Introduction to Educational and Psychological Measurement Using R

Anthony D. Albano

March 21, 2016

Contents

Preface

This book provides an introduction to the theory and application of measurement in education and psychology. Topics include test development, item writing, item analysis, reliability, dimensionality, and item response theory. These topics come together in overviews of validity and, finally, test evaluation.

Validity and test evaluation are based on both qualitative and quantitative analysis of the properties of a measure. This book addresses the qualitative side using a simple argument-based approach. The quantitative side is addressed using descriptive and inferential statistical analyses, all of which are presented and visualized within the statistical environment R (R Core Team 2015).

The intended audience for this book includes advanced undergraduate and graduate students, practitioners, researchers, and educators. Knowledge of R is not a prerequisite to using this book. However, familiarity with data analysis and introductory statistics concepts, especially ones used in the social sciences, is recommended.

Perspectives on testing

Testing has become a controversial topic in the context of education. Consider this summary by Nelson (2013) from a study on the costs of educational testing in school districts in the US:

> Testing has spiraled out of control, and the related costs are unacceptably high and are taking their educational toll on students, teachers, principals and schools.

The conclusions of this study reflect a sentiment that is shared by many in the educational community, that we often rely too heavily on testing in schools, and that we do so to the detriment of students.

Those critical of educational testing in the US highlight two main problems with assessment mandated from the top down. The first is an over reliance on tests in decision making at all levels, including decisions that impact students, teachers, administrators, and other stakeholders. There are too many tests, given too often, with too high a cost, financially and in terms of lost instructional time. Nelson (2013, 3) notes that "if testing were abandoned altogether, one school district in this study could add from 20 to 40 minutes of instruction to each school day for most grades."

The second problem is a reliance on tests which are not relevant enough to what is being taught and learned in the classroom. There is a lack of cohesion, a mismatch in content, where teachers, students, and tests are not on the same page. As a result, the tests become frustrating, unpleasant, and less meaningful to the people who matter most, those who take them and those who administer or oversee them at the classroom level.

Both of these problems identified in our testing agenda have to do with what is typically regarded as the pinnacle of quality testing, the all-encompassing, all-powerful *validity*. Commenting over 50 years ago on its status for the test developer, Ebel (1961) concluded with some religious imagery that "validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few" [p. 640]. Arguably, nothing is more important in the testing process than validity, the extent to which test scores accurately represent what they're intended to measure.

However, even today, it may be that "relatively few testing programs give validation the attention it deserves" (Brennan 2013, 81).

Establishing validity evidence for a test has traditionally been the responsibility of the test developer, who may only have limited interaction with test users and secondhand familiarity with the issues they face. Yet, Kane (2013) notes that from early on in the testing movement in the US, the appropriateness of testing applications at the student level was a driving force in bringing validity to prominence. He cites Kelley (1927), who observed:

> The question of validity would not be raised so long as one man uses a test or examination of his own devising for his private purposes, but the purposes for which schoolmasters have used tests have been too intimately connected with the weal of their pupils to permit the validity of a test to go unchallenged. The pupil... is the dynamic force behind the validity movement. ... Further, now that the same tests are used in widely scattered places and that many very different tests all going by the same name are gently recommended by their respective authors, even the most complacent schoolmen, the most autocratic, and the least in touch with pupils, are beginning to question the real fitness of a test. [pp. 13-14]

Motivation for this book

It appears that Kelley (1927) recognized in the early 1900s the same issue that we're dealing with today. Tests are recommended or required for a variety of applications, and we (even the most complacent and autocratic) often can only wonder about their fitness. Consumers and other stakeholders in testing need access to information and tools, provided by a test author, that allow them to understand and evaluate the quality of a test. Consumers need to be informed. In a way, Kelley (1927) was promoting accessible training in educational and psychological measurement.

As a researcher and psychometrician, someone who studies methods for building and using measurement tools, I admire a good test, much like a computer programmer admires a seamless data access layer or an engineer admires a sturdy truss system. Testing can provide critical information to the systems it supports. It is essential to measuring key outcomes in both education and psychology, and can be used to enhance learning and encourage growth (e.g., Black and Wiliam 1998, Meyer and Logan (2013), Roediger et al. (2011)). However, test content and methods can also become outdated and out of touch, and, as a result, they can waste time and even produce erroneous, biased, and damaging information (Santelices and Wilson 2010).

Rather than do away with testing, we need to refine and improve it. We need to clarify its function and ensure that it is fulfilling its purpose. In the end, there is clearly disagreement between the people at each end of the testing process, that is, those creating the tests and those taking them or witnessing firsthand their results, in terms of the validity or effectiveness of the endeavor. This disconnect may never go away completely, especially since an education system has many roles and objectives, inputs and outputs, not all of which are unanimously endorsed. However, there is definitely room for change and growth.

This book will prepare you to contribute to the discussion by giving you a broad overview of the test development and evaluation processes. Given this scope, some deeper topics, like dimensionality and item response theory, are covered only superficially. The purpose of this book is not to make you an expert in every topic, but to help you:

1. recognize what makes a test useful and understand the importance of valid testing; and
2. gain some of the basic skills needed to build and evaluate effective tests.

Note that this book addresses affective, non-cognitive, non-educational testing as well as cognitive, educational testing. Affective testing includes, for example, testing used to measure psychological attributes, perceptions, and behaviors, ones that are typically the target of clinical support rather than ones that are the target of instruction. This form of testing is just as relevant, but is also less controversial than educational testing, so it didn't make for as strong an opening to the book.

Structure of this book

The book is divided into ten chapters. Most chapters involve analysis of data in R, and item writing activities to be completed with the Proola web app available at proola.org. An introduction to R and Proola is given in Chapter 1, and examples are provided throughout the book.

1. Introduction
   An intro to the resources we'll be using in this book, with an overview in R of some introductory statistical topics.
2. Measurement, Scales, Scoring
   See what measurement really means, and look at the different procedures used to carry it out.
3. Testing Applications
   Review features of the testing process and terms used to describe and categorize tests, with examples of popular measures.
4. Test Development
   Background on cognitive and noncognitive test development, with an overview of item writing guidelines.
5. Reliability
   A key topic in measurement, introduced via classical test theory.
6. Item Analysis
   Classical descriptive analysis of item-level data, with an emphasis on difficulty, discrimination, and contribution to internal consistency.
7. Item Response Theory
   Also known as modern measurement theory, a latent variable modeling approach to item analysis and test construction.
8. Dimensionality
   A brief overview of exploratory and confirmatory factor analysis, with applications.
9. Validity
   The pinnacle of the test development process, helping us evaluate the extent to which a test measures what it is intended to measure.
10. Test Evaluation
    A summary of Chapters 1 through 9 applied to the evaluation and comparison of educational and psychological tests.

Learning objectives

Each chapter in this book is structured around a set of learning objectives appearing at the end of the corresponding chapter. These learning objectives capture, on a superficial and summary level, all the material you'd be expected to master in a course accompanying this book. Any assignments, quizzes, or class discussions should be built around these learning objectives, or a subset of them.

Exercises

Each chapter also ends with self assessments and other activities for stretching and building understanding. These exercises include discussion questions and analyses in R. As this is a book on assessment, you'll also write your own questions, with the opportunity to submit them to an online learning community for feedback.

Chapter 1

Introduction

This chapter introduces two resources we'll be using throughout the book, including the assessment development tools at Proola.org, and the statistical software R. The chapter ends with an introductory statistics review.

1.1 Proola

Proola is a web application for collaborative assessment development. It was created specifically for individuals looking for practice and support in the item writing process. There's a simple interface for creating assessment items, and features for reviewing and commenting on the items of your peers.

You need to set up a free account at proola.org/users/sign_up before you can complete the assignments in later chapters. Once you have an account, you can start writing and commenting on items.

Some things to keep in mind as you get started:

- The site itself is still under development, with new features on the way. Report bugs or send suggestions to contact@proola.org.
- Everything you share is public. Don't post copyrighted items, images, or other information, and don't share items you need to keep secure.
- You can learn a lot from the successes and failures of others. Search the bank for items related to your content area, and then see where people struggle and what they do well.

The Proola item writing process is broken down into four general steps (see proola.org/learn_more).

First, create a new item by clicking "Share" in the top navbar. This takes you to a series of text boxes and drop-downs where you'll provide basic information about your item. See Figure 1.1 for an example. First, give it a short but descriptive title. If you're writing to a specific learning objective, which is highly recommended, you can describe it in your title, or find a Common Core standard that meets your needs. Next, choose from a variety of intended subject areas and grade levels. The item itself then consists of a stem, where the question statement or prompt resides, and a response format, whether selected-response constructed-response. Any other background information you'd like to share can go in the comments.

Second, get feedback from peers and assessment specialists. Once an item is saved as a draft, anyone can view and comment on it. Wait patiently for comments, or recruit peers in your grade level, subject area, department, school, district, and get them to sign up and leave feedback. Assessment specialists include faculty and graduate students with training in assessment development who contribute to the Proola community. Comments are listed below your item. After signing up for an account, see proola.org/items/439 and proola.org/items/296 for examples.

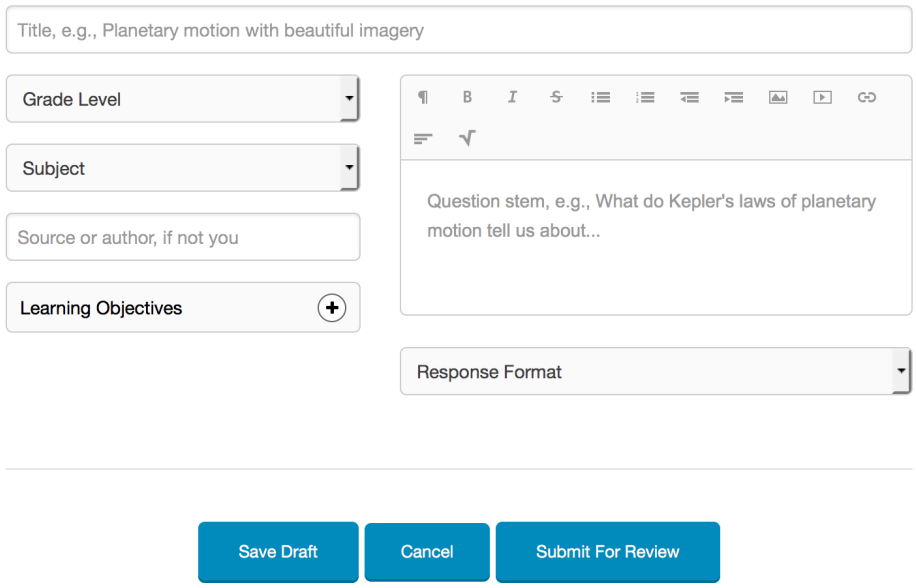


Figure 1.1: Proola interface for drafting a new item.

When commenting or responding to comments, remember to be constructive. Comments will naturally flow toward the limitations of an item. Remember to highlight strengths as well as weaknesses, and always provide suggestions for improvement. Always reference the item writing guidelines, as flawed items tend to miss one or more of them. Finally, comment on the scope of an item, that is, how well it addresses the intended learning objective(s) at the appropriate depth of knowledge.

Third, edit your item based on feedback. After input from peers, edit and improve your item. Focus on the item writing guidelines and your selected learning objective(s). Aim for the highest depth of knowledge. Double-check for clarity and correct spelling and grammar.

Finally, submit for review to share with the community. After submitting an item, there's no turning back. You and others can still comment, but edits are locked until a formal peer review is complete. When revisions are recommended, edits can only be made to a new version of the item. Previous versions are viewable but not editable.

Revise and repeat this process as needed up to four times. Once approved, your item can be saved, printed, and exported by other users.

The test development and item writing processes will be covered in detail in Chapters 3 and 4. For now, you should sign up for an account and become familiar with how the site is organized. At the end of this chapter are some item writing activities to try.

1.2 R

R is both a programming language and software environment for statistical analysis. It differs from other software like SPSS in three key ways. First, R is free, no strings (or warranties) attached. Download it at cran.r-project.org The popular editor RStudio is also available for free at rstudio.com. Second, R is open-source, with thousands of active contributors sharing add-on packages. See the full list at cran.r-project.org/web/packages (there are currently over 8,000). Third, R is accessed primarily through code,

rather than by pointing and clicking through drop-down menus and dialog windows. This third point is a road block to some, but it ends up being a strength in the long run.

At this point you should download and install R and RStudio using the links above. The internet abounds with helpful tips on installation and getting started. Here are a few pointers.

- R is the software that runs all your analyses. RStudio is an *Integrated Development Environment* or IDE that simplifies your interaction with R. RStudio isn't essential, but it gives you nice features for saving your R code, organizing the output of your analyses, and managing your add-on packages.
- As noted above, R is accessed via code, primarily in the form of commands that you'll type or paste in the R console. The R console is simply the window where your R code goes, and where output will appear. Note that RStudio will present you with multiple windows, one of which will be the R console. That said, when instructions here say to run code in R, this applies to R via RStudio as well.
- When you type or paste code directly into the R console, any previous code you've already entered gets pushed up on the screen. In the console, you can scroll through your old code by hitting the up arrow once your cursor is in front of the R prompt `>`. In RStudio, you can also view and scroll through a list of previous commands by holding one of the control/command buttons on your keyboard while hitting up.
- Only type directly in the console for simple and quick calculations that you don't care about forgetting. Otherwise, type all your code in a text file that is separate from the console itself. In R and RStudio, these text files are called R scripts. They let you save your code in a separate document, so you always have a structured record of what you've done. Remember that R scripts are only used to save code and any comments annotating the code, not data or results.

1.2.1 Code

We'll start our tour of R with a summary of how R code is used to interact with R via the console. In this book, blocks of example R code are offset from the main text as shown below. Comments within code blocks start with a single hash `#`, the code itself has nothing consistent preceding it, and output from my R console is preceded by a double hash `##`. You can copy and paste example code directly into your R console. Anything after the `#` will be ignored.

```
# This is a comment within a block of R code. Comments start with the
# hash sign and are ignored by R. The code below will be interpreted
# by R once you paste or type it into the console.
x <- c(4, 8, 15, 16, 23, 42)
mean(x)  # Only code after the hash is ignored
## [1] 18
sd(x)
## [1] 13.49074
```

In the code above, we're creating a short vector of scores in `x` and calculating its mean and standard deviation. You should paste this code in the console and verify that you get the same results. Note that code you enter at the console is preceded by the R prompt `>`, whereas output printed in your console is not.

The first thing to notice about the example code above is that the functions that get things done in R have names, and to use them, we simply call them by name with parentheses enclosing any required information or instructions for how the function should work. Whenever functions are discussed in this book, you'll recognize them by the parentheses. For example, we used the function `c()` to combine a set of numbers into a "vector" of scores. The information supplied to `c()` consisted of the scores themselves, separated by commas. `mean()` and `sd()` are functions for obtaining the mean and standard deviation of vectors of scores, like the scores in `x`.

The second thing to notice in the example above is that data and results are saved to "objects" in R using the assignment operator `<-`. We used the concatenate function to stick our numbers together in a set, `c(4, 8, 15, 16, 23, 42)`, and then we assigned the result to have the name `x`. Objects created in this way can be accessed later on by their assigned name, for example, to find a mean or standard deviation. If we wanted to access it later, we could also save the mean of `x` to a new object.

```
# Calculate the mean of x
mx <- mean(x)
# Print x and the mean of x
x
## [1]  4  8 15 16 23 42
mx
## [1] 18
print(mx)
## [1] 18
```

When the output from a function is stored in an R object, you typically don't see the output printed to the console. If we type an object name at the console, R does its best to print out the contents, as shown above for `mx` and `x`. This is simply a shortcut for using the `print()` function on the object.

Note that for larger objects, like `data.frame`s with lots of rows and columns, viewing all the data at once isn't very helpful. In this book we'll analyze data from the Programme for International Student Assessment (PISA), with dozens of variables measured for thousands of students. Printing all the PISA data would flood the console with information. This brings us to the third thing to notice about the example code above, that the console isn't the best place to view results. The console is functional and efficient, but it isn't pretty or well organized. Fortunately, R offers other mediums besides fixed-width text for visualizing output, discussed below.

1.2.2 Packages

When you install R on your computer, you get a variety of functions and example data sets by default as part of the base packages that come with R. For example, `mean()` and `print()` come from the base R packages. Commonly used procedures like simple linear regression, via `lm()`, and t-testing, via `t.test()`, are also included in the base packages. Additional functionality comes from add-on packages written and shared online by a community of R enthusiasts.

The examples in this book rely on a few different R packages. The book itself is compiled using the bookdown and knitr packages (Xie, n.d., Xie (2016)), and some knitr code is shown for formatting output tables. The ggplot2 package (Wickham 2009) is used for plotting. In this chapter we also need the devtools package (Wickham and Chang 2016), which allows us to install R packages directly from the code sharing website github.com. Finally, throughout the book we'll then be using a package called epmr, which contains functions and data used in many of the examples.

Packages published to the official R website CRAN are initially installed using `install.packages()`. You only need to do this once per package. Each time you open a new R session, you load the package with `library()` to use it. Alternatively, you can use individual functions without explicitly loading a package by referencing both the package and function at each use, separated by `::`.

devtools, knitr, and ggplot2 are on CRAN. The development version of epmr is not yet on CRAN and must be installed using `devtools::install_github()`.

```
# Install required packages - you only need to do this once
install.packages("devtools")
install.packages("knitr")
install.packages("ggplot2")
```

```
# Install epmr from github
devtools::install_github("talbano/epmr")
## Skipping install for github remote, the SHA1 (cc1e1460) has not changed since last install.
##   Use `force = TRUE` to force installation
# Load required packages - do this every time you restart R
library("epmr")
library("ggplot2")
```

After a package is installed, and you've run `library()`, you have access to the functionality of the package. Here, we've loaded epmr so we can type functions directly, without referencing the package name each time.

1.2.3 Getting help

Help files in R are easily accessed by entering a function name preceded by a question mark, for example, `?c`, `?mean`, or `?devtools::install_github`. Parentheses aren't necessary. The question mark approach is a shortcut for the `help()` function, where, for example, `?mean` is the same as `help("mean")`. Either approach provides direct access to the R documentation for a function.

The documentation for a function should at least give you a brief description of the function, definitions for the arguments that the function accepts, and examples of how the function can be used.

At the console, you can also perform a search through all of the available R documentation using two question marks. For example, `??"regression"` will search the R documentation for the term "regression." This is a shortcut for the function `help.search()`. Finally, you can browse through the R documentation with `help.start()`. This will open a web browser with links to manuals and other resources.

If you're unsatisfied with the documentation internal to R, an online search can be surprisingly effective for finding function names or instructions on running a certain procedure or analysis in R.

1.2.4 Data

Data can be entered directly into the console by using any of a variety of functions for collecting information together into an R object. These functions typically give the object certain properties, such as length, rows, columns, variable names, and factor levels. In the code above, we created `x` as a vector of quantitative scores using `c()`. You could think of `x` as containing test scores for a sample of `length(x)` 6 test takers, with no other variables attached to those test takers.

We can create a factor by supplying a vector of categorical values, as quoted text, to the `factor()` function.

```
# Create a factor variable
classroom <- c("A", "B", "C", "A", "B", "C")
classroom <- factor(classroom)
classroom[c(1, 4)]
## [1] A A
## Levels: A B C
```

In this code, the `classroom` object is first assigned a vector of letters, which might represent labels for three different classrooms. The classroom object is then converted to a factor, and assigned to an object of the same name, which essentially overwrites the first assignment. Reusing object names like this is not recommended. This is just to show that a name in R can't be assigned two separate objects at once.

The code above also demonstrates a simple form of indexing. Square brackets are used after some R objects to select subsets of their elements, for example, the first and fourth values in `classroom`. The vector `c(1, 4)`

is used as an indexing object. Take a minute to practice indexing with `x` and `classroom`. Can you print the last three classrooms? Can you print them in reverse order? Can you print the first score in `x` three times?

The `data.frame()` function combines multiple vectors into a set of variables that can be viewed and accessed as a matrix with rows and columns.

```
# Combine variables as columns in a data frame
mydata <- data.frame(scores = x, classroom)
mydata[1:4, ]
##   scores classroom
## 1      4         A
## 2      8         B
## 3     15         C
## 4     16         A
```

We can index both the columns and rows of a matrix. The indexing objects we use must be separated by a comma. For example, `mydata[1:3, 2]` will print the first three rows in the second column. `mydata[6, 1:2]` prints both columns for the sixth row. `mydata[, 2]`, with the rows index empty, prints all rows for column two. Note that the comma is still needed, even if the row or column index object is omitted. Also note that a colon `:` was used here as a shortcut function to obtain sequences of numbers, where, for example, `1:3` is equivalent to typing `c(1, 2, 3)`.

Typically, we'll import or load data into R, rather than enter it manually. Importing is most commonly done using `read.table()` and `read.csv()`. Each one takes a "file path" as its first argument. See the help documentation for instructions on their use and the types of files they require. Data and any other objects in the console can also be saved directly to your computer using `save()`. This creates an "rda" file, which can then be loaded back in to R using `load()`. Finally, some data are already available in R and can be loaded into our current session with `data()`. The PISA data, referenced above, are loaded with `data(PISA09)`. Make sure the epmr package has been loaded with `library()` before you try to load the data.

The `PISA09` object is a `data.frame` containing demographic, noncognitive, and cognitive variables for a subset of questions and a subset of students from the PISA 2009 study ((nces.ed.gov/surveys/pisa)[https://nces.ed.gov/surveys/pisa/]). It is stored within the epmr R package that accompanies this book. After loading the data, we can print a few rows for a selection of variables, and the first 10 ages.

```
# Load the PISA dataset and print subsets of it
data(PISA09)
PISA09[c(1, 10000, 40000), c(1, 6, 7, 38)]
##       cnt   age grade  cstrat
## 1     AUS 16.00    11 -0.5485
## 10000 CAN 15.58     9  0.7826
## 40000 RUS 16.25     9  0.2159
PISA09$age[1:10]
##  [1] 16.00 16.08 16.08 15.42 16.00 15.67 16.17 15.83 15.92 15.33
```

The dollar sign `$` is used to access a single variable by name within a `data.frame`. Here, we've printed `age`, measured in years, for the first ten students in the data set. The different variables in `PISA09` will be described in later chapters. For a quick overview, see the help file. Documentation for data sets is accessed in the same way as documentation for functions.

1.3 Intro stats

With the basics of R under your belt, you're now ready for a review of the introductory statistics that are prerequisite for the analyses that come later in this book.

Many people are skeptical of statistics, and for good reasons. We often encounter statistics that contradict one another or that are appended to fantastic claims about the effectiveness of a product or service. At their heart, statistics are pretty innocent and shouldn't be blamed for all the confusion and misleading. Statistics are just numbers designed to summarize and capture the essential features of larger amounts of data or information.

> Facts are stubborn things, but statistics are pliable. — *Mark Twain*

Statistics are important in measurement because they allow us to score and summarize the information collected with our tests and instruments. They're used to describe the reliability, validity, and predictive power of this information. They're also used to describe how well our test covers a domain of content or a network of constructs, including in relation to other content areas or constructs. We rely heavily on statistics in Chapters 2 and 5.3.1 through 7.

1.3.1 Some terms

We'll begin this review with some basic statistical terms. First, a *variable* is a set of values that can differ for different people. For example, we often measure variables such as *age* and *gender*. These are italicized here to denote them as statistical variables, as opposed to words. The term *variable* is synonymous with quality, attribute, trait, or property. Constructs are also variables. Really, a variable is anything assigned to people that can potentially take on more than just a single constant value. As noted above, variables in R can be contained within simple vectors, for example, `x`, or they can be grouped together in a `data.frame`.

Generic variables will be labeled in this book using capital letters, usually X and Y. Here, X might represent a generic test score, for example, the total score across all the items in a test. It might also represent scores on a single item. Both are considered variables. The definition of a generic variable like X depends on the context in which it is defined.

Indices can also be used to denote generic variables that are part of some sequence of variables. Most often this will be scores on items within a test, where, for example, X_1 is the first item, X_2 is the second, and X_J is the last, with J being the number of items in the test and j representing any given item. Subscripts can also be used to index individual people on a single variable. For example, test scores for a group of people could be denoted as X_1, X_2, ..., X_n, where n is the number of people and i represents a generic person. Combining people and items, X_{ij} would be the score for person i on item j.

The number of people is denoted by n or sometimes N. Typically, the lowercase n represents sample size and the uppercase N represents the population, however, the two are often used interchangeably. Greek and Arabic letters are used for other sample and population statistics. The sample mean is denoted by m and the population mean by μ, the standard deviation is s or σ, variance is s^2 or σ^2, and correlation is r or ρ. Note that the mean and standard deviation are sometimes abbreviated as M and SD. Note also that distinctions between sample and population values often aren't necessary, in which case the population terms are used. If a distinction is necessary, it will be identified.

Finally, you may see named subscripts added to variable names and other terms, for example, $M_{control}$ might denote the mean of a control group. These subscripts depend on the situation and must be interpreted in context.

1.3.2 Descriptive and inferential

Descriptive statistics, or descriptives, are any statistics used simply to describe certain features of distributions, rather than to make inferences about unobserved parameters or population distributions. Descriptives are typically used to explore the structure of a variable or data set, or the relationships among variables. They are not typically used to answer research questions or inform decision making. Inferential statistics are more appropriate for these less exploratory and more confirmatory analyses.

Table 1.1: Descriptive Statistics for Three PISA Variables

	mean	median	sd	skew	kurt	min	max	n	na
elab	-0.12	0.04	1.03	-0.18	3.38	-2.41	2.76	44265	0
cstrat	0.05	-0.04	1.02	-0.41	4.69	-3.45	2.50	44265	0
memor	0.00	0.10	1.03	-0.22	4.23	-3.02	2.69	44265	0

Inferential statistics involve an inference to a parameter or a population value. The quality of this inference is gauged using statistical tests that index the error associated with our estimates. In this review we're focusing on descriptive statistics. Later we'll consider some inferential applications.

The `dstudy()` function in the epmr package returns some commonly used univariate descriptive statistics, including the mean, median, standard deviation (SD), skewness (skew), kurtosis (kurt), minimum (min), and maximum (max). Each is discussed further below. Descriptives for three PISA variables are shown here straight from the R console, and formatted for HTML.

```
dstudy(PISA09[, c("elab", "cstrat", "memor")])
##
## Descriptive Study
##
##            mean  median   sd   skew kurt   min  max     n na
## elab   -0.12099  0.0385 1.03 -0.181 3.38 -2.41 2.76 44265  0
## cstrat  0.04951 -0.0411 1.02 -0.408 4.69 -3.45 2.50 44265  0
## memor   0.00151  0.1036 1.03 -0.218 4.23 -3.02 2.69 44265  0
knitr::kable(dstudy(PISA09[, c("elab", "cstrat", "memor")]),
  digits = 2, caption = "Descriptive Statistics for Three PISA Variables")
```

1.3.3 Distributions

A variable for a given sample can be summarized by counting up the numbers of people having the same values. The result is a *frequency distribution*, where each total number of people is a *frequency* denoted by f. For example, the categorical variable X representing eye color may have three distinct values in a classroom of $n = 20$ students: blue, brown, and green. A frequency distribution would simply total up the number of students with each color, for example, $f_{blue} = 8$.

A frequency can be converted to a *proportion* by dividing it by the sample size. For example, $p_{blue} = f_{blue}/n$. Multiplying by 100 then gives you a percentage. Frequencies, proportions, and percentages all describe the same information for values within a distribution, but in slightly different ways.

Lets look at some frequencies within the PISA data. The first variable in `PISA09`, named `cnt`, is classified as a factor in R, and it contains the country that each student was tested in. We can use a frequency `table()` to see the number of students from each country.

```
class(PISA09$cnt)
## [1] "factor"
table(PISA09$cnt)
##
##  AUS  BEL  CAN  DEU  ESP  GBR  HKG  ITA  JPN  RUS  SGP  USA
## 4334 2524 7125 1420 7984 3717 1484 9551 1852 1638 1638 1611
```

To convert the frequencies of students by country into percentages, we can divide the table output by the number of students, that is, the number of rows, in `PISA09`, and then multiply by 100. We can then `round()`

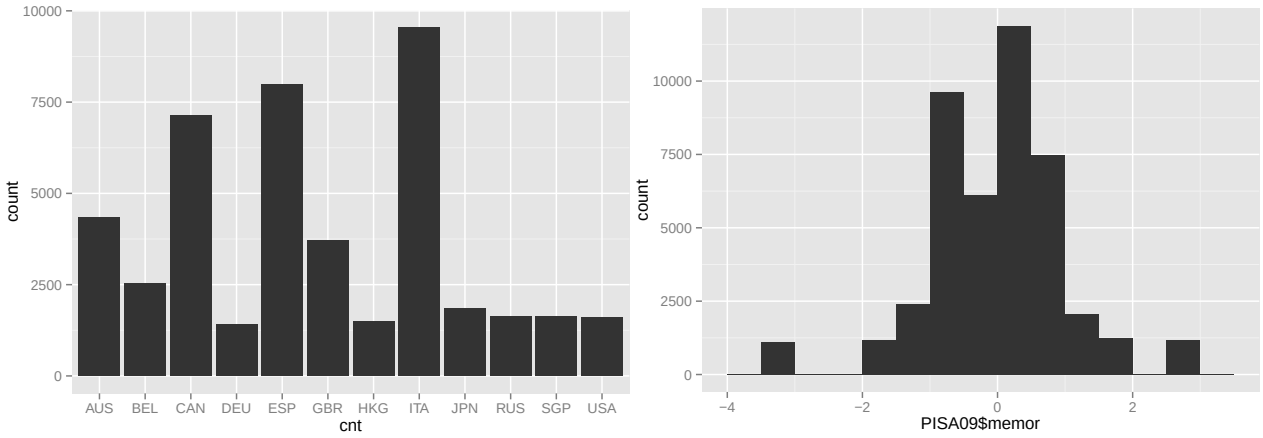


Figure 1.2: A bar plot of student frequencies by country in the first plot, and a histogram of memor scores in the second.

the result to 2 decimal places. This tells us the percentage of students in each country, relative to all $N = 44878$ students.

```
cntpct <- table(PISA09$cnt) / nrow(PISA09) * 100
round(cntpct, 2)
##
##   AUS   BEL   CAN   DEU   ESP   GBR   HKG   ITA   JPN   RUS   SGP   USA
##  9.66  5.62 15.88  3.16 17.79  8.28  3.31 21.28  4.13  3.65  3.65  3.59
```

Using what you've learned so far, find the frequency distribution for `PISA09$grade`. Then, see what happens when you put both `PISA09$cnt` and `PISA09$grade` into `table()`, with a comma separating them. Note that the PISA study is only intended for students who are 15 years old, regardless of their grade.

