Research, 367–394.

and M*plus* software packages, (c) to outline the manner by which each of these programs addresses two critically important issues in SEM application – analysis of data that are non-normally distributed, and analysis of data that are of a categorical nature (nominal, dichotomous, ordinal), (d) to highlight perceived strengths and weaknesses of each program, and (e) to illustrate and compare an example SEM application that addresses the issues of both categorical and non-normal data based on the EQS (Version 6.2) and M*plus* (Version 6.1) programs.

The material in this chapter is presented in five sections. I begin in Section 1, by discussing what I consider to be essential aspects of SEM software in creating the best match between a program and the reader's own particular needs. Section 2 presents a general description of each program within the context of its most recent version (as this volume goes to press), followed by notation of its particularly distinctive elements and features. In Section 3, I describe the approach taken by each of the programs in addressing the issues of non-normal and categorical data. Section 4 reviews what I perceive to be the strengths and possible weaknesses of each program. Finally, in Section 5, I present an illustrative comparison of the EQS and Mplus programs with respect to components of the input file and selected analytic results in the output file.

MATCHING SEM PROGRAMS TO PROFESSIONAL AND PERSONAL NEEDS

Although the development of all SEM programs is rooted in the statistical and theoretical rudiments of this methodology, they can and do vary widely in their particular approach to diverse practical applications and their related analytic procedures. Furthermore, SEM programs differ in their capabilities for analyzing certain SEM models. Given these aberrations, it is important that program selection be tailored to one's own professional and personal needs. Accordingly, I believe that program selection should be guided by the consideration of at least three factors: (a) how knowledgeable the user is with the basic concepts and applications of SEM, (b) the types of models he or she is likely to be testing, and (c) whether the user prefers to work within a graphical versus a textual framework with respect to model specification.

Hopefully, by the time that you have worked your way through the various program descriptors in this chapter, you will have a fairly good idea which of the four programs might be most suitable to your needs. Based on my own knowledge of, and familiarity with each of the AMOS, EQS, LISREL, and Mplus programs, I offer the following recommendations. For readers who may be fairly new to SEM, I believe you will likely find the EQS program the most informative and easiest with which to work. My rationale in making this endorsement stems from its interactive facility, which makes this program the most user-friendly of the four programs reviewed here. Although the EQS program also has excellent graphical facility, some researchers may prefer to work through all applications entirely from a graphical, rather than from a textual perspective. If this aspect of a program is

important to you, then I would suggest that the AMOS program would likely serve your needs best as it provides users with a comprehensive toolbox of icons for which all analytic actions are linked. On the other hand, readers who are familiar with the use and application of SEM and/or are interested in and need to work with more advanced models may prefer to work with the EQS, LISREL, or Mplus programs. In general, however, many researchers and practitioners will likely work with at least two or a combination of SEM programs thereby capitalizing on the unique aspects and strengths of each. For a comprehensive, yet non-mathematical introduction to basic SEM concepts and applications within the framework of the specific program notation of AMOS, EQS, LISREL, or Mplus, readers are referred to Byrne (2010, 2006, 1998, 2011), respectively. In addition to introducing readers to the essential elements of SEM, each of these books literally "walks" readers through many different single- and multiple-group applications by detailing aspects of both the input and output files, in addition to their related graphical model representations.

PROGRAM DESCRIPTIONS AND UNIQUE CHARACTERISTICS

I now present a general description of the AMOS, EQS, LISREL, and Mplus programs based on their current version (as this volume goes to press). In addition, I note particular features considered to be unique to each.

AMOS (Version 18.0)

Most people tend to equate AMOS with a truly graphical approach to SEM modeling. Indeed, this perception seemingly remains as strong today as it was 25 years or so ago and likely derives from the ease, speed, and wide array of modeling tools with which the program enables the building and testing of models. Thus, it is not surprising that by and large, the majority of users base their analyses on the AMOS Graphics interface. In this regard, AMOS provides users with all the tools they will ever need in creating and working with SEM path diagrams. Each tool is represented by an icon (or button) and performs one particular function; there are 42 from which to choose. Whereas some icons represent drawing functions, others represent specific aspects of the modeling process itself. A few examples of the latter are as follows: Data icon (selects and reads data files), Calculate Estimates icon (calculates default and/or requested estimates), Analysis Properties icon (allows for additional calculations such as, for example, modification indices, squared multiple correlations), Multiple Groups icon (enables analyses of multiple groups), and Bayesian icon (enables particular analyses based on Bayesian statistics). Immediately upon opening the program, this toolbox of icons appears to the left of a blank workspace. For a step by step guide to using these icons in building and testing a wide variety of SEM models, readers are referred to Byrne (2010).

Despite the popularity of its Graphics module, AMOS does provide for an alternative approach to SEM modeling that operates within a programming interface

complete with a built-in editor. As such, the user specifies a model via equation-like statements. Initially, this programming mode was termed AMOS Basic. (For a review of model input files based on AMOS Basic, see Byrne, 2001.) However, with the introduction of Version 6.0 in 2005, AMOS Basic was subsequently replaced by the VB.net (Visual Basic) and C# languages. Nonetheless, given apparent user preference for the AMOS Graphics facility over the equations format, no example files based on VB.net or C# are included in the second edition of Byrne (2010).

Regardless of which approach to structuring model input is preferred, all options related to the analyses are available from drop-down menus, and all parameters can be presented in text format. In addition, AMOS Graphics allows for the estimates to be displayed graphically in a path diagram. Thus, the choice between these two AMOS approaches to building and testing SEM models ultimately boils down to one's comfort level in working within a graphical versus a mathematically-oriented framework.

In contrast to the other three programs considered here, AMOS does not accept data in ASCII (.dat) format; rather, only data formats that include: Microsoft Excel (.xls), SPSS (.sav), and text delimited files (.txt) are supported. Estimation methods include: ULS, GLS, ML, ADF[1], and scale-free least squares (SLS). In addition to providing for the testing of CFA, full SEM path, structured means, multiple-group, and latent growth curve models, AMOS can now (as of Version 16) facilitate the analysis of mixture models. A second recent addition to the AMOS program is the capability to handle categorical data. Of import, however, is that analyses based on categorical data in general, and analysis of mixture models in particular, are enabled only through use of the AMOS Bayesian statistics facility. (For a step-by-step illustration of the Bayesian approach to analyses based on categorical data, readers are referred to Byrne, 2010.)

Perhaps one of the most unique features of the AMOS program is its Specification Search facility. Although the AMOS Graphics interface is most typically used in the analysis of hypothesized SEM models, the program has extensive capabilities for the conduct of exploratory analyses enabled via the Specification Search function. In using this facility, the researcher allocates certain paths in the model diagram to be optional. The program then fits the model to the data using every possible subset of paths and the models are subsequently sorted according to their fit to the data based on particular fit statistics. An important caveat concerning this feature, however, is that the final best-fitting model must be substantiated by theory and other empirical research. A second feature of AMOS that may be considered unique lies with its broad array of bootstrapping capabilities in that it is able to: (a) generate bootstrapped standard errors and confidence intervals for all parameter and effect estimates, as well as for sample means, variances, covariances, and correlations, (b) implement percentile intervals and bias-corrected percentile estimates (Stine, 1989), and (c) perform the bootstrap approach to model testing proposed by Bollen and Stine's (1992). Two final unique aspects of the AMOS program are: (a) its Bayesian approach to the analysis of categorical data, and (b) its specification and testing of

multiple-group models, which can be accomplished by means of either a graphical approach or an icon-generated approach.

Contact information: SPSS an IBM Company
(General): http://www.SPSS.com
(Technical Support): http://support.spss.com

EQS (Version 6.2)

To the best of my knowledge, EQS is the only SEM program capable of performing comprehensive data management tasks, pre-analytic data screening, and exploratory non-SEM analyses, over and above its full range of SEM modeling capabilities thereby bypassing the need of other statistical packages for these purposes. Given that EQS is a Windows-based program, all procedures are selected from drop-down menus with subsequent operations being completed interactively. A few of the many management tasks include conditional case selection, variable labeling and transformation, group formation, data merging, and coding of missing data. Examples of data screening capabilities include graphical presentations in the form of either charts or plots such as histograms, pie charts, scatterplots, boxplots, and error bar plots, with color coding preferences being available for each. Finally, exploratory analyses available to users are: descriptive statistics (this information is also provided in all SEM output files), frequencies, t-tests, analysis of variance (ANOVA), cross-tabulations, factor analyses, correlations, non-parametric tests, multiple regression (standard, stepwise, and hierarchical), and intraclass correlations. One additional set of exploratory analyses pertinent to the detection of missing data are worthy of separate mention. In this instance, the user is provided with options related to: (a) display of missing data (e.g., exclusion of cases having a specified percentage of missing data), (b) search for evidence of missing values and if detected, a thorough diagnosis and reporting of these findings, and (c) imputation of missing values based on means, group means, regression or unstructured expectation maximization (EM).

In the EQS program, the specification of all models is based on the Bentler-Weeks (1980) representation system. As such, all variables in a model, regardless of whether measured (i.e. observed) or unmeasured (i.e., latent), can be categorized as either dependent or independent variables. Any variable having a single-headed arrow pointing towards it represents a dependent variable; if no arrow is present, it represents an independent variable. This simple representation system therefore makes it very easy to conceptualize and specify any SEM model. Likewise, the EQS notation is equally straightforward. All measured variables are designated as V's and constitute the actual data of a study. All other (unmeasured) variables are hypothetical and represent the structural network of the phenomenon under investigation. There are three such variables: (a) the latent construct, regarded generally as a factor in EQS and designated as F, (b) a residual (i.e., error term) associated with any measured variable, designated as E, and (c) a residual associated with the prediction of each factor, designated as D (disturbance term).

Model specification in EQS can be conducted in one of three ways – manually, interactively, and graphically. Based on the *manual approach*, users write out the model specifications consistent with its hypothesized parameterization, thereby creating the model input file. Despite the simplicity and straightforwardness of this approach, it is nonetheless the most time-consuming. Nonetheless, the manual approach can be invaluable in helping users make the link between a schematic portrayal of the model under study and the equivalent equation statements expressed in EQS language. The *interactive approach* to EQS model specification is implemented through use of a feature labeled "Build_EQS". This method presents the user with a series of dialog boxes, each of which relates to a particular section of the input file. Completion of each dialog box results in a line by line automatic building of the input file. Finally, the EQS *graphical approach* to model specification is enabled through of the "Diagrammer" facility. For a comprehensive review of the EQS notation and input file, together with a step by step walk-through in building an EQS file based on each of the three approaches noted above, readers are referred to Byrne (2006).

