As part of their research activities, researchers in all areas of education develop measuring instruments, design and conduct experiments and surveys, and analyze data resulting from these activities. Educational research has a strong tradition of employing state-of-the-art statistical and psychometric (psychological measurement) techniques. Commonly referred to as quantitative methods, these techniques cover a range of statistical tests and tools. Quantitative research is essentially about collecting numerical data to explain a particular phenomenon of interest. Over the years, many methods and models have been developed to address the increasingly complex issues that educational researchers seek to address.

This handbook serves to act as a reference for educational researchers and practitioners who desire to acquire knowledge and skills in quantitative methods for data analysis or to obtain deeper insights from published works. Written by experienced researchers and educators, each chapter in this handbook covers a methodological topic with attention paid to the theory, procedures, and the challenges on the use of that particular methodology. It is hoped that readers will come away from each chapter with a greater understanding of the methodology being addressed as well as an understanding of the directions for future developments within that methodological area.
Handbook of Quantitative Methods for Educational Research
Handbook of Quantitative Methods for Educational Research

Edited by
Timothy Teo
University of Auckland, New Zealand
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section 1: Measurement Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Psychometrics</td>
</tr>
<tr>
<td><em>Mark Wilson &amp; Perman Gochyyev</em></td>
</tr>
<tr>
<td>2. Classical Test Theory</td>
</tr>
<tr>
<td><em>Ze Wang &amp; Steven J. Osterlind</em></td>
</tr>
<tr>
<td>3. Item Response Theory</td>
</tr>
<tr>
<td><em>Xitao Fan &amp; Shaojing Sun</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section 2: Methods of Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Multiple Regression</td>
</tr>
<tr>
<td><em>Ken Kelley &amp; Jocelyn Holden</em></td>
</tr>
<tr>
<td>5. Cluster Analysis</td>
</tr>
<tr>
<td><em>Christine DiStefano &amp; Diana Mindrila</em></td>
</tr>
<tr>
<td>6. Multivariate Analysis of Variance: With Discriminant Function Analysis Follow-up</td>
</tr>
<tr>
<td><em>Lisa L. Harlow &amp; Sunny R. Duerr</em></td>
</tr>
<tr>
<td>7. LoGistic Regression</td>
</tr>
<tr>
<td><em>Brian F. French, Jason C. Immekus &amp; Hsiao-Ju Yen</em></td>
</tr>
<tr>
<td>8. Exploratory Factor Analysis</td>
</tr>
<tr>
<td><em>W. Holmes Finch</em></td>
</tr>
<tr>
<td>9. A Brief Introduction to Hierarchical Linear Modeling</td>
</tr>
<tr>
<td><em>Jason W. Osborne &amp; Shevaun D. Neupert</em></td>
</tr>
<tr>
<td>10. Longitudinal Data Analysis</td>
</tr>
<tr>
<td>11. Meta-Analysis</td>
</tr>
<tr>
<td><em>Spyros Konstantopoulos</em></td>
</tr>
<tr>
<td>12. Agent Based Modelling</td>
</tr>
<tr>
<td><em>Mauricio Salgado &amp; Nigel Gilbert</em></td>
</tr>
</tbody>
</table>
### TABLE OF CONTENTS

   *James Hall & Pamela Sammons*  
   267

#### Section 3: Structural Equation Models

   *Matthew W. Gallagher & Timothy A. Brown*  
   289

15. Testing Measurement and Structural Invariance: Implications for Practice  
   *Daniel A. Sass & Thomas A. Schmitt*  
   315

16. Mixture Models in Education  
   *George A. Marcoulides & Ronald H. Heck*  
   347

17. Selecting SEM Computer Programs: Considerations and Comparisons  
   *Barbara Byrne*  
   367

About the Authors  
395
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

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
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.”
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. 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: 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. 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”.
](image-url)
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.

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

<table>
<thead>
<tr>
<th>Location</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5th grade</td>
<td>Student is able to put the motions of the Earth and Moon into a complete description of rotation in the Solar System which explains:</td>
</tr>
<tr>
<td></td>
<td>the day/night cycle</td>
</tr>
<tr>
<td></td>
<td>the phases of the Moon (including the illumination of the Moon by the Sun)</td>
</tr>
<tr>
<td></td>
<td>the seasons</td>
</tr>
<tr>
<td>4th grade</td>
<td>Student is able to coordinate apparent and actual motion of objects in the sky. Student knows that:</td>
</tr>
<tr>
<td></td>
<td>the Earth is both orbiting the Sun and rotating on its axis</td>
</tr>
<tr>
<td></td>
<td>the Earth orbits the Sun once per year</td>
</tr>
<tr>
<td></td>
<td>the Earth rotates on its axis once per day, causing the day/night cycle and the appearance that the Sun moves across the sky</td>
</tr>
<tr>
<td></td>
<td>the Moon orbits the Earth once every 28 days, producing the phases of the Moon</td>
</tr>
<tr>
<td></td>
<td>COMMON ERROR: Students are confused by the changing distance between the Earth and Sun.</td>
</tr>
<tr>
<td></td>
<td>COMMON ERROR: The phases of the Moon are caused by a shadow of the planets, the Sun, or the Earth falling on the moon.</td>
</tr>
<tr>
<td>3rd grade</td>
<td>Student knows that:</td>
</tr>
<tr>
<td></td>
<td>the Earth orbits the Sun</td>
</tr>
<tr>
<td></td>
<td>the Moon orbits the Earth</td>
</tr>
<tr>
<td></td>
<td>the Earth rotates on its axis</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>COMMON ERROR: It gets dark at night because the Earth goes around the Sun once a day.</td>
</tr>
<tr>
<td>2nd grade</td>
<td>Student recognizes that:</td>
</tr>
<tr>
<td></td>
<td>the Sun appears to move across the sky every day</td>
</tr>
<tr>
<td></td>
<td>the observable shape of the Moon changes every 20 days</td>
</tr>
<tr>
<td></td>
<td>Student may believe that the Sun moves around the Earth.</td>
</tr>
<tr>
<td></td>
<td>COMMON ERROR: All motion in the sky is due to the Earth spinning on its axis.</td>
</tr>
<tr>
<td></td>
<td>COMMON ERROR: It gets dark at night because the Sun goes around the Earth once a day.</td>
</tr>
<tr>
<td></td>
<td>COMMON ERROR: The Earth is the center of the universe.</td>
</tr>
<tr>
<td>1st grade</td>
<td>Student does not recognize the systematic nature of the appearance of objects in the sky. Student may not recognize that the Earth is spherical.</td>
</tr>
<tr>
<td></td>
<td>COMMON ERROR: It gets dark at night because something (e.g., clouds, the atmosphere, &quot;darkness&quot;) covers the Sun.</td>
</tr>
<tr>
<td></td>
<td>COMMON ERROR: The phases of the Moon are caused by clouds covering the Moon.</td>
</tr>
<tr>
<td></td>
<td>COMMON ERROR: The Sun goes below the Earth at night.</td>
</tr>
<tr>
<td>0th grade</td>
<td>No evidence or off-track</td>
</tr>
</tbody>
</table>