Bar charts and histograms are visual representations of frequency distributions, where a bar chart shows a bar for each frequency and a histogram may collapse some bars to improve interpretation with continuous variables, that is, variables where people don't fit into a discrete set of categories. A bar chart works well with a categorical factor like country. For a somewhat continuous variable like `PISA09$memor`, a histogram is more appropriate. `memor` is one of the scale scores produced by PISA. It measures the extent to which students report to use memorization strategies when reading to understand text.

```
ggplot(PISA09, aes(cnt)) + geom_bar(stat = "bin")
qplot(PISA09$memor, binwidth = .5)
```

Note that these plotting functions come from the ggplot2 package. The base package contains `barplot()` and `hist()` which will also get the job done.

Some distributions have notable shapes. For example, a distribution with the same or very similar amounts of people having each value is referred to as a *uniform distribution*. Plot a histogram of `PISA09$age` and you'll see a good example. A distribution is called a *normal distribution* when certain amounts of it fall within a central midpoint. For example, in a normal distribution, 68% of scores should be found within 1 standard deviation the mean, and frequencies should decrease and taper off as they get further away from the center. A distribution that tapers off to the left but not the right is described as negatively skewed, whereas tapering to the right but not left is positive skew. Finally, a distribution that is more peaked than normal is called leptokurtic, with high kurtosis, and a distribution that is less peaked than normal is platykurtic, with low kurtosis.

1.3.4 Central tendency

Central tendency provides statistics that describe the middle, most common, or most normal value in a distribution. The mean, which is technically only appropriate for interval or ratio scaled variables, is the score that is closest to all other scores. The mean also represents the balancing point in a distribution, so that the further a score is from the center, the more pull it will have on the mean in a given direction. The mean for a variable X is simply the sum of all the X scores divided by the sample size:

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}. \tag{1.1}$$

The median is the middle score in a distribution, the score that divides a distribution into two halves with the same number of people on either side. The mode is simply the score or scores with the largest frequencies.

The mean is by far the most popular measure of central tendency, in part because it forms the basis of many other statistics, including standard deviation, variance, and correlation. As a result, the mean is also the basis for regression and ANOVA.

In R, we can easily find the `mean()` and `median()` for a vector of scores. There is no base function for the mode. Instead, we can examine a frequency table to find the most common value(s).

1.3.5 Variability

Variability describes how much scores are spread out or differ from one another in a distribution. Some simple measures of variability are the minimum and maximum, which together capture the range of scores for a variable.

```
min(PISA09$age)
## [1] 15.17
max(PISA09$age)
## [1] 16.33
range(PISA09$age)
## [1] 15.17 16.33
```

Variance and standard deviation are much more useful measures of variability as they tell us *how much* scores vary. Both are defined based on variability around the mean. As a result, they too are are technically only appropriate with variables measured on interval and ratio scales. The variance is the mean squared distance for each score from the mean, or the sum of squared distances from the mean divided by the sample size minus 1:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n - 1}. \tag{1.2}$$

Because it is expressed as a squared value, the metric of a variance is the squared score metric, which typically does not have much practical use. As a result, variance is not often examined or reported as a standalone statistic. Instead, the square root is taken to obtain the standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n - 1}}. \tag{1.3}$$

The standard deviation is interpreted as the average distance from the mean, and it is expressed in the raw score metric making it more easy to interpret. The standard deviation of `sd(PISA09$age)` 0.292078 tells us that students in `PISA09` vary on average by about 0.3 years, or 3.5 months.

1.3.6 Correlation

Covariability, similar to variability, describes how much scores are spread out or differ from on another, but it takes into account how similar these changes are for each person from one variable to the other. As changes are more consistent across people from one variable to the other, covariability estimates increase. Covariability is most often estimated using the covariance and correlation.

Covariability is calculated using two score distributions, which are refereed to as a *bivariate score distribution*. The covariance then is the bivariate equivalent of the variance for a univariate distribution, and it is calculated in much the same way:

$$\sigma_{XY} = \frac{\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)}{n-1}. \tag{1.4}$$

Note that we now have two different variables, X and Y, and our means are labeled accordingly. Covariance is often denoted by σ_{XY}.

Like the variance, the covariance isn't very useful in and of itself because it is expressed in terms of products of scores, rather than in the more familiar raw score metric. However, square rooting the covariance doesn't help us because there are two raw score metrics involved in the calculation. The correlation solves this problem by removing, or dividing by, these metrics entirely:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{1.5}$$

By dividing the covariance by the product of the standard deviations for X and Y we obtain a measure of the relationship between them that does not have an interpretable metric. The correlation coefficient is expressed as the shared variability between two variables relative to all the available variability between and within two variables. In this ratio form, the correlation is bound by -1 and 1. A correlation of 0 indicates that none of the available variability in two variables is shared commonly between them, whereas a -1 or 1 indicates that all of the variability is shared.

In R, `cov()` and `cor()` are used to obtain covariances and correlations. When we only have two variables, we separate them with a comma. To get the covariance or correlation for different combinations of variables, we can provide a matrix or data.frame.

```
cov(PISA09$age, PISA09$grade, use = "complete")
## [1] 0.02163678
cor(PISA09[, c("elab", "cstrat", "memor")], use = "complete")
##             elab    cstrat     memor
## elab   1.0000000 0.5126801 0.3346299
## cstrat 0.5126801 1.0000000 0.4947891
## memor  0.3346299 0.4947891 1.0000000
```

The argument `use = "complete"` is used to obtain covariances or correlations only based on individuals with data on all variables.

Correlations are commonly used to index the strength of linear relationship between two variables. The small to moderate correlations between the three learning strategy scales tell us that there are some linear relationships between them. Scatter plots help us visualize these relationships. Figure 1.3 shows a scatter plot for two learning strategy variables for students in the US. The remaining countries are excluded for simplicity.

```
# Scatter plot for two learning strategies variables
# geom_point() adds the points to the plot
# position_jitter() wiggles them around a little, to uncover the densities
# of points that would otherwise overlap
```

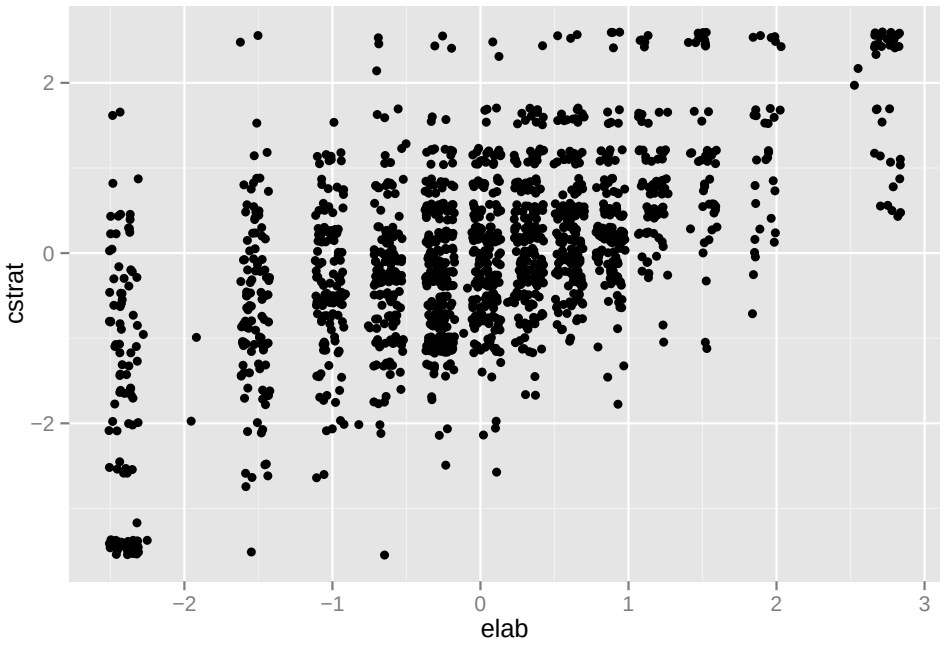


Figure 1.3: Scatter plot for the elab and cstrat PISA learning strategy scales for the US.

```
ggplot(PISA09[PISA09$cnt == "USA", ], aes(elab, cstrat)) +
  geom_point(position = position_jitter(w = 0.1, h = 0.1))
```

Positive correlations indicate that, overall, as scores on one variable increase they also tend to increase on the other variable. Stronger correlations then indicate, overall, the extent to which this is true. In Figure 1.3, scores for one variable differ substantially for a given value on the other. For example, cstrat scores span nearly the entire y-axis for elab scores below -1. This represents inconsistency in scoring, relative to the mean, from one variable to the other.

Note that `PISA09$cnt == "USA"` is used in the code above as an indexing object to select only rows for students in the US. You can create indexing objects in R using comparison operators. For example, `==` asks for each value on the left if it is equal to a single value on the right. The result is a vector of class `logical`, containing `TRUE` and `FALSE` values. With a `logical` variable as an index, any row receiving `FALSE` is omitted, whereas `TRUE` is kept. Other comparisons include `>` for greater than, and `<=` for less than or equal to. See `?Comparison` for details.

1.3.7 Rescaling

Variables are often modified to have certain properties, including smaller or larger score intervals, different midpoints, and different variability. A common example is the z-score scale, which is defined to have a mean of 0 and SD of 1. Any variable having a mean and SD can be converted to z-scores, which express each score in terms of distances from the mean in SD units. Once a scale has been converted to the z-score metric, it can then be transformed to have any midpoint, via the mean, and any scaling factor, via the standard deviation.

To convert a variable Y from its original score scale to the z-score scale, we subtract out μ_Y, the mean on Y, from each score, and then divide by σ_Y, the SD of Y. The resulting z transformation of Y, labeled as Y_z, is:

$$Y_z = \frac{Y - \mu_Y}{\sigma_Y}. \tag{1.6}$$

Having subtracted the mean from each score, the mean of our new variable Y_z is 0, and having divided each score by the SD, the SD of our new variable is 1. We can now multiply Y_z by any constant s, and then add or subtract another constant value m to obtain a linearly transformed variable with mean m and SD equal to s. The new rescaled variable is labeled here as Y_r:

$$Y_r = Y_z \mathrm{s} + \mathrm{m}. \tag{1.7}$$

The linear transformation of any variable Y from its original metric, with mean and SD of μ_Y and σ_Y, to a scale defined by a new mean and standard deviation, is obtained via the combination of these equations, as:

$$Y_r = (Y - \mu_Y)\frac{\mathrm{s}}{\sigma_Y} + \mathrm{m}. \tag{1.8}$$

Scale transformations are often employed in testing for one of two reasons. First, transformations can be used to express a variable in terms of a familiar mean and SD. For example, IQ scores are traditionally expressed on a scale with mean of 100 and SD of 15. In this case, Equation 1.8 is used with m = 100 and s = 15. Another popular score scale is referred to as the *t*-scale, with m = 50 and s = 10. Second, transformations can be used to express a variable in terms of a new and unique metric. When the GRE (www.ets.org/gre) was revised in 2011, a new score scale was created, in part to discourage direct comparisons with the previous version of the exam. The former quantitative and verbal reasoning GRE scales ranged from 200 to 800, and the revised versions range from 130 to 170.

Let's transform `PISA09$age` to z-scores. Then, we'll rescale the result to have a new mean and standard deviation. You should check the mean and SD of each variable below, and compare them to the original values. Note that R sometimes represents zero as an infinitely small number, using scientific notation, instead of just reporting 0.

```
# Convert age to z-scores, then rescale to have a new mean and SD
# You should check the mean and SD of both zage and newage
# Also, see setmean() and setsd() from epmr, and scale() from base R
zage <- (PISA09$age - mean(PISA09$age)) / sd(PISA09$age)
newage <- (zage * 150) + 500
```

Rescaling using addition/subtraction and division/multiplication, as shown here, is referred to as *linear transformation*. When we linearly transform a variable, the shape of its distribution does not change. Linear transformations only affect the values on the x-axis. Thus, they're typically only used for interpretation purposes.

```
ggplot(PISA09, aes(factor(round(zage, 2)))) + geom_bar()
ggplot(PISA09, aes(factor(round(newage, 2)))) + geom_bar()
```

1.4 Summary

This chapter introduced two resources that will be used throughout this book, Proola and R. Some introductory statistics were also reviewed. These will support our discussions of reliability, item analysis, item response theory, dimensionality, and validity. Before moving forward, you should get an account at Proola and browse around the item bank. You should also install R and optionally RStudio on your own computer, and make sure you are comfortable explaining frequency distributions, central tendency, variability, correlation, and rescaling, including what they are and how they're used.

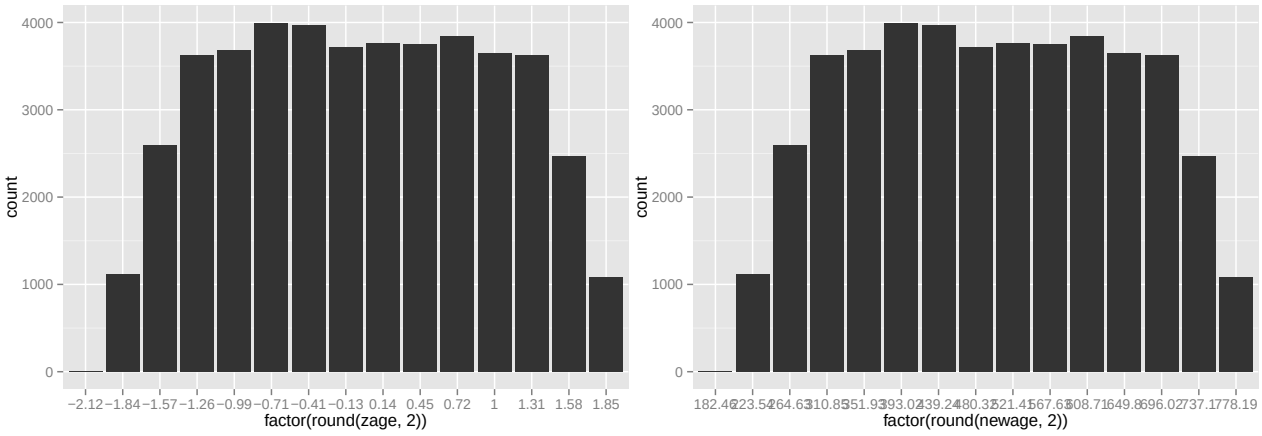


Figure 1.4: Histograms of age in the first plot, and age rescaled to have a mean of 500 and SD of 150 in the second.

1.4.1 Learning objectives

1. Identify and use the following notation: X, n and N, μ, σ, σ^2, ρ and r.
2. Calculate and interpret frequencies, proportions, and percentages.
3. Use a frequency distribution plot and histogram to describe the central tendency, variability, and shape of a distribution.
4. Calculate and interpret measures of central tendency and describe what they represent, including the mean, median, and mode.
5. Calculate and interpret measures of variability and describe what they represent, including the standard deviation and variance.
6. Apply and explain the process of rescaling variables using means and standard deviations for linear transformations.
7. Calculate and interpret the correlation between two variables and describe what it represents in terms of the shape, direction, and strength of the linear relationship between the variables.
8. Create and interpret a scatter plot, explaining what it indicates about the shape, direction, and strength of the linear relationship between two variables.

1.4.2 Exercises

1. What percentage of PISA students are in 10th grade?
2. How do the countries JPN and RUS compare in terms of their mean memor scores? Note that the international average across all countries is 0.
3. The variable `PISA09$bookid` identifies which of four test booklets, or test forms, students were given in this subsample of the full PISA data set. Was each booklet given to an equal number of students?
4. Which of the learning strategy scales has the most normal score distribution? What information supports your choice?
5. `PISA09$r414q02s` and `PISA09$r414q11s` contain scores on two of the PISA reading items, with 1 coded as a correct response and 0 as incorrect. Find and interpret the correlation between these two variables.

Chapter 2

Measurement, Scales, and Scoring

> Measurement is never better than the empirical operations by which it is carried out, and operations range from bad to good.
> — Stanley Stevens, *On the Theory of Scales of Measurement*

The Preface to this book introduced a few perspectives on testing, with an emphasis on validity as a measure of the effectiveness of test scores. Validity is an overarching issue that encompasses all stages in the test development and administration processes, from blueprint to bubble sheet, including the stage wherein we choose the empirical operations that will assign numbers or labels to test takers based on their performance or responses.

In this chapter, we examine the measurement process at its most fundamental or basic level, the measurement level. We'll define the three requirements for measurement, and consider the simplicity of physical measurement in comparison to the complexities of educational and psychological measurement where the thing we measure is often intractable and best represented using item sets and composite scores. Along the way, we'll describe the four types of measurement scales that are available, with examples from the PISA data set, and we'll look into why Stevens (1946) concluded that not all scales are created equal. Last are scoring and score referencing, including examples of norm and criterion referencing.

```
# R setup for this chapter
# Required packages are assumed to be installed - see chapter 1
library("epmr")
library("ggplot2")
# Functions we'll use in this chapter
```

2.1 What is measurement?

2.1.1 How do we define it?

We usually define the term *measurement* as the assignment of values to objects according to some system of rules. This definition originates with Stevens (1946), who presented what have become the four traditional scales or types of measurement. We'll talk about these shortly. For now, let's focus on the general measurement process, which involves giving an *object of measurement*, the person or thing for whom we're measuring, a value that represents something about it.

Measurement is happening all the time, all around us. Daily, we measure what we eat, where we go, and what we do. For example, drink sizes are measured using categories like tall, grande, and venti. A jog or a commute is measured in miles or kilometers. We measure the temperature of our homes, the air pressure in

our tires, and the carbon dioxide in our atmosphere. The wearable technology you might have strapped to your wrist could be monitoring your lack of movement and decreasing heart rate as you doze off reading this sentence. After you wake up, you might check your watch and measure the length of your nap in minutes or hours.

These are all examples of physical measurement. In each example, you should be able to identify 1) the object of measurement, 2) the property or quality that's being measured for it, and 3) the kinds of values used to represent amounts of this property or quality. The property or quality that's being measured for an object is called the *variable*. The kinds of values we assign to an object, for example, grams or degrees Celsius or beats per minute, are referred to as the *units of measurement* that are captured within that variable.

Let's look at some examples of measurement from the `state` data sets in R. The object `state.x77` contains data on eight variables, with the fifty US states as the objects of measurement. For details on where the data come from, see the help file `?state`.

```
# Load state data
data(state)
# Print first 6 rows, all columns
head(state.x77)
##            Population Income Illiteracy Life Exp Murder HS Grad Frost
## Alabama          3615   3624        2.1    69.05   15.1    41.3    20
## Alaska            365   6315        1.5    69.31   11.3    66.7   152
## Arizona          2212   4530        1.8    70.55    7.8    58.1    15
## Arkansas         2110   3378        1.9    70.66   10.1    39.9    65
## California      21198   5114        1.1    71.71   10.3    62.6    20
## Colorado         2541   4884        0.7    72.06    6.8    63.9   166
##              Area
## Alabama     50708
## Alaska     566432
## Arizona    113417
## Arkansas    51945
## California 156361
## Colorado   103766
```

Take a minute to consider what the variables in `state.x77` are measuring, and what the units of measurement are for these variables. For example, `state.x77[, "Population"]` contains population estimates from 1975 for each state, expressed as thousands. So, `state.x77["Wisconsin", "Population"]` gives us 4589, or a population of 4589 thousand people. What other variables from `state.x77` are measured as simple counts? To practice what you learned in Chapter 1, try to convert the illiteracy rates in `state.x77[, "Illiteracy"]` from proportions to counts *for each state*.

Note that `state.x77` is a matrix, which means we can't index columns using `$`. That only works with a `list` or `data.frame`. However, the rows and columns of a matrix can have names, accessed with `rownames()` and `colnames()`, and we can use these names to index the matrix, as shown above.

2.1.2 From physical to intangible

With most physical measurements, the property that we're trying to represent or capture with our values can be clearly defined and consistently measured. For example, amounts of food are commonly measured in grams. A cup of cola has about 44 grams of sugar in it. When you see that number printed on your can of soda pop or fizzy water, the meaning is pretty clear, and there's really no need to question if its accurate. Cola has a lot of sugar in it.

But, just as often, we take a number like the amount of sugar in our food and use it to represent something abstract or intangible like how healthy or nutritious the food is. A food's healthiness isn't as easy to define as

its mass or volume. A measurement of healthiness or nutritional value might account for the other ingredients in the food and how many calories they boil down to. Furthermore, different foods can be more or less nutritional for different people, depending on a variety of factors. Healthiness, unlike physical properties, is intangible and difficult to measure.

Most of the variables in `state.x77` are relatively easy to measure, as they involve observable quantities, such as numbers of dollars for `state.x77[, "Income"]`, years for `state.x77[, "Life Exp"]`, and days for `state.x77[, "Frost"]`. On the other hand, illiteracy rates are not as easily measured. A variable such as illiteracy is not countable or directly observable, which makes it subject to measurement error.

The social sciences of education and psychology typically focus on the measurement of *constructs*, intangible and unobservable qualities, attributes, or traits that we assume are causing certain observable behavior or responses. In this book, our objects of measurement are typically people, and our goal is to give these people numbers or labels that tell us something meaningful about qualities such as their intelligence, reading ability, or social anxiety. Constructs such as these are difficult to measure. That's why we need an entire book to discuss how to best measure them.

A good question to ask at this point is, how can we measure and provide values for something that's unobservable? How do we score a person's math ability if we can't observe it directly? What we need is an *operationalization* of our construct, an observable behavior or response that increases or decreases as a person moves up or down on the construct. With math ability, that operationalization might be the number of math questions a person answers correctly out of 20. With social anxiety, it might be the frequency of feeling anxious over a given period of time. With illiteracy, the operationalization might be the number of words read correctly from a passage of text. Or it could be a simple "yes" or "no," in response to the question, "Can you read?"

When using a proxy for our construct, we have to assume or *infer* that the operationalization we're actually observing and measuring accurately represents the underlying quality or property that we're interested in. This brings us to the overarching question addressed in this book.

2.1.3 What makes measurement good?

In the last year of my undergraduate work in psychology I conducted a research study on the constructs of aggression, sociability, and victimization with Italian preschoolers (Nelson et al. 2010). I spent about four weeks collecting data in preschools. Data collection involved covering a large piece of cardboard with pictures of all the children in a classroom, and then asking each child, individually, questions about their peers.

To measure sociability, we asked three simple questions: "who is fun to talk to?" "who is fun to do pretend things with?"" and "who has many friends?" Kids with lots of peer nominations on these questions received a higher score, indicating that they were more sociable. After asking these and other questions to about 300 preschoolers, and then tallying up the scores, I wondered how well we were actually measuring the constructs we were targeting. Were these scores any good? Was three or five questions enough? Maybe we were missing something important? Maybe some of these questions, which had to be translated from English into Italian, meant different things on the coast of the Mediterranean than they did in the Midwest US?

This project was my first experience on the measuring side of measurement, and it fascinated me. The questions that I asked then are the same questions that we'll ask and answer in this book. How consistently and accurately are we measuring what we intend to measure? What can we do to improve our measurement? And how can we identify instruments that are better or worse than others? These questions all have to do with what makes measurement good.

Many different things make measurement good, from writing high-quality items to adherence to established test development guidelines. For the most part, the resulting scores are considered good, or effective, when they consistently and accurately describe a target construct. Consistency and accuracy refer to the *reliability* and *validity* of test scores, that is, the extent to which the same scores would be obtained across repeated administrations of a test, and the extent to which scores fully represent the construct they are intended to measure.

Table 2.1: Intended Uses for Some Common Types of Standardized Tests

Test Type	Intended Use
Accountability	Hold various people responsible for student learning
Admissions	Selection for entrance to an educational institution
Employment	Help in hiring and promotion of employees
Exit Testing	Check for mastery of content required for graduation
Licensing	Verify that candidates are fit for practice
Placement	Selecting coursework or instructional needs

These two terms, reliability and validity, will come up many times throughout the book. The second one, validity, will help us clarify our definition of measurement in terms of its purpose. Of all the considerations that make for effective measurement, the first to address is purpose.

2.1.4 What is the purpose?

Measurement is useless unless it is based on a clearly articulated purpose. This purpose describes the goals of administering a test or survey, including what will be measured, for whom, and why? We've already established the "what?" as the variable or construct, the property, quality, attribute, or trait that our numbers or values represent. We've also established the "for whom?" as the object, in our case, people, but more specifically perhaps students, patients, or employees. Now we need to establish the "why?"

The purpose of a test specifies its intended application and use. It addresses how scores from the test are designed to be interpreted. A test without a clear purpose can't be effective.

Suppose someone asks you to create a measure of students' financial savvy, that is, their understanding of money and how it is used in finance. You've got here a simple construct, understanding of finance, and the object of measurement, students. But before you can develop this test you'd need to know how it is going to be used. Its purpose will determine key features like what specific content the test contains, the level of difficulty of the questions, the types of questions used, and how it is administered. If the test is used as a final exam in a finance course, it should capture the content of that course, and it might be pretty rigorous. On the other hand, if it's used with the general student body to see what students know about balancing budgets and managing student loans, the content and difficulty might change. Clearly, you can't develop a test without knowing its purpose. Furthermore, a test designed for one purpose may not function well for another.

Take a minute to think about some of the tests you've used or taken in the past. How would you express the purposes of these tests? When answering this question, be careful to avoid simply saying that the purpose of the test is to measure something. A statement of test purpose should clarify what can be done with the resulting scores. For example, scores from placement tests are used to determine what courses a student should take or identify students in need of certain instructional resources. Scores on admissions tests inform the selection of applicants for entrance to a college or university. Scores on certification and licensure exams are used to verify that examinees have the knowledge, skills, and abilities required for practice in a given profession. Table 2.1 includes these and a few more examples. In each case, scores are intended to be used in a specific way.

Here's one more example. Some of my research is based on a type of standardized placement testing that is used to measure student growth over a short period of time. In addition to measuring growth, scores are also used to evaluate the effectiveness of intervention programs, where effective interventions lead to positive results for students. My latest project in this area of assessment involved measures of early literacy called IGDIs (Bradfield et al. 2014). A brochure for the measures from www.myigdis.com states,

> myIGDIs are a comprehensive set of assessments for monitoring the growth and development of young children. myIGDIs are easy to collect, sensitive to small changes in children's achievement,

and mark progress toward a long-term desired outcome. For these reasons, myIGDIs are an excellent choice for monitoring English Language Learners and making more informed Special Education evaluations.

This summary contains specific claims regarding score use, including monitoring growth, and sensitivity to small changes in achievement. Validity evidence is needed to demonstrate that scores can effectively be used in these ways.

The point of these examples is simply to clarify what goes into a statement of purpose, and why a well articulated purpose is an essential first step to measurement. We'll come back to validation of test purpose in Chapter 9. For now, you just need to be familiar with how a test purpose is phrased and why it's important.

2.1.5 Summary

To summarize this section, the measurement process allows us to capture information about individuals that can be used to describe their standing on a variety of constructs, from educational ones, like math ability and vocabulary knowledge, to psychological ones, like sociability and aggression. We measure these properties by operationalizing our construct, for example, in terms of the number of items answered correctly or the number of times individuals exhibit a certain behavior. These operational variables are then assumed to represent our construct of interest. Finally, our measures of these constructs can then be used for specific purposes, such as to inform research questions about the relationship between sociability and aggression, or to measure growth in early literacy.

So, measurement involves a construct that we don't directly observe and an operationalization of it that we do observe. Our measurement is said to be effective when there is a strong connection between the two, which is best obtained when our measurement has a clear purpose. In the next two sections, on measurement scales and scoring, we'll focus on how to handle the operational side of measurement. Then, with measurement models, we'll consider the construct side. Finally, in the section on score referencing, we'll talk about additional labels that we use to give meaning to our scores.

2.2 Measurement scales

Now that we've established what measurement is, and some key features that make the measurement process good, we can get into the details of how measurement is carried out. As defined by Stevens (1946), measurement involves the assignment of values to objects according to certain rules. The rules that guide the measurement process determine the type of measurement scale that is produced and the statistics that can be used with that scale.

Measurement scales are grouped into four different types. These differ in the meaning that is given to the values that are assigned, and the relationship between these values for a given variable.