EQS can read raw ASCII (.dat), covariance matrix, and SPSS (.sav) data, albeit all imported data are automatically converted to EQS format (.ess). Although some might initially perceive this transition as a possible hindrance, nothing could be further from reality. Indeed, this transition to an EQS system file is more appropriately regarded as an advantage in that all information related to the data file (e.g., labels, scaling, sample size) is retained by the program in this .ess file, thereby facilitating any subsequent structuring of files and/or graphical displays. Estimation methods include: ULS, GLS, ML, robust ML including residual-based chi-square and F tests, and asymptotic distribution-free (ADF) including corrected ADF tests; as well as their counterparts for correlation structure models.[2] Finally, EQS tests a full range of SEM models for both continuous and categorical data that include latent regression, CFA, full SEM path, structured means, multiple-group, latent growth curve and multilevel models with the latter capable of being based on three different analytic estimation approaches: (a) ML, (b) Muthén's (1994) approximate estimator (MUML), and hierarchical linear-like modeling (HLM). In addition, EQS provides for the conduct of easy-to-use bootstrapping procedures and simulation analyses with summaries.

At least two aspects of EQS are very unique among its competitors. First, its multifaceted data management facility as described earlier is an unparalleled program bonus that is worth its weight in gold. Second, its very efficient and easy-to-use graphical interface enables users to effortlessly transfer any model schema into Word for ease of publication purposes.

Contact information: Multivariate Software Inc.
(General): http://www.mvsoft.com
(Technical Support): support@mvsoft.com

LISREL (Version 8.8)

LISREL is packaged as a suite of programs that includes its initial companion program, PRELIS 2, with both programs being compatible with Windows 7. The LISREL program is structured around what was commonly referred to in the 70's and early 80's as the "LISREL" model and serves as the workhorse for all SEM analyses.

PRELIS (Jöreskog & Sörbom, 1996) was designed specifically to serve as a preprocessor for LISREL and hence the acronym PRElis. Nonetheless, because it can be used effectively to manipulate and save data files, as well as to provide an initial descriptive overview of data, it can also function as a stand-alone program. In addition to providing descriptive statistics and graphical displays of the data, PRELIS can prepare the correct matrix to be read by LISREL when the data: (a) are continuous, ordinal, censored, or any combination thereof, (b) are severely non-normally distributed, and/or (c) have missing values. In addition, PRELIS can be used for a variety of data management functions such as recoding and transformation of variables, case selection, computation of new variables, and creation and merger of raw data files. Finally, PRELIS can be used to generate bootstrapped samples and estimates, as well as to conduct simulation studies with variables specified to have particular distributional characteristics.

Model specification in LISREL can take one of three distinctive forms, each of which is conducted interactively. First, and most complex, is via the original LISREL commands, the syntax of which requires an understanding of matrix algebra in general, and matrices pertinent to specification of CFA and full SEM path analytic models in particular. In LISREL, these matrices are denoted by upper case Greek letters while their elements, which represent the model parameters, are denoted by lower-case Greek letters. Unlike earlier versions of LISREL, however, the Windows interface enables an interactive approach whereby the user is prompted for model and data information; once this information is entered, the related command syntax is completed automatically. The second and less complex form of model specification is through use of SIMPLIS, a command language introduced in the initial version of LISREL 8 (Jöreskog & Sörbom, 1993a). This approach requires only that the user name the observed and latent (if any) variables, together with the estimated regression paths. Indeed, Jöreskog & Sörbom (1993b, p. i), in their introduction of this new command language stated "It is not necessary to be familiar with the LISREL model or any of its submodels. No Greek or matrix notations are required". The third and most intuitive approach to model specification in LISREL is via the graphical interface. In this instance, the user simply constructs the path diagram on the screen and then identifies the regression paths to be estimated.

Of the three approaches to model specification, it would appear (at least from a review of reported SEM analyses), that the SIMPLIS command language is the most popular; use of the original LISREL command language seems limited to researchers familiar with its earlier versions and/or having a strong statistical background.

Only raw data in ASCII (.dat) format can be read directly by LISREL. Other data forms such as SPSS (.sav) and Excel (.xls) must first be imported to PRELIS whereupon the file is then converted to covariance matrix format. Estimation methods include: unweighted least squares (ULS), generalized least squares (GLS), maximum likelihood (ML), robust ML, weighted least squares (WLS), and diagonally weighted least squares (DWLS). Finally, LISREL enables tests applicable to a broad range of SEM models that include CFA, full SEM path, structured means, multiple-group, multilevel, latent growth curve, and mixture models.

In my view, two features in particular distinguish LISREL from the other three programs reviewed in this chapter: (a) its continued use of the original LISREL command language, and (b) its companion preprocessor program, PRELIS. With respect to the original LISREL matrix-linked syntax, I consider this feature to be a highly constructive in the sense that it compels users to think through their model specification within the framework of its related parameter matrices. As such, this approach can be extremely instructive in assisting those new to SEM to develop a solid understanding of the SEM methodology. With respect to my second point, I consider the need for a separate preprocessor program to be somewhat limiting when compared, for example, with the EQS program, which is capable of performing these same tasks, in addition to many more, without the need of an additionally supporting program.

Contact information: Scientific Software International
(General): http://www.ssicentral.com/lisrel
(Technical Support): lisrel@ssicentral.com

Mplus (Version 6.12)

Mplus provides for the specification and testing of many different models based on a wide choice of estimators and algorithms for analyses of data that are continuous, ordered categorical (ordinal), unordered categorical (nominal), censored, and binary and any combination thereof. The program is divided into a basic program and three optional add-on modules. The basic unit, (Mplus Base), provides for the estimation of regression, CFA, EFA, SEM, and latent growth models, as well as discrete- and continuous –time survival models. Module 1 and Module 2 support the estimation of a wide variety of mixture and multilevel models, respectively. Module 3 contains all features of the first two Modules and, in addition, enables the estimation of several advanced models that combine the features of both. Consistent with the other three programs, Mplus is Windows-based with drop-down menus allowing for the full range of usual editing procedures, as well as for model execution. In addition, Mplus provides for specification of graphical displays of observed data and analysis results based on a post-processing graphics model. Unlike the other three programs, as this volume goes to press, Mplus does not have a graphical interface and works solely within the framework of a programming interface. As a result, the building of its input files is always equation-based.

Model specification in Mplus is relatively straightforward and entails a maximum of 10 command statements. As might be expected, however, each of these commands provides for several options that can further refine model specification and desired outcome information. Given that, typically, only a small subset of these 10 commands and their options is needed, even specification of very complex models requires only minimal input. This minimization of input structure has been made possible largely as a consequence of numerous programmed defaults chosen on the basis of models that are the most commonly tested in practice.

To assist with the building of input files, Mplus provides for the optional use of its Language Generator, an interactive facility that leads users through a series of screens that prompt for information pertinent to their data and analyses. One limitation of this feature, however, is that its functioning terminates at the point where details related to the model under test must be specified. As such, this information must be added manually to the input file. Additionally, commands related to the transformations of variables (DEFINE), post analysis graphical displays (PLOT), simulation analyses (MONTECARLO) and any features added to the program following Version 2.0 must be manually inserted. New users of Mplus, in particular, will find this language generating resource to be extremely helpful as not only does it reduce the time involved in structuring the file, but it also ensures the correct formulation of commands and their options. (For a step-by-step illustration of the Mplus Language Generator as applied to the model, see Byrne, 2011.)

Mplus reads data in ASCII (.dat) format only. Thus, if the data to be analyzed are saved in another format (e.g., SPSS .sav), conversion to ASCII format is essential. Furthermore, any variable labels appearing on the first line of the data file must be deleted. Although estimation methods in the main include: ULS, GLS, WLS, and ML, robust and other variants related to the ML estimator (MLM, MLMV, MLR,), and weight variants related to both the ULS (ULSMV) and WLS (WLSM, WLSMV) estimators provide for a wide variety from which to choose. However, selection of the most appropriate estimator is further conditioned by both the type of model under test and the type of outcome variables being analyzed (i.e., all continuous, at least one binary or ordered categorical variable, at least one censored, unordered categorical, or count variable). Furthermore, certain conditions (missing data) can apply.

Models capable of being tested in Mplus are many and can be classified according to whether they include continuous latent variables (e.g., CFA, full SEM, and latent growth models), categorical latent variables (e.g., path analysis mixture, loglinear, and multiple-group models), or a combination of both (e.g., factor mixture, SEM mixture, and latent growth mixture models). In addition, Mplus provides for two approaches to the analysis of complex survey data. Whereas the first approach takes into account stratification, non-independence of observations resulting from cluster sampling, and/or unequal probability of selection, the second approach (commonly known as multilevel modeling) allows for the modeling of non-independence of observations (due to clustering) at each level of the data. Beyond these modeling capabilities, Mplus has wide-ranging Monte Carlo simulation capabilities for both

data generation and data analysis, and in addition, allows for standard, as well as the Bollen-Stine (1992) residual bootstrapping facilities.

I consider the unique features of the Mplus program to be twofold. The first of these relates to its capability to analyze a wide array of specified models based on a variety of data scaling formats. A second very unique feature of the Mplus program is its exceptionally comprehensive and informative website, which is updated on a regular basis. In addition to an extensive source of material pertinent to the program itself, users can seek out and access papers on particular topics, review a backlog of questions and answers on specific analytic issues, and find details related to upcoming training courses.

Contact information: Muthén and Muthén
(General): http://www.statmodel.com
(Technical Support): support@statmodel.com

ANALYSIS OF NON-NORMAL AND CATEGORICAL DATA: COMPARATIVE APPROACHES

Two areas of concern in SEM are (a) analyses of data that are non-normally distributed, and (b) analyses of data that are either nominally or ordinally-scaled. Although not all SEM programs are capable of addressing these issues, the four considered in this chapter are, albeit they differ in their individual approaches to both. We now review these differences.