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 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 occurring 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

![Figure 3. A sample OMC item based upon ESS construct map. (L indicates location on construct map.)](image-url)
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):

\[ X = T + E, \]

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. 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-focused 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 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. 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 ($\theta$) and item location ($\delta$). Mathematically, this statement is represented as

$$\text{Probability(correct} \mid \theta, \delta) = f(\theta - \delta)$$ (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:

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

(3)

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

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

(4)

where $\Phi$ 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, $\delta = 0.0$).](image-url)
In the Rasch model, the total score of the correct (endorsed) items is monotonically (but not linearly) related to the estimated ability. 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.

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

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.](image)

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.
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

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

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).

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.

Note that 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.

The natural logarithm of the odds ratio.

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.

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.

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


M. WILSON & P. GOCHYYEV


28


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).
Z. WANG & S. J. OSTERLIND

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 \( t_{gp} \) 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, \( X_{gp} - t_{gp} \)). Under CTT, \( t_{gp} \) 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 standpoint the mathematical model for CTT can be deduced, and consists of two equations:

\[
\begin{align*}
t_{gp} &= E(X_{gp}) \quad (1) \\
E_{gp} &= X_{gp} - t_{gp} \quad (2)
\end{align*}
\]

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, \( \Gamma_g \) is a random variable over repeated sampling of persons in a population, which takes a specific value \( \tau_g \) 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,

\[
X = \Gamma + E \quad (3)
\]

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 \( \sigma_x^2 \) 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:

\[
\sigma^2_X = \sigma^2_\Gamma + \sigma^2_E. \tag{4}
\]

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., \( X = \Gamma + E \)), \( \Gamma \) 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: \( \sigma_X^2 = \sigma_T^2 + \sigma_E^2 \). 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

\[
\sigma_T^2 \leq \sigma_X^2 \quad \text{and} \quad \sigma_E^2 \leq \sigma_X^2.
\]

The ratio of these variances is expressed as:

\[
\rho_X = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}
\] (6)

This ratio quantifies the reliability of using observed scores to describe the traits of a population of individuals and \( \rho_X \) 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, \( X, \Gamma, \) 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

\[
\rho_{XX'} = \rho_X = \rho_{XX'}^2
\]  

(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:

$$ss\rho_X = \frac{2\rho_{X_{1}X_{2}}}{1 + \rho_{X_{1}X_{2}}}$$  \hspace{1cm} (8)

where $\rho_{X_{1}X_{2}}$ 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:

$$\alpha\rho_X = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_X^2 - \sum_{f} \sigma_{X_f}^2}{\sigma_X^2}\right)$$  \hspace{1cm} (9)

where $n$ is the number of parts, $\sigma_X^2$ 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):

$$20\rho_X = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum_{f} \phi_f (1-\phi_f)}{\sigma_X^2}\right)$$  \hspace{1cm} (10)

where $\phi_f$ is the proportion of scores of 1 on item $f$.
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,

$$\text{SEM} = \sigma_x = \sigma_x (1 - \rho_x)$$  \hspace{1cm} (11)

where $\rho_x$ 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 $\tau_p$ 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 $\tau_p$ 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 $\sigma^2_x$ is normally distributed. With this additional assumption, the interval estimation of $\tau_p$ is $x_p \pm z\sigma_x$, 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 $\tau_p$ depends on an additional assumption that, with a random sample of multiple persons on whom test scores are observed, the random variables $\Gamma$ and $X$ follow a bivariate normal distribution. With this assumption, a point estimate of an individual’s true score is $\rho_x(x - \mu_x) + \mu_x$, where $\rho_x$ is the reliability coefficient, and $\mu_x$ 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 $\tau_p$ is
\[ \left[ \rho_x (x - \mu_x) + \mu_x \right] \pm z \sigma_x \sqrt{\rho_x} \]. 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.
CLASSICAL TEST THEORY

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 examinee, item analysis, and some understanding of error in the measurement. For these reasons CTT remains a most popular approach to measuring mental processes.

REFERENCES


Table 1. An example of item and test scores

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

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 examinee, item analysis, and some understanding of error in the measurement. For these reasons CTT remains a most popular approach to measuring mental processes.
Z. WANG & S. J. OSTERLIND