2.2.1 Nominal

The most basic measurement scale is really the absence of a scale, because the values used are simple categories or names, rather than quantities of a variable. For this reason it is referred to as a *nominal scale*, where our objects of measurement are grouped qualitatively, for example by gender or political party. The nominal scale can also represent variables such as zip code or eye color, where multiple categories are present. So, identifying (ID) variables such as student last name or school ID are also considered nominal.

Only frequencies, proportions, and percentages (and related nonparametric statistics) are permitted with nominal variables. Means and standard deviations (and related parametric statistics) do not work. It would be meaningless to calculate something like an average gender or eye color, because nominal variables lack any inherent ordering or quantity in their values.

What variables from `PISA09` would be considered nominal? Nominal variables often have a specific class in R, which you can check with `class()`, as we did in Chapter 1. The `class` of an R object doesn't map directly to its measurement scale, but it can provide helpful information about features of the scale. `mode()` can also be informative regarding how a variable is measured.

Nominal variables are often coded in R using text strings. Our `classroom` variable from Chapter 1 is an example. After converting this character object to the factor class, R identifies each unique value within the object as a level, or unit of measurement. Other functions in R will then be able to recognize the object as a factor, and will interpret its contents differently. To see this interpretation in action, try to calculate the mean of our classroom variable, or any of the nominal variables in `PISA09`, and you'll receive a warning from R.

```
# Nominal classroom labels as character
roomnumber <- c("1", "2", "3", "1", "2", "3")
class(roomnumber)
## [1] "character"
# Nominal classroom labels as factor
roomnumber <- factor(roomnumber)
class(roomnumber)
## [1] "factor"
```

2.2.2 Ordinal

The dominant feature of the *ordinal scale* is order, where values do have an inherent ordering that cannot be removed without losing meaning. Common examples of ordinal scales include ranks (e.g., first, second, third, etc.), the multi-point rating scales seen in surveys (e.g., strongly disagree, disagree, etc.), and level of educational attainment.

The distance between the ordered categories in ordinal scale variables (i.e., the interval) is never established. So, the difference between first and second place does not necessarily mean the same thing as the difference between second and third. In a swimming race, first and second might differ by a matter of milliseconds, whereas second and third differ by minutes. We know that first is faster than second, and second is faster than third, but we don't know *how much* faster. Note that the construct we're measuring here is probably swimming ability, which is actually operationalized on a ratio scale, in terms of speed, but it is simplified to an ordinal scale when giving out awards.

What variables in `PISA09` are measured on ordinal scales? To identify a variable as being measured on an ordinal scale, we really need to know how data were collected for the variable, and what the scale values are intended to represent. Whereas the choice of nominal is relatively simple, the choice of ordinal over interval or ratio is often requires a judgement call, as discussed below.

In theory, statistics which rely on interval level information, such as the mean, standard deviation, and all mean-based statistical tests, are still not allowed with an ordinal scale. Statistics permitted with ordinal variables include the median and any other statistics based on percentiles.

2.2.3 Interval

Interval scales include ordered values where the distances, or intervals, between them are meaningful. Whereas an ordinal scale describes one category only as greater than, less than, or equal to another, with an interval scale the difference between categories is quantified in scale points that have a consistent meaning across the scale. With interval scales we can finally use means, standard deviations, and related parametric statistical tests.

One common example of an interval scale is test score based on number correct, where each item in a test is worth the same amount when calculating the total. When treating test scores as interval variables, we make the assumption that a difference in score points reflects a consistent difference in the construct no matter

where we are on the scale. This can sometimes be problematic. A test of vocabulary could be measured on an interval scale, where each correctly defined word contributes the same amount to the total score. However, in this case we assume that each correct definition is based on the same amount of construct, vocabulary knowledge. That is, the vocabulary words need to be similar in difficulty for the students we're testing. Otherwise, scale intervals will not have a consistent meaning. Instead, an increase in number correct will depend on the word that is answered correctly. Item response theory, described in Chapter 7, seeks to address this issue.

Another common example of an interval scale is temperature as measured in degrees centigrade or Fahrenheit. These temperature scales both have meaningful intervals, where a given increase in heat, for example, produces the same increase in degrees no matter where you are on the scale. However, a zero on the Fahrenheit or centigrade scales does not indicate an absence of the variable we are measuring, temperature. This is the key distinction between an interval and a ratio scale.

2.2.4 Ratio

The ratio scale is an interval scale with a meaningful absolute zero, or a point at which there is an absence of the variable measured. Whereas an interval scale describes differences between scale values in scale points, a ratio scale can compare values by ratios. A simple example is time, where 1 hour is equivalent to 2/3 hours + 1/3 hours. Other examples include counts of observations or occurrences, such as the number of aggressive or prosocial behaviors per hour, or the frequency of drug use in the past month.

Note that we often reference ratio scales when operationalizing constructs, in which case we may lose our meaningful zero point. For example, zero prosocial behaviors does in fact indicate that nothing noticeably prosocial occurred for a student over a certain period of time. However, this may not mean that a student is completely void of prosociability. In the same way, zero aggressive behaviors does not necessarily indicate an absence of aggression. Thus, when a ratio variable is used to operationalize a construct, it may necessarily lose its ratio properties.

All statistics are permitted with ratio scales, though the only ones we talk about, in addition to those available with interval scales, are statistics that let you make comparisons in scores using ratios. For example, a two hour test is twice as long as a one hour test, and five aggressive episodes is half as many as ten. However, as before, if our scale is assumed to reference some underlying construct, five aggressive episodes may not indicate twice as much aggression as ten.

2.2.5 Comparing scales

Progressing from nominal to ratio, the measurement scales become more descriptive of the variable they represent, and more statistical options become available. So, in general, the further from a nominal scale the better, as once a variable is measured its scale can only be downgraded. Consider the variable *age*, which could be represented in the following four ways:

1. number of days spent living, from 0 to infinity;
2. day born within a given year, from 1 to 365;
3. degree of youngness, including toddler, adolescent, adult, etc.; or
4. type of youngness, such as the same as Mike, or the same as Ike.

The first of these four, a ratio scale, is the most versatile and can be converted into any of the scales below it. However, once age is defined based on a classification, such as "same as Mike," no improvement can be made. For this reason a variable's measurement scale should be considered in the planning stages of test design, ideally when we identify the purpose of our test.

In the social sciences, measurement with the ratio scale is difficult to achieve because our operationalizations of constructs typically don't have meaningful zeros. So, interval scales are considered optimal, though they

too are not easily obtained. Consider the sociability measure described above. What type of scale is captured by this measure? Does a zero score indicate a total absence of sociability? This is required for ratio. Does an incremental increase at one end of the scale mean the same thing as an incremental increase at the other end of the scale? This is required for interval.

Upon close examination, it is difficult to measure sociability, and most other constructs in the social sciences, with anything more than an ordinal scale. Unfortunately, an interval or ratio scale is required for the majority of statistics that we'd like to use. Along these lines, Stevens (1946) concluded:

> Most of the scales used widely and effectively by psychologists are ordinal scales. In the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these scales, for these statistics imply a knowledge of something more than the relative rank-order of data. On the other hand, for this "illegal" statisticizing there can be invoked a kind of pragmatic sanction: In numerous instances it leads to fruitful results. While the outlawing of this procedure would probably serve no good purpose, it is proper to point out that means and standard deviations computed on an ordinal scale are in error to the extent that the successive intervals on the scale are unequal in size. When only the rank-order of data is known, we should proceed cautiously with our statistics, and especially with the conclusions we draw from them. [p. 679]

Based on this argument, a mean sociability score is only as useful as the scale itself is interval. The less meaningful the intervals between sociability scores, the less meaningful our mean estimate will be. Thus, when designing an instrument, we need to be aware of this limitation, and do our best to improve the intervalness of our scales. When stating the purpose of a test, we need to be aware of how our construct and operationalization of it will impact our resulting scale. Finally, we need to acknowledge the limitations of our scales, especially when utilizing potentially incorrect statistics.

Let's look at the `PISA09` reading scores as a final example. All of the scored item responses in `PISA09` have only two possible values, 0 and 1, representing incorrect and correct responses. These scored item responses could all be considered at least on ordinal scales, as 1 represents more of the measured construct than 0. The total score across all reading items could also be considered at least ordinal.

```
# Names of all reading items, to be used as an indexing object
ritems <- c("r414q02", "r414q11", "r414q06", "r414q09", "r452q03",
  "r452q04", "r452q06", "r452q07", "r458q01", "r458q07", "r458q04")
# Paste an "s"" to the end of each name, for scored items
rsitems <- paste0(ritems, "s")
# apply() applies to a data set (the first argument) across either
# rows or columns (the second argument) the function named (in the
# third argument). See also rowSums(). Here, we treat missings as
# 0s, by excluding them from the sum.
PISA09$rtotal <- apply(PISA09[, rsitems], 1, sum, na.rm = TRUE)
dstudy(PISA09$rtotal)
##
## Descriptive Study
##
##   mean median   sd    skew kurt min max     n na
## 1 5.46      6 2.92 -0.0914 1.98   0  11 44878  0
```

The challenge in identifying a measurement scale is interpreting the meaningfulness of the intervals between scale values, for an interval scale, and the meaningfulness of a zero value, for a ratio scale. Do increments in the total reading score have a consistent meaning, relative to the construct of reading ability, across the scale? The answer depends on our interpretation of the items themselves and the mechanisms used to score them.

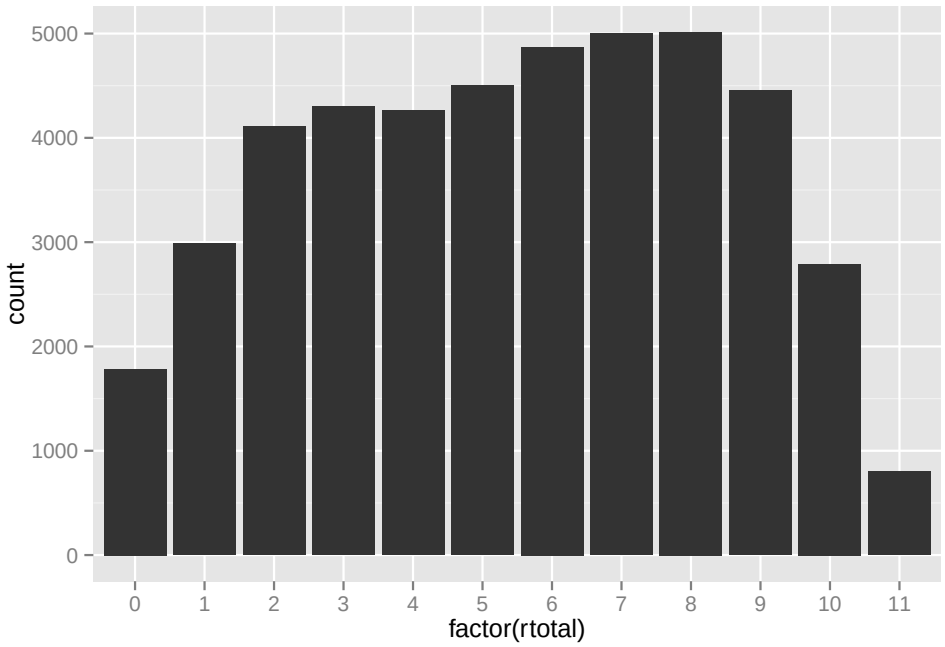


Figure 2.1: A bar plot of total scores on PISA09 scored reading items.

```
# Bar plot of reading totals, which are, ironically, converted to
# a factor before plotting, so ggplot treats them as a discrete
# scale rather than continuous. Continuous defaults to a histogram
# which isn't as useful.
ggplot(PISA09, aes(factor(rtotal))) + geom_bar()
```

When we talk about measurement scales, and next scoring, keep in mind that we're focusing on the x-axis of a plot like the one shown in Figure 2.1. The distribution itself doesn't necessarily help us interpret the measurement scale or scoring process. Instead, we're examining how individual items and combinations of them capture differences in the underlying construct.

2.3 Scoring

This book focuses on cognitive and noncognitive, that is, affective, test scores as operationalizations of constructs in education and psychology. As noted above, these test scores often produce ordinal scales with some amount of meaning in their intervals. The particular rules for assigning values within these scales depend on the type of scoring mechanisms used. Here, we'll cover the two most common scoring mechanisms or rules, dichotomous and polytomous scoring, and we'll discuss how these are used to create rating scales and composite scores.

2.3.1 Dichotomous scoring

Dichotomous scoring refers to the assignment of one of two possible values based on a person's performance or response to a test question. A simple example is the use of correct and incorrect to score a cognitive item response. These values are mutually exclusive, and describe the correctness of a response in the simplest terms possible, as completely incorrect or completely correct.

Multiple-choice questions, discussed further in Chapter 4, are usually scored dichotomously. Most cognitive tests involve at least some dichotomously scored items. The PISA 2009 study implemented both dichotomously and polytomously scored items. However, all of the reading items in `PISA09` are dichotomously scored.

```
# The apply() function again, used to iterate through reading items,
# and for each column (hence the 2), run a frequency table.
# Then, divide the result by the number of students, and round
# to 2 decimal places.
rtab <- apply(PISA09[, rsitems], 2, table)
round(rtab / nrow(PISA09), 2)
##   r414q02s r414q11s r414q06s r414q09s r452q03s r452q04s r452q06s r452q07s
## 0     0.50     0.61     0.44     0.34     0.86     0.35     0.48     0.51
## 1     0.48     0.37     0.53     0.64     0.13     0.64     0.50     0.47
##   r458q01s r458q07s r458q04s
## 0     0.44     0.43     0.41
## 1     0.55     0.57     0.59
```

Dichotomous scoring is also used in affective measures, such as attitude surveys, behavior checklists, and personality tests. The most common example is scoring that represents a response of either "yes" or "no." Statements are written to capture some feature of the construct, such as adventurousness, and individuals then indicate whether or not the statements are characteristic of them.

> I would enjoy being a bounty hunter.
> * No
> * Yes

Alternatively, measures like the Myers-Briggs Type Indicator (MBTI; Myers et al. 1998), discussed further in Chapter 4, present a dichotomous choice in the form of sentence completion. One option must then be *keyed* to indicate a higher score on the construct being measured.

> I would prefer to hunt bounties
> * Alone
> * In a group

If "alone" were keyed positively, and "in a group" were keyed zero, how would you describe the construct that it measures? What if "in a group" were scored positively, and "alone" negatively?

2.3.2 Polytomous scoring

Polytomous scoring simply refers to the assignment of three or more possible values for a given test question or item. In cognitive testing, a simple example is the use of rating scales to score written responses such as essays. In this case, score values may still describe the correctness of a response, but with differing levels of correctness, for example, incorrect, partially correct, and fully correct.

Polytomous scoring with cognitive tests can be less straightforward and less objective than dichotomous scoring, primarily because it usually requires the use of human raters with whom it is difficult to maintain consistent meaning of assigned categories such as partially correct. The issue of interrater reliability will be discussed in Chapter 5.3.1.

Polytomous scoring with affective or non-cognitive measures most often occurs with the use of rating scales. For example, individuals may use a rating scale to describe how much they identify with a statement, or how well a statement represents them, rather than simply saying "yes" or "no." Such rating scales measure multiple levels of agreement (e.g., from disagree to agree) or preference (e.g., from dislike to like). In this

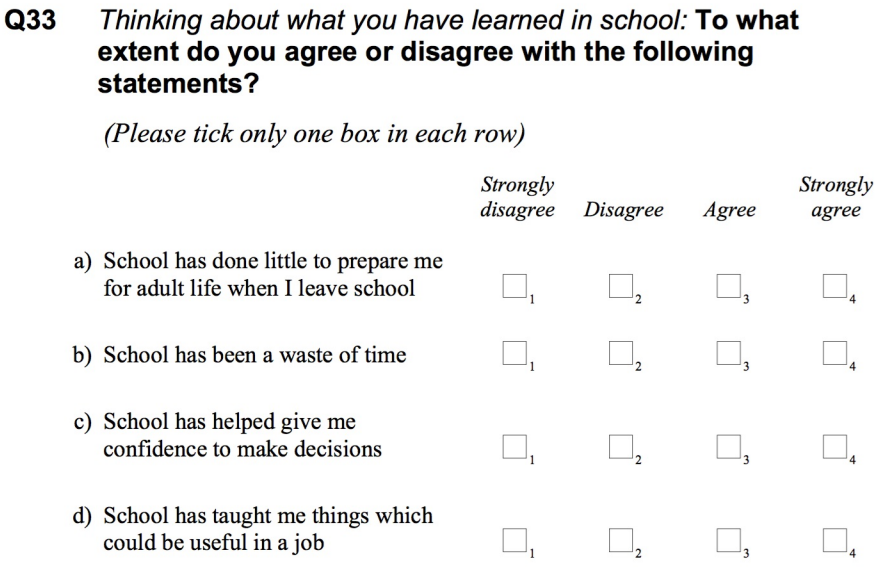


Figure 2.2: PISA 2009 student survey items measuring attitude toward school.

case, because individuals provide their own responses, subjectivity in scoring is not an issue as it is with polytomous scoring in cognitive tests. Instead, the challenge with rating scales is in ensuring that individuals interpret the rating categories in the same way. For example, strongly disagree could mean different things to different people, which will impact how the resulting scores can be compared across individuals.

The student survey variables in `PISA09$st27q01` through `PISA09$st42q05` are polytomously scored rating scale items that contribute to four different composite scales. One of these scales measures students attitude toward school, shown in Figure 2.2.

Notice that the first two items in the scale are phrased negatively, and the second two are phrased positively. Items `PISA09$st33q01` and `PISA09$st33q02`, labeled a and b in Figure 2.2, need to be reverse coded if we want higher scores to represent more positive attitude, as in c and d. `recode()` from the empr package will automatically reverse score a numeric variable.

```
PISA09$st33q01r <- recode(PISA09$st33q01)
PISA09$st33q02r <- recode(PISA09$st33q02)
```

Having recoding the negatively worded attitude items, we can consider ways of combining information across all four items to get an overall measure for each student. Except in essay scoring and with some affective measures, an individual question is rarely used alone to measure a construct. Instead, scores from multiple items are combined to create composite scores or rating scale scores.

2.3.3 Rating scales

When I was in graduate school, the professor for my introductory measurement class would chastise students when they referred to multipoint rating scales as "Likert scales." Likert (1932) did not invent the rating scale. Instead, he detailed two methods for combining scores across multiple rating scale items to create a composite score that would be, in theory, a stronger measure of the construct than any individual item. One of these methods, which has become a standard technique in affective measurement, is to assign ordinal numerical values to each rating scale category, and then calculate a sum or average across a set of these rating scale items.

The scaling technique demonstrated by Likert (1932) involves, first, the scoring of individual rating scale items using polytomous scales. For example, response options for one set of survey questions in Likert (1932) included five categories, ranging from strongly disapprove to undecided to strongly approve, similar to the PISA survey items. These categories were assigned score values of 1 through 5. Then, a total score was obtained across all items in the set. Low scores were interpreted as indicating strong disapproval and high scores were interpreted as indicating strong approval. This process could be referred to as Likert scaling. But in this book we'll simply refer to it as scaling, or creating a total or average score across multiple items.

```
# Names of all attitude toward school items, to be used as an indexing object
# Note the added "r" in the names for the first two items
atsitems <- c("st33q01r", "st33q02r", "st33q03", "st33q04")
# Calculate total scores again using apply(), and look at descriptives
PISA09$atotal <- apply(PISA09[, atsitems], 1, sum, na.rm = TRUE)
dstudy(PISA09$atotal)
##
## Descriptive Study
##
##   mean median   sd  skew kurt min max     n na
## 1 12.3     13 2.71 -1.66 8.12   0  16 44878  0
```

In Chapter 4 we will address rating scales in more detail. We'll cover issues in the construction and administration of rating categories. Here, we are more concerned with the benefits of using composite scale scores.

2.3.4 Composites versus components

A composite score is simply the result of some combination of separate subscores, referred to as components. Thus far, we have created total scores in R. Most often, we will deal with either total scores or *factor scores* on a test, where individual items make up the components. Factor scores refer to scores obtained from some measurement model, such as a classical test theory model, discussed in Chapter 5.3.1, or an item response theory model, discussed in Chapter 7. We will also encounter composite scores based on totals and means from rating scale items. In each case, the composite is going to be preferable to any individual component for a number of reasons.

Composite scores are preferable from a statistical standpoint because they tend to provide a more reliable and valid measure of our construct. Composites are more reliable and valid because they combine information from multiple smaller, repeated measures of the construct. These smaller components may each be limited in certain ways, or may only present a small piece of the big picture, and when combined, the resulting score is more comprehensive and more easily reproduced in subsequent measurements. In Chapter 5.3.1, we'll learn more about why reliability is expected to increase, in theory, as we increase the number of items in our composite.

For example, when measuring a construct such as attitude toward animal rights, a single item would only provide information about a specific instance of the issue. Consider the example survey items presented by @mathews1997personality:

> The Animal Attitude Scale (AAS) assesses individual differences in attitudes toward the treatment of animals... It is composed of 29 items which subjects rate on a five-point Likert scale (strongly agree to strongly disagree). Sample items include, "I do not think that there is anything wrong with using animals in medical research," "It is morally wrong to hunt wild animals just for sport," and "I would probably continue to use a product that I liked even though I know that its development caused pain to laboratory animals." [p. 171]

By themselves, any one of these items may not reflect the full construct that we are trying to measure. A person may strongly support animal rights, except in the case of medical research. Or a person may define the phrase "that I liked," from the third example question, in different ways so that this individual question would produce different results for people who might actually be similar in their regard for animals. A composite score will help wash out the limitations of individual items. (Side note from this study: a regression model showed that 25% of the variance in attitude toward animals was accounted for by gender and a personality measure of sensitivity.)

The simpler methods for creating composites, by averaging and totaling across items, are used with smaller-scale instruments to facilitate scoring and score reporting. A major drawback to simple averages and totals is the lack of interval properties in the resulting scales. To avoid this and other drawbacks, many instruments, including large-scale educational tests and psychological measures, are scaled using measurement models.

2.4 Measurement models

Whereas a simple sum or average over a set of items lets each item contribute the same amount to the overall score, more complex measurement models can be used to estimate the different contributions of individual items to the underlying construct. These contributions can be examined in a variety of ways, as discussed in Chapters 5.3.1, 8, and 7. Together, they can provide useful information about the quality of a measure, as they help us understand the relationship between our operationalization of the construct, in terms of individual items, and the construct itself.

Measurement models represent an unobservable construct by formally incorporating a measurement theory into the measurement process. We will review two theories in this book. The first, presented in 5.3.1, is called classical test theory, and the second is item response theory (for a comparison, see Hambleton and Jones 1993). For now, we'll just look at the basics of what a measurement model does.

Figure 2.3 contains a visual representation of a simple measurement model where the underlying construct of sociability, shown in an oval, causes, in part, the observed responses in a set of three questions, shown in rectangles as Item 1, Item 2, and Item 3. Unobservable quantities in a measurement model are typically represented by ovals, and observable quantities by rectangles. Causation is then represented by the arrows which point from the construct to the item responses. The numbers over each arrow from the construct are the scaled factor loadings reported in Nelson et al. (2010), which represent the strength of the relationship between the items and the construct which they together define. As with a correlation coefficient, the larger the factor loading, the stronger the relationship. Thus, item 1 has the strongest relationship with the sociability factor, and item 3 has the weakest.

The other unobserved quantities in Figure 2.3 are the error terms, in the circles, which also impact responses on the three items. Without arrows linking the error terms from one to another, the model assumes that errors are independent and unrelated across items. In this case, any influence on a response that does not come from the common factor of sociability is attributed to measurement error.

Models such as the one in Figure 2.3 are referred to as confirmatory factor analysis models, because we propose a given structure for the relationships between constructs, error, and observations, and seek to confirm it by placing certain constraints on the relationships we estimate. In Chapter 8, we'll discuss these and exploratory models were we aren't certain how many underlying constructs are causing responses.

2.5 Score referencing

Now that we've discussed the measurement process we can go over some common methods for giving meaning to the scores that our measures produce. These methods are referred to as norm and criterion score referencing. Each is discussed briefly below, with examples.

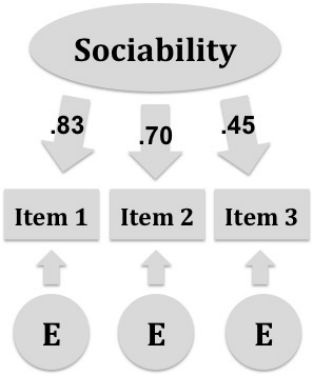


Figure 2.3: A simple measurement model for sociability with three items, based on results from D. A. Nelson et al. (2010). Numbers are factor loadings and E represents unique item error.

2.5.1 Norm referencing

Norm referencing gives meaning to scores by comparing them to values for a specific norm group. For example, when my kids bring home their standardized test results from school, their scores in each subject area, math and reading, are given meaning by comparing them to the distribution of scores for students across the state. A score of 22 means very little to a parent who does not have access to the test itself. However, a percentile score of 90 indicates that a student scored at or above 90% of the students in the norming group, regardless of what percentage of the test questions they answered correctly.

Norms are also frequently encountered in admissions testing. If you took something like the ACT or SAT, college admissions exams used in the US, or the GRE, an admissions test for graduate school, you're probably familiar with the ambiguous score scales these exams use in reporting. Each scale is based on a conversion of your actual test scores to a scale that is intentionally difficult or impossible to understand. In a way, the objective in this rescaling of scores is to force you to rely on the norm referencing provided in your score report. The ACT scales range from 1 to 36, but a score of 20 on the math section doesn't tell you a lot about how much math you know or can do. Instead, by referencing the published norms, a score of 20 tells you that you scored around the 50th percentile for all test takers.

The two examples above involve simple percentile norms, where scores are compared to the full score distribution for a given norm group. Two other common types of norm referencing are grade and age norms, which are obtained by estimating the typical or average performance on a test by grade level or age. For example, we can give meaning to `PISA09` reading scores by comparing them to the medians by grade.

```
# tapply() is like apply() but instead of specifying rows or columns of
# a matrix, we provide an index variable. Here, median() will be applied
# over subsets of rtotal by grade. with() is used to subset only German
# students.
with(PISA09[PISA09$cnt == "DEU", ],
  tapply(rtotal, grade, median, na.rm = TRUE))
##    7    8    9   10   11   12
##  1.5  3.0  5.0  7.0 10.0  9.0
# Most German students are in 9th grade. Medians aren't as useful for
# grades 7, 11, and 12.
table(PISA09$grade[PISA09$cnt == "DEU"])
##
##   7   8   9  10  11  12
##  12 155 771 476   5   1
```

A reading score of 6, for example, would be given a grade norm of 9, as it exceeds the median score for 9th

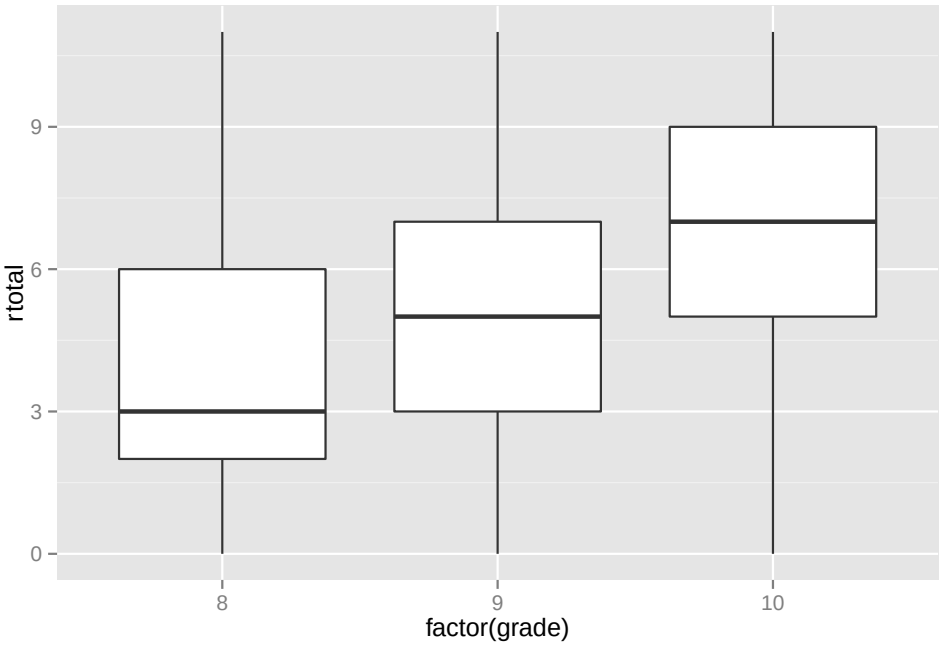


Figure 2.4: Bar plots of total scores on PISA09 scored reading items for Germany by grade.

graders but not 10th graders. In practice, grade norms are reported using decimals that capture the month within the school year as well, For example, a 9.6 would indicate that a student's reading score is at the median performance of students in their sixth month of 9th grade. These normative scores by grade are referred to as grade equivalents. Age norms and age equivalents are calculated in the same way, but using age as the indexing variable.