Analysis of Non-normal Data

By default, all SEM programs are based on maximum likelihood (ML) estimation. However, a critically important assumption in these analyses is that the data have a multivariate normal distribution in the population. Violation of this assumption can seriously invalidate statistical hypothesis-testing with the result that the normal theory test statistic (χ^2) may not reflect an adequate evaluation of the model under study, thereby leading to results that may be seriously misleading (Hu, Bentler, & Kano, 1992).

AMOS (Version 18.0)

One approach to working with data that are non-normally distributed is use of asymptotic (large sample) distribution-free (ADF) estimators for which normality assumptions are not required. The early work of Browne (1984) was instrumental in the development of this methodology. One approach to the analysis of non-normal data in AMOS is based on this ADF estimator which can be selected from the Estimation tab of the Analysis Properties icon or drop-down View menu of AMOS Graphics. However, one major limitation associated with this approach in addressing non-normality has been its excessively demanding sample-size requirement. It is

now well known that unless sample sizes are extremely large (1000 to 5,000 cases; West, Finch, & Curran, 1995), the ADF estimator performs very poorly and can yield severely distorted estimated values and standard errors (Curran, West, & Finch, 1996; Hu et al., 1992; West et al., 1995). More recently, statistical research has suggested that, at the very least, sample sizes should be greater than 10 times the number of estimated parameters, otherwise the results from the ADF method generally cannot be trusted (Raykov & Marcoulides, 2000).

A second viable approach to analyses of non-normal data is based on a procedure known as "the bootstrap" (Efron. 1979; West et al., 1995; Yung & Bentler, 1996; Zhu, 1997). This approach is the one most commonly taken by AMOS users in addressing the issue of non-normality. The bootstrapping technique enables the researcher to compare the extent to which the ML estimates deviate across the total number bootstrapped samples. (For an example application of this bootstrapping approach, see Byrne, 2010.)

EQS (Version 6.2)

Although other estimation methods have been developed for use when the normality assumption does not hold (e.g., ADF; elliptical; heterogeneous kurtotic), Chou, Bentler, and Satorra (1991), and Hu et al. (1992) have argued that it may be more appropriate to correct the test statistic, rather than use a different method of estimation. Satorra and Bentler (1988) developed such a statistic that incorporates a scaling correction for the χ^2 statistic when distributional assumptions are violated; its computation takes into account the model, the estimation method, and the sample kurtosis values. The resulting Satorra-Bentler (corrected) chi-square value (S-Bχ^2) and standard errors are said to be "robust," meaning that their computed values are valid, despite violation of the normality assumption underlying the estimation method. The S-Bχ^2 has been shown to be the most reliable test statistic for evaluating mean and covariance structure models under various distributions and sample sizes (Hu et al., 1992; Curran et al., 1996). In addition, EQS computes robust versions of the CFI, RMSEA, and the 90% C.I. for the latter.

In addition to its usual application with continuous data, the S-Bχ^2 can be used with non-normal categorical data. Although it basically treats the ordered data as if they were continuous, DiStefano (2002) has reported this scaled approach to be beneficial in yielding standard errors that are more precise than ML estimates for non-normally distributed data having as few as three ordered categories.

In SEM, data are often incomplete, as well as non-normally distributed. When this condition holds, correction based on the S-Bχ^2 is not appropriate. Rather, analyses should be based on the Yuan-Bentler (2000) scaled statistic (Y-Bχ^2) which corrects both the test statistics and standard errors when the input file specifies the use of robust statistics and indicates the presence of missing data.

In addition to these scaled statistics, EQS has three distribution-free statistics based on the distribution of residuals; robust versions of these test statistics are

automatically computed when this option is specified. The first of these, the residual-based statistic, is of a type developed by Browne (1984). As noted earlier, however, use of this statistic is curtailed by the fact that its interpretation is meaningful only when sample size is very large. In contrast, the Yuan-Bentler residual-based statistic (Yuan & Bentler, 1998) represents an extension of Browne's (1984) residual-based test such that it can be used with smaller samples. Of particular note, however, is that in addition to performing better in small samples than the original residual-based statistic, it does so without any loss of its large-sample properties (Bentler, 2005). Finally, the Yuan-Bentler residual-based F-statistic (Yuan & Bentler, 1998), designed to take sample size into account more adequately, represents a more extensive modification of Browne's (1984) statistic and is considered by Bentler (2005) to be the best available residual-based test at this time.

LISREL (Version 8.8)

This ADF approach noted earlier is the one embraced by the LISREL program in dealing with this non-normality issue. Implementation of this strategy, however, involves a two-step process. Step 1 involves use of the PRELIS companion package in recasting the data into asymptotic matrix form, whereas Step 2 focuses on analysis of this restructured matrix based on use of LISREL with weighted least squares (WLS) estimation. However, given the restrictions of sample size noted earlier, this option is typically of little use to most practical researchers.

Mplus (Version 6.12)

As with the EQS program, treatment of non-normal data in Mplus is addressed via estimators that can yield corrected test statistics and standard errors. However, these robust estimators vary according to measurement scale of the data as well as to whether the data are complete or incomplete. For data with outcome variables that are continuous, non-normally distributed, and complete, the MLM estimator is most appropriately used; it provides for correction to the estimates, standard errors and a mean-adjusted chi-square statistic that is reportedly equivalent to the S-Bχ^2 statistic (Muthén & Muthén, 2007–2010). Likewise, the MLMV estimator, although computationally more intensive than MLM, is similarly robust albeit that the chi-square statistic is both mean and variance-adjusted. In the event that continuous data are both non-normal and incomplete, it is most appropriate to base analyses on the MLR estimator. Muthén and Muthén (2007–2010) posit that the MLR estimator yields a corrected chi-square statistic that is asymptotically equivalent to the Y-B χ^2.

When data are both categorical and non-normally distributed, Mplus provides for the use of two robust weighted least squares (WLS) estimators. Although both the WLSM and WLSMV estimators use a diagonal weight matrix with standard errors that use a full weight matrix, the chi-square test statistic for the WLSM estimator is mean-adjusted whereas this statistic is mean- and variance-adjusted for the WLSMV

estimator. Importantly, these estimators are not appropriate for use with data that are incomplete and involve censored, unordered, or count dependent variables.

Analysis of Categorical Data

In conducting SEM with categorical data, analyses must be based on the correct correlation matrix. Where the correlated variables are both of an ordinal scale, the resulting matrix will comprise polychoric correlations; where one variable is of an ordinal scale, while the other is of a continuous scale, the resulting matrix will comprise polyserial correlations. If two variables are dichotomous, this special case of a polychoric correlation is called a tetrachoric correlation. If a polyserial correlation involves a dichotomous, rather than a more general ordinal variable, the polyserial correlation is also called a biserial correlation.

AMOS (Version 18.0)

The methodological approach to analysis of categorical variables in AMOS differs substantially from that of the other above-noted programs. In lieu of ML or ADF estimation, AMOS analyses are based on Bayesian estimation. As with other analyses based on AMOS Graphics, all analyses are initiated either through selection of the appropriate icon from the toolbox, or from the appropriate pull-down menu.

Because Bayesian analyses require the estimation of all observed variable means and intercepts, the first step in the process is to request this information via the Analysis Properties dialog box. Once you have the appropriately specified model (i.e., the means and intercepts are specified as freely estimated), you are ready to move on to Step 2, which invokes the Bayesian analyses. To activate these analyses, again you either click on the Bayesian icon in the toolbox, or pull down the Analyze menu and select Bayesian Estimation. Once you do this, you will be presented with the Bayesian SEM window where you will note a column of numbers in which the latter are constantly changing. The reason for these ongoing number changes is because as soon as you request Bayesian estimation, the program immediately initiates the steady drawing of random samples based on the joint Posterior distribution. This random sampling process is accomplished in AMOS via an algorithm termed the Markov Chain Monte Carlo (MCMC) algorithm. The basic idea underlying this ever-changing number process is to identify, as closely as possible, the true value of each parameter in the model. This process will continue until you halt the process by clicking on the Pause button. For a thoroughly detailed walk-through of an illustrated application based on this AMOS Bayesian approach to the analysis of categorical data, readers are referred to Byrne (2010).

EQS (Version 6.2)

Until recently (see Mplus text that follows below), two primary approaches to the analysis of categorical data (Jöreskog 1990, 1994; Muthén 1984) have dominated the

estimation of polychoric and polyserial correlations, followed by a type of asymptotic distribution-free (ADF) methodology for the structured model. Unfortunately, the positive aspects of these categorical variable methodologies have been offset by the ultra-restrictive assumptions noted above and which, for most practical researchers, are both impractical and difficult to meet. In light of these limitations, Bentler (2005) has argued that it may make more sense to correct the test statistic using a mode of estimation that works well with not-too-large samples. The use of an improved estimator of polychoric and polyserial correlations (Lee, Poon, & Bentler, 1995), together with "robust" methodologies, distinguishes the EQS approach to the analysis of categorical data from that of other SEM programs.

Consistent with the traditions of Muthén (1984), Jöreskog (1994), and Lee, Poon, & Bentler (1990, 1992), EQS follows a 3-step sequential approach to estimation. Univariate statistics such as thresholds are estimated first, followed by estimation of bivariate statistics such as correlations; estimation of the SEM model is completed using a method like ML, followed by "robust" computations based on an appropriate weight matrix. (For technical details related to this three-stage approach, readers are referred to Bentler, 2005, and to the original articles.) It is important to note that, although the correlation estimates and weight matrices in EQS are similar to those of Muthén (1984) and Jöreskog (1994), they are not identical.

From the perspective of sample size, at least, the EQS approach to analysis of categorical data is more practical than the one based on full estimation. Whereas sample size requirements for both the Muthén (1984) and Jöreskog (1994) methodological strategies have been reported as substantial (see e.g., Dolan, 1994; Lee, Poon, & Bentler, 1995), those associated with the ML ROBUST approach in EQS are much less so. Indeed, Bentler (2005) contends that the ROBUST methodology allows for the attainment of correct statistics, which are quite stable even in relatively small samples. Although the ML estimator is not asymptotically optimal when used with categorical variables, the inefficiency is small, and certainly offset by improved performance in smaller samples. The Satorra-Bentler scaled χ^2 and ROBUST standard errors provide trustworthy statistics.