Again, norms give meaning to a score by comparing it to the score distribution for a particular norming group. Box plots can be used to visualize a score distribution based on the 25th, 50th, and 75th percentiles, along with any outliers.

```
ggplot(PISA09[PISA09$cnt == "DEU" & PISA09$grade %in% 8:10, ],
  aes(x = factor(grade), y = rtotal)) +
  geom_boxplot()
```

2.5.2 Criterion referencing

The main limitation of norm referencing is that it only helps describe performance relative to other test takers. Criterion score referencing does the opposite. Criterion referencing gives meaning to scores by comparing them to values directly linked to the test content itself, regardless of how others perform on the content (Popham and Husek 1969).

Educational tests supporting instructional decision making are often criterion referenced. For example, classroom assessments are used to identify course content that a student has and has not mastered, so that deficiencies can be addressed before moving forward. The vocabulary test mentioned above is one example. Others include tests used in student placement and exit testing.

Standardized state test results, which were presented above as an example of norm referencing, are also given meaning using some form of criterion referencing. The criteria in state tests are established, in part, by a panel of teachers and administrators who participate in what is referred to as a *standard setting*. State test standards are chosen to reflect different levels of mastery of the test content. In Nebraska, for example, two

cut-off scores are chosen per test to categorize students as below the standards, meets the standards, and exceeds the standards. These categories are referred to as performance levels. Student performance can then be evaluated based on the description of typical performance for their level. Here is the performance level description for grade 5 science, meets the standard, as of 2014:

> Overall student performance in science reflects satisfactory performance on the standards and sufficient understanding of the content at fifth grade. A student scoring at the Meets the Standards level generally draws on a broad range of scientific knowledge and skills in the areas of inquiry, physical, life, and Earth/space sciences.

The Nebraska performance categories and descriptions are available online at www.education.ne.gov/assessment. Performance level descriptions are accompanied by additional details about expected performance for students in this group on specific science concepts. Again for grade 5 science, meets the standard:

> A student at this level generally:
> 1. Identifies testable questions,
> 2. Identifies factors that may impact an investigation,
> 3. Identifies appropriate selection and use of scientific equipment,
> 4. Develops a reasonable explanation based on collected data,
> 5. Describes the physical properties of matter and its changes.

The performance levels and descriptors used in standardized state tests provide general information about how a test score relates to the content that the test is designed to measure. Given their generality, these results are of limited value to teachers and parents. Instead, performance level descriptors are used for accountability purposes, for example, to assess performance at the school, district, and even the state levels in terms of the numbers of students meeting expectations.

The Beck Depression Inventory (BDI; Beck et al. 1961) is an example of criterion referencing in psychological testing. The BDI includes 21 items representing a range of depressive symptoms. Each item is scored polytomously from 0 to 3, and a total score is calculated across all of the items. Cutoff scores are then provided to identify individuals with minimal, mild, moderate, and severe depression, where lower scores indicate fewer depressive symptoms and higher scores indicate more severe depressive symptoms.

2.5.3 Comparing referencing methods

Although norm and criterion referencing are presented here as two distinct methods of giving meaning to test scores, they can sometimes be interrelated and thus difficult to distinguish from one another. The myIGDI testing program described above is one example of score referencing that combines both norms and criteria. These assessments were developed for measuring growth in early literacy skills in preschool and kindergarten classrooms. Students with scores falling below a cut-off value are identified as potentially being at risk for future developmental delays in reading. The cut-off score is determined in part based on a certain percentage of the test content (criterion information) and in part using mean performance of students evaluated by their teachers as being at-risk (normative information).

Norm and criterion referencing serve different purposes. Norm referencing is typically associated with tests designed to rank order test takers and make decisions involving comparisons among individuals, whereas criterion referencing is associated with tests designed to measure learning or mastery and make decisions about individuals and programs (e.g., Bond 1996; Popham and Husek 1969). These different emphases are relevant to the purpose of the test itself, and should be considered in the initial stages of test development, as discussed in Chapters 3 and 4.

2.6 Summary

This chapter provides an overview of what measurement is, how measurement is carried out in terms of scaling and scoring, and how measurement is given additional meaning through the use of score referencing and scale transformation. Before moving on to the next chapter, make sure you can respond to the learning objectives for this chapter, and complete the exercises below.

2.6.1 Learning objectives

1. Define the process of measurement.
2. Define the term *construct* and describe how constructs are used in measurement, with examples.
3. Compare and contrast measurement scales, including nominal, ordinal, interval, and ratio, with examples, and identify their use in context.
4. Compare and contrast dichotomous and polytomous scoring.
5. Describe how rating scales are used to create composite scores.
6. Explain the benefits of composites over component scores.
7. Create a generic measurement model and define its components.
8. Define norm referencing and identify contexts in which it is appropriate.
9. Compare three examples of norm referencing: grade, age, and percentile norms.
10. Define criterion referencing and identify contexts in which it is appropriate.
11. Describe how standards and performance levels are used in criterion referencing with standardized state tests.
12. Compare and contrast norm and criterion score referencing, and identify their uses in context.

2.6.2 Exercises

1. Teachers often use brief measures of oral reading fluency to see how many words students can read correctly from a passage of text in one minute. Describe how this variable could be modified to fit the four different scales of measurement.
2. Examine frequency distributions for each attitude toward school item, as was done with the reading items. Try converting counts to percentages.
3. Plot a histogram and describe the shape of the distribution of attitude toward school scores.
4. What country has the most positive attitude toward school?
5. Describe how both norm and criterion referencing could be helpful in an exam used to screen applicants for a job.
6. Describe how norm and criterion referencing could be used in evaluating variables outside the social sciences, for example, with the physical measurement applications presented at the beginning of the chapter.
7. Provide details about a measurement application that interests you.
    a. How would you label your construct? What terms can be used to define it?
    b. With whom would you measure this construct? Who is your object of measurement?
    c. What are the units of measurement? What values are used when assigning scores to people? What type of measurement scale will these values produce?
    d. What is the purpose in measuring your construct? How will scores be used?
    e. How is your construct commonly measured? Are there existing measures that would suit your needs? If you're struggling to find a measurement application that interests you, you can start with the construct addressed in this book. As a measurement student, you possess an underlying construct that will hopefully increase as you read, study, practice, and contribute to discussions and assignments. This construct could be labeled *assessment literacy* (Stiggins 1991).

Chapter 3

Testing Applications

> If my future were determined just by my performance on a standardized test, I wouldn't be here.
> — Michelle Obama

Standardized tests are often used to inform high-stakes decisions, including decisions that limit access to educational programs, career paths, and other valuable opportunities. Ideally, test scores supplement other relevant information in these decisions, from sources such as interviews, recommendations, observations, and portfolios of work. In many situations, standardized test scores are considered essential because they provide a common objective measure of knowledge, achievement, and skills.

Unfortunately, standardized test scores can be misused when the information they provide is inaccurate or when they have an undue influence on these decisions. A number of studies and reports refer to bias in standardized tests and an over-reliance on scores in high-stakes decision making (e.g., Santelices and Wilson 2010). As noted in the quote above, some people do not perform well on standardized tests, despite having what it takes to succeed.

This chapter gives an overview of the various different types of tests, including standardized and unstandardized ones, that are used to support decision making in education and psychology. Chapter 2 referred to a test's purpose as the starting point for determining its quality or effectiveness. In this chapter we'll compare types of tests in terms of their purposes. We'll examine how these purposes are associated with certain features in a test, and we'll look again at how the quality or validity of a test score can impact the effectiveness of score interpretations.

3.1 Tests and decision making

As mentioned in Chapter 2, tests are designed for different purposes, for example, to inform decisions impacting accountability, admissions, employment, graduation, licensing, and placement (see Table 2.1). Test results can also impact decisions regarding treatment in mental health and counseling settings, interventions in special education, and policy and legal issues.

3.1.1 Educational decisions

Educational tests support decision making in both *low-stakes* and *high-stakes* situations. These terms refer to the consequences and impact of test results for those involved, where low-stakes testing has low impact and high-stakes can have large or lasting impact. Low-stakes tests address decision making for instructional planning and student placement. The myIGDI testing program described in Chapter 2 involves low-stakes decisions regarding the selection of instructional interventions to support student learning. PISA could be

considered a low-stakes test, at least for the student, as scores do not impact decisions made at the student level. Teacher-made classroom assessments and other tools for measuring student growth are also considered low-stakes tests.

Let's look at one of the oldest and best known standardized tests as an example of high-stakes decision making. Development of the first college admissions test began in the late 1800s when a group of universities in the US came together to form the College Entrance Examination Board (now called the College Board). In 1901, this group administered the original version of what would later be called the SAT. This original version consisted only of essay questions in select subject areas such as Latin, history, and physics. Some of these questions resemble ones we might find in standardized tests today. For example, from the physics section:

> A steamer is moving eastward at the rate of 240 meters per minute. A man runs northward across her deck at the rate of 180 meters per minute. Show by a drawing his actual path and compute his actual velocity in centimeters per second.

The original test was intended only for limited use within the College Board. However, in 1926, the SAT was redesigned to appeal to institutions across the US. The 1926 version included nine content areas: analogies, antonyms, arithmetic, artificial classification, language, logical inference, number series, reading, and word definitions. It was based almost entirely on multiple-choice questions. For additional details, see sat.collegeboard.org.

The College Board notes that the SAT was initially intended to be a universal measure of preparation for college. It was the first test to be utilized across multiple institutions, and it provided the only common metric with which to evaluate applicants. In this way, the authors assert it helped to level the playing field for applicants of diverse socio-economic backgrounds, to "democratize access to higher education for all students" (College Board 2012, 3). For example, those who may have otherwise received preferential treatment because of connections with alumni could be compared directly to applicants without legacy connections.

Since it was formally standardized in 1926, the SAT has become the most widely used college admissions exam, with over 1.5 million administrations annually (as of 2015). The test itself has changed substantially over the years; however, its stated purpose remains the same (College Board 2012):

> Today the SAT serves as both a measure of students' college readiness and as a valid and reliable predictor of college outcomes. Developed with input from teachers and educators, the SAT covers core content areas presented as part of a rigorous high school curriculum and deemed critical for success in college ? critical reading, mathematics and writing. The SAT measures knowledge and skills that are considered important by both high school teachers and college faculty.

As we'll see in Chapter 9, test developers such as the College Board are responsible for backing up claims such as these with validity evidence. However, in the end, colleges must evaluate whether or not these claims are met, and, if not, whether admission decisions can be made without a standardized test. Colleges are responsible for choosing how much weight test scores have in admissions decisions, and whether or not minimum cutoff scores are used.

Criticism of the SAT, primarily regarding perceived bias in item content and scoring (e.g., Santelices and Wilson 2010), has led a number of colleges to drop it as an admissions requirement. These colleges base admissions decisions on other information, such as in-person interviews with applicants (Miller and Stassun 2014).

A 2009 survey of 246 colleges in the US found that 73% used the SAT in admissions decisions (Briggs 2009). Of those colleges using the SAT, 78% reported using scores holistically, that is, as supporting information within a portfolio of work contained in an application. On the other hand, 31% reported using SAT scores for quantitative rankings of applicants, and 21% reported further to have defined cut-off scores below which applicants would be disqualified for admission.

Another controversial high-stakes use of educational testing involves accountability decisions within the US education system. For reviews of these issues in the context of the No Child Left Behind Act of 2001 (NCLB;

Table 3.1: Original Clinical Scales of the MMPI

Scale	What is measured	Items
Hypochondriasis	Concern with bodily symptoms	32
Depression	Depressive symptoms	57
Hysteria	Awareness of problems and vulnerabilities	60
Psychopathic Deviate	Conflict, struggle, anger, respect for rules	50
Masculinity/Femininity	Stereotypical interests and behaviors	56
Paranoia	Level of trust, suspiciousness, sensitivity	40
Psychasthenia	Worry, anxiety, tension, doubts, obsessiveness	48
Schizophrenia	Odd thinking and social alienation	78
Hypomania	Level of excitability	46
Social Introversion	People orientation	69

Public Law 107-110), see Hursh (2005) and Linn, Baker, and Betebenner (2002). Abedi (2004) discusses validity implications of NCLB for English language learners.

3.1.2 Psychological decisions

As an example of decision making in the context of psychology, we'll look at one of the most widely used standardized personality tests. In the 1930s and 1940s, two researchers at the University of Minnesota pioneered an empirical and atheoretical method for developing personality and pathology scales. This method involved administering hundreds of short items to patients with known diagnoses. Items measuring specific personality traits and pathologies were identified based on consistent patterns of response for patients with those traits and pathologies. For example, patients diagnosed with depression responded in similar ways across certain items. Regardless of the content of the items (hence the atheoretical nature of the method), if depressed individuals responded in consistent ways, the items were assumed to measure depression. Table 3.1 contains names and descriptions for the original clinical scales of the Minnesota Multiphasic Personality Inventory (MMPI; for details, see wikipedia.org).

The MMPI and other personality and psychopathology measures are used to support decisions in a variety of clinical settings, including diagnosis and treatment and therapy planning and evaluation. They are also used in personnel selection, where certain personality traits have been shown to correlate strongly with certain aspects of job performance. However, because of its emphasis on pathology and implications regarding mental illness, the MMPI can only be used in the US in employment decisions for high-risk or security-related positions such as police officer and fire fighter. Measures such as the MMPI can also be used in forensics, criminal investigations, and in court (Pope, Butcher, and Seelen 2006). Given their impact on mental health outcomes, career choices, and legal proceedings, these could all be considered high-stakes decisions.

3.2 Test types and features

Over the past hundred years numerous terms have been introduced to distinguish different tests from one another based on certain features of the tests themselves and what they are intended to measure. Educational tests have been described as measuring constructs that are the focus of instruction and learning, whereas psychological tests measure constructs that are not the focus of teaching and learning. Thus, educational and psychological tests differ in the constructs they measure. Related distinctions are made between cognitive and affective tests, and achievement and aptitude tests. Other distinctions include summative versus formative, mastery versus growth, and knowledge versus performance.

3.2.1 Cognitive and affective

Cognitive testing refers to testing where the construct of interest is a cognitive ability. Cognition includes mental processes related to knowledge, comprehension, language acquisition and production, memory, reasoning, problem solving, and decision making. Intelligence tests, achievement tests, and aptitude tests are all considered cognitive tests because they assess constructs involving cognitive abilities and processing. Other examples of cognitive tests include educational admissions tests and licensure and certification tests.

In affective testing, the construct of interest relates to psychological attributes not involving mental processing. Affective constructs include personality traits, psychopathologies, interests, attitudes, and perceptions, as discussed below. Note that cognitive measures are often used in both educational and psychological settings. However, affective measures are more common in psychological settings.

3.2.2 Achievement and aptitude

Achievement and aptitude describe two related forms of cognitive tests. Both types of tests measure similar cognitive abilities and processes, but typically for slightly different purposes. Achievement tests are intended to describe learning and growth, for example, in order to identify how much content students have mastered in a unit of study. Accountability tests required by NCLB are achievement tests built on the educational curricula for states in the US. State curricula are divided into what are called *learning standards* or *curriculum standards*. These standards operationalize the curriculum in terms of what proficient students should know or be able to do.

In contrast to achievement tests, aptitude tests are typically intended to measure cognitive abilities that are predictive of future performance. This future performance could be measured in terms of the same or a similar cognitive ability, or in terms of performance on other constructs and in other situations. For example, intelligence or IQ tests are used to identify individuals with developmental and learning disabilities and to predict job performance (e.g., Carter 2002). The Stanford Binet Intelligence Scales, originally developed in the early 1900s, were the first standardized aptitude tests. Others include the Wechsler Scales and the Woodcock-Johnson Psycho-Educational Battery.

Intelligence tests and related measures of cognitive functioning have traditionally been used in the US to identify students in need of special education services. However, an over-reliance on test scores in these screening and placement decisions has led to criticism of the practice. A federal report (US Department of Education 2002, 25) concluded that,

> Eliminating IQ tests from the identification process would help shift the emphasis in special education away from the current focus, which is on determining whether students are eligible for services, towards providing students the interventions they need to successfully learn.

The myIGDI testing program discussed in Chapter 2 is one example of a special education screening and placement measure that focuses on intervention to support student learning. Note that the emphasis on student learning in this context has resulted in tests that measure both aptitude and achievement, as they predict future performance while also describing student learning. Thus, tests for screening and placement of students with disabilities are now designed for multiple purposes.

Tests that distinctly measure either achievement or aptitude usually differ in content and scope as well as purpose. Achievement tests are designed around a well defined content domain using a *test outline* (discussed further in Chapter **??**(development)). The test outline presents the content areas and learning standards or objectives needed to represent the underlying construct. An achievement test includes test questions that map directly to these content areas and objectives. As a result, a given question on an achievement test should have high *face validity*, that is, it should be clear what content the question is intended to measure. Furthermore, a correct response to such a question should depend directly on an individual's learning of that content.

On the other hand, aptitude tests, which are not intended to assess learning or mastery of a specific content domain, need not be restricted to specific content. They are still designed using a test outline. However, this outline captures the abilities and skills that are most related to or predictive of the construct, rather than content areas and learning objectives. Aptitude tests typically measure abilities and skills that generalize to other constructs and outcomes. As a result, the content of an aptitude question is not as important as the cognitive reasoning or processes used to respond correctly. Aptitude questions may not have high face validity, that is, it may not be clear what they are intended to measure and how the resulting scores will be used.

3.2.3 Other distinctions

In the 1970s and 1980s, researchers in the areas of education and school psychology began to highlight a need for educational assessments tied directly to the curriculum and instructional objectives of the classroom in which they would be used. The term *performance assessment* was introduced to describe a more authentic form of assessment requiring students to demonstrate skills and competencies relevant to key outcomes of instruction (Mehrens 1992, Stiggins (1987)). Various types of performance assessment were developed specifically as alternatives to the summative tests that were then often created outside the classroom. For example, curriculum-based measurement (CBM), developed in the early 1980s (Deno 1985), involved brief, performance-based measures used to monitor student progress in core subject areas such as reading, writing, and math. Reading CBM assessed students' oral reading fluency in the basal texts for the course; the content of the reading assessments came directly from the readings that students would encounter during the academic year. These assessments produced scores that could be used to model growth, and predict performance on end-of-year tests (Deno et al. 2001, Fuchs and Fuchs (1999)).

Although CBM and other forms of performance assessment remain popular today, the term *formative assessment* is now commonly used as a general label for the classroom-focused alternative to the traditional *summative* or end-of-year test. The main distinction between formative and summative is in the purpose or intended use of the resulting test score. Formative assessments are described as measuring incrementally, where the purpose is to directly encourage student growth (Black and Wiliam 1998). They can be spread across multiple administrations, or used in conjunction with other versions of an assessment, so as to monitor and promote progress. Thus, formative assessments are designed to inform teaching and form learning. They seek to answer the question, "how are we doing?" Wiliam and Black (1996) further assert that in order to be formative, an assessment must provide information that is used to address a need or a gap in the student's understanding; in other words, beyond an intent, there must be an attempt to apply the results. Note that it is less appropriate to label a test as formative, and preferable to instead label the process or use of scores as formative.

On the other hand, summative assessments, or assessments used summatively, measures conclusively, usually at a single time point, where the intention is to describe current status. Summative assessment encourages growth only indirectly. In contrast to formative, it is designed to sum or summarize, to answer the question, "how did we do?" Cizek (2010) describes summative as an assessment that is administered at the end of an instructional unit the purpose of which is "primarily to categorize the performance of a student or system." Wiliam and Black (1996) go further by saying that summative assessments cannot be formative, by definition.

Despite debate over what specifically constitutes a formative assessment (e.g., Bennett 2011), numerous studies have documented at least some positive impact resulting from the use of assessments that inform instruction during the school year (Black and Wiliam 1998). Formative assessments have become a key component in many educational assessment systems (Militello, Schweid, and Sireci 2010).

3.2.4 Summary

Although a variety of terms are available for describing educational and psychological tests, many tests can be described in different ways. The myIGDI measures was mentioned as an example of how achievement and aptitude tests can be difficult to distinguish from one another. Summative and formative tests often overlap as well, where a test can be used to both summarize and inform learning. In the end, the purpose of the test

should be the main source of information for determining what type of test you're dealing with and what that test is intended to do.

3.3 Finding test information

Information for commercially available tests can usually be found by searching online, with test publishers sharing summaries and technical documentation for their tests. Repositories of test information are also available online. ETS provides a searchable data base of over 25,000 tests at www.ets.org/test_link. A search for the term "creativity" returns 213 records, including the Adult Playfulness Scale, "a personality measure which assesses the degree to which an individual tends to define an activity in an imaginative, nonserious or metaphoric manner so as to enhance intrinsic enjoyment, involvement, and satisfaction," and the Fantasy Measure, where "Children complete stories in which the main character is a child under stress of failure." In addition to a title and abstract, the data base includes basic information on publication date and authorship, sometimes with links to publisher websites.

The Buros Center for Testing buros.org also publishes a comprehensive data base of educational and psychological measures. In addition to descriptive information, they include peer evaluations of the psychometric properties of tests in what is called the *Mental Measurements Yearbook*. Buros peer reviews are available through university library subscriptions, or can be accessed online for a fee.

3.4 Summary

This chapter provides an overview of how different types of tests are designed to inform a variety of decisions in education and psychology. For the most part, tests are designed merely to *inform* decision making processes, and test authors are often careful to clarify that no decision should be made based solely on test scores. Online data bases provide access to descriptive summaries and peer reviews of tests.

3.4.1 Learning objectives

1. Provide examples of how testing supports low-stakes and high-stakes decision making in education and psychology.
2. Describe the general purpose of aptitude testing and some common applications.
3. Identify the distinctive features of aptitude tests and the main benefits and limitations in using aptitude tests to inform decision-making.
4. Describe the general purpose of standardized achievement testing and some common applications.
5. Identify the distinctive features of standardized achievement tests and the main benefits and limitations in using standardized achievement tests to inform decision-making.
6. Compare and contrast different types of tests and test uses and identify examples of each, including summative, formative, mastery, and performance.

3.4.2 Exercises

1. Is it appropriate for colleges and graduate programs to have minimum cutoffs when reviewing standardized test scores in the admissions process? How would you recommend that scores from admissions tests be incorporated into admissions decisions?
2. Describe the challenges involved in using a single test for multiple purposes, such as to measure both achievement and aptitude, or both status and growth, or both formative and summative information.
3. Would you describe admissions tests like the SAT and GRE as aptitude or achievement tests? Explain your reasoning.

4. Conduct an online search for information on one of the tests referenced in this chapter or in Chapter 2. Look for details on the publication date, authors, and accessibility of the test. From the available information, summarize the test using the terms presented in this chapter.

Chapter 4

Test Development

> A careful review of any testing program will identify poorly worded test items, written by persons with minimal training and inadequate insights into their audience. We need to do much more work to produce quality test items.
> — Mark Reckase, *2009 NCME Presidential Address*

Good items are the building blocks of good tests, and the validity of test scores can hinge on the quality of individual test items. Unfortunately, test makers, both in low-stakes and high-stakes settings, often presume that good items are easy to come by. As noted above, item writing is often not given the attention it deserves. Research shows that effective item writing is a challenging process, and even the highest-stakes of tests include poorly written items (Haladyna and Rodriguez 2013).

This chapter summarizes the main stages of both cognitive and noncognitive test construction, from conception to development, and the main features of commonly used item formats. Cognitive test development is covered first. The cognitive item writing guidelines presented in Haladyna, Downing, and Rodriguez (2002) are summarized, along with the main concepts from the style guides used by testing companies. Next, noncognitive personality test development is discussed. Noncognitive item writing guidelines are reviewed, along with strategies for reducing the impact of response sets.

4.1 Validity and test purpose

As is the case throughout this book, we will begin this chapter on test development with a review of validity and test purpose. Recall from Chapters 1 through 3 that validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of a test. In other words, validity indexes the extent to which test scores can be used for their intended purpose. These are generic definitions of validity that apply to any type of educational or psychological measure.

In this chapter we focus first on cognitive tests, where the purpose of the test is to produce scores that can inform decision making in terms of aptitude and achievement, presumably of students. So, we need to define validity in terms of these more specific test uses. Let's use a midterm exam from an introductory measurement course as an example. We could say that validity refers to the degree to which the content coverage of the exam (as specified in the outline, based on the learning objectives) supports the use of scores as a measure of student learning for topics covered in the first part of the course. Based on this definition of validity, what would you say is the purpose of the exam? Note how test purpose and validity are closely linked.

Construction of a valid test begins with a test purpose. You need to be able to identify the three components of a test purpose, both when presented with a well defined purpose, and when presented with a general description of a test. Later in the course you'll be reviewing information from test reviews and technical

documentation which may or may not include clear definitions of test purpose. You'll have to take the available information and identify, to the best of your ability, what the test purpose is. Here are some verbs to look for: assess, test, and measure (obviously), but also describe, select, identify, examine, and gauge, to name a few.

Do your best to distill the lengthy description below into a one-sentence test purpose. This should be pretty straightforward. The information is all there. This description comes from the technical manual for the 2011 California Standards Test, which is part of what was formerly known as the Standardized Testing and Reporting (STAR) program for the state of California (see www.cde.ca.gov). These are more recent forms of the state tests that I took in school in the 1980s.

> California Standards Tests (CSTs) are produced for California public schools to assess the California content standards for ELA, mathematics, history/social science, and science in grades two through eleven.
> A total of 38 CSTs form the cornerstone of the STAR program. The CSTs, given in English, are designed to show how well students in grades two through eleven are performing with respect to California's content standards. These standards describe what students should know and be able to do at each grade level in selected content areas.
> CSTs carry the most weight in school and district Academic Performance Index (API) calculations. In addition, the CSTs for ELA and mathematics (grades two through eight) are used in determining Adequate Yearly Progress (AYP), which is used to meet the requirement of the federal Elementary and Secondary Education Act (ESEA) that all students score at the proficient level or above by 2014.

You should have come up with something like this for the CST test purpose: the CST measures ELA, mathematics, history/social science, and science for students in grades two through eleven to show how well they are performing with respect to California's content standards, and to help determine AYP.

4.2 Learning objectives

To keep the description of the CSTs brief, I omitted details about the content standards. California, like all other states, has detailed standards or *learning objectives* defining what content/skills/knowledge/information/etc. must be covered by schools in the core subject areas. The standards specify what a student should know and be able to do after their educational experience. They establish the overarching goals for teaching and learning. Teachers, schools, and districts, to some extent, are then free to determine the best way to teach the standards.

In this chapter, educational standards are presented as a form of learning objective, which identify the goals or purposes of instruction. Here's a simplified example of a learning objective for this chapter: write and critique test items. This objective is extremely simple and brief. Can you describe why it would be challenging to assess proficiency or competency for this objective? How could the objective be changed to make it easier to assess?

Learning objectives that are broadly or vaguely defined lead to low-quality, unfocused test questions. The simple item-writing objective above does not include any qualifiers specifying how it is achieved or obtained or appropriately demonstrated. In state education systems, the standards are very detailed. For example, Nebraska defines more than 75 science standards in grade 11 alone (for details, see www.education.ne.gov/academicstandards). From the Nebraska State Standards, Grade 11 Abilities to do Scientific Inquiry:

> 1. Design and conduct investigations that lead to the use of logic and evidence in the formulation of scientific explanations and models.

2. Formulate a testable hypothesis supported by prior knowledge to guide an investigation.

3. Design and conduct logical and sequential scientific investigations with repeated trials and apply findings to new investigations.

Note that these standards reflect specific things students should be able to do, and some conditions for how students can do these things well. Such specific wording greatly simplifies the item writing process because it clarifies precisely the knowledge, skills, and abilities that should be measured.

Notice also that the simplest way to assess the first science objective listed above would be to simply ask students to design and conduct an investigation that leads to the use of logic and evidence in the formulation of scientific explanations and models. The standard itself is almost worded as a test question. This is often the case with well-written standards. Unfortunately, the standardized testing process includes constraints, like time limits, that make it difficult or impossible to assess standards so directly. Designing and conducting an experiment requires time and resources. Instead, in a test we might refer students to an example of an experiment and ask them to identify correct or incorrect procedures, or we might ask students to use logic when making conclusions from given experimental results. In this way, we use individual test questions to indirectly assess different components of a given standard.

4.3 Features of cognitive items

4.3.1 Depth of knowledge

In addition to being written to specific standards or learning objectives, cognitive test items are also written to assess at a specific depth of knowledge (DOK). The DOK of an item indicates its level of complexity in terms of the knowledge and skills required to obtain a correct response. Bloom and Krathwohl (1956) presented the original framework for categorizing depth of knowledge in cognitive assessments. However, the majority of achievement tests nowadays use some version of the DOK categories presented by Webb (2002). These DOK differ somewhat by content area, but are roughly defined in levels of increasing complexity as 1) recall and reproduction, 2) skills and concepts, 3) strategic thinking, and 4) extended thinking.

These simple DOK categories can be modified to meet the needs of a particular testing program. For example, here is the description of Level 1 DOK used in writing items for the standardized science tests in Nebraska:

> *Level 1 Recall and Reproduction* requires recall of information, such as a fact, definition, term, or a simple procedure, as well as performing a simple science process or procedure. Level 1 only requires students to demonstrate a rote response, use a well-known formula, follow a set procedure (like a recipe), or perform a clearly defined series of steps. A "simple" procedure is well-defined and typically involves only one-step.
> Verbs such as "identify," "recall," "recognize," "use," "calculate," and "measure" generally represent cognitive work at the recall and reproduction level. Simple word problems that can be directly translated into and solved by a formula are considered Level 1. Verbs such as "describe" and "explain" could be classified at different DOK levels, depending on the complexity of what is to be described and explained.