LISREL (Version 8.8)

As with the analysis of non-normal data, those involving categorical outcomes are locked into a two-step process. Here again, the PRELIS program is used to generate the correct correlation matrix for the SEM analyses to be based on LISREL. Accordingly, a polychoric correlation matrix is computed for the analysis of ordinal variables, and a tetrachoric correlation matrix for dichotomous variables. Estimation is typically based on the WLS estimator, which as noted earlier, demands exceptionally large sample sizes. Indeed, in a study of this estimator with small and moderate sample sizes, Muthén and Kaplan (1992) found an oversensitivity of the $\chi 2$ statistics, as well as increased negative bias of the standard errors with increased model complexity. These findings have led Flora and Curran (2004) to conclude that WLS is not a good estimator of

categorical data. Alternatively, analyses can be conducted using the DWLS estimator, which represents a mathematically simpler version of the WLS estimator that Kline (2011) suggests may be better when the sample size is somewhat moderate.

Mplus (Version 6.12)

Despite attempts to resolve difficulties associated with SEM analyses of categorical data over the past few years, there appear to be only three primary estimators: unweighted least squares (ULS), WLS, and diagonally weighted least squares (DWLS) (Byrne, 2011). Corrections to the estimated means and/or means and variances based on only ULS and DWLS estimation yield their related robust versions as follows: ULSMV (correction to means and variances of ULS estimates, WLSM (correction to means of DWLS estimates, and WLSMV (correction to means and variances of DWLS estimates. Of these, Brown (2006) contends that the WLSMV estimator performs best in the CFA modeling of categorical data. Mplus currently offers 7 estimators (see Muthén & Muthén, 2007–2010) for use with data comprising at least one binary or ordered categorical indicator variable.

Of particular note with the Mplus program is its use of the WLSMV estimator, which is default for the analyses of categorical data based on CFA and SEM analyses. Developed by Muthén, du Toit, and Spisic (1997), it was based on earlier robustness research reported by Satorra and Bentler (1986, 1988, 1990) and intended for use with small and moderate sample sizes (at least in comparison with those needed for use with the WLS estimator). The parameter estimates derive from use of a diagonal weight matrix (W) and robust standard errors and mean- and variance-adjusted χ^2 statistic (Brown, 2006). Thus, the robust goodness-of-fit test of model fit can be considered analogous to the Satorra-Bentler scaled χ^2 statistic. Subsequent simulation research related to the WLSMV estimator has shown it to yield accurate test statistics, parameter estimates, and standard errors under both normal and non-normal latent response distributions across sample sizes ranging from 100 to 1,000, as well as across four different CFA models (1-factor with 5 and 10 indicators; 2-factor with 5 and 10 indicators; see Flora & Curran, 2004). As this edited volume goes to press, the WLSMV estimator is available only in Mplus.

PROGRAM STRENGTHS AND WEAKNESSES

It is important that I preface this section by noting that the program strengths and weaknesses cited here are solely my own perceptions based on my extensive use of each within the frameworks of teaching, research, and publication.

AMOS (Version 18.0)

Although AMOS distinguishes itself from the other three programs considered here with respect to its Specification Search function, this capability may generally not be perceived as a particular strength of the program. In my view, AMOS' primary

forte unquestionably lies with the Graphics interface in terms of its model building, model specification, and model execution capabilities. In addition, however, the ease with which one can access and use an extensive selection of bootstrapping capabilities has long been popular with researchers who may or may not have a solid understanding of SEM methodology.

In my view, one of the most notable weaknesses of the AMOS program has been its lack of a viable and efficient technical support facility. Given its recent change of ownership, it remains to be seen whether such customer service takes a turn for the better. A second weakness that I personally, find to be very frustrating is the program's inability to directly address the issue of non-normality by simply using the appropriate robust estimator. Finally, a third weakness, which, admittedly, may not be perceived by others as a strength rather than a weakness, is the need to use a Bayesian approach in testing models based on data that are of a categorical nature.

EQS (Version 6.2)

Particular strengths of the EQS program are numerous, albeit restriction of space allows for only a succinct summary here. First, the ease with which a user can build an input file, manage and explore data, and graphically construct models of publication-acceptable quality is incomparable among its program contemporaries. Second, the large number of different statistical methods available in the analysis of diverse types of non-normal data is again unrivalled. Third, EQS provides for a wide and varied range of graphical and analytic procedures in the detection, diagnosis, and estimation of missing data. Fourth, EQS has established statistical methods for testing samples that may be of less than optimal size. Fifth, unlike other SEM programs in calculating only an alpha coefficient of reliability, EQS, in addition, reports maximal, model-based, and greatest lower-bound reliability coefficients. Finally, the EQS support facility is efficient and operates within a very fast turnaround time frame.

Although not a particularly major issue, one possible limitation of the EQS program at the present time is the unavailability of online access to both the manual and the User's Guide. However, this issue is expected to be resolved once Version 7 becomes available in 2014.

LISREL (Version 8.8)

In my view, one of the major strengths of the LISREL program is that, despite the availability of the more user-friendly SIMPLIS version, it has continued to maintain its original LISREL matrix-linked syntax. As noted earlier, I consider this aspect of the program to be extremely constructive as it compels users to think through their model specification within the framework of its related parameter matrices. As such, this approach is extremely helpful in assisting those new to SEM to develop a solid understanding of the SEM methodology. On the other hand, a major weakness of the LISREL program is its need for a separate preprocessor program in order to be able to address issues such as non-normality and use of categorical data.

Mplus (Version 6.12)

Although Mplus has many superior model-specific capabilities, limitations of space permit only reflections of the program in general. In my view, the most outstanding aspect of Mplus is its capacity to estimate an absolutely phenomenal number of different models based on an equally expansive variety of data types. In this regard, it is unquestionably in a class all its own. Of course, a second outstanding strength of Mplus is its longstanding capabilities in dealing with categorical data. Finally, although not specific to the program itself, the Mplus website serves as an exceptionally rich and invaluable source of information in the form of articles, training materials, group discussions, and program updates, all of which bear on various aspects of both SEM and particular program applications. This resource, together with regular email notifications and almost instantaneous technical support assistance, in my view, is certainly one of its most valuable assets.

Perceived limitations of the Mplus program, I believe, can be linked to the user's degree of knowledge of, and experience with the application of SEM methodological procedures. In particular, the extensive number of programmed defaults in the program can be very confusing and somewhat daunting for someone new to both the concepts and application of SEM. A second perceived limitation of the program could be its lack of a graphical interface as the user is left to his/her own devices in producing related schema related to models under study.

COMPARATIVE APPICATION OVERVIEW OF EQS AND MPLUS

The intent of this section is to give readers some essence of how programs can differ in terms of (a) specification of the model as documented in the input file, and (b) the reporting of selected results in the output file. Given the popularity and efficiency of both the EQS and Mplus programs, together with the soundness of their theoretical and methodological underpinnings, I consider their comparison to be most appropriate. Although necessarily limited by chapter space restrictions, this brief overview provides at least a quick glimpse into the extent to which the two programs are similar, as well as dissimilar. Readers interested in more extensive program details related to this application are referred to Byrne (2006) and Byrne (2011), with respect to EQS and Mplus, respectively.

The illustrative example presented here examines a first-order CFA model designed to test factorial validity of the Maslach Burnout Inventory (MBI; Maslach & Jackson, 1986) for use with teachers. The application is taken from a study by Byrne (1994) the primary purpose of which was to test for both the validity and invariance of this measuring instrument across calibration and validation samples and, subsequently, across gender. The example here is limited only to tests for its validity across male elementary teachers. A schematic portrayal of this model within the framework of both EQS and Mplus is presented in Figure 1.

Of the two graphical representations shown in Figure 1, it is evident that the EQS Diagrammer-produced model is much more explicit in the identification of parameters than is the case for the same model as typically presented for Mplus analyses. Although the model shown here is exactly specified for each program, the EQS-labeled one visually informs the reader that the first factor loading of each congeneric set of indicator variables (i.e., MBI items) is constrained to a value of 1.00 for purposes of statistical identification, with asterisks indicating that the remaining loadings are freely estimated, along with the factor covariances and error variances (termed residuals in Mplus). Although no asterisks are automatically produced for the factor variances, they too are freely estimated.[3] Specifically demarcated only in the EQS figure, the regression path leading from each observed variable to its related error term is automatically fixed to 1.00 in all SEM programs as only the error variance is of interest. Finally, EQS automatically assigns a V-label to all observed variables, as well as a number in accordance with their data entry placement. Analogously, error variances are assigned an E-label that is numerically consistent with its related observed variable.

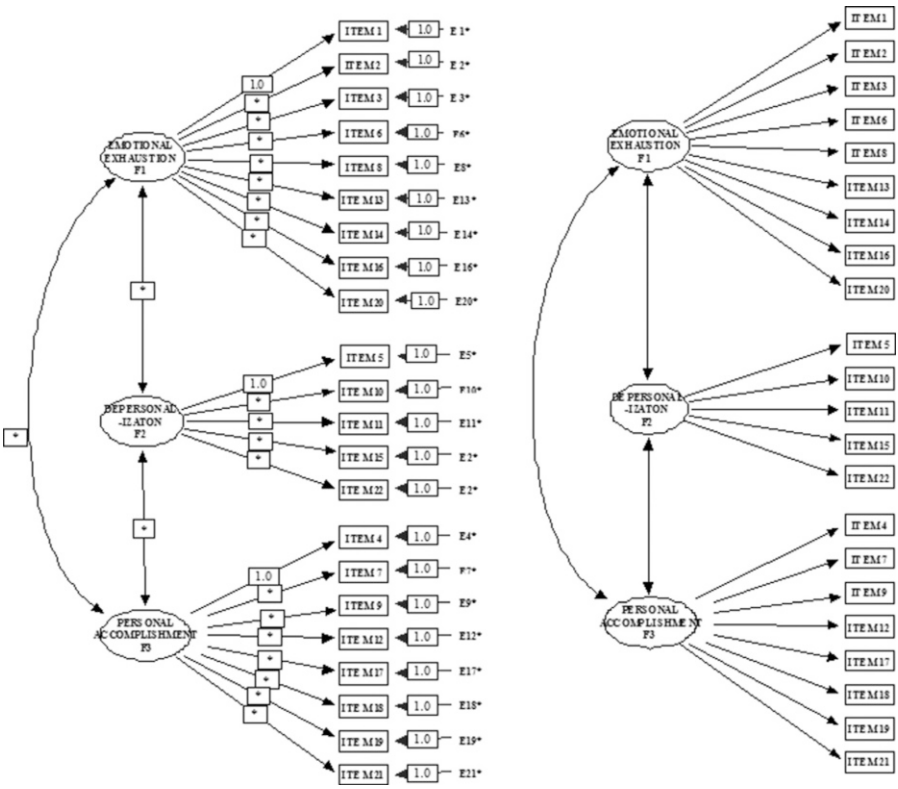


Figure 1. Hypothesized factorial structure of the Maslach Burnout Inventory within the frameworks of the EQS (Bentler, 2005) and mplus (Muthén & Muthén, 2007–2010) programs.

Let's turn now to Table 1 in which the combined input files for EQS and Mplus are presented. A review of these files again reveals a substantial difference between the two programs in terms of information specified, with the EQS program denoting considerably more detail. The primary factor contributing to this difference lies with the numerous defaults implemented by Mplus. For example, whereas specification of the model in EQS requires three paragraphs of input, each of which describes the pattern of factor loadings (EQUATIONS), the variances (factors and errors), and covariances (factors), respectively, Mplus needs to know only the variables to be

Table 1. Input files for hypothesized CFA model

EQS

```
/TITLE
CFA of MBI for Male Elementary Tchrs (Calbrn Grp) "MBIELM11"
Initial Model
/SPECIFICATIONS
DATA='c:\eqs62\files\books\runs\elemm1.ess';
VARIABLES=22; CASES=372;
METHOD=ML,ROBUST; ANALYSIS=COVARIANCE;
MATRIX=RAW;/LABELS
V1=ITEM 1; V2=ITEM 2; V3=ITEM 3; V4=ITEM 4; V5=ITEM 5;
V6=ITEM 6; V7=ITEM 7; V8=ITEM 8; V9=ITEM 9; V10=ITEM 10;
V11=ITEM 11; V12=ITEM 12; V13=ITEM 13; V14=ITEM 14; V15=ITEM 15;
V16=ITEM 16; V17=ITEM 17; V18=ITEM 18; V19=ITEM 19; V20=ITEM 20;
V21=ITEM 21; V22=ITEM 22;
F1=EE; F2=DP; F3=PA;
/EQUATIONS
V1 = 1F1 + E1;
V2 = *F1 + E2;
V3 = *F1 + E3;
V6 = *F1 + E6;
V8 = *F1 + E8;
V13 = *F1 + E13;
V14 = *F1 + E14;
V16 = *F1 + E16;
V20 = *F1 + E20;
V5 = 1F2 + E5;
V10 = *F2 + E10;
V11 = *F2 + E11;
V15 = *F2 + E15;
V22 = *F2 + E22;
```

Table 1. Input files for hypothesized CFA model (continued)

```
V4 = 1F3 + E4;
V7 = *F3 + E7;
V9 = *F3 + E9;
V12 = *F3 + E12;
V17 = *F3 + E17;
V18 = *F3 + E18;
V19 = *F3 + E19;
V21 = *F3 + E21;
/VARIANCES
F1 to F3 = *;
E1 to E22 = *;
/COVARIANCES
F1 to F3 = *;
/PRINT
FIT=ALL;
/LMTEST
SET=PEE,GVF;
/END
```

MPLUS

```
TITLE: CFA of MBI for Male Elementary Tchrs (Calibrn Group)
Initial Model - MLM Estimation

DATA:
FILE IS "C:\Mplus\Files\elemm1.dat";
FORMAT IS 22F1.0;

VARIABLE:
NAMES ARE ITEM1 - ITEM22;
USEVARIABLES ARE ITEM1 - ITEM22;

ANALYSIS:
ESTIMATOR = MLM;

MODEL:
F1 by ITEM1 - ITEM3 ITEM6 ITEM8 ITEM13 ITEM14 ITEM16 ITEM20;
F2 by ITEM5 ITEM10 ITEM11 ITEM15 ITEM22;
F3 by ITEM4 ITEM7 ITEM9 ITEM12 ITEM17 - ITEM19 ITEM21;

OUTPUT: SAMPSTAT MODINDICES;
```

used in the analysis (USE VARIABLES ARE) and the variables loading on each factor as defined with a BY statement.

A second important distinction between the two programs in terms of model specification is the requirement that analyses take into account the non-normality of the data. Given that pre-analysis of the data for the current application revealed evidence of substantially high multivariate kurtosis, it was essential that this non-normality be taken into account. In EQS, this issue is addressed under the SPECIFICATION command by adding the term "ROBUST" to the METHOD=ML statement, with a comma separating the two terms. Accordingly, this expanded method command will yield the corrected S-Bχ^2 statistic and standard errors as noted earlier in this chapter. The parallel directive in Mplus is found under the ANALYSIS command with the specification of MLM estimation. Accordingly, the resulting corrected χ^2 statistic represents the equivalent S-B χ^2 value and relatedly corrected standard errors.

A final key component of both input files is notation related to possible model misspecification. Whereas EQS takes a multivariate approach to the detection of misspecified parameters through use of the Lagrange Multiplier Test (LMTest), Mplus takes a univariate approach based on the Modification Index (MI; Sörbom, 1989). The EQS input file paragraph labeled LMTest addresses this issue, though it limits the search to possible cross-loadings (GVF) and error covariances (PEE). The Mplus input file requests computation of modification indices (MODINDICES) in the OUTPUT command; in the present case, sample statistics (SAMPSTAT) are also requested; this latter information is automatically included in the EQS output.

Table 2 presents model goodness-of-fit statistics as reported in the EQS and Mplus output files. As you will readily note, the information reported in terms of both estimated values and model fit criteria, varies minimally, albeit there are several different optional goodness-of-fit statistics reported across the two programs. Worthy of particular note is that the EQS fit statistics reported here represent the complete list when FIT=ALL is specified in the PRINT paragraph of the input file (see Table 1); in the absence of this command results include only a few key fit statistics. A second difference of note is the resulting output when the researcher has specified robust estimation. In EQS, this command results in two sets of fit statistics being reported: (a) those based on the ML estimator, and (b) those based on the robust ML estimator; only the robust goodness-of-fit statistics are included in Table 2. In contrast, Mplus reports only the robust statistics requested. For readers who may be unfamiliar with both the EQS and Mplus programs, I assure you that, although the numbers are not exactly the same (likely due to computational rounding errors), the information conveyed is definitely consistent in revealing exceptionally poor model fit to the sample data. Finally, although the caveat concerning calculation of chi-square difference values noted in the Mplus output applies to both programs, this admonition is cited only in Mplus. To assist you in your comparison of the two sets of results, I have assigned matching parenthesized numerals to each of the three key model fit statistics typically reported in the SEM literature.[4]

For a final comparison of EQS and Mplus, we turn to Table 3 where results for tests of possible model misspecification are reported. Given the extremely poor fit

Table 2. Output files: goodness-of-fit statistics for hypothesized model (model 1)

EQS

GOODNESS OF FIT SUMMARY FOR METHOD = ROBUST

ROBUST INDEPENDENCE MODEL CHI-SQUARE = 2919.314 ON 231 DEGREES OF FREEDOM

INDEPENDENCE AIC = 2457.314 INDEPENDENCE CAIC = 1321.050
MODEL AIC = 157.275 MODEL CAIC = -856.018

(1) SATORRA-BENTLER SCALED CHI-SQUARE = 569.2745 ON 206 DEGREES OF FREEDOM
PROBABILITY VALUE FOR THE CHI-SQUARE STATISTIC IS 0.00000

MEAN- AND VARIANCE-ADJUSTED CHI-SQUARE = 196.316 ON 71 D.F.
PROBABILITY VALUE FOR THE CHI-SQUARE STATISTIC IS 0.00000

RESIDUAL-BASED TEST STATISTIC = 913.978
PROBABILITY VALUE FOR THE CHI-SQUARE STATISTIC IS 0.00000

YUAN-BENTLER RESIDUAL-BASED TEST STATISTIC = 263.380

PROBABILITY VALUE FOR THE CHI-SQUARE STATISTIC IS 0.00424

YUAN-BENTLER RESIDUAL-BASED F-STATISTIC = 1.985
DEGREES OF FREEDOM = 206, 166
PROBABILITY VALUE FOR THE F-STATISTIC IS 0.00000

FIT INDICES

BENTLER-BONETT NORMED FIT INDEX = 0.805
BENTLER-BONETT NON-NORMED FIT INDEX = 0.848

---

(2) COMPARATIVE FIT INDEX (CFI) = 0.865
BOLLEN'S (IFI) FIT INDEX = 0.866
MCDONALD'S (MFI) FIT INDEX = 0.614
(3) ROOT MEAN-SQUARE ERROR OF APPROXIMATION (RMSEA) = 0.069
90% CONFIDENCE INTERVAL OF RMSEA (0.062, 0.076)

MPLUS

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

(1) Value 588.869*

Table 2. Output files: goodness-of-fit statistics for hypothesized model (model 1) (continued)

Degrees of Freedom 206
P-Value 0.0000
Scaling Correction Factor 1.181
for MLM

* The chi-square value for MLM, MLMV, MLR, ULSMV, WLSM and WLSMV cannot be used for chi-square difference tests. MLM, MLR and WLSM chi-square difference testing is described in the Mplus Technical Appendices at www.statmodel.com.
See chi-square difference testing in the index of the Mplus User's Guide.