DOK descriptions such as this are used to categorize items in the item writing process, and thereby ensure that the items together support the overall DOK required in the purpose of the test. Typically, higher DOK is preferable. However, lower levels of DOK are sometimes required to assess certain objectives, for example, ones that require students to recall or reproduce definitions, steps, procedures, or other key information. Furthermore, constraints on time and resources within the standardized testing process often make it impossible to assess the highest level of DOK, which requires extended thinking and complex cognitive demands.

4.3.2 Item types

Cognitive test items come in a variety of types that differ in how material is presented to the test taker, and how responses are then collected. Most cognitive test questions begin with a *stem* or question statement. The stem should present all of the information needed to answer an item correctly. Optionally, an item, or a set of items, can refer to a figure, table, text, or other presentation of information that students must read, interpret, and sometimes interact with before responding. These are referred to as *prompts*, and they are typically presented separately from the item stem.

Selected-response (SR) items collect responses from test takers using two or more *response options*. The classic multiple-choice question is an SR item with a stem ending in a question or some direction that the test taker must choose one or more of options.

> In general, what is the optimal number of response options in a cognitive multiple-choice test question?
> A. Two
> B. Three
> C. Four

Research shows that the optimal number of questions in a multiple-choice item is three (Rodriguez 2005). Tradition leads many item-writers consistently to use four options. However, a feasible fourth option is often difficult to write, leading test takers to easily discount it, and thus making it unnecessary.

A variety of SR item types are available. More popular types include: 1. true/false, where test takers simply indicate whether a statement is true or false; 2. multiple correct or select all that apply, where more than one option can be selected as correct; 3. multiple true/false, a simplified form of multiple correct where options consist of binary factual statements (true/false) and are preceded by a prompt or question statement linking them together in some way; 4. matching, where test takers select for each option in one list the correct match from a second list; 5. complex multiple-choice, where different combinations of response options can be selected as correct, resembling a constrained form of multiple correct (e.g., options A and B, A and C, or all of the above); and 6. evidence-based, which can be any form of SR item where a follow-up question requires test takers to select an option justifying their response to the original item.

Evidence-based questions are becoming more popular in standardized achievement testing, as, test makers claim, they can be used to assess more complex reasoning. This is achieved via the nesting of content from one question inside the follow-up. Here's a simple evidence-based question on DOK.

> Part I. In a constructed-response science question, students are given a hypothesis and must then describe with an essay an experiment that could be used to test the hypothesis. In their description they must identify the key components of the experiment and justify the importance of each component in testing the hypothesis.
> What depth of knowledge level does this science question assess?
> A. 1
> B. 2
> C. 3
> D. 4

> Part II. What task from the science question in Part I best supports the answer for Part I?
> A. Describe an experiment.
> B. Identify the key components of an experiment.
> C. Justify the importance of each component.

A constructed-response (CR) item does not present options to the test taker. As the name implies, a response must be constructed. Constructed-response items include short-answer, fill-in-the-blank, graphing,

manipulation of information, and essays. Standardized performance assessments, for example, reading fluency measures, can also be considered CR tasks.

The science question itself within Part I of the evidence-based DOK item above is an example of a simple essay question. Note that this science question could easily be converted to a SR question with multiple correct answers, where various components of an experiment, some correct and some incorrect, could be presented to the student. Parts I and II from the evidence-based DOK question could also easily be converted to a single CR question, where test takers identify the correct DOK for the science question, and then provide their own supporting evidence.

There are some key advantages and disadvantages to multiple-choice or SR items and CR items. In terms of advantages, SR items are typically easy to administer and score, and are more objective and reliable than CR items. They are also more efficient, and can be used to cover more test content in a shorter period of time. Finally, SR items can provide useful diagnostic information about specific misconceptions that test takers might have.

Although they are more efficient and economical, SR items are more difficult to write well, they tend to focus on lower-order thinking and skills, such as recall and reproduction, and they are more susceptible to test-wiseness and guessing. Constructed-response items address each of these issues. They are easier to write, especially for higher-level thinking, and they eliminate the potential for simple guessing.

The main benefit of CR questions is they can be used to test more practical, authentic, and realistic forms of performance and tasks, including creative skills and abilities. The downside is that these types of performance and tasks require time to demonstrate and are then complex and costly to score.

4.3.3 Performance assessment

CR items can be considered a subset of what is commonly referred to as performance assessment. Performance assessment focuses on the measurement of *processes and products*. Performance assessments require individuals to generate a response, demonstrate a skill, or perform a task. The key feature of a performance assessment is the requirement that an individual do something, i.e., perform in some way, to obtain a score.

In addition to CR items, some less common examples of performance assessment are objective structured clinical examinations (OSCE) and portfolio or body-of-work assessments. OSCE are often used in medical fields to assess a student's ability to perform certain tasks in a real-life scenario. For example, the real-life scenario might involve an actor pretending to have some ailment, and the student is rated on how well they respond in terms of their diagnosis and the treatment they prescribe. Portfolios often contain a sample of products that are used to represent an individual's accomplishments in a given field or area of expertise. For example, professors might be required to record publications, awards, and recognitions over the course of multiple years. The collection of work, as a portfolio, could then be used in an evaluation for promotion.

The GRE analytical writing section is a familiar form of performance assessment. This section of the GRE now includes two 30-minute essay questions. Below is an example of one essay prompt from the GRE website. The question is open-ended, requiring students to develop and support a position on the idea that reliance on technology will deteriorate our ability to think for ourselves.

> As people rely more and more on technology to solve problems, the ability of humans to think for themselves will surely deteriorate.
> Discuss the extent to which you agree or disagree with the statement and explain your reasoning for the position you take. In developing and supporting your position, you should consider ways in which the statement might or might not hold true and explain how these considerations shape your position.

Take a minute to think about the construct that this question is intended to measure. "Analytical writing" is a good start, but you should try to be more specific. What knowledge, skills, and abilities are required to do well on the question? What are the benefits of performance assessments such as these?

As with CR items, the main benefit of performance assessment is that it is considered more authentic than traditional mastery assessment, because it allows us to assess directly what we're trying to measure. For example, the GRE essay question above measures the construct of analytical writing by asking examinees to write analytically. This improves the validity of the resulting score as an indicator of the construct itself. Performance assessments, because they require individuals to generate their own response, rather than select a response from a list of options, are also able to assess higher order thinking and skills like synthesis and evaluation. These skills are not easily assessed with simple selected-response questions.

Two of the drawbacks of performance assessments result from the fact that humans are involved in the scoring process. First, as noted above, performance assessments are less practical, because they require substantially more time and resources to develop and score. Second, the scoring process becomes subjective, to some extent. A third drawback to performance assessment is that content, though it may be assessed deeply, for example, using more depth of knowledge, it is not assessed broadly.

4.3.4 Rubrics

Of these three drawbacks to performance assessment, subjectivity in scoring can be addressed so as to limit its negative effects. Subjectivity in scoring is reduced by using standardized scoring criteria within a *rubric*. More objective scores are achieved by training judges to correctly apply scoring rubrics. Scoring rubrics outline the possible scores that an individual could receive on an assessment, and the levels of performance that must be demonstrated for each score to be given. The scoring process is standardized when different judges can consistently apply the correct scores to the corresponding performance levels.

Here is an example from the GRE analytical writing rubric, available online. GRE essays are scored on a scale from 1 to 6, and the description below is for one of the six possible score categories. The first sentence describes the overall quality of the essay as demonstrating "some competence" but also being "obviously flawed." Then, a list of common features for this category of essay is provided.

> Response demonstrates some competence in addressing the specific task directions, in analyzing the issue and in conveying meaning, but is obviously flawed. A typical response in this category exhibits one or more of the following characteristics:
> 1. is vague or limited in addressing the specific task directions and in presenting or developing a position on the issue or both,
> 2. is weak in the use of relevant reasons or examples or relies largely on unsupported claims,
> 3. is limited in focus and/or organization,
> 4. has problems in language and sentence structure that result in a lack of clarity,
> 5. contains occasional major errors or frequent minor errors in grammar, usage or mechanics that can interfere with meaning.

After reading through this portion of the rubric, try to guess which of the six score categories it applies to.

Rubrics are typically described as either analytic or holistic. An analytic rubric breaks down a response into characteristics or components, each of which can be present or correct to different degrees. For example, an essay response may be scored based on its introduction, body, and conclusion. A required feature of the introduction might be a clear thesis statement. Rubrics that analyze components of a response are more time consuming to develop and use. However, they can provide a more detailed evaluation than rubrics that do not analyze the components of a response, that is, holistic rubrics. A holistic rubric provides a single score based on an overall evaluation of a response. Holistic rubrics are simpler to develop and use. However, they do not provide detailed information about the strengths or weaknesses in a response.

Figure 4.1 contains a prompt used in the PISA 2009 reading test, with items `r414q02`, `r414q11`, `r414q06`, and `r414q09`. The full text is also included in Appendix @ref(#appendixa). This prompt presented students with information on whether cell phones are dangerous, along with recommendations for safe cell phone use. Some key points were also summarized in the left margin of the prompt. Question `PISA09$r414q06` required that students interpret information from a specific part of the prompt. It read, "Look at Point 3 in the No

Table 4.1: Simple Example Test Outline

Scale	Learning Objective	DOK	Items
Reading	Define key vocabulary	1	12
	Select the most appropriate word	2	10
Writing	Write a short story	3	1
	Evaluate an argument and construct a rebuttal	4	2
Math	Solve equations with two unknowns	4	8
	Run a linear regression and interpret the output	4	5

column of the table. In this context, what might one of these 'other factors' be? Give a reason for your answer."

The rubric for scoring this CR item is relatively simple. It would be considered a holistic rubric, since a breakdown of points by components of the response was not provided. Correct or incorrect scores were given based on short scoring guidelines. If a student response demonstrated partial understanding of the issue, the rater would have to use their best judgement when designated it as correct or incorrect. As will be discussed in Chapter 5.3.1, ambiguity and subjectivity in the rating process introduce measurement error into scores, which decreases reliability.

Correct
Answers which identify a factor in modern lifestyles that could be related to fatigue, headaches, or loss of concentration. The explanation may be self-evident, or explicitly stated.
Incorrect
Answers which give an insufficient or vague response.
Fatigue. [Repeats information in the text.]
Tiredness. [Repeats information in the text.]
Answers which show inaccurate comprehension of the material or are implausible or irrelevant.

4.3.5 Test outline

In its simplest form, a test outline is a table that summarizes how the items in a test are distributed in terms of key features such as content areas or subscales (e.g., quantitative reasoning, verbal reasoning), standards or objectives, item types, and depth of knowledge. Table 4.1 contains a simple example for a cognitive test with three content areas.

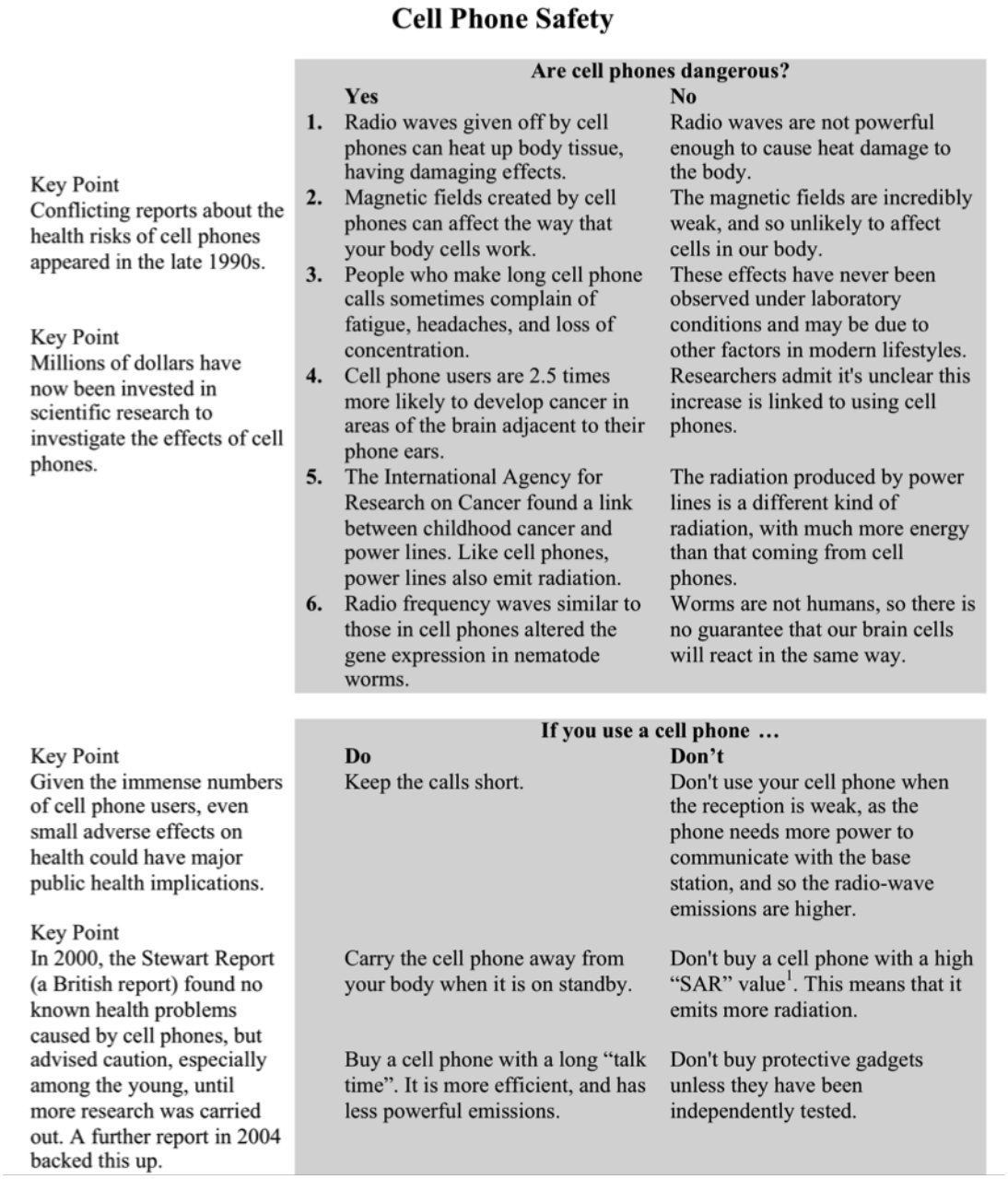
A test outline is used to ensure that a test measures the content areas captured by the tested construct, and that these content areas are measured in the appropriate ways. Notice that in Table 4.1 we're only assessing reading using the first two levels of DOK. Perhaps scores from this test will be used to identify struggling readers. The test purpose would likely need to include some mention of reading comprehension, which would then be assessed at a deeper level of knowledge.

The learning objectives in Table 4.1 are intentionally left vague. How can they be improved to make these content areas more testable? Consider how qualifying information could be included in these objectives to clarify what would constitute high-quality performance or responses.

4.4 Cognitive item writing

The item writing guidelines presented in Haladyna, Downing, and Rodriguez (2002) are paraphrased here for reference. The guidelines are grouped into ones addressing content concerns, formatting concerns, style concerns, issues in writing the stem, and issues in writing the response options.

Content concerns



Cell Phone Safety

	Are cell phones dangerous?	
	Yes	**No**
Key Point Conflicting reports about the health risks of cell phones appeared in the late 1990s.	**1.** Radio waves given off by cell phones can heat up body tissue, having damaging effects.	Radio waves are not powerful enough to cause heat damage to the body.
	2. Magnetic fields created by cell phones can affect the way that your body cells work.	The magnetic fields are incredibly weak, and so unlikely to affect cells in our body.
Key Point Millions of dollars have now been invested in scientific research to investigate the effects of cell phones.	**3.** People who make long cell phone calls sometimes complain of fatigue, headaches, and loss of concentration.	These effects have never been observed under laboratory conditions and may be due to other factors in modern lifestyles.
	4. Cell phone users are 2.5 times more likely to develop cancer in areas of the brain adjacent to their phone ears.	Researchers admit it's unclear this increase is linked to using cell phones.
	5. The International Agency for Research on Cancer found a link between childhood cancer and power lines. Like cell phones, power lines also emit radiation.	The radiation produced by power lines is a different kind of radiation, with much more energy than that coming from cell phones.
	6. Radio frequency waves similar to those in cell phones altered the gene expression in nematode worms.	Worms are not humans, so there is no guarantee that our brain cells will react in the same way.

	If you use a cell phone …	
	Do	**Don't**
Key Point Given the immense numbers of cell phone users, even small adverse effects on health could have major public health implications.	Keep the calls short.	Don't use your cell phone when the reception is weak, as the phone needs more power to communicate with the base station, and so the radio-wave emissions are higher.
Key Point In 2000, the Stewart Report (a British report) found no known health problems caused by cell phones, but advised caution, especially among the young, until more research was carried out. A further report in 2004 backed this up.	Carry the cell phone away from your body when it is on standby.	Don't buy a cell phone with a high "SAR" value[1]. This means that it emits more radiation.
	Buy a cell phone with a long "talk time". It is more efficient, and has less powerful emissions.	Don't buy protective gadgets unless they have been independently tested.

Figure 4.1: PISA 2009 prompt for reading items r414q02, r414q11, r414q06, and r414q09.

1. Every item should reflect specific content and a single specific mental behavior, as called for in test specifications (two-way grid, test outline).
2. Base each item on important content to learn; avoid trivial content.
3. Use novel material to test higher level learning. Paraphrase textbook language or language used during instruction when used in a test item to avoid testing for simply recall.
4. Keep the content of each item independent from content of other items on the test.
5. Avoid over specific and over general content when writing multiple-choice (MC) items.
6. Avoid opinion-based items.
7. Avoid trick items.
8. Keep vocabulary simple for the group of students being tested.

Formatting concerns

9. Use the question, completion, and best answer versions of the conventional MC, the alternate choice, true-false, multiple true-false, matching, and the context-dependent item and item set formats, but AVOID the complex MC (Type K) format.
10. Format the item vertically instead of horizontally.

Style concerns

11. Edit and proof items.
12. Use correct grammar, punctuation, capitalization, and spelling.
13. Minimize the amount of reading in each item.

Writing the stem

14. Ensure that the directions in the stem are very clear.
15. Include the central idea in the stem instead of the choices.
16. Avoid window dressing (excessive verbiage).
17. Word the stem positively, avoid negatives such as NOT or EXCEPT. If negative words are used, use the word cautiously and always ensure that the word appears capitalized and boldface.

Writing the choices

18. Develop as many effective choices as you can, but research suggests three is adequate.
19. Make sure that only one of these choices is the right answer.
20. Vary the location of the right answer according to the number of choices.
21. Place choices in logical or numerical order.
22. Keep choices independent; choices should not be overlapping.
23. Keep choices homogeneous in content and grammatical structure.
24. Keep the length of choices about equal.
25. None-of-the-above should be used carefully.
26. Avoid All-of-the-above.
27. Phrase choices positively; avoid negatives such as NOT.
28. Avoid giving clues to the right answer, such as
    a. Specific determiners including always, never, completely, and absolutely.
    b. Clang associations, choices identical to or resembling words in the stem.
    c. Grammatical inconsistencies that cue the test-taker to the correct choice.
    d. Conspicuous correct choice.
    e. Pairs or triplets of options that clue the test-taker to the correct choice.
    f. Blatantly absurd, ridiculous options.
29. Make all distractors plausible.
30. Use typical errors of students to write your distractors.
31. Use humor if it is compatible with the teacher and the learning environment.

4.4.1 Construct irrelevant variance

Rather than review each item writing guideline, we'll just summarize the main theme that they all address. This theme has to do with the intended construct that a test is measuring. Each guideline targets a different source of what is referred to as construct irrelevant variance that is introduced in the testing process.

For example, consider guideline 8, which recommends that we "keep vocabulary simple for the group of students being tested." When vocabulary becomes unnecessarily complex, we end up testing vocabulary knowledge and related constructs in addition to our target construct. The complexity of the vocabulary should be appropriate for the audience and should not interfere with the construct being assessed. Otherwise, it introduces variability in scores that is irrelevant or confounding with respect to our construct.

Another simple example is guideline 17, which recommends that we "word the stem positively" and "avoid negatives such as NOT or EXCEPT." The use of negatives, and worse yet, double negatives, introduces a cognitive load into the testing process that may not be critical to the construct we want to assess.

4.5 Personality

> If personality is an unbroken series of successful gestures, then there was something gorgeous about him, some heightened sensitivity to the promise of life... an extraordinary gift for hope, a romantic readiness such as I have never found in any other person and which it is not likely I shall ever find again.
> — F. Scott Fitzgerald, *The Great Gatsby*

Personality is defined as the combination of qualities or attributes that make up a person's character. These qualities or attributes are sometimes referred to as traits, or enduring characteristics, as opposed to states, which are temporary or transitory characteristics. Together, these characteristics come together to define who we are.

In general, personality theories state that our personalities are manifested in themes or patterns in our preferences and behavior, our habits and responses to our experiences. Behaviors and responses are key in the measurement of personality. The general purpose of personality testing is to describe our personalities and other constructs related to them as they are operationalized through our responses to items on a test. As with any form of testing, this operationalization requires that we make an inference from our test scores to the underlying construct assumed to cause or precede them.

The term *affective* refers broadly to non-cognitive constructs. The vast majority of affective or non-cognitive constructs examined in research and in practice are personality characteristics, and the examples we will consider in this chapter all focus on some aspect of personality. Other related non-cognitive constructs include moods, attitudes, and preferences, which may or may not be used as indicators of specific personality traits.

4.6 Validity and test purpose

Validity and test purpose are once again at the start of the test construction process. As with cognitive test construction, the valid use of affective test scores requires a clearly articulated test purpose. This purpose tells us about the construct we intend to measure, for whom we measure it, and for what reasons. Item writing then directly supports this purpose.

Affective tests are used in a variety of contexts. For example, test results can support evaluations of the effectiveness of clinical or counseling interventions. They can also inform clinical diagnosis. Affective measures can also be used for research purposes, for example, to examine relationships between patterns of thought or behavior. See Chapter 3 for example applications in the areas of mental health and job placement.

The Myers-Briggs Type Indicator (MBTI), first mentioned in Chapter 2, is a popular but somewhat controversial personality test based on the work of Carl Jung, who is famous, in part, for his research on *psychological archetypes*. Jung's original archetypes were defined by the extraversion-introversion and perception-judgment dichotomies. For each of these dichotomies, Jung claimed that people tend to find themselves at one end, for example, extraversion, more than the other, introversion.

The MBTI seeks to measure combinations of these original archetypes with the additional dichotomies of sensing-intuition and thinking-feeling. It does do using simple questions such as the following (from Myers et al. 1998):

> I am most comfortable being
> * Spontaneous
> * A planner
>
> Change for me is
> * Difficult
> * Easy
>
> I prefer to work
> * Alone
> * In a team
>
> I consider myself to be
> * Social
> * Private

The main criticism of the MBTI is that there is insufficient evidence supporting its reliability and validity. The test is used widely in counseling settings and employment settings for personnel selection and professional development. However, these uses may not be validated. For example, Gardner and Martinko (1996) found that relationships between MBTI types and variables such as managerial effectiveness, which would provide validity evidence for the use of scores in this setting, were weak or not well described. Pittenger (2005) concluded that overuse of the MBTI, where support for it is lacking, may be due to the simplicity of the measure and the MBTI publisher's marketing strategy.

At this point you may be wondering, what does the MBTI actually claim to do? What is its intended purpose? Consider the following broad disclaimer provided by the publisher within the test manual:

> All types are equal. The purpose of taking the MBTI is to recognize your strengths and weaknesses as well as those of others. The MBTI was created in order to facilitate an understanding and appreciation of differences among human beings. No type is better than another.
> The Myers-Briggs Type Indicator does not measure ability, traits, or character. Unlike other personality assessments, the MBTI does not do any of the above. Carl Jung and Isabel Briggs-Myers believed that preferences are inborn while traits are not. Someone can improve upon a trait (e.g. working on their public speaking) but they cannot change their preference (e.g. preferring to work alone than with a group in general).
> Your type does not dictate who you are as a person. Ethical use of the MBTI is being able to discern and understand your results. However, your type does not truly represent who you are. You are your own person. Myers believed that all individuals are unique in their own way. Being assigned a type does not mean you are every little detail outlined in the description. You should make your own reasonable judgment and verify your own preferences.

Contrast this with the variety of potential uses described on the publisher's website www.myersbriggs.org. For example,

> The MBTI instrument is a popular training tool for professional development and organizational improvement in all kinds of organizations. MBTI results give people in training programs helpful type feedback about themselves and how they are different from others. In organizations and workplaces the Indicator is particularly useful with teams, for conflict management and performance improvement, for employee coaching, for management development, or for executive coaching.

Note that there are now multiple versions of the MBTI for different applications. One version, called Step III, is described as being

> Designed for anyone who wants to increase awareness about the specific and unique ways they use their type in making life choices. In-depth sessions with a trained Step III professional help the client gain insight necessary for becoming more effective in the natural use of their type. The MBTI Step III Interpretive Report is written directly to the client using "non-type" language. The counselor helps the client process the statements generated to address their current level of self-confidence, approach to difficulties, sources of enjoyment, and more. The Step III instrument and Interpretive Report are available for purchase and use only by professionals who have successfully completed the MBTI Step III Certification Program.

In the end, it seems that the purpose of the MBTI is simply to inform individuals about their profile of types. The publisher then claims that type scores can be used in employment or counseling settings, or by individuals, to inform "life choices" and help them "become more effective in the natural use of their type." Note that the goal is not to change types, or work on deficiencies, but to capitalize on strengths.

4.7 Noncognitive test construction

Personality tests like the MBTI, BDI (introduced in Chapter 2), and MMPI (introduced in Chapter 3) are developed using one of two approaches or strategies. These strategies serve the same purpose as the test outline in cognitive test construction, that is, they identify the underlying structure or framework for the test. The two strategies used with affective and personality tests are called *deductive* and *empirical*.

4.7.1 Deductive

Deductive approaches to test construction are based on some belief or theory about how a construct can be operationalized in terms of item responses. Ideally, an established theory guides this process. For example, the BDI is based on the theory that the cognitive symptoms of depression, or the manifestation of depression in terms of thought processes, *precedes* the affective symptoms of depression, that is, their manifestation in terms of feelings. This cognitive theory of depression lends itself to a depression inventory that measures individuals' thought processes. Thus, theory determines the content of test questions and the mechanism for response.

Theory can also guide the selection of subscales or content areas within an overarching construct. Continuing with the example of depression, research indicates that depression is evidenced by anxiousness, restlessness, irritability, and changes in eating habits. Each of these symptoms of depression could be represented by a subset of items within a depression inventory. Theory could also indicate that other symptoms, such as obsessive-compulsive tendencies, should not be represented in the test because they do not constitute an important content area within the construct.

With simple constructs that are more easily operationalized than constructs such as depression, logic may be used in place of a specific theory to frame the content of a test. For example, logic suggests that the presence of an eating disorder could be measured using questions about behaviors corresponding to the disorder. To measure anorexia (an emotional disorder characterized by an obsessive desire to lose weight by abstaining

from eating), we can simply measure for the presence of anorexic behaviors. Note that the simplicity of this approach stems from the simplicity of the construct itself. When a construct is easily operationalized in terms of observable behaviors, logic may be sufficient for determining test content.

4.7.2 Empirical

Empirical approaches to test construction are based on statistical analysis of data from an administration of test items. Thus, items are written and administered, and then the data are examined for relationships among item responses and patterns of response for certain groups of people.

Examining relationships among item responses is a *factor analytic approach* to test construction. Factor analysis is a statistical method for exploring and confirming the dimensionality of a set of item responses (more on this in Chapter 8). Items that are strongly and positively correlated with one another can be estimated to share a common cause, referred to as an unobserved factor. This factor is the same as the construct defined in Chapter 1 in terms of measurement models. In fact, a measurement model, where one or more constructs are assumed to cause variability in item responses, can be considered a type of factor analysis.

The factor analytic approach to test construction typically begins with the administration of a large number of test items, written based on logic or theory, which can then be reduced to a smaller set by removing items which do not relate strongly to or "load on" the intended factor. Factor loadings estimate the relationship between the item response and the underlying factor. Items with low loadings are assumed to be less related to the construct that the items are written to assess.

Another empirical approach to test construction was described in Chapter 3 for the MMPI. This approach resembles norm referencing, in that items are chosen as being representative of the construct when examinees with known characteristics respond to them in consistent ways. This is referred to as the *criterion-group approach* to test construction. A criterion group of examinees is purposively selected because they are known to represent the construct in some way.

The main drawbacks to the both of these empirical approaches to test construction are: 1) they require a large and representative sample of the test population, and can lead to biased results when the examinee sample is not representative; and 2) they may lead to test content that lacks face validity and that contradicts both logic and theory. The first drawback is addressed through appropriate sampling techniques, and via replication and cross-validation studies. However, the second often becomes an accepted feature of the test. For example, using the criterion-group approach, if a group of schizophrenic individuals agreed in a consistent way to the statement "I am able to read other people's minds," there would be empirical support for using this question to measure schizophrenia. The actual content of the item is unimportant, as long as the data come from representative and cross-validated samples of people with schizophrenia.

4.8 Features of personality items

As with cognitive test items, noncognitive test items come in a variety of types that differ in how material is presented to the test taker, and how responses are then collected. Aside from simple differences in style, most personality questions begin with a simple stem, statement, or scenario, and then include one or more options for response in the form of a rating scale.