Chi-Square Test of Model Fit for the Baseline Model

Value 2922.960
Degrees of Freedom 231
P-Value 0.0000

CFI/TLI

(2) CFI 0.858
TLI 0.841

Loglikelihood

H0 Value -12811.043
H1 Value -12463.184

Information Criteria

Number of Free Parameters 69
Akaike (AIC) 25760.087
Bayesian (BIC) 26030.490
Sample-Size Adjusted BIC 25811.575
(n* = (n + 2) / 24)
RMSEA (Root Mean Square Error Of Approximation)

(3) Estimate 0.071

SRMR (Standardized Root Mean Square Residual)

Value 0.070

WRMR (Weighted Root Mean Square Residual)

Value 1.730

Table 3. Output files: modification indices

EQS

MULTIVARIATE LAGRANGE MULTIPLIER TEST BY SIMULTANEOUS PROCESS IN STAGE 1

CUMULATIVE MULTIVARIATE STATISTICS					UNIVARIATE INCREMENT			
							HANCOCK'S SEQUENTIAL	
STEP	PARAMETER	CHI-SQUARE	D.F.	PROB.	CHI-SQUARE	PROB.	D.F.	PROB.
1	**E16,E6**	**91.039**	**1**	**0.000**	**91.039**	**0.000**	**206**	**1.000**
2	**E2,E1**	**169.787**	**2**	**0.000**	**78.748**	**0.000**	**205**	**1.000**
3	V12,F1	211.189	3	0.000	41.402	0.000	204	1.000
4	**E11,E10**	**249.137**	**4**	**0.000**	**37.948**	**0.000**	**203**	**1.000**
5	E21,E7	281.829	5	0.000	32.692	0.000	202	1.000
6	E7,E4	319.388	6	0.000	37.558	0.000	201	1.000
7	V1,F3	344.576	7	0.000	25.188	0.000	200	1.000
8	E21,E4	365.373	8	0.000	20.797	0.000	199	1.000
9	E6,E5	382.042	9	0.000	16.669	0.000	198	1.000
10	E3,E1	397.776	10	0.000	15.734	0.000	197	1.000

MPLUS

MODEL MODIFICATION INDICES

```
Minimum M.I. value for printing the modification index 10.000
          M.I.    E.P.C.     Std      E.P.C.     StdYX     E.P.C.

BY Statements

F1        BY      ITEM12    35.141    -0.313    -0.400    -0.335
F2        BY      ITEM12    11.992    -0.329    -0.276    -0.232
F3        BY      ITEM1     24.320     0.872     0.383     0.231
F3        BY      ITEM2     10.741     0.565     0.248     0.161
F3        BY      ITEM13    10.712    -0.583    -0.256    -0.152

WITH Statements

ITEM2     WITH    ITEM1     69.786     0.613     0.613     0.549
ITEM6     WITH    ITEM5     14.552     0.354     0.354     0.232
ITEM7     WITH    ITEM4     28.298     0.209     0.209     0.324
ITEM11    WITH    ITEM10    32.234     0.580     0.580     0.525
ITEM12    WITH    ITEM3     13.129    -0.255    -0.255    -0.225
ITEM15    WITH    ITEM5     13.190     0.313     0.313     0.243
ITEM16    WITH    ITEM6     77.264     0.733     0.733     0.529
ITEM18    WITH    ITEM7     10.000    -0.145    -0.145    -0.211
ITEM19    WITH    ITEM18    15.749     0.250     0.250     0.285
ITEM20    WITH    ITEM8     12.029     0.230     0.230     0.240
ITEM20    WITH    ITEM13    11.059     0.237     0.237     0.214
ITEM21    WITH    ITEM4     11.090     0.201     0.201     0.201
ITEM21    WITH    ITEM7     28.380     0.263     0.263     0.326
```

of this model to the data, we can expect to see several large modification indices that can lead to its improvement. Although results for a greater number of parameters are reported in the EQS output, only the first 10 are shown here; in Mplus, by default, no MI values <10.00 are reported.[5]

In EQS, one looks for the point at which there is a substantial drop in the Univariate Chi Square Increment values. However, along with this observation, one also must be able to argue for incorporation of these additional parameters into the model on the basis of substantive meaningfulness. In the present case, the first four highlighted values shown in Table 3 have been found consistently in other research, to be the parameters contributing most to model misspecification. They represent three error covariances (E16,E6; E2,E1; E11,E10) and the cross-loading of item 12 on Factor 1 (V12,F1).

In Mplus, the program separates the BY from the WITH statements, with the former representing factor loadings and the latter representing error covariances. As we see here, the four largest values replicate the results of the EQS program. Thus, both programs suggest that if the error variances between Items 16 and 6, Items 2 and 1, and Items 11 and 10 were allowed to covary, and Item 12 allowed to cross-load onto F1 (see in Table 1 its hypothesized loading onto Factor 2) the cumulative drop in the chi square value would be approximately 249.137 (see the fourth chi-square value at Step 4 of the EQS output). In the Mplus output, these chi-square drop-values are presented separately for each MI; in total, they represent a cumulative value of approximately 214.425. Following the incorporation of these new parameters to the model based on both their substantive meaningfulness and previous empirical replication, the final model fit in EQS yielded a corrected CFI value of 0.937, and in Mplus, of 0.934.

NOTES

[1] Brown (2006) notes that the ADF estimator in AMOS is actually the WLS estimator.

[2] Termed arbitrary generalized least squares (AGLS) in EQS (Brown, 2006).

[3] In Mplus, the first factor loading in each congeneric set is automatically fixed to 1.00 by default. Likewise, the factor variances and covariances pertinent to independent factors in a model such as the CFA model discussed here are estimated by default. Importantly, however, all defaults in Mplus can be overridden. These same defaults hold for EQS when the model is specified graphically, rather than interactively and manually.

[4] Although the 90% Confidence Interval for the RMSEA value should also be reported, this information is not reported in the Mplus output and therefore is included in the matching parenthesized values here.

[5] Although it may appear that EQS considers substantially more parameters to be misspecified than Mplus, this is not so. Rather, this discrepancy derives from the Mplus default in reporting no MI values less than 10.00 (see Table 3). Alternatively, EQS input could likewise be tailored to limit potentially misspecified parameters to those having values equal to or greater than 10.00.

REFERENCES

Arbuckle, J. L. (2009). *AMOS18 User's Guide*. Chicago, IL: SPSS.

Bentler, P. M. (2005). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., & Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika*, *45*, 289–308.

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods and Research*, *21*, 205–229.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

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

Byrne, B. M. (1994). Testing for the factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research*, *29*, 289–311.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

Byrne, B. M. (2001). Structural equation modeling with AMOS: Basic concepts, applications, and programming. Mahwah, NJ: Erlbaum.

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Erlbaum.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York, NY: Taylor & Francis/Routledge.

Byrne, B. M. (2011). *Structural equation modeling with Mplus: Basic concepts, applications and programming*. New York, NY: Taylor & Francis/Routledge.

Chou, C.-P, Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for nonnormal data in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, *44*, 347–357.

Curran, P. J., West. S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16–29.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, *9*, 327–346.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*, 309–326.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. Psychological Methods, 9, 466–491.

Hu, L-T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*, 351–362.

Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, *24*, 387–404.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*, 381–389.

Jöreskog, K. G., & Sörbom, D. (1993a). *New features in LISREL 8*. Chicago, IL: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1993b). LISREL 8. *Structural equation modeling with the SIMPLIS command language*. Chicago, IL: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1996). PRELIS 2. Chicago, IL: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.8. for Windows*. Lincolnwood, IL: Scientific Software International.

Lee, S. -Y., Poon, W. -Y., & Bentler, P. M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, *48*, 339–358.

Maslach, C., & Jackson, S. E. (1986). *Maslach burnout inventory manual* (2nd ed.). Palo Alto, CA: Consulting Psychologists Press.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376–398.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-nonnormal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology*, *45*, 19–30.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Technical report. UCLA.

Muthén, L. K., & Muthén, B. O. (2007–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Authors.

Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling*. Mahwah, NJ: Erlbaum.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi square statistics in covariance structure analysis. *American statistical association 1988 Proceedings of the business and economic sections* (pp. 308–313). Alexandria, VA: American Statistical Association.

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*, 371–384.

Stine, R. A. (1989). An introduction to bootstrap methods: Examples and ideas. *Sociological Methods and Research*, *18*, 243–291.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology*, *51*, 289–309.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In M. P. Becker (Ed.), *Sociological methodology 2000* (pp. 165–200). New York: Wiley-Blackwell.

Yung, Y-F., & Bentler, P. M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 195–226). Mahwah NJ: Erlbaum.

Zhu, W. (1997). Making bootstrap statistical inferences: A tutorial. *Research Quarterly for Exercise and Sport*, *68*, 44–55.

ABOUT THE AUTHORS

Jocelyn H. Bolin is an Assistant Professor of Educational Psychology at Ball State University, where she teaches graduate level applied statistics classes. Dr. Holden's main line of research investigates advances and optimal use of statistical classification analysis. She also studies applications and developments in multilevel modeling for the social sciences. In addition, she is a member of the American Educational Research Association as well as the American Statistical Association.

Timothy A. Brown is a Professor in the Department of Psychology at Boston University (BU), and Director of Research at BU's Center for Anxiety & Related Disorders. He has published extensively in the areas of the classification of anxiety and mood disorders, psychometrics, and methodological advances in social sciences research. In addition to his own grant-supported research, Dr. Brown has served as a statistical investigator or consultant on numerous federally funded research projects. He has been on the editorial boards of several scientific journals, including a current appointment as Associate Editor to the *Journal of Abnormal Psychology*.

Barbara M. Byrne is a Professor Emeritus in the School of Psychology, University of Ottawa. She has authored 7 introductory books on structural equation modeling (SEM), one book addressing the measurement of self-concept across the lifespan, and over 100 scholarly journal articles and book chapters. Dr. Byrne is the recipient of three Distinguished Teaching Awards presented by the Canadian Psychological Association, the American Psychological Association (APA), and the APA Division 5 (Jacob Cohen Award). She is a Fellow of the APA and an elected member of the Society of Multivariate Experimental Psychology. Dr. Byrne's research interests focus on psychometric issues, with particular specialization in the area of SEM.

Christine DiStefano is an Associate Professor of Educational Research and Measurement at the University of South Carolina where she teaches courses in educational assessment, measurement theory, and structural equation modeling. Her research interests include structural equation modeling with ordinal data, cluster analysis, using Rasch modeling for scale development, and mixture modeling.

Sunny Duerr received a Master of Science in Applied Statistics and Research Methods from the University of Northern Colorado in 2010, and a Ph.D. in 2012 from the Behavioral Science program at the University of Rhode Island. He is currently the Coordinator for Assessment and Accreditation at the School of Education, State University of New York at New Paltz. His academic interests include the analysis of large data sets, multivariate statistics, structural equation modeling,

education research (specifically math, science, and reading), and the psychology of video games. Dr. Duerr's dissertation involved cluster analyses of data from over 50 countries and assessing the invariance of structural equation models of math achievement, across and within country clusters, using data from the 2007 Trends in International Mathematics and Science Study (TIMSS). He has presented at the annual conferences for, the American Psychological Association (APA), the Association for Psychological Science (APS), and the American Educational Research Association (AERA). Whether as an instructor for first-year college students or as an invited lecturer discussing the nuances of confirmatory factor analysis with advanced graduate students, Sunny enjoys leading people to increased statistical and methodological understanding. Sunny was awarded the Exceptional Accommodation of Persons with Disabilities Award and the (student nominated) Outstanding Faculty or Staff Award while serving as an instructor for the University of Northern Colorado. While at the University of Rhode Island, he was honored with the Beaupre Hope and Heritage Award and the Graduate Student Excellence in Behavioral Science Award.