4.8.1 The stem or prompt

In the traditional personality item, examinees evaluate a statement in terms of how well it represents them or is characteristic of them. The MMPI contains over 500 statements such as these, and examinees respond with either a "yes" or "no" to indicate whether or not the statement is true for them. Big Five Inventories (BFI) often take a similar approach. For example, the BFI administered within the Synthetic Aperture Personality

Table 4.2: BFI Agreeableness Scale

	1	2	3	4	5	6
Am indifferent to the feelings of others						
Inquire about others' well-being						
Know how to comfort others						
Love children						
Make people feel at ease						

Assessment (SAPA; sapa-project.org) contains twenty five statements, with five statements per factor, and responses collected with a six-point rating scale. Table 4.2 contains five items assessing agreeableness. The response scale for the SAPA BFI measures the extent to which statements accurately represent examinees, with scores of 1 for "very inaccurate", 2 for "moderately inaccurate", 3 for "slightly inaccurate", 4 for "slightly accurate", 5 for "moderately accurate", and 6 for "very accurate".

Multiple statements can also be used in place of a rating scale that references a single statement in a single item. This is the approach taken in the BDI (Beck et al. 1961). For example, a BDI question on social withdrawal reads,

0. I have not lost interest in other people.

1. I am less interested in other people now than I used to be.

2. I have lost most of my interest in other people and have little feeling for them.

3. I have lost all my interest in other people and don't care about them at all.

Here, statements are grouped together to represent increasing levels of social withdrawal in a particular context, and scores range from 0 to 3. A similar approach is taken in the MBTI, as shown above, where examinees essentially fill in the blanks for statements such as "I consider myself to be Social/Private." An alternative to these approaches would be to present one statement regarding social withdrawal or sociability, and then request a rating or a yes/no regarding how representative or true the statement is for examinees.

4.8.2 Response scales

As noted above, in addition to differing in the structure of the statement to which examinees respond, personality items also differ in the types of response scales used. These scales seek to measure agreement, preference, or frequency of occurrence along a continuum. Scale anchors are used as the response options on this continuum. A simple example was presented above for the MMPI, where anchors are only provided for the ends of the continuum, with yes/no, or, similarly, true/false. Yes/no and true/false responses can be used to measure agreement and preference. Frequency can be represented dichotomously as never/always or never and some other nonzero amount. A dichotomous scale could also include anchors for like/dislike to measure preference.

Different *degrees* of agreement, preference, and frequency are measured by including additional scale anchors, typically spaced evenly between the bottom and top of the continuum. The BFI mentioned above utilizes a six-point scale to measure agreement in terms of accuracy. Likert (1932) utilized a five-point scale to measure approval, with anchors ranging from strongly disapprove to strongly approve.

Three main issues arise in the construction of a rating scale for gathering responses. First, the length or number of scale anchors must be determined. An effective rating scale is only as long as necessary to capture meaningful variability in responses. Consider the basic pain assessment scale often used in doctors' offices, where a patient rates the pain they are experiencing on a scale from 0, labeled "no hurt" or "no pain," to 10,

Figure 4.2: A ten-point rating scale used to measure pain.

labeled "hurts worst" or "worst pain." Intermediate levels of pain may also be labeled as in Figure 4.2, taken from Hockenberry and Wilson (2012).

How well can a patient really distinguish between subtle differences pain, such as what appears to be a "little bit" of pain denoted for a score of 2 verses a "little more" for a score of 4? Is that a measurable and meaningful amount of pain? This depends on the patient, of course. However, an item analysis, discussed in Chapter 6, might reveal that patients are only able to consistently use a subset of the ten pain points. Perhaps we could reduce the ten-point scale to a four-point scale that includes categories such as none, some, lots, and most. Additional scale points can potentially lead to increases in score variability. However, in many measurement applications this variability reflects inconsistent use of the scale, that is, measurement error, rather than meaningful differences between individuals.

Although pain is subjective, and the anchors themselves seem arbitrary, the pain scale is useful in two contexts. Numerical anchors and visual cues, via the faces, are helpful when verbal communication is challenging. And as long as an individual interprets the anchors consistently over time, the scale can be used to identifying changes in pain within the individual.

A second issue in rating scale construction is whether or not to include a central, neutral option. Although examinees may prefer to have it, especially when responding to controversial topics where commitment one way or the other is difficult, the neutral option rarely provides useful information regarding the construct being measured, and should be avoided whenever possible (Kline 1986).

4.9 Personality item writing

The item writing guidelines presented in Spector (1992) and Kline (1986) are paraphrased here for reference. Regarding these guidelines, Kline (1986) notes,

> Much of what I shall say is obvious and little more than common sense. Nevertheless, examination of many published tests and tests used for internal selection by large organizations has convinced this author that these things need to be said. Too often, test constructors, blinded by the brilliance of the technology of item analysis, forget the fact that a test can be no better (but it can be worse) than its items.

As with the cognitive item writing guidelines, the affective guidelines mainly address issues related to clarity and conciseness of expression. Item writing, whether for cognitive or affective tests, is a form of writing, and together the item writing guidelines encourage efficient and effective writing.

The key difference between writing and item writing is that writing involves communication of information in one direction, to the reader, whereas item writing involves communication intended to elicit a response

from the reader. Thus, it is helpful to consider the guidelines in the context of testing, and how they support measurement of our construct of interest. Following the list of guidelines below is an overview of the main types of construct irrelevant variance that tend to influence affective measurement. These types are referred to as *response sets*.

4.9.1 Guidelines

1. Reduce insight: in general, the less examinees know about the construct being measured, the more authentic or genuine their responses are likely to be.
2. Encourage immediate response: related to insight, the more an examinee reflects on an item, in general, the less likely they are to respond genuinely.
3. Write clearly, specifically, and unambiguously: reliable and valid measurement requires that examinees consistently interpret and understand what is asked of them. Conflicting, confusing, uninterpretable, or excessive information introduces measurement error.
4. Reference behaviors: feelings, like pleasure and pain, mean different things to different people. Refer to easily identified symptoms, experiences, or behaviors, rather than feelings or interpretations of them.

4.9.2 Response sets

Response sets describe patterns of response that introduce bias into the process of measuring noncognitive constructs via self-reporting. Bias refers to systematic error that has a consistent and predictable impact on responses. The main response sets include social desirability, acquiescence, extremity, and neutrality.

Social desirability refers to a tendency for examinees to respond in what appears to be a socially desirable or favorable way. Examinees tend to under-report or de-emphasize constructs that carry negative connotations, and over-report or overemphasize constructs that carry positive connotations. For example, examinees tend to be less likely to endorse or identify with items that are perceived to measure stigmatized constructs such as depression, anxiety, and aggression, or behaviors such as procrastination and drug use. On the other hand, examinees are more likely to endorse or identify with items measuring desirable traits such as kindness, resilience, and generosity. Social desirability can be reduced by reducing insight, encouraging immediate response, and limiting the use of contexts that have obvious negative or positive connotations.

Acquiescence refers to a tendency for examinees to agree with items regardless of their content. The pattern may result from an underlying examinee disinterest and lack of involvement, or from a desire simply to respond in the affirmative. Whatever the cause, the result is consistent endorsement of items. One way to identify and potentially reduce acquiescence is to use both positively and negatively worded items. Examinees who acquiesce may notice the shift in emphasis from positive to negative and respond more accurately. Examinees who endorse both positively and negatively worded items will have inconsistent scores that can be used to identify them as invalid.

Extremity and neutrality refer to a tendency to over-exaggerate and under-exaggerate response. In both cases, the underlying problem is an inconsistent interpretation and use of a rating scale across examinees. To reduce extremity and neutrality, Kline (1986) simply recommends the use of dichotomous response options, for example, yes/no, where only the extremes of the scale are available.

4.10 Summary

This chapter provides an overview of cognitive and noncognitive test construction and item writing. Effective cognitive tests have a clear purpose and are structured around well-defined learning objectives within a test outline, or around a theoretical or empirical framework. The outline for a cognitive test describes key features of the test, such as content areas or subscales, the depth of knowledge assessed, and the types of items used. Together, these features specify the number of types of items that must be developed to adequately

address the test purpose. Noncognitive personality measures may also be structured around a test outline that describes key features of the test, such as subscales or subtypes of the construct assessed, and the types of items and response categories used. Again, these features specify the number and types of items that must be developed to adequately address the test purpose.

4.10.1 Learning objectives

Cognitive

1. Describe the purpose of a cognitive learning objective or learning outcome statement, and demonstrate the effective use of learning objectives in the item writing process.
2. Describe how a test outline or test plan is used in cognitive test development to align the test to the content domain and learning objectives.
3. Compare items assessing different cognitive levels or depth of knowledge, for example, higher-order thinking such as synthesizing and evaluating information versus lower-order thinking such as recall and definitional knowledge.
4. Identify and provide examples of SR item types (multiple-choice, true/false, matching) and CR item types (short-answer, essay).
5. Compare and contrast SR and CR item types, describing the benefits and limitations of each type.
6. Identify the main theme addressed in the item writing guidelines, and how each guideline supports this theme.
7. Create and use a scoring rubric to evaluate answers to a CR question.
8. Write and critique cognitive test items that match given learning objectives and depths of knowledge and that follow the item writing guidelines.

Noncognitive

9. Define affective measurement and contrast it with cognitive measurement in terms of applications and purposes, the types of constructs assessed, and the test construction process.
10. Compare affective test construction strategies, with examples of their use, and the strengths and limitations of each.
11. Compare and contrast item types and response types used in affective measurement, describing the benefits and limitations of each type, and demonstrating their use.
12. Define the main affective response sets, and demonstrate strategies for reducing their effects.
13. Write and critique effective affective items using empirical guidelines.

4.10.2 Exercises

1. Consider the advantages and disadvantages for the different forms of the DOK question above, and the science question within it. Would the limitations of the selected response forms be worth the gains in efficiency? Or would the gains in authenticity and DOK justify the use of the CR forms?
2. Evaluate the cognitive items presented in this chapter in terms of the DOK they assess and the extent to which they follow the item writing guidelines.
3. Evaluate the released PISA 2009 reading items in `PISA09` in terms of the DOK they assess and the extent to which they follow the item writing guidelines.
4. Describe common features of cognitive items that allow them to assess effectively higher DOK.
5. Evaluate the noncognitive items presented in this chapter in terms of the guidelines and response sets. Consider the constructs measured by each item, and how failure to follow the guidelines and address the response sets could reduce the quality of the results.
6. Evaluate the PISA 2009 survey items in `PISA09` in terms of the guidelines and response sets. Consider the constructs measured by each item, and how failure to follow the guidelines and address the response sets could reduce the quality of the results.

Chapter 5

Reliability

> Too much consistency is as bad for the mind as it is for the body. Consistency is contrary to nature, contrary to life. The only completely consistent people are the dead.
> — Aldous Huxley

> Consistency is the hallmark of the unimaginative.
> — Oscar Wilde

From the perspective of the creative thinker or innovator, consistency can be viewed as problematic. Consistent thinking leads to more of the same, as it limits diversity and change. On the other hand, inconsistent thinking or thinking outside the box produces new methods and ideas, inventions and breakthroughs, leading to innovation and growth.

Standardized tests are designed to be consistent, and, by their very nature, they are poor measures of creative thinking. In fact, the construct of creativity is one of the more elusive in educational and psychological testing. Although published tests of creative thinking and problem solving exist, administration procedures are complex and consistency in the resulting scores can be, unsurprisingly, very low. Creativity seems to involve an inconsistency in thinking and behavior that is challenging to measure reliably.

From the perspective of the test maker or test taker, which happens to be the perspective we take in this book, consistency is critical to valid measurement. An inconsistent or unreliable test produces unreliable results that only inconsistently support the intended inferences of the test. Nearly a century of research provides us with a framework for examining and understanding the reliability of test scores, and, most importantly, how reliability can be estimated and improved.

This chapter introduces reliability within the framework of the classical test theory (CTT) model, which is then extended to generalizability (G) theory. In Chapter 7, we'll learn about reliability within the item response theory model. These theories all involve measurement models, sometimes referred to as latent variable models, which are used to describe the construct or constructs assumed to underlie responses to test items.

This chapter starts with a general definition of reliability in terms of consistency of measurement. The CTT model and assumptions are then presented in connection with statistical inference and measurement models, which were introduced in Chapter 1. Reliability and unreliability, that is, the standard error of measurement, are discussed as products of CTT, and the four main study designs and corresponding methods for estimating reliability are reviewed. Finally, reliability is discussed for situations where scores come from raters. This is called interrater reliability, and it is best conceptualized using G theory.

```
# R setup for this chapter
# Required packages are assumed to be installed - see chapter 1
```

```r
library("epmr")
library("ggplot2")
# Functions we'll use in this chapter
# set.seed(), rnorm(), and runif() to simulate data
# paste0(), rowSums(), and data.frame() for prepping data
# rsim() and setrange() from epmr to simulate and modify scores
# ggplot(), aes(), geom_point(), and geom_abline() for plotting
```

5.1 Consistency of measurement

In educational and psychological testing, reliability refers to the precision of the measurement process, or the consistency of scores produced by a test. Reliability is a prerequisite for validity. That is, for scores to be valid indicators of the intended inferences or uses of a test, they must first be reliable or precisely measured. However, precision or consistency in test scores does not necessarily indicate validity.

A simple analogy may help clarify the distinction between reliability and validity. If the testing process is represented by an archery contest, where the test taker, an archer, gets multiple attempts to hit the center of a target, each arrow could be considered a repeated measurement of the construct, archery ability. Imagine someone whose arrows all end up within a few millimeters of one another, tightly bunched together, but all stuck in the trunk of a tree standing behind the target itself. This represents consistent but inaccurate measurement. On the other hand, consider another archer whose arrows are scattered around the target, with one hitting close to the bulls-eye and the rest spread widely around it. This represents inconsistent and inaccurate measurement. Reliability and validity are both present only when the arrows are all close to the center of the target. In that case, we're consistently measuring what we intend to measure.

A key assumption in this analogy is that our archers are actually skilled, and any errors in their shots are due to the measurement process itself. Instead, consistently hitting a nearby tree may be evidence of a reliable test given to someone who is simply missing the mark. In reality, if someone scores systematically off target or above or below their true underlying ability, we have a hard time attributing this to bias in the testing process versus a true difference in ability. A key point here is that evidence supporting the reliability of a test can be based on results from the test itself. However, evidence supporting the validity of the test must come, in part, from external sources. The only ways to determine that consistently hitting a tree represents low ability are to a) confirm that our test is unbiased and b) conduct a separate test. These are validity issues, which will be covered in Chapter 9.

Consider the reliability of other familiar physical measurements. One common example is measuring weight. How can measurements from a basic floor scale be both unreliable and invalid? Think about the potential sources of unreliability and invalidity. For example, consider measuring the weight of a young child every day before school. What is the variable we're actually measuring? How might this variable change from day to day? And how might these changes be reflected in our daily measurements? If our measurements change from day to day, how much of this change can be attributed to actual changes in weight versus extraneous factors such as the weather or glitches in the scale itself?

In both of these examples, there are multiple interrelated sources of potential inconsistency in the measurement process. These sources of score change can be grouped into three categories. First, the construct itself may actually change from one measurement to the next. For example, this occurs when practice effects or growth lead to improvements in performance over time. Archers may shoot more accurately as they calibrate their bow or readjust for wind speed with each arrow. In achievement testing, students may learn or at least be refreshed in the content of a test as they take it.

Second, the testing process itself could differ across measurement occasions. Perhaps there is a strong cross-breeze one day but not the next. Or maybe the people officiating the competition allow for different amounts of time for warm-up. Or maybe the audience is rowdy or disagreeable at some points and supportive at others. These factors are tied to the testing process itself, and they may all lead to changes in scores.

Finally, our test may simply be limited in scope. Despite our best efforts, it may be that arrows differ from one another to some degree in balance or construction. Or it may be that archers' fingers occasionally slip for no fault of their own. By using a limited number of shots, scores may change or differ from one another simply because of the limited nature of the test. Extending the analogy to other sports, football and basketball each involve many opportunities to score points that could be used to represent the ability of a player or team. On the other hand, because of the scarcity of goals in soccer, a single match may not accurately represent ability, especially when the referees have been bribed!

5.2 Classical test theory

5.2.1 The model

Now that we have a definition of reliability, with examples of the inconsistencies that can impact test scores, we can establish a framework for estimating reliability based on the changes or variability that occur in a set of test scores. Recall from Chapter 1 that statistics allow us to make an inference from a) the changes we observe in our test scores to b) what we assume is the underlying cause of these changes, the construct. Reliability is based on and estimated within a simple measurement model that decomposes an observed test score X into two parts, truth T and error E:

$$X = T + E. \tag{5.1}$$

Note that X in Equation 5.1 is a composite consisting of two component scores. The true score T is the construct we're intending to measure. We assume that T is impacting our observation in X. The error score E is everything randomly unrelated to the construct we're intending to measure. Error also directly impacts our observation.

To understand the CTT model in Equation 5.1, we have to understand the following question: how would an individual's observed score X vary across multiple repeated administrations of a test? If you took a test every day for the next 365 days, and somehow you forgot about all of the previous administrations of the test, would your observed score change from one testing to the next? And if so, why would it change?

CTT answers this question by making two key assumptions about the variability and covariability of T and E. First, your true underlying ability T is assumed to be constant for an individual. If your true ability can be expressed as 20 out of 25 possible points, every time you take the test your true score will consistently be 20. It will not change, according to CTT. Second, error that influences an observed score at a given administration is assumed to be completely random, and thus unrelated to your true score and to any other error score for another administration.

So, at one administration of the test, some form of error may cause your score to decrease by two points. Maybe you weren't feeling well that day. In this case, knowing that $T = 20$, what is E in Equation 5.1, and what is X? At another administration, you might guess correctly on a few of the test questions, resulting in an increase of 3 based solely on error. What is E now? And what is X?

Solving for E in Equation 5.1 clarifies that random error is simply the difference between the true score and the observation, where a negative error always indicates that X is too low and a positive error always indicates that X is too high:

$$E = X - T. \tag{5.2}$$

So, having a cold produces $E = -2$ and $X = 18$, compared to your true score of 20. And guessing correctly produces $E = 3$ and $X = 23$.

According to CTT, over infinite administrations of a test without practice effects, your true score will always be the same, and error scores will vary completely randomly, some being positive, others being negative, but

being on average zero. Given these assumptions, what should your average observed score be across infinite administrations of the test? And what should be the standard deviation of your observed score over these infinite observed scores? In responding to these questions, you don't have to identify specific values, but you should instead reference the means and standard deviations that you'd expect T and E to have. Remember that X is expressed entirely as a function of T and E, so we can derive properties of the composite from its components.

```
# Simulate a constant true score, and randomly varying error scores from
# a normal population with mean 0 and SD 1
# set.seed() gives R a starting point for generating random numbers
# so we can get the same results on different computers
# You should check the mean and SD of E and X
# Creating a histogram of X might be interesting too
set.seed(160416)
myt <- 20
mye <- rnorm(1000, mean = 0, sd = 1)
myx <- myt + mye
```

Here is an explanation of the questions above. We know that your average observed score would be your true score. Error, because it varies randomly, would cancel itself out in the long run, and your mean X observed score would simply be T. The standard deviation of these infinite observed scores X would then be entirely due to error. Since truth does not change, any change in observed scores must be error variability. This standard deviation is referred to as the *standard error of measurement* (SEM), discussed more below. Although it is theoretically impossible to obtain the actual SEM, since you can never take a test an infinite number of times, we can estimate SEM using data from a sample of test takers. And, as we'll see below, reliability will be estimated as the opposite of measurement error.

Figure 5.1 demonstrates the parts of the CTT model using `PISA09` total reading scores for students from Belgium. We're transitioning here from lots of X and E scores for an individual with a constant T, to an X, T, and E score per individual within a sample. Total reading scores on the x-axis represent X, and simulated T scores are on the y-axis. The solid line represents what we'd expect if there were no error, in which case $X = T$. As a result, the horizontal scatter in the plot represents E. Note that T are simulated to be continuous and to range from 0 to 11. X scores are discrete, but they've been "jittered" slightly left to right to reveal the densities of points in the plot.

```
# Calculate total reading scores, as in Chapter 2
ritems <- c("r414q02", "r414q11", "r414q06", "r414q09", "r452q03",
  "r452q04", "r452q06", "r452q07", "r458q01", "r458q07", "r458q04")
rsitems <- paste0(ritems, "s")
xscores <- rowSums(PISA09[PISA09$cnt == "BEL", rsitems], na.rm = TRUE)
# Simulate error scores based on known SEM of 1.4, which we'll calculate
# later, then create true scores
# True scores are truncated to fall bewteen 0 and 11 using setrange()
escores <- rnorm(length(xscores), 0, 1.4)
tscores <- setrange(xscores - escores, y = xscores)
# Combine in a data frame and create a scatterplot
scores <- data.frame(x1 = xscores, t = tscores, e = escores)
ggplot(scores, aes(x1, t)) +
  geom_point(position = position_jitter(w = .3)) +
  geom_abline(col = "blue")
```

Consider individuals with a true score of $T = 6$. The perfect test would measure true scores perfectly, and produce observed scores of $X = T = 6$, on the solid line. In reality, lots of people scored at or around $T = 6$, but their actual scores on X varied from about $X = 3$ to $X = 9$. Again, any horizontal distance from the blue line for a given true score represents E.

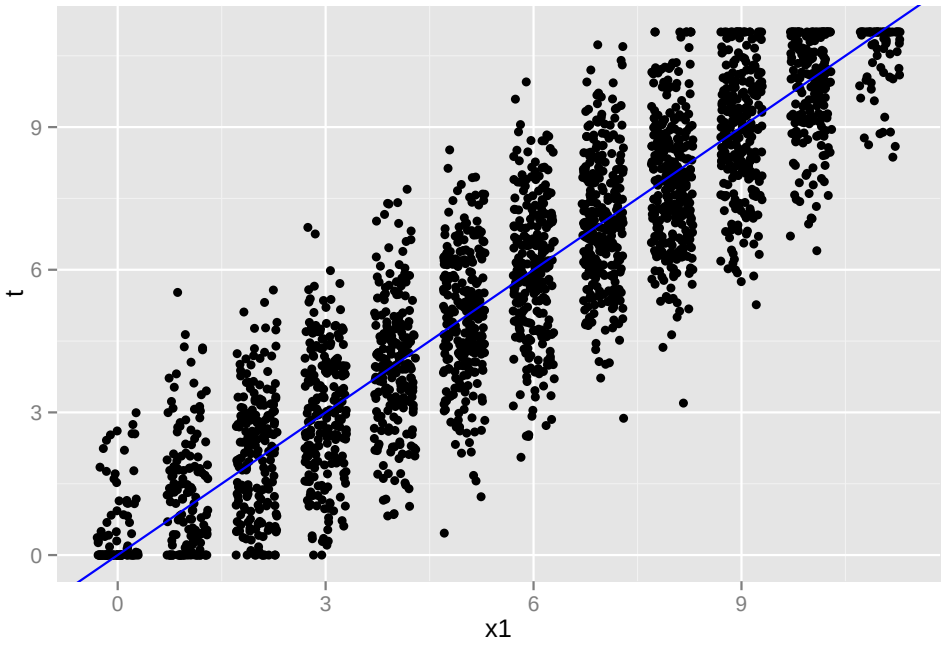


Figure 5.1: PISA total reading scores with simulated error and true scores based on CTT.

5.2.2 Applications of the model

Let's think about some specific examples now of the classical test theory model. Consider a construct that interests you, how this construct is operationalized, and the kind of measurement scale that results from it. Consider the possible score range, and try to articulate X and T in your own example.

Next, let's think about E. What might cause an individual's observed score X to differ from their true score T in this situation? Think about the conditions in which the test would be administered. Think about the population of students, patients, individuals that you are working with. Would they tend to bring some form of error or unreliability into the measurement process?

Here's a simple example involving preschoolers. As I mentioned in earlier chapters, some of my research involves measures of early literacy. In this research, we test children's phonological awareness by presenting them with a target image, for example, an image of a star, and asking them to identify the image among three response options that rhymes with the target image. So, we'd present the images and say, "Which one rhymes with star?" Then, for example, children might point to the image of a car.

Measurement error is problematic in a test like this for a number of reasons. First of all, preschoolers are easily distracted. Even with standardized one-on-one test administration apart from the rest of the class, children can be distracted by a variety of seemingly innocuous features of the administration or environment, from the chair they're sitting in, to the zipper on their jacket. In the absence of things in their environment, they'll tell you about things from home, what they had for breakfast, what they did over the weekend, or, as a last resort, things from their imagination. Second of all, because of their short attention span, the test itself has to be brief and simple to administer. Shorter tests, as mentioned above in terms of archery and other sports, are less reliable tests; fewer items makes it more difficult to identify the reliable portion of the measurement process. In shorter tests, problems with individual items have a larger impact on the test as a whole.

Think about what would happen to E and the standard deviation of E if a test were very short, perhaps including only five test questions. What would happen to E and its standard deviation if we increased the number of questions to 200? What might happen to E and its standard deviation if we administered the test outside? These are the types of questions we will answer by considering the specific sources of measurement error and the impact we expect them to have, whether systematic or random, on our observed score.

5.2.3 Systematic and random error

A systematic error is one that influences a person's score in the same way at every repeated administration of a test. A random error is one that could be positive or negative for a person, one that changes randomly by administration. In the preschooler literacy example, as students focus less on the test itself and more on their surroundings, their scores might involve more guessing, which introduces random error if the guessing is truly random. Interestingly, we noticed in pilot studies of early literacy measures that students tended to choose the first option when they didn't know the correct response. This resulted in a systematic change in their scores based on how often the correct response happened to be first.

Distinguishing between systematic and random error can be difficult. Some features of a test or test administration can produce both types of error. A popular example of systematic versus random error is demonstrated by a faulty floor scale. Revisiting the example from above, suppose I measure my oldest son's weight every day for two weeks as soon as he gets home from school. For context, my oldest is ten years old at the time of writing this. Suppose also that his average weight across the two weeks was 60 pounds, but that this varied with a standard deviation of 5 pounds. Think about some reasons for having such a large standard deviation. What could cause my son's weight, according to a floor scale, to differ from his true weight at a given measurement? What about his clothing? Or how many toys are in his pockets? Or how much food he ate for lunch?

What type of error does the standard deviation *not* capture? Systematic error doesn't vary from one measurement to the next. If the scale itself is not calibrated correctly, for example, it may overestimate or underestimate weight consistently from one measure to the next. The important point to remember here is that only one type of error is captured by E in CTT: the random error. Any systematic error that occurs consistently across administrations will become part of T, and will not reduce our estimate of reliability.

5.3 Reliability and unreliability

5.3.1 Reliability

Figure 5.2 contains a plot similar to the one in Figure 5.1 where we identified X, T, and E. This time, we have scores on two reading test forms, with the first form is now called X_1 and second form is X_2, and we're going to focus on the overall distances of the points from the line that goes diagonally across the plot. Once again, this line represents truth. A person with a true score of 11 on X_1 will score 11 on X_2, based on the assumptions of the CTT model.

Although the solid line represents what we'd expect to see for true scores, we don't actually know anyone's true score, even for those students who happen to get the same score on both test forms. The points in Figure 5.2 are all observed scores. The students who score the same on both test forms do indicate more consistent measurement. However, it could be that their true score still differs from observed. There's no way to know. To calculate truth, we would have to administer the test an infinite number of times, and then take the average, or simply simulate it, as in Figure 5.1.

```
# Simulate scores for a new form of the reading test called y
# rho is the made up reliability, which is set to 0.80
# x is the original reading total scores
# Form y is slightly easier than x with mean 6 and SD 3
xysim <- rsim(rho = .8, x = scores$x1, meany = 6, sdy = 3)
scores$x2 <- round(setrange(xysim$y, scores$x1))
ggplot(scores, aes(x1, x2)) +
  geom_point(position = position_jitter(w = .3, h = .3)) +
  geom_abline(col = "blue")
```

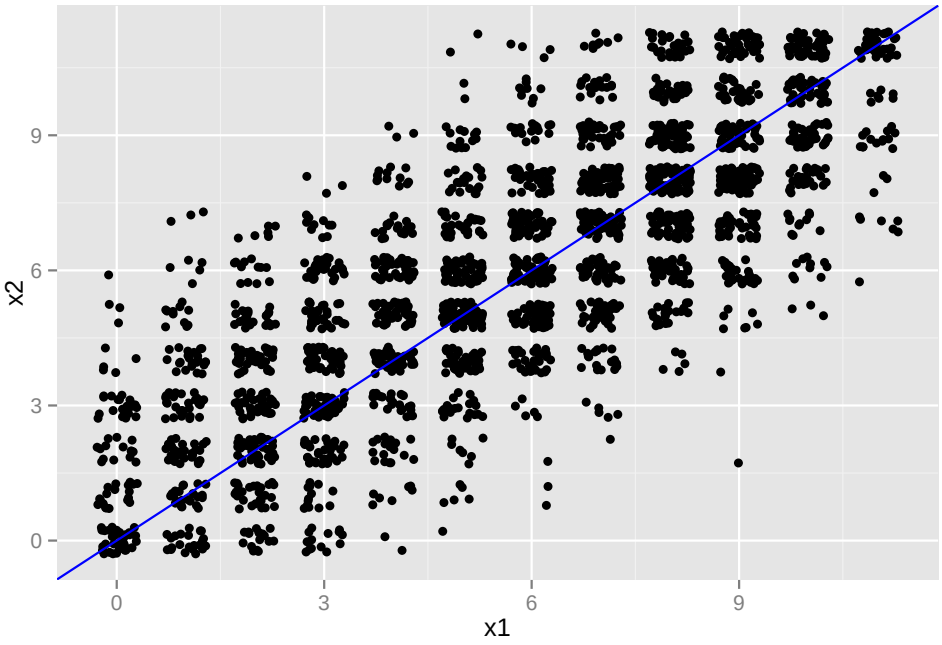


Figure 5.2: PISA total reading scores and scores on a simulated second form of the reading test.