Xitao Fan is Chair Professor and Dean, Faculty of Education, University of Macau, China. Prior to his current appointment, he was an associate, full, and endowed chair professor (Curry Memorial Professor of Education) in the Curry School of Education, University of Virginia, USA, and was assistant/associate professor in the Psychology Department, Utah State University, USA. He obtained his Ph.D. (1993) from Texas A&M University in educational psychology. He is a quantitative methodologist in education, with research and teaching interests in applied multivariate methods, especially in structural equation modeling, longitudinal analysis of change, model fit assessment, reliability and validity issues in measurement and assessment, and inter-disciplinary education research. He has published widely in both methodological and substantive areas related to reliability and validity issues in measurement, to multivariate statistical techniques in general, and structural equation modeling and growth modeling in particular. He served as the associate editor (2000-2004) and editor (2004-2010) for *Educational and Psychological Measurement* (Sage Publications, USA). He is an AERA Fellow (2012; American Educational Research Association).

Holmes Finch is a Professor of Educational Psychology at Ball State University where he has been since 2003. He received his PhD from the University of South Carolina in 2002. Professor Finch teaches courses in factor analysis, structural equation modeling, categorical data analysis, regression, multivariate statistics and measurement to graduate students in psychology and education. His research interests are in the areas of latent variable modeling, methods of prediction and classification, and nonparametric multivariate statistics.

Brian F. French is a Professor of Educational Psychology and Director of the Learning and Performance Research Center at Washington State University in

Pullman, Washington. Dr. French completed his graduate work at Purdue University and an internship at ACT, Inc. Dr. French teaches courses in measurement/ psychometrics, statistics, research methods, and advanced quantitative methods. Dr. French's research is collaborative and focuses on educational and psychological measurement issues. The first area concerns the application of psychometric methods to gather score validity evidence for a variety of instruments. The second area, informed by the first, is the use of methodological studies to evaluate and improve methods in terms of efficiency and accuracy used to gather test score validity evidence. A sample of topics of interest include: Measurement Invariance, Structural Equation Modeling, Item Response Theory, Classical Test Theory, Monte Carlo studies. His work and work with others has been funded from various agencies such as the National Science Foundation, Institute of Education Sciences, and National Institutes of Health. His work appears in journals such as *Structural Equation Modeling, Educational and Psychological Measurement, Child Development, Journal of Experimental Education,* and the *Journal of Educational Measurement.*

Matthew W. Gallagher is currently a staff research psychologist at the Behavioral Sciences Division of the National Center for PTSD at the Boston VA and a research assistant professor at Boston University. He received his PhD in Clinical and Quantitative Psychology from the University of Kansas and completed a postdoctoral fellowship at the Center for Anxiety and Related Disorders at Boston University. His research focuses on how hope, self-efficacy and other positive thinking constructs promote well-being and provides resilience to PTSD and other anxiety disorders. He has received awards from the Association for Psychological Science and the Society for Multivariate Experimental Psychology for his research, and has served as statistical consultant on research funded by NIMH and the Robert Wood Johnson Foundation.

Nigel Gilbert is a Professor at the University of Surrey's Department of Sociology. His main research interests are processual theories of social phenomena, the development of computational sociology and the methodology of computer simulation, especially agent-based modelling. He is Director of the Centre for Research in Social Simulation (CRESS). He is the author or editor of several textbooks on sociological methods of research and statistics, including *Simulation for the Social Scientist*, *Researching Social Life*, and *Understanding Social Statistics*. He is also editor of the *Journal of Artificial Societies and Social Simulation*.

Perman Gochyyev is a doctoral student in the Quantitative Methods and Evaluation (QME) program at the Graduate School of Education, University of California, Berkeley, and graduate student researcher at the Berkeley Evaluation and Assessment Research (BEAR) Center. His research interests include multidimensional and explanatory item response models and latent class models.

James Hall is a Research Fellow at the Department of Education, University of Oxford. His research lies at the intersection of Developmental Psychology and Education where he focuses upon at-risk populations, biopsychosocial mechanisms, and determining the potential for early intervention.

John Hattie is a Professor and has Director of Melbourne Education Research Institute at the University of Melbourne since November 2010. Professor Hattie's work is internationally acclaimed. His influential 2008 book Visible Learning: A synthesis of over 800 Meta-Analyses Relating to Achievement is believed to be the world's largest evidence-based study into the factors which improve student learning. He is a Member of the Faculty of Education and Director of Visible Learning Labs at the University of Auckland, Prof. Hattie regularly advises governments in New Zealand, Australia and the US. He has authored or co-authored 12 books and more than 500 papers.

Lisa Harlow is a Professor of Psychology at University of Rhode Island. She has over 75 scholarly publications plus 6 books (e.g., *What if there were no significance tests?* and *The Essence of Multivariate Thinking*). Dr. Harlow obtained over $8 million in grant funding, most recently from the National Science Foundation on Advancing Women in Science, and Quantitative Training for Underrepresented groups (QTUG). Since 2004, she has served as co-director of QTUG where over 250 students and early PhDs from underrepresented groups attended conferences, gaining confidence and understanding in state-of-the-art methods. Dr. Harlow served as 2010-2011 president of the Society of Multivariate Experimental Psychology (SMEP) that has 65 elected members from across and outside the US. Since 1995, she is the Editor for the SMEP Multivariate Applications Book Series that published 20+ books covering among the most exciting quantitative methods (e.g., item response theory, meta-analysis, mediation analysis, multilevel modeling, multivariate thinking, structural equation modeling). Lisa is currently Editor for *Psychological Methods* and past Associate editor of the *Structural Equation Modeling Journal*; and is on editorial boards for *Multivariate Behavioral Research* and *European Methodology* journals. She received numerous honors including: Past-President of American Psychological Association (APA) Division 5; Fellow in APA Divisions 1 (General), 2 (Teaching), 5 (Statistics), 38 (Health), and 52 (International); the Jacob Cohen Distinguished Teaching and Mentoring Award from Division 5 of APA; Distinguished Fellowship at the Institute for Advanced Study, Latrobe University, Australia; and a Fulbright Award at York University, Toronto, Canada.

Ronald H. Heck is a Professor of educational administration and policy at the University of Hawaii at Manoa. His research interests include school effects on student learning and the evaluation of educational policies. Recent publications include *Studying Educational and Social Policy: Theoretical Concepts and Research Methods* and *An Introduction to Multilevel Modeling Techniques* (with Scott L. Thomas).

Jason C. Immekus is an Associate Professor in the Department of Educational Research & Administration at California State University, Fresno. Dr. Immekus completed his graduate studies at Purdue University and a one year Post-Doctoral research position at the Center of Health Statistics at the University of Illinois at Chicago. His research interests focus on the use model-based approaches to examining test score validity issues (e.g., item bias, multidimensionality). Dr. Immekus teaches graduate-level courses in measurement, statistics, research methods, and program evaluation. His work has appeared at national conferences and in such journals as Educational & Psychological Measurement and Psychiatric Services. He is actively involved in translational research to promote the use of psychometric research findings to applied testing contexts within school districts and community-based organizations.

Ken Kelley is the Viola D. Hank Associate Professor of Management at the University of Notre Dame. Dr. Kelley's research involves the development, improvement, and evaluation of quantitative methods, especially as they relate to statistical and measurement issues in applied research. More specifically, Dr. Kelley's research focuses on issues of sample size planning, effect size estimation, and confidence interval formation. He is the developer of the MBESS R package, associate editor of *Psychological Methods*, and a member of the Statistics and Modeling Scientific Review Panel for the US Department of Education's Institute of Education Sciences.

Spyros Konstantopoulos is an Associate Professor and program director of measurement and quantitative methods at the department of counseling, educational psychology, and special education at the College of Education at Michigan State University. He received his MS in statistics and his Ph.D. in research methods from the University of Chicago. His research interests include the extension and application of statistical methods to issues in education, social science, and policy studies. His methodological work involves statistical methods for quantitative research synthesis (i.e., meta-analysis) and mixed effects models with nested structure (i.e., multilevel linear models). His substantive work encompasses research on class size effects, teacher and school effects, program evaluation, labor market performance of young adults, and the social distribution of academic achievement. He serves as the associate editor of the Journal of Research on Educational Effectiveness and the Journal of Research Synthesis Methods.

John P. Madura has a B.A. in Mathematics (Logic and Computability) from Boston University and an M.A. in History and Education from Teachers College, Columbia University. John taught secondary mathematics for four years, and then entered the Measurement, Evaluation and Assessment doctoral program at the Neag School of Education at the University of Connecticut. John's research interests center on aspects of interpersonal perception that occur in school settings. His substantive interests mainly address issues that involve developmental learning relationships

and how they change over time. His methodological research focuses broadly on multilevel structural equation modeling (MSEM), latent growth/change models, and factor analysis in the context of instrument design in the affective domain.

George A. Marcoulides is a Professor of Research Methods & Statistics in the Graduate School of Education and in the Interdepartmental Graduate Program in Management (IGPM) in the A. Gary Anderson Graduate School of Management at the University of California, Riverside. He is a Fellow of the American Educational Research Association, a Fellow of the Royal Statistical Society, and a member of the Society of Multivariate Experimental Psychology. He is currently editor of the journals *Structural Equation Modeling* and *Educational and Psychological Measurement*, editor of the Quantitative Methodology Book Series, and on the editorial board of numerous other scholarly journals.

D. Betsy McCoach is an Associate Professor in the Measurement, Evaluation and Assessment program at the University of Connecticut. She has extensive experience in longitudinal data analysis, hierarchical linear modeling, instrument design, factor analysis, and structural equation modeling. Betsy has published over 60 journal articles and book chapters, and she co-edited the volume *Multilevel Modeling of Educational Data* with Ann O'Connell. Betsy served as the founding co-editor for the *Journal of Advanced Academics*, and she is the incoming co-editor of *Gifted Child Quarterly*. Betsy is the current Director of the DATIC, where she teaches summer workshops in Hierarchical Linear Modeling and Structural Equation Modeling, and she is the founder and conference chair of the Modern Modeling Methods conference, held at UCONN every May. Betsy serves as a Co-Principal Investigator and research methodologist on several federally funded research grants, including *Project Early Vocabulary Intervention,* funded by IES, and *School Structure and Science Success: Organization and Leadership Influences on Student Success,* funded by NSF. In addition, she has served as the Research Methodologist for the National Research Center on the Gifted and Talented for the last 5 years.

Diana Mindrila is an Assistant Professor of Educational Research at the University of West Georgia, where she teaches graduate courses in research methodology. Her research interests include comparing multivariate classification methods and using multilevel classification techniques to construct typologies of school behavior and school climate.

Shevaun D. Neupert is an Associate Professor at North Carolina State University. She earned her Ph.D. in Family Studies and Human Development with a minor in Statistics from the University of Arizona. During her time in graduate school she had the opportunity to attend an intensive summer workshop on HLM taught by Stephen Raudenbush. She has been using the method in her research ever since. She teaches a graduate course on HLM each spring and her research focuses on three

primary areas: (1) Daily stressors and their associations with affect, physical health, and memory across the adult lifespan using daily diary designs; (2) Socioeconomic and personality differences in reactivity to various types of stressors (e.g., naturally-occurring, laboratory-induced); and (3) Statistical techniques and methods for examining change and intraindividual variability.

Ann A. O'Connell is a Professor of Quantitative Research, Evaluation and Measurement in the College of Education and Human Ecology at The Ohio State University. She specializes in regression, multivariate, and multilevel modeling; generalized linear mixed models; and categorical data analysis. Her collection of published work emphasizes these and other quantitative methodologies in the field of educational evaluation and for understanding the impact of health and HIV prevention interventions. She is a research affiliate of the Children's Learning Research Collaborative, and the faculty director for the methodology core of the International Poverty Solutions Collaborative, both at OSU. Dr. O'Connell has received research support from the American Educational Research Association (AERA), the Centers for Disease Control and Prevention (CDC), and the National Institutes for Health (NIH). Her work has appeared in peer-reviewed journals including *Evaluation and the Health Professions*, *Women and Health*, *Measurement and Research in Counseling and Development*, *MMWR*, and the *Journal of Modern Applied Statistical Methods*; she also published a book with Sage on *Logistic Regression Models for Ordinal Response Variables*, and co-edited a book on *Multilevel Modeling of Educational Data* with her colleague, D. Betsy McCoach. She serves as the evaluator and methodologist on several ongoing multi-year initiatives at OSU including an HIV disclosure intervention; a reading comprehension evaluation and clinical trial; and a professional development evaluation for pre-school educators. She is currently developing a study of access to and retention in HIV care for women living in poverty in Ohio.

Jason W. Osborne is a Professor and Chair of Educational and Counseling Psychology at the University of Louisville, in Louisville KY USA. He teaches and publishes on best practices in quantitative and applied research methods. Jason also publishes on identification with academics and on issues related to social justice and diversity. He is author of three books on statistics and research methods from Sage, over 60 peer-reviewed journal articles, and has presented over 70 times at national and international conferences. He has been PI or co-PI on over $9,000,000.00 in externally funded projects spanning topics such as K-12 public education, instructional technology, higher education, nursing and health care, and medicine and medical training. He is Specialty Chief Editor of *Frontiers in Quantitative Psychology and Measurement* and *Frontiers in Educational Psychology*, as well as being involved in several other journals.

Steven J. Osterlind is a Professor Emeritus of Measurement and Statistics, University of Missouri. His expertise is in psychometrics, tests and measurement and statistical

modeling. He also holds a joint adjunct appointment in the Department of Statistics at MU. He teaches courses in statistics and measurement, including hierarchical linear modeling, multivariate methods, regression, as well as specialized topics specialized topics like Item Response Theory. From 1985 to 1996 he was director of the Center for Education Assessment, an MU research unit with responsibility for developing and administering the state of Missouri's public school assessment program. He was a Fulbright Senior Scholar in Ireland in 1004, and from 2004 to 2006 he held Distinguished Professor title at National College of Ireland. He earned his Ph.D. in 1976 from the University of Southern California, followed by an American Scholar's Fellow at Yale University in 1977-79. He has taught at USC (University of Southern California), UC Berkeley, and since 1985, at the University of Missouri. He has written five books, most recently Modern Measurement: Theory, Principles, and Applications of Mental Appraisal, 2nd ed. (Prentice- Hall 2010); as well as more than 60 juried articles, 100 technical reports, and three major tests, including College BASE, a major standardized achievement test for collegians.

Karen E. Rambo-Hernandez is an assistant professor at Colorado State University in the School of Education and the School of Teacher Education and Principal Preparation, where she teaches courses in teacher preparation and educational statistics. Karen graduated from the University of Connecticut with a Ph.D. in Educational Psychology with a specialization in Measurement, Evaluation, and Assessment. As a former mathematics teacher, she is particularly interested in the assessment of student learning. Karen's research focuses on the assessment of academic growth and the modeling of student growth data.

Mauricio Salgado is a Research Fellow and Associate Professor of sociology, formalisation and simulation of social dynamics at the School of Sociology in Universidad Andres Bello (Chile). He is a sociologist and completed his PhD in Computational Sociology at the Centre for Research in Social Simulation (CRESS), Department of Sociology, University of Surrey (The United Kingdom). His work on agent-based modelling has been published in Advances in Complex Systems, the Journal of Artificial Societies and Social Simulation and Discrete Dynamics in Nature and Society.

Pamela Sammons is a Professor of Education at the Department of Education, University of Oxford and a Senior Research Fellow at Jesus College, Oxford. Previously she was a Professor at the University of Nottingham, and at the Institute of Education, University of London. Her research focuses on school effectiveness and improvement, the early years and equity in education. She is a Principal Investigator for the longitudinal Effective Pre-school, Primary &Secondary Education study, investigating children's development from age 3 to 16+ years (EPPSE 3-16+). She is also a Principal Investigator on the Evaluation of Children's Centres in England.

Daniel A. Sass is Associate Professor and Director of the Statistical Consulting Center at the University of Texas and at San Antonio in the Department of Management Science and Statistics department. His research interests include methodological issues related to multivariate statistics and psychometrics, with a central focus on factor analysis and structural equation modeling. Within these modeling frameworks, his research specifically focuses on model estimation, model fit, model misspecification, multidimensionality, measurement invariance, and structural invariance. He is also interested in topics related to test development and validation.

Thomas A. Schmitt is the Co-Founder and CEO of Equastat, which is a statistical and methodological consulting company. His specialization is in applied statistics and measurement as related to education and psychology. His research interests include methodological issues related to latent variable modeling such as factor analysis, structural equation modeling, mixture modeling, latent class analysis, multilevel modeling, item response theory, adaptive testing, and test and instrument construction.

Shaojing Sun is an Associate Professor, School of Journalism, Fudan University, China. Prior to his current appointment, he was assistant professor in the Department of Communication, Weber State University, USA, and adjunct assistant professor in the Department of Communication, University of Maryland, College Park, USA. He is a quantitative methodologist in communication research, with research and teaching interests in structural equation modeling, item response theory, longitudinal data analysis, social network analysis, etc. He has conducted methodological and substantive studies related to measurement bias, item response theory, and applications of multivariate statistical techniques. He had his Ph.D. degree (2003) from Kent State University, majoring in communication studies and Ph.D. degree (2006) from University of Virginia majoring in quantitative research methodology.

Timothy Teo is Professor of Education at the University of Macau. His research interests are multi-disciplinary in nature and include both substantive and methodological areas. These are *Educational Psychology* (Self-efficacy-teachers and students; Beliefs about teaching and learning; Meta-cognition), *ICT in Education* (Technology acceptance and adoption; e-learning), *Music Education* (Psychological processes of music teaching and learning), and *Quantitative Methods* (Psychometrics; Instrument development and validation; cross-cultural measurement; issues in survey development and administration; structural equation modeling; multilevel modeling; latent growth modeling; Item Response Theory modeling). Timothy is chief editor of two international journals, *The Asia-Pacific Education Researcher* (TAPER) and *International Journal of Quantitative Research in Education* (IJQRE) while sitting in 20 journal editorial boards at the same time. As an author, he has edited three books, many book chapters and conference papers, and published over 100 peer-reviewed journal articles, many of which in highly-ranked SSCI journals.

Ze Wang is an Assistant Professor of Measurement and Statistics at the University of Missouri. Her research interests include statistical modeling and psychometrics, particularly their applications to large-scale complex assessment data. She teaches quantitative methods and advanced statistics courses at the University of Missouri. She has published in nine different research journals including Educational Psychology, The Journal of Experimental Education, Journal of Psychoeducational Assessment, International Journal of Science and Mathematics Education, and International Journal of Testing. She has co-authored one book and three book chapters, and had more than 40 conference presentations, talks and workshops.

Megan E. Welsh is an Assistant Professor in the Neag School of Education at the University of Connecticut, where she teaches courses in assessment, evaluation, and educational measurement. Megan graduated from the University of Arizona with a Ph.D. in Educational Psychology and a specialization in Measurement and Research Methods. A former elementary teacher and school district administrator, her interests include the assessment of student learning and how assessments affect educator practice.

Mark Wilson is a professor in the Graduate School of Education, University of California, Berkeley, and Director of the Berkeley Evaluation and Assessment Research (BEAR) Center. He has research interests in psychometrics, educational assessment, and applied statistics.

Hsiao-Ju Yen earned a Ph.D in Educational Psychology with an emphasis in psychometrics and statistics at Washington State University in Pullman, Washington. She is a research assistant in Learning and Performance Research Center LPRC under the guidance of Dr. Brian F. French. Her research interests include Structural Equation Modeling, Measurement Invariance, Item Response Theory, Classical Test Theory, and Cognitive Diagnostic Modeling. Her work in latent variable modeling with diverse groups (e.g., English language learners, Autism) using large-scale national datasets has appeared at conferences including the American Educational Research Association and the National Council on Measurement in Education