The assumptions of CTT make it possible for us to estimate the reliability of scores using a sample of individuals. Figure 5.2 shows scores on two test forms, and the overall scatter of the scores from the solid line gives us an idea of the linear relationship between them. There appears to be a strong, positive, linear relationship. Thus, people tend to score similarly from one form to the next, with higher scores on one form corresponding to higher scores on the other. The correlation coefficient for this data set, `cor(scores$x, scores$y)` $= 0.795$, gives us an estimate of how similar scores are, on average from X_1 to X_2. Because the correlation is positive and strong for this plot, we would expect a person's score to be pretty similar from one testing to the next.

Imagine if the scatter plot were instead nearly circular, with no clear linear trend from one test form to the next. The correlation in this case would be near zero. Would we expect someone to receive a similar score from one test to the next? On the other hand, imagine a scatter plot that falls perfectly on the line. If you score, for example, 10 on one form, you also score 10 on the other. The correlation in this case would be 1. Would we expect scores to remain consistent from one test to the next?

We're now ready for a statistical definition of reliability. In CTT, reliability is defined as the proportion of variability in X that is due to variability in true scores T:

$$r = \frac{\sigma_T^2}{\sigma_X^2}. \tag{5.3}$$

Note that true scores are assumed to be constant in CTT *for a given individual*, but not across individuals. Thus, reliability is defined in terms of variability in scores for a population of test takers. Why do some individuals get higher scores than others? In part because they actually have higher abilities or true scores than others, but also, in part, because of measurement error. The reliability coefficient in Equation 5.3 tells us *how much* of our observed variability in X is due to true score differences.

5.3.2 Estimating reliability

Unfortunately, we can't ever know the CTT true scores for test takers. So we have to estimate reliability indirectly. One indirect estimate made possible by CTT is the correlation between scores on two forms of the

same test, as represented in Figure 5.2:

$$r = \rho_{X_1 X_2} = \frac{\sigma_{X_1 X_2}}{\sigma_{X_1} \sigma_{X_2}}. \tag{5.4}$$

This correlation is estimated as the covariance, or the shared variance between the distributions on two forms, divided by a product of the standard deviations, or the total available variance within each distribution.

There are other methods for estimating reliability from a single form of a test. The only ones presented here are split-half reliability and coefficient alpha. Split-half is only presented because of its connection to what's called the Spearman-Brown reliability formula. The split-half method predates coefficient alpha, and is computationally simpler. It takes scores on a single test form, and separates them into scores on two halves of the test, which are treated as separate test forms. The correlation between these two halves then represents an indirect estimate of reliability, based on Equation 5.3.

```
# Split half correlation, assuming we only had scores on one test form
# With an odd number of reading items, one half has 5 items and the
# other has 6
xsplit1 <- rowSums(PISA09[PISA09$cnt == "BEL", rsitems[1:5]])
xsplit2 <- rowSums(PISA09[PISA09$cnt == "BEL", rsitems[6:11]])
cor(xsplit1, xsplit2, use = "complete")
## [1] 0.624843
```

The Spearman-Brown formula was originally used to correct for the reduction in reliability that occurred when correlating two test forms that were only half the length of the original test. In theory, reliability will increase as we add items to a test. Thus, Spearman-Brown is used to estimate, or predict, what the reliability would be if the half-length tests were made into full-length tests.

```
# sbr() in the epmr package uses the Spearman-Brown formula to estimate how
# reliability would change when test length changes by a factor k
# If test length were doubled, k would be 2
sbr(r = cor(xsplit1, xsplit2, use = "complete"), k = 2)
## [1] 0.7691119
```

The Spearman-Brown formula also has other practical uses. Today, it is most commonly used during the test development process to predict how reliability would change if a test form were reduced or increased in length. For example, if you are developing a test and you gather pilot data on 20 test items with a reliability estimate of 0.60, Spearman-Brown can be used to predict how this reliability would go up if you increased the test length to 30 or 40 items. You could also pilot test a large number of items, say 100, and predict how reliability would decrease if you wanted to use a shorter test.

The Spearman-Brown reliability, r_{new}, is estimated as a function of what's labeled here as the old reliability r_{old} and the factor by which the length of X is predicted to change, k:

$$r_{new} = \frac{k r_{old}}{(k-1) r_{old} + 1}. \tag{5.5}$$

Again, k is the factor by which the test length is increased or decreased. It is equal to the number of items in the new test divided by the number of items in the original test. Multiply k by the old reliability, and then divided the result by $(k-1)$ times the old reliability, plus 1. For the example mentioned above, going from 20 to 30 items, we have $(30/20 \times 0.60)$ divided by $(30/20 - 1) \times 0.60 + 1 = .69$. Going to 40 items, we have a new reliability of 0.75. The epmr package contains `sbr()`, a simple function for estimating the Spearman-Brown reliability.

Alpha is arguably the most popular form of reliability. Many people refer to it as "Chronbach's alpha," but Chronbach himself never intended to claim authorship for it and in later years he regretted the fact that it

was attributed to him (see Cronbach and Shavelson 2004). The popularity of alpha is due to the fact that it can be calculated using scores from a single test form, rather than two separate administrations or split halves. Alpha is defined as

$$r = \alpha = \left(\frac{J}{J-1}\right)\left(\frac{\sigma_X^2 - \sum \sigma_{X_j}^2}{\sigma_X^2}\right), \tag{5.6}$$

where J is the number of items on the test, σ_X^2 is the variance of observed total scores on X, and $\sum \sigma_{X_j}^2$ is the sum of variances for each item j on X. To see how it relates to the CTT definition of reliability in Equation 5.3, consider the top of the second fraction in Equation 5.6. The total test variance σ_X^2 captures all the variability available in the total scores for the test. We're subtracting from it the variances that are unique to the individual items themselves. What's left over? Only the shared variability among the items in the test. We then divide this shared variability by the total available variability. Within the formula for alpha you should see the general formula for reliability, true variance over observed.

```
# epmr includes rstudy() which estimates alpha and a related form of
# reliability called omega, along with corresponding SEM
# You can also use alpha() to obtain coefficient alpha directly
rstudy(PISA09[, rsitems])
##                r      sem
## alpha 0.7596071 1.402265
## omega 0.7632029 1.391738
```

Keep in mind, alpha is an estimate of reliability, just like the correlation is. So, any equation requiring an estimate of reliability, like SEM below, can be computed using either a correlation coefficient or an alpha coefficient. Students often struggle with this point: correlation is one estimate of reliability, alpha is another. They're both estimating the same thing, but in different ways based on different reliability study designs.

5.3.3 Unreliability

Now that we've defined reliability in terms of the proportion of observed variance that is true, we can define *unreliability* as the portion of observed variance that is error. This is simply 1 minus the reliability:

$$1 - r = \frac{\sigma_E^2}{\sigma_X^2}. \tag{5.7}$$

Typically, we're more interested in how the unreliability of a test can be expressed in terms of the available observed variability. Thus, we multiply the unreliable proportion of variance by the standard deviation of X to obtain the SEM:

$$SEM = \sigma_X \sqrt{1 - r}. \tag{5.8}$$

The SEM is the average variability in observed scores attributable to error. As any statistical standard error, it can be used to create a confidence interval (CI) around the statistic that it estimates, that is, T. Since we don't have T, we instead create the confidence interval around X to index how confident we are that T falls within it for a given individual. For example, the verbal reasoning subtest of the GRE is reported to have a reliability of 0.93 and an SEM of 2.2, on a scale that ranges from 130 to 170. Thus, an observed verbal reasoning score of 155 has a 95% confidence interval of about ± 4.4 points. At $X = 155$, we are 95% confident that the true score falls somewhere between 150.8 and 159.2. (Note that scores on the GRE are actually estimated using IRT.)

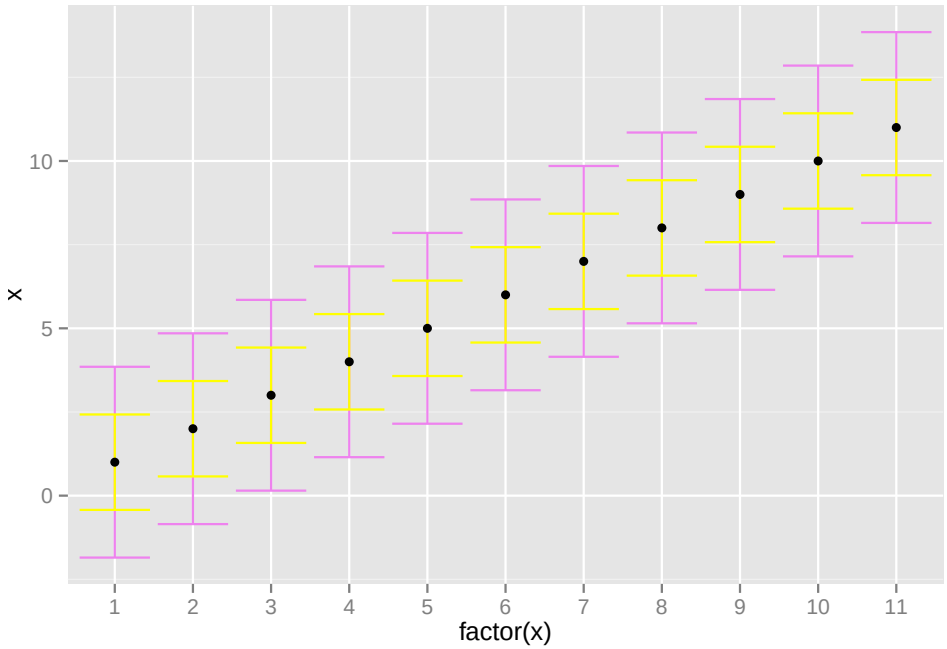


Figure 5.3: The PISA09 reading scale shown with 68 and 95 percent confidence intervals around each point.

Confidence intervals for `PISA09` can be estimated in the same way. First, we choose a measure of reliability, find the SD of observed scores, and obtain the corresponding SEM. Then, we can find the CI, which gives us the expected amount of uncertainty in our observed scores due to random measurement error. Here, we're calculating SEM and the CI using alpha, but other reliability estimates would work as well. Figure 5.3 shows the 11 possible `PISA09` reading scores in order, with error bars based on SEM for students in Belgium.

```
# Get alpha and SEM for students in Belgium
bela <- alpha(PISA09[PISA09$cnt == "BEL", rsitems])
belsem <- sem(r = bela, sd = sd(scores$x1, na.rm = T))
# Plot the 11 possible total scores against themselves
# Error bars are shown for 1 SEM, giving a 68% confidence interval
# and 2 SEM, giving the 95% confidence interval
# x is converted to factor to show discrete values on the x-axis
beldat <- data.frame(x = 1:11, sem = belsem)
ggplot(beldat, aes(factor(x), x)) +
  geom_errorbar(aes(ymin = x - sem * 2, ymax = x + sem * 2), col = "violet") +
  geom_errorbar(aes(ymin = x - sem, ymax = x + sem), col = "yellow") +
  geom_point()
```

Figure 5.3 helps us visualize the impact of unreliable measurement on score comparisons. For example, note that the top of the 95% confidence interval for X of 2 extends nearly to 5 points, and thus overlaps with the CI for adjacent scores 3 through 7. It isn't until X of 8 that the CI no longer overlap. With a CI of `belsem` 1.425, we're 95% confident that students with observed scores differing at least by `belsem * 4` 5.7 have different true scores. Students with observed scores closer than `belsem * 4` may actually have the same true scores.

5.3.4 Interpreting reliability and unreliability

There are no agreed-upon standards for interpreting reliability coefficients. Reliability is bound by 0 on the lower end and 1 at the upper end, because, by definition, the amount of true variability can never be less or

Table 5.1: General Guidelines for Interpreting Reliability Coefficients

Reliability	High Stakes Interpretation	Low Stakes Interpretation
≥ 0.90	Excellent	Excellent
$0.80 \leq r < 0.90$	Good	Excellent
$0.70 \leq r < 0.80$	Acceptable	Good
$0.60 \leq r < 0.70$	Borderline	Acceptable
$0.50 \leq r < 0.60$	Low	Borderline
$0.20 \leq r < 0.50$	Unacceptable	Low
$0.00 \leq r < 0.20$	Unacceptable	Unacceptable

more than the total available variability in X. Higher reliability is clearly better, but cutoffs for acceptable levels of reliability vary for different fields, situations, and types of tests. The stakes of a test are an important consideration when interpreting reliability coefficients. The higher the stakes, the higher we expect reliability to be. Otherwise, cutoffs depend on the particular application.

In general, reliabilities for educational and psychological tests can be interpreted using scales like the ones presented in Table 5.1. With medium-stakes tests, a reliability of 0.70 is sometimes considered minimally acceptable, 0.80 is decent, 0.90 is quite good, and anything above 0.90 is excellent. High stakes tests should have reliabilities at or above 0.90. Low stakes tests, which are often simpler and shorter than higher-stakes ones, often have reliabilities as low as 0.70. These are general guidelines, and interpretations can vary considerably by test. Remember that the cognitive measures in PISA would be considered low-stakes at the student level.

A few additional considerations are necessary when interpreting coefficient alpha. First, alpha assumes that all items measure the same single construct. Items are also assumed to be equally related to this construct, that is, they are assumed to be parallel measures of the construct. When the items are not parallel measures of the construct, alpha is considered a lower-bound estimate of reliability, that is, the true reliability for the test is expected to be higher than indicated by alpha. Finally, alpha is not a measure of dimensionality. It is frequently claimed that a strong coefficient alpha supports the unidimensionality of a measure. However, alpha does not index dimensionality. It is impacted by the extent to which all of the test items measure a single construct, but it does not necessarily go up or down as a test becomes more or less unidimensional.

5.3.5 Reliability study designs

Now that we've established the more common estimates of reliability and unreliability, we can discuss the four main study designs that allow us to collect data for our estimates. These designs are referred to as internal consistency, equivalence, stability, and equivalence/stability designs. Each design produces a corresponding type of reliability that is expected to be impacted by different sources of measurement error.

The four standard study designs vary in the number of test forms and the number of testing occasions involved in the study. Until now, we've been talking about using two test forms on two separate administrations. This study design is found in the lower right corner of Table 5.2, and it provides us with an estimate of equivalence (for two different forms of a test) and stability (across two different administrations of the test). This study design has the potential to capture the most sources of measurement error, and it can thus produce the lowest estimate of reliability, because of the two factors involved. The more time that passes between administrations, and as two test forms differ more in their content and other features, the more error we would expect to be introduced. On the other hand, as our two test forms are administered closer in time, we move from the lower right corner to the upper right corner of Table 5.2, and our estimate of reliability captures less of the measurement error introduced by the passage of time. We're left with an estimate of the equivalence between the two forms.

As our test forms become more and more equivalent, we eventually end up with the same test form, and we move to the first column in Table 5.2, where one of two types of reliability is estimated. First, if we administer

Table 5.2: Four Main Reliability Study Designs

	1 Form	2 Forms
1 Occasion	Internal Consistency	Equivalence
2 Occasions	Stability & Equivalence	Equivalence & Stability

the same test twice with time passing between administrations, we have an estimate of the stability of our measurement over time. Given that the same test is given twice, any measurement error will be due to the passage of time, rather than differences between the test forms. Second, if we administer one test only once, we no longer have an estimate of stability, and we also no longer have an estimate of reliability that is based on correlation. Instead, we have an estimate of what is referred to as the internal consistency of the measurement. This is based on the relationships among the test items themselves, which we treat as miniature alternate forms of the test. The resulting reliability estimate is impacted by error that comes from the items themselves being unstable estimates of the construct of interest.

Internal consistency reliability is estimated using either coefficient alpha or split-half reliability. All the remaining cells in Table 5.2 involve estimates of reliability that are based on correlation coefficients.

Table 5.2 contains four commonly used reliability study designs. There are others, including designs based on more than two forms or more than two occasions, and designs involving scores from raters, discussed below.

5.4 Interrater reliability

> It was like listening to three cats getting strangled in an alley.
> — Simon Cowell, disparaging a singer on *American Idol*

Interrater reliability extends the concepts of CTT to applications of testing that rely on scores from raters or judges. These applications most often involve some form of performance assessment, for example, stage performance within a singing competition. Judgment and scoring of performance by raters introduces additional error into the measurement process. Interrater reliability allows us to examine the negative impact of this error on our scores.

5.4.1 Proportion agreement

The proportion of agreement is the simplest measure of interrater reliability. It is calculated as the total number of times a set of ratings agree, divided by the total number of units of observation that are rated. The strengths of proportion agreement are that it is simple to calculate and it can be used with any type of *discrete* measurement scale. The major drawbacks are that it doesn't account for chance agreement between ratings, and it only utilizes the nominal information in a scale, that is, any ordering of values is ignored.

To see the effects of chance, let's simulate scores from two judges where ratings are completely random, as if scores of 0 and 1 are given according to the flip of a coin. In this case, we would expect two raters to agree a certain proportion of the time by chance alone. The `table()` function creates a cross-tabulation of frequencies, also known as a crosstab. Frequencies for agreement are found in the diagonal cells, from upper left to lower right, and frequencies for disagreement are found everywhere else. Using the results in the crosstab, find the proportion agreement.

```
# Simulate random coin flips for two raters
# runif() generates random numbers from a uniform distribution
flip1 <- round(runif(30))
flip2 <- round(runif(30))
table(flip1 == flip2)
```

Table 5.3: Crosstab of Scores From Rater 1 in Rows and Rater 2 in Columns

	0	1	2	3	4	5	6
0	2	0	0	0	0	0	0
1	3	2	0	0	0	0	0
2	2	2	4	1	0	0	0
3	0	3	7	4	1	0	0
4	0	1	10	14	12	0	0
5	0	0	0	6	9	4	0
6	0	0	0	0	3	5	5

```
##
## FALSE  TRUE
##    19    11
```

Data for the next few examples were simulated to represent scores given by two raters with a certain correlation, that is, a certain reliability. Thus, agreement here isn't simply by chance. In the population, scores from these raters correlated at 0.90. The score scale ranged from 0 to 6 points, with means set to 4 and 3 points for the raters 1 and 2, and SD of 1.5 for both. We'll refer to these as essay scores, much like the essay scores on the analytical writing section of the GRE. Scores were also dichotomized around a hypothetical cut score of 3, resulting in either a "Fail" or "Pass."

```
# Simulate essay scores from two raters with a population correlation of
# 0.90, and slightly different mean scores, with score range 0 to 6
# Note the capital T is an abbreviation for TRUE
essays <- rsim(100, rho = .9, meanx = 4, meany = 3, sdx = 1.5, sdy = 1.5,
  to.data.frame = T)
colnames(essays) <- c("r1", "r2")
essays <- round(setrange(essays, to = c(0, 6)))
# Use a cut off of greater than or equal to 3 to determine pass versus
# fail scores
# ifelse() takes a vector of TRUEs and FALSEs as its first argument, and
# returns here "Pass" for TRUE and "Fail" for FALSE
essays$f1 <- factor(ifelse(essays$r1 >= 3, "Pass", "Fail"))
essays$f2 <- factor(ifelse(essays$r2 >= 3, "Pass", "Fail"))
table(essays$f1, essays$f2)
##
##        Fail Pass
##   Fail   15    1
##   Pass   21   63
```

The upper left cell in the `table()` output above shows that for 15 individuals, the two raters both gave "Fail." In the lower right cell, the two raters both gave "Pass" 63 times. Together, these two totals represent the agreement in ratings, 78 . The other cells in the table contain disagreements, where one rater said "Pass" but the other said "Fail." Disagreements happened a total of 22 times. Based on these totals, what is the proportion agreement in the pass/fail ratings?

Table 5.3 shows the full crosstab of raw scores from each rater, with scores from rater 1 (`essays$r1`) in rows and rater 2 (`essays$r2`) in columns. The bunching of scores around the diagonal from upper left to lower right shows the tendency for agreement in scores.

Proportion agreement for the full rating scale, as shown in Table 5.3, can be calculated by summing the agreement frequencies within the diagonal elements of the table, and dividing by the total number of people.

```
# Pull the diagonal elements out of the crosstab with diag(), sum
# them, and divide by the number of people
sum(diag(table(essays$r1, essays$r2))) / nrow(essays)
## [1] 0.33
```

Finally, let's consider the impact of chance agreement between one of the hypothetical human raters and a monkey who randomly applies ratings, regardless of the performance that is demonstrated, as with a coin toss.

```
# Randomly sample from the vector c("Pass", "Fail"), nrow(essays) times,
# with replacement
# Without replacement, we'd only have 2 values to sample from
monkey <- sample(c("Pass", "Fail"), nrow(essays), replace = TRUE)
table(essays$f1, monkey)
##       monkey
##        Fail Pass
##   Fail    7    9
##   Pass   43   41
```

The results show that the hypothetical rater agrees with the monkey 48 percent of the time. Because we know that the monkey's ratings were completely random, we know that this proportion agreement is due entirely to chance.

5.4.2 Kappa agreement

Proportion agreement is useful, but because it does not account for chance agreement, it should not be used as the only measure of interrater consistency. Kappa agreement is simply an adjusted form of proportion agreement that takes chance agreement into account.

Equation @ref{eq:kappa1} contains the formula for calculating kappa for two raters.

$$\kappa = \frac{P_o - P_c}{1 - P_c} \tag{5.9}$$

To obtain kappa, we first calculate the proportion of agreement, P_o, as we did with the proportion agreement. This is calculated as the total for agreement divided by the total number of people being rated. Next we calculate the chance agreement, P_c, which involves multiplying the row and column proportions (row and column totals divided by the total total) from the crosstab and then summing the result, as shown in Equation 5.10.

$$P_c = P_{first-row}P_{first-col} + P_{next-row}P_{next-col} + \cdots + P_{last-row}P_{last-col} \tag{5.10}$$

You do not need to commit Equations 5.9 and 5.10 to memory. Instead, they're included here to help you understand that kappa involves removing chance agreement from the observed agreement, and then dividing this observed non-chance agreement by the total possible non-chance agreement, that is, $1 - P_c$.

The denominator for the kappa equation contains the maximum possible agreement beyond chance, and the numerator contains the actual observed agreement beyond chance. So, the maximum possible kappa is 1.0. In theory, we shouldn't ever observe less agreement than than expected by chance, which means that kappa should never be negative. Technically it is possible to have kappa below 0. When kappa is below 0, it indicates that our observed agreement is below what we'd expect due to chance. Kappa should also be no larger than proportion agreement. In the example data, the proportion agreement decreased from 0.33 to 0.185 for kappa.

A weighted version of the kappa index is also available. Weighted kappa let us reduce the negative impact of partial disagreements relative to more extreme disagreements in scores, by taking into account the ordinal nature of a score scale. For example, in Table 5.3, notice that only the diagonal elements of the crosstab measure perfect agreement in scores, and all other elements measure disagreements, even the ones that are close together like 2 and 3. With weighted kappa, we can give less weight to these smaller disagreements and more weight to larger disagreements such as scores of 0 and 6 in the lower left and upper right of the table. This weighting ends up giving us a higher agreement estimate.

Here, we use the function `astudy()` from epmr to calculate proportion agreement, kappa, and weighted kappa indices. Weighted kappa gives us the highest estimate of agreement. Refer to the documentation for `astudy()` to see how the weights are calculated.

```
# Use the astudy() function from epmr to measure agreement
astudy(essays[, 1:2])
##     agree     kappa    wkappa
## 0.3300000 0.1850140 0.4746356
```

5.4.3 Pearson correlation

The Pearson correlation coefficient, introduced above for CTT reliability, improves upon agreement indices by accounting for the ordinal nature of ratings without the need for explicit weighting. The correlation tells us how consistent raters are in their rank orderings of individuals. Rank orderings that are closer to being in agreement are automatically given more weight when determining the overall consistency of scores.

The main limitation of the correlation coefficient is that it ignores *systematic differences* in ratings when focusing on their rank orders. This limitation has to do with the fact that correlations are oblivious to linear transformations of score scales. We can modify the mean or standard deviation of one or both variables being correlated and get the same result. So, if two raters provide consistently different ratings, for example, if one rater is more forgiving overall, the correlation coefficient can still be high as long as the rank ordering of individuals does not change.

This limitation is evident in our simulated essay scores, where rater 2 gave lower scores on average than rater 1. If we subtract 1 point from every score for rater 2, the scores across raters will be more similar, as shown in Figure 5.4, but we still get the same interrater reliability.

```
cor(essays$r1, essays$r2)
## [1] 0.8399124
dstudy(essays[, 1:2])
##
## Descriptive Study
##
##    mean median   sd   skew kurt min max   n na
## r1 3.89      4 1.41 -0.572 3.05   0   6 100  0
## r2 3.00      3 1.50 -0.125 2.57   0   6 100  0
cor(essays$r1, essays$r2 + 1)
## [1] 0.8399124
```

A systematic difference in scores can be visualized by a consistent vertical or horizontal shift in the points within a scatter plot. Figure 5.4 shows that as scores are shifted higher for rater 2, they are more consistent with rater 1 in an absolute sense, despite the fact that the underlying linear relationship remains unchanged.

```
ggplot(essays, aes(r1, r2)) +
  geom_point(position = position_jitter(w = .1, h = .1)) +
  geom_abline(col = "blue")
```

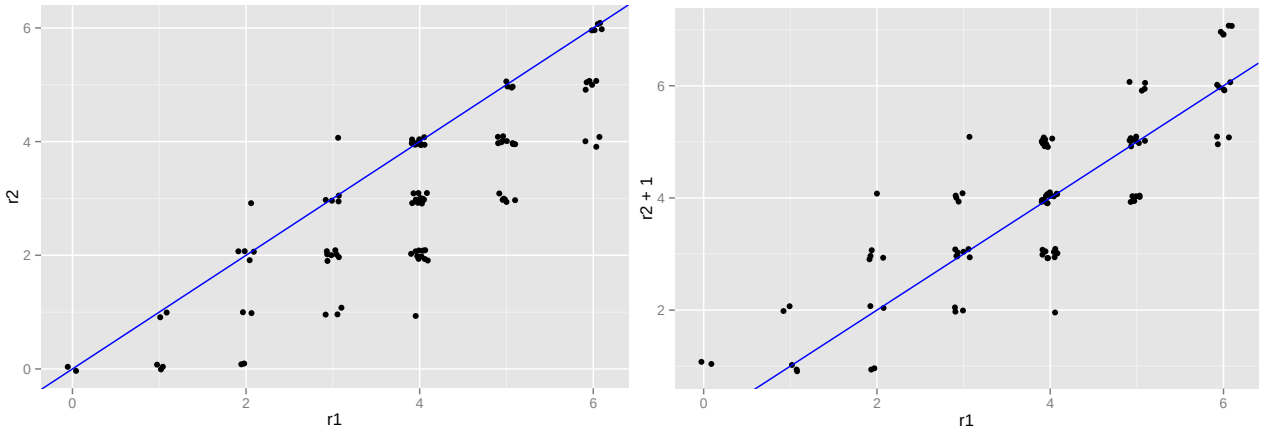


Figure 5.4: Scatter plots of simulated essay scores with a systematic difference around 0.5 points.

```
ggplot(essays, aes(r1, r2 + 1)) +
  geom_point(position = position_jitter(w = .1, h = .1)) +
  geom_abline(col = "blue")
```

Is it a problem that the correlation ignores systematic score differences? Can you think of any real-life situations where it wouldn't be cause for concern? A simple example is when awarding scholarships or giving other types of awards or recognition. In these cases consistent rank ordering is key and systematic differences are less important because the purpose of the ranking is to identify the top candidate. There is no absolute scale on which subjects are rated. Instead, they are rated in comparison to one another. As a result, a systematic difference in ratings would technically not matter.

5.5 Generalizability theory

Generalizability (G) theory addresses the limitations of other measures of reliability by providing a framework wherein systematic score differences can be accounted for, as well as the interactions that may take place in more complex reliability study designs. The G theory framework builds on the CTT model.

A brief introduction to G theory is given here, with a discussion of some key considerations in designing a G study and interpreting results. Resources for learning more include an introductory didactic paper by Brennan (1992), and textbooks by Shavelson and Webb (1991) and Brennan (2001).

5.5.1 The model

Recall from Equation 5.1 that in CTT the observed total score X is separated into a simple sum of the true score T and error E. Given the assumptions of the model, a similar separation works with the total score variance:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \tag{5.11}$$

Because observed variance is simply the sum of true and error variances, we can rewrite the reliability coefficient in Equation 5.3 entirely in terms of true scores and error scores:

$$r = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}. \tag{5.12}$$

Reliability in G theory, as in CTT, is considered to be the proportion of reliable score variance σ_T^2 to reliable plus unreliable variance $\sigma_T^2 + \sigma_E^2$, that is, all score variance σ_X^2.

In contrast to CTT, G theory breaks down the reliable variability into smaller pieces based on *facets* in the study design. Each facet is a feature of our data collection that we expect will lead to estimable variability in our total score X. In CTT, there is only one facet, people. But G theory also allows us to account for reliable variability due to raters, among other things. As a result, Equations 5.1 and 5.11 are expanded or generalized in 5.13 and 5.14 so that the true score components are expressed as a combination of components for people P and raters R.

$$X = P + R + E \tag{5.13}$$

$$\sigma_X^2 = \sigma_P^2 + \sigma_R^2 + \sigma_E^2 \tag{5.14}$$

5.5.2 Estimating generalizability

The breakdown of T into P and R in Equations 5.13 and 5.14 allows us to estimate reliability with what is referred to as a generalizability coefficient g:

$$g = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_R^2 + \sigma_E^2}. \tag{5.15}$$

The G coefficient is an extension of CTT reliability in Equation 5.3, where the reliable variance of interest, formerly σ_T^2, now comes from the people being measured, σ_P^2. Reliable variance due to raters is then separated out in the denominator, and it is technically treated as a systematic component of our error scores. Note that, if there are no raters involved in creating observed scores, R disappears, T is captured entirely by P, and Equation 5.15 becomes 5.3.

Generalizability coefficients can be estimated in a variety of ways. The epmr package includes functions for estimating a few different types of g using multilevel modeling via the lme4 package (Bates et al. 2015).

5.5.3 Applications of the model

If all of these equations are making your eyes cross, you should return to the question we asked earlier: why would a person score differently on a test from one administration to the next? The answer: because of differences in the person, because of differences in the rating process, and because of error. The goal of G theory is to compare the effects of these different sources on the variability in scores.

The definition of g in Equation 5.15 is a simple example of how we could estimate reliability in a person by rater study design. As it breaks down score variance into multiple reliable components, g can also be used in more complex designs, where we would expect some other facet of the data collection to lead to reliable variability in the scores. For example, if we administer two essays on two occasions, and at each occasion the same two raters provide a score for each subject, we have a fully crossed design with three facets: raters, tasks (i.e., essays), and occasions (e.g., Goodwin 2001). When developing or evaluating a test, we should consider the study design used to estimate reliability, and whether or not the appropriate facets are used.

In addition to choosing the facets in our reliability study, there are three other key considerations when estimating reliability with g. The first consideration is whether we need to make *absolute* or *relative* score interpretations. Absolute score interpretations account for systematic differences in scores, whereas relative score interpretations do not. As mentioned above, the correlation coefficient considers only the relative consistency of scores, and it is not impacted negatively by systematic, that is, absolute, differences in scores. With g, we may decide to adjust our estimate of reliability downward to account for these differences. The absolute reliability estimate from G theory is sometimes called the dependability coefficient d.

The second consideration is the number of *levels* for each facet in our reliability study that will be present in operational administrations of the test. By default, reliability based on correlation is estimated for a single level of each facet. When we correlate scores on two test forms, the result is an estimate of the reliability of test itself, not some combination of two forms of the test. In the G theory framework, an extension of the Spearman-Brown formula can be used to predict increases or decreases in g based on increases or decreases in the number of levels for a facet. When Spearman-Brown was introduced in Equation 5.5, we considered changes in test length, where an increase in test length increased reliability. Now, we can also consider changes in the number of raters, or occasions, or any facet of our study. These predictions are referred to in the G theory literature as decision studies. The resulting g coefficients are labeled as *average* when more than one level is involved for a facet, and *single* otherwise.

The third consideration is the type of sampling used for each facet in our study, resulting in either *random* or *fixed* effects. Facets can be treated as either random samples from a theoretically infinite population, or as fixed, that is, not sampled from any larger population. Thus far, we have by default treated all facets as random effects. The people in our reliability study represent a broader population of people, as do raters. Fixed effects would be appropriate for facets with units or levels that will be used consistently and exclusively across operational administrations of a test. For example, in a study design with two essay questions, scored by two raters, we would treat the essay questions as fixed if we never intend to utilize other questions besides these two. Our questions are not samples from a population of potential essay questions. If instead our questions are simply two of many possible ones that could have been written or selected, it would be more appropriate to treat this facet in our study as a random effect. Fixing effects will increase reliability, as the error inherent in random sampling is eliminated from g. Specifying random effects will then reduce reliability, as the additional error from random sampling is accounted for by g.

The `gstudy()` function in the epmr package estimates variances and relative, absolute, single, and average g coefficients for reliability study designs with multiple facets, including fixed and random effects. Here, we're simulating essay scores for a third rater in addition to the other two from the previous examples, and we're examining the interrater reliability of this hypothetical testing example for all three raters at once. Without G theory, correlations or agreement indices would need to be computed for each pair of raters.

```
essays$r3 <- rsim(100, rho = .9, x = essays$r2,
  meany = 3.5, sdy = 1.25)$y
essays$r3 <- round(setrange(essays$r3, to = c(0, 6)))
gstudy(essays[, c("r1", "r2", "r3")])
##
## Generalizability Study
##
##
## Call:
## gstudy.merMod(x = mer)
##
## Model Formula:
## score ~ 1 + (1 | person) + (1 | rater)
##
## Reliability:
##                         g    sem
## Relative Average 0.9340 0.3431
## Relative Single  0.8251 0.5942
## Absolute Average 0.8998 0.4308
## Absolute Single  0.7495 0.7461
##
## Variance components:
##          variance  n1 n2
## person     1.6657 100  1
## rater      0.2036   3  3
```

```
## residual   0.3531  NA NA
##
## Decision n:
## person  rater
##    100      3
```

When given a `data.frame` object such as `essays`, `gstudy()` automatically assumes a design with one facet, comprising the columns of the data, where the units of observation are in the rows. Both rows and columns are treated as random effects. To estimate fixed effects for one or more facets, the formula interface for the function must instead be used. See the `gstudy()` documentation for details.

Our simulated ratings across three essays produce a relative average g of 0.934. This is interpreted as the consistency of scores that we'd expect in an operational administration of this test where an average essay score is taken over three raters. The raters providing scores are assumed to be sampled from a population of raters, and any systematic score differences over raters are ignored. The relative single g of 0.825 is interpreted as the consistency of scores that we'd expect if we decided to reduce the scoring process to just one rater, rather than all three. This single rater is still assumed to be sampled from a population, and systematic differences in scores are still ignored. The absolute average g 0.900 and absolute single g 0.750 are interpreted in the same way, but systematic score differences are taken into account, and the reliability estimates are thus lower than the relative ones.

5.6 Summary

This chapter provided an overview of reliability within the frameworks of CTT, for items and test forms, and G theory, for reliability study designs with multiple facets. After a general definition of reliability in terms of consistency in scores, the CTT model was introduced, and two commonly used indices of CTT reliability were discussed: correlation and coefficient alpha. Reliability was then presented as it relates to consistency of scores from raters. Interrater agreement indices were compared, along with the correlation coefficient, and an introduction to G theory was given.

5.6.1 Learning objectives

CTT reliability

1. Define reliability, including potential sources of reliability and unreliability in measurement, using examples.
2. Describe the simplifying assumptions of the classical test theory (CTT) model and how they are used to obtain true scores and reliability.
3. Identify the components of the CTT model (X, T, and E) and describe how they relate to one another, using examples.
4. Describe the difference between systematic and random error, including examples of each.
5. Explain the relationship between the reliability coefficient and standard error of measurement, and identify how the two are distinguished in practice.
6. Calculate the standard error of measurement and describe it conceptually.
7. Compare and contrast the three main ways of assessing reliability, test-retest, parallel-forms, and internal consistency, using examples, and identify appropriate applications of each.
8. Compare and contrast the four reliability study designs, based on 1 to 2 test forms and 1 to 2 testing occasions, in terms of the sources of error that each design accounts for, and identify appropriate applications of each.
9. Use the Spearman-Brown formula to predict change in reliability.
10. Describe the formula for coefficient alpha, the assumptions it is based on, and what factors impact it as an estimate of reliability.

11. Estimate different forms of reliability using statistical software, and interpret the results.
12. Describe factors related to the test, the test administration, and the examinees, that affect reliability.

Interrater reliability

13. Describe the purpose of measuring interrater reliability, and how interrater reliability differs from traditional reliability.
14. Describe the difference between interrater agreement and interrater reliability, using examples.
15. Calculate and interpret indices of interrater agreement and reliability, including proportion agreement, Kappa, Pearson correlation, and intraclass correlation.
16. Identify appropriate uses of each interrater index, including the benefits and drawbacks of each.
17. Describe the three main considerations involved in using intraclass correlations.

5.6.2 Exercises

1. Explain the CTT model and its assumptions using the archery example presented at the beginning of the chapter.
2. Use the R object `scores` to find the average variability in `x1` for a given value on `t`. How does this compare to the SEM?
3. Use `PISA09` to calculate split-half reliabilities with different combinations of reading items in each half. Then, use these in the Spearman-Brown formula to estimate reliabilities for the full-length test. Why do the results differ?
4. Suppose you want to reduce the SEM for a final exam in a course you are teaching. Identify three sources of measurement error that could contribute to the SEM, and three that could not. Then, consider strategies for reducing error from these sources.
5. Estimate the internal consistency reliability for the attitude toward school scale. Remember to reverse code items as needed.
6. Dr. Phil is developing a measure of relationship quality to be used in counseling settings with couples. He intends to administer the measure to couples multiple times over a series of counseling sessions. Describe an appropriate study design for examining the reliability of this measure.
7. More and more TV shows lately seem to involve people performing some talent on stage and then being critiqued by a panel of judges, one of whom is British. Describe the "true score" for a performer in this scenario, and identify sources of measurement error that could result from the judging process, including both systematic and random sources of error.
8. With proportion agreement, consider the following questions. When would we expect to see 0% or nearly 0% agreement, if ever? What would the counts in the table look like if there were 0% agreement? When would we expect to see 100% or nearly 100% agreement, if ever? What would the counts in the table look like if there were 100% agreement?
9. What is the maximum possible value for kappa? And what would we expect the minimum possible value to be?
10. Given the strengths and limitations of correlation as a measure of interrater reliability, with what type of score referencing is this measure of reliability most appropriate?
11. Compare the interrater agreement indices with interrater reliability based on Pearson correlation. What makes the correlation coefficient useful with interval data? What does it tell us, or what does it do, that an agreement index does not?
12. Describe testing examples where different combinations of the three ICC considerations are appropriate. For example, when would we want an ICC that captures reliability for an average across 2 raters, where raters are a random effect, ignoring systematic differences?

Chapter 6

Item Analysis

Chapter 5.3.1 covered topics that rely on statistical analyses of data from educational and psychological measurements. These analyses are used to examine the relationships among scores on one or more test forms, in reliability, and scores based on ratings from two or more judges, in interrater reliability. Aside from coefficient alpha, all of the statistical analyses introduced so far focus on composite scores. Item analysis focuses instead on statistical analysis of the items themselves that make up these composites.

As discussed in Chapter 4, test items make up the most basic building blocks of an assessment instrument. Item analysis lets us investigate the quality of these individual building blocks, including in terms of how well they contribute to the whole and improve the validity of our measurement.

This chapter extends concepts from Chapters 2 and 5.3.1 to analysis of item performance within a CTT framework. The chapter begins with an overview of item analysis, including some general guidelines for preparing for an item analysis, entering data, and assigning score values to individual items. Some commonly used item statistics are then introduced and demonstrated. Finally, two additional item-level analyses are discussed, differential item functioning analysis and distractor analysis.

```
# R setup for this chapter
# Required packages are assumed to be installed - see Chapter 1
library("epmr")
library("ggplot2")
# Functions we'll use in this chapter
# paste0(), rowSums(), and data.frame() for prepping data
# rsim() and setrange() from epmr to simulate and modify scores
# ggplot(), aes(), geom_point(), and geom_abline() for plotting
```

6.1 Preparing for item analysis

6.1.1 Item quality

As noted above, item analysis lets us examine the quality of individual test items. Information about individual item quality can help us determine whether or not an item is measuring the content and construct that it was written to measure, and whether or not it is doing so at the appropriate ability level. Because we are discussing item analysis here in the context of CTT, we'll assume that there is a single construct of interest, perhaps being assessed across multiple related content areas, and that individual items can contribute or detract from our measurement of that construct by limiting or introducing *construct irrelevant variance* in the form of *bias* and *random measurement error*.

Bias represents a systematic error with an influence on item performance that can be attributed to an interaction between examinees and some feature of the test. Bias in a test item leads examinees having a known background characteristic, aside from their ability, to perform better or worse on an item simply because of this background characteristic. For example, bias sometimes results from the use of scenarios or examples in an item that are more familiar to certain gender or ethnic groups. Differential familiarity with item content can make an item more relevant, engaging, and more easily understood, and can then lead to differential performance, even for examinees of the same ability level. We identify such item bias primarily by using measures of item difficulty and *differential item functioning* (DIF), discussed below and again in Chapter 7.

Bias in a test item indicates that the item is measuring some other construct besides the construct of interest, where systematic differences on the other construct are interpreted as meaningful differences on the construct of interest. The result is a negative impact on the validity of test scores and corresponding inferences and interpretations. Random measurement error on the other hand is not attributed to a specific identifiable source, such as a second construct. Instead, measurement error is inconsistency of measurement at the item level. An item that introduces measurement error detracts from the overall internal consistency of the measure, and this is detected in CTT, in part, using item analysis statistics.

6.1.2 Piloting

The goal in developing an instrument or scale is to identify bias and inconsistent measurement at the item level *prior to* administering a final version of our instrument. As we talk about item analysis, remember that the analysis itself is typically carried out in practice using pilot data. Pilot data are gathered prior to or while developing an instrument or scale. These data require at least a preliminary version of the educational or psychological measure. We've written some items for our measure, and we want to see how well they work.

Ferketich (1991) and others recommend that the initial pilot "pool" of candidate test items should be at least twice as large as the final number of items needed. So, if you're dreaming up a test with 100 items on it, you should pilot at least 200 items. That may not be feasible, but it is a best-case scenario, and should at least be followed in large-scale testing. By collecting data on twice as many items as we intend to actually use, we're acknowledging that, despite our best efforts, many of our preliminary test items may either be low quality, for example, biased or internally inconsistent, and they may address different ability levels or content than intended.

Ferketich (1991) also recommends that data should be collected on at least 100 individuals from the population of interest. This too may not be feasible, however, it is essential if we hope to obtain results that will generalize to other samples of individuals. When our sample is not representative, for example, when it is a convenience sample or when it contains fewer than 100 people, our item analysis results must be interpreted with caution. This goes back to inferences made based on any type of statistic: small samples leads to erroneous results. Keep in mind that every statistic discussed here has a standard error and confidence interval associated with it, whether it is directly examined or not. Note also that bias and measurement error arise in addition to this standard error or sampling error, and we cannot identify bias in our test questions without representative data from our intended population. Thus, adequate sampling in the pilot study phase is critical.

The item analysis statistics discussed here are based on the CTT model of test performance. In Chapter 7 we'll discuss the more complex item response theory (IRT) and its applications in item analysis.

6.1.3 Data entry

After piloting a set of items, raw item responses are organized into a data frame with test takers in rows and items in columns. The `str()` function is used here to summarize the structure of the unscored items on the `PISA09` reading test. Each unscored item is coded in R as a factor with four to eight factor levels. Each factor level represents different information about a student's response.

```
# Recreate the item name index and use it to check the structure
# of the unscored reading items
# The strict.width argument is optional, making sure the results fit
# in the console window
ritems <- c("r414q02", "r414q11", "r414q06", "r414q09", "r452q03",
  "r452q04", "r452q06", "r452q07", "r458q01", "r458q07", "r458q04")
str(PISA09[, ritems], strict.width = "cut")
## 'data.frame':    44878 obs. of  11 variables:
##  $ r414q02: Factor w/ 7 levels "1","2","3","4",..: 2 1 4 1 1 2 2 2 3 2 ...
##  $ r414q11: Factor w/ 7 levels "1","2","3","4",..: 4 1 1 1 3 1 1 3 1 1 ...
##  $ r414q06: Factor w/ 5 levels "0","1","8","9",..: 1 4 2 1 2 2 2 4 1 1 ...
##  $ r414q09: Factor w/ 8 levels "1","2","3","4",..: 3 7 4 3 3 3 3 5 3 3 ...
##  $ r452q03: Factor w/ 5 levels "0","1","8","9",..: 1 4 1 1 1 2 2 1 1 1 ...
##  $ r452q04: Factor w/ 7 levels "1","2","3","4",..: 4 6 4 3 2 2 2 1 2 2 ...
##  $ r452q06: Factor w/ 4 levels "0","1","9","r": 1 3 2 1 2 2 2 2 1 2 ...
##  $ r452q07: Factor w/ 7 levels "1","2","3","4",..: 3 6 3 1 2 4 4 4 2 4 ...
##  $ r458q01: Factor w/ 7 levels "1","2","3","4",..: 4 4 4 3 4 4 3 4 3 3 ...
##  $ r458q07: Factor w/ 4 levels "0","1","9","r": 1 3 2 1 1 2 1 2 2 2 ...
##  $ r458q04: Factor w/ 7 levels "1","2","3","4",..: 2 3 2 3 2 2 2 3 3 4 ...
```

In addition to checking the structure of the data, it's good practice to run frequency tables on each variable. An example is shown below for a subset of `PISA09` reading items. The frequency distribution for each variable will reveal any data entry errors that resulted in incorrect codes. Frequency distributions should also match what we expect to see for correct and incorrect response patterns and missing data.

`PISA09` items that include a code or factor level of "0" are constructed-response items, scored by raters. The remaining factor levels for these CR items are coded "1" for full credit, "7" for not administered, "9" for missing, and "r" for not reached, where the student ran out of time before responding to the item. Selected-response items do not include a factor level of "0." Instead, they contain levels "1" through up to "5," which correspond to multiple-choice options one through five, and then codes of "7" for not administered, "8" for an ambiguous selected response, "9" for missing, and "r" again for not reached.

```
# Subsets of the reading item index for constructed and selected items
# Check frequency tables by item (hence the 2 in apply) for CR items
critems <- ritems[c(3, 5, 7, 10)]
sritems <- ritems[c(1:2, 4, 6, 8:9, 11)]
apply(PISA09[, critems], 2, table, exclude = NULL)
##      r414q06 r452q03 r452q06 r458q07
## 0       9620   33834   10584   12200
## 1      23934    5670   22422   25403
## 9      10179    4799   11058    6939
## r       1145     575     814     336
## <NA>       0       0       0       0
```

In the piloting and data collection processes, response codes or factor levels should be chosen carefully to represent all of the required response information. Responses should always be entered in a data set in their most raw form. Scoring should then happen after data entry, through the creation of new variables, whenever possible.

6.1.4 Scoring

In Chapter 2, which covered measurement, scales, and scoring, we briefly discussed the difference between dichotomous and polytomous scoring. Each involves the assignment of some value to each possible observed

response to an item. This value is taken to indicate a difference in the construct underlying our measure. For dichotomous items, we usually assign a score of 1 to a correct response, and a zero otherwise. Polytomous items involve responses that are correct to differing degrees, for example, 0, 1, and 2 for incorrect, somewhat correct, and completely correct.

In noncognitive testing, we replace "correctness" from the cognitive context with "amount" of the trait or attribute of interest. So, a dichotomous item might involve a yes/no response, where "yes" is taken to mean the construct is present in the individual, and it is given a score of 1, whereas "no" is taken to mean the construct is not present, and it is given a score of 0. Polytomous items then allow for different amounts of the construct to be present.

Although it seems standard to use dichotomous 0/1 scoring, and polytomous scoring of 0, 1, 2, ect., these values should not be taken for granted. The score assigned to a particular response determines how much a given item will contribute to any composite score that is later calculated across items. In educational testing, the typical scoring schemes are popular because they are simple. Other scoring schemes could also be used to given certain items more or less weight when calculating the total.

For example, a polytomous item could be scored using partial credit, where incorrect is scored as 0, completely correct is given 1, and levels of correctness are assigned decimal values in between. In psychological testing, the center of the rating scale could be given a score of 0, and the tails could decrease and increase from there. For example, if a rating scale is used to measure levels of agreement, 0 could be assigned to a "neutral" rating, and -2 and -1 might correspond to "strongly disagree" and "disagree," with 1 and 2 corresponding to "agree" and "strongly agree." Changing the values assigned to item responses in this way can help improve the interpretation of summary results.

Scoring item responses also requires that direction, that is, decreases and increases, be given to the correctness or amount of trait involved. Thus, at the item level, we are at least using an ordinal scale. In cognitive testing, the direction is simple: increases in points correspond to increases in correctness. In psychological testing, reverse scoring may also be necessary.

`PISA09` contains examples of scoring for both educational and psychological measures. First, we'll check the scoring for the CR and SR reading items. A crosstab for the raw and scored versions of an item shows how each code was converted to a score. Note students not reaching an item, with an unscored factor level "r", were given an `NA` for their score.

```
# Indices for scored reading items
rsitems <- paste0(ritems, "s")
crsitems <- paste0(critems, "s")
srsitems <- paste0(sritems, "s")
# Tabulate unscored and scored responses for the first CR item
# exclude = NULL shows us NAs as well
# raw and scored are not arguments to table, but are used simply
# to give labels to the printed output
table(raw = PISA09[, critems[1]], scored = PISA09[, crsitems[1]],
  exclude = NULL)
##       scored
## raw        0     1  <NA>
##   0     9620     0     0
##   1        0 23934     0
##   8        0     0     0
##   9    10179     0     0
##   r        0     0  1145
##   <NA>     0     0     0
# Create the same type of table for the first SR item
```

For a psychological example, we revisit the attitude toward school items presented in Figure 2.2. In `PISA09`, these items were coded during data entry with values of 1 through 4 for "Strongly Disagree" to "Strongly

Table 6.1: Comparing Terms in Cognitive vs Noncognitive Item Analysis

General Term	Cognitive	Noncognitive
Construct	Ability	Trait
Levels on construct	Correct and incorrect	Keyed and unkeyed
Item performance	Difficulty	Endorsement

Agree." We could utilize these as the scored responses in the item analyses that follow. However, we first need to recode the two items that were worded in the opposite direction as the others. Then, higher scores on all four items will represent more positive attitudes toward school.

```
# Check the structure of raw attitude items
str(PISA09[, c("st33q01", "st33q02", "st33q03", "st33q04")])
## 'data.frame':    44878 obs. of  4 variables:
##  $ st33q01: num  3 3 2 1 2 2 2 3 2 3 ...
##  $ st33q02: num  2 2 1 1 2 2 2 2 1 2 ...
##  $ st33q03: num  2 1 3 3 3 3 1 3 1 3 ...
##  $ st33q04: num  3 3 4 3 3 3 3 2 1 3 ...
# Rescore two items
PISA09$st33q01r <- recode(PISA09$st33q01)
PISA09$st33q02r <- recode(PISA09$st33q02)
```

6.2 Traditional item statistics

Three statistics are commonly used to evaluate the items within a scale or test. These are item difficulty, discrimination, and alpha-if-item-deleted. Each is presented below with examples based on `PISA09`.

6.2.1 Item difficulty

Once we have established scoring schemes for each item in our test, and we have applied them to item-response data from a sample of individuals, we can utilize some basic descriptive statistics to examine item-level performance. The first statistic is item difficulty, or, how easy or difficult each item is for our sample. In cognitive testing, we talk about easiness and difficulty, where test takers can get an item correct to different degrees, depending on their ability or achievement. In noncognitive testing, we talk instead about endorsement or likelihood of choosing the keyed response on the item, where test takers are more or less likely to endorse an item, depending on their level on the trait. In the discussions that follow, ability and trait can be used interchangeably, as can correct/incorrect and keyed/unkeyed response, and difficulty and endorsement. See Table 6.1 for a summary of these terms.

In CTT, the item difficulty is simply the mean score for an item. For dichotomous 0/1 items, this mean is referred to as a p-value, since it represents the proportion of examinees getting the item correct or choosing the keyed response. With polytomous items, the mean is simply the average score. When testing noncognitive traits, the term p-value may still be used. However, instead of item difficulty we refer to endorsement of the item, with proportion correct instead becoming proportion endorsed.

Looking ahead to IRT, item difficulty will be estimated as the predicted mean ability required to have a 50% chance of getting the item correct or endorsing the item.

Here, we calculate p-values for the scored reading items, by item type. Item `PISA09$r452q03`, a CR item, stands out from the rest as having a very low p-value of 0.13. This tells us that only 13% of students who took this item got it right. The next lowest p-value was 0.37.

```
round(colMeans(PISA09[, crsitems], na.rm = T), 2)
## r414q06s r452q03s r452q06s r458q07s
##     0.55     0.13     0.51     0.57
round(colMeans(PISA09[, srsitems], na.rm = T), 2)
## r414q02s r414q11s r414q09s r452q04s r452q07s r458q01s r458q04s
##     0.49     0.37     0.65     0.65     0.48     0.56     0.59
```

For item `r452q03`, students read a short description of a scene from *The Play's the Thing*, shown in Appendix A. The question then is, "What were the characters in the play doing just before the curtain went up?" This question is difficult, in part, because the word "curtain" is not used in the scene. So, the test taker must infer that the phrase "curtain went up" refers to the start of a play. The question is also difficult because the actors in this play are themselves pretending to be in a play. For additional details on the item and the rubric used in scoring, see Appendix A.

Although difficult questions may be frustrating for students, sometimes they're necessary. Difficult or easy items, or items that are difficult or easy to endorse, may be required given the purpose of the test. Recall that the purpose of a test describes: the construct, what we're measuring; the population, with whom the construct is being measured; and the application or intended use of scores. Some test purposes can only be met by including some very difficult or very easy items. PISA, for example, is intended to measure students along a continuum of reading ability. Without difficult questions, more able students would not be measured as accurately. On the other hand, a test may be intended to measure lower level reading skills, which many students have already mastered. In this case, items with high p-values would be expected. Without them, low ability students, who are integral to the test purpose, would not be able to answer any items correctly.

This same argument applies to noncognitive testing. To measure the full range of a clinical disorder, personality trait, or attitude, we need items that can be endorsed by individuals all along the continuum for our construct. Consider again the attitude toward school scale. All four items have mean scores above 2. On average, students agree with these attitude items, after reverse coding the first two, more often than not. If we recode scores to be dichotomous, with disagreement as 0 and agreement as 1, we can get p-values for each item as well. These p-values are interpreted as agreement rates. They tell us that at least 70% of students agreed with each attitude statement.

```
# Index for attitude toward school items, with the first two items
# recoded
atsitems <- c("st33q01r", "st33q02r", "st33q03", "st33q04")
# Check mean scores
round(colMeans(PISA09[, atsitems], na.rm = T), 2)
## st33q01r st33q02r  st33q03  st33q04
##     3.04     3.38     2.88     3.26
# Convert polytomous to dichotomous, with any disagreement coded as 0
# and any agreement coded as 1
ats <- apply(PISA09[, atsitems], 2, recode,
  list("0" = 1:2, "1" = 3:4))
round(colMeans(ats, na.rm = T), 2)
## st33q01r st33q02r  st33q03  st33q04
##     0.78     0.92     0.77     0.89
```

Given their high means and p-values, we might conclude that these items are not adequately measuring students with negative attitudes toward school, assuming such students exist. Perhaps if an item were worded differently or were written to ask about another aspect of schooling, such as the value of homework, more negative attitudes would emerge. On the other hand, it could be that students participating in PISA really do have overall positive attitudes toward school, and regardless of the question they will tend to have high scores.

This brings us to one of the major limitations of CTT item analysis: the item statistics we compute are dependent on our sample of test takers. For example, we assume a low p-value indicates the item was difficult,

but it may have simply been difficult for individuals in our sample. What if our sample happens to be lower on the construct than the broader population? Then, the items would tend to have lower means and p-values. If administered to a sample higher on the construct, item means would be expected to increase. Thus, the difficulty of an item is dependent on the ability of the individuals taking it.

Because we estimate item difficulty and other item analysis statistics without accounting for the ability or trait levels of individuals in our sample, we can never be sure of how sample-dependent our results really are. This sample dependence in CTT will be addressed in IRT.

6.2.2 Item discrimination

Whereas item difficulty tell us the mean level of performance on an item, across everyone taking the item, *item discrimination* tells us how item difficulty changes for individuals of different abilities. Discrimination extends item difficulty by describing mean item performance in relation to individuals' levels of the construct. Highly discriminating cognitive items are easier for high ability students, but more difficult for low ability students. Highly discriminating noncognitive items are endorsed less frequently by test takers low on the trait, but more frequently by test takers high on the trait. In either case, a discriminating item is able to identify levels on the construct of interest, because scores on the item itself are positively related to the construct.

Item discrimination is measured by comparing performance on an item for different groups of people, where groups are defined based on some measure of the construct. In the early days of item analysis, these groups were simply defined as "high" and "low" using a cutoff on the construct to distinguish the two. If we knew the true abilities for a group of test takers, and we split them into two ability groups, we could calculate and compare p-values for a given item for each group. If an item were highly discriminating, we would expect it to have a higher p-value in the high ability group than in the low ability group. We would expect the discrepancy in p-values to be large. On the other hand, for an item that doesn't discriminate well, the discrepancy between p-values would be small.

```
# Get total reading scores and check descriptives
PISA09$rtotal <- rowSums(PISA09[, rsitems])
dstudy(PISA09$rtotal)
##
## Descriptive Study
##
##   mean median   sd   skew kurt min max     n na
## 1 5.57      6 2.86 -0.106    2   0  11 43628  0
# Compare CR item p-values for students below vs above the
# median total score
round(colMeans(PISA09[PISA09$rtotal <= 6, crsitems],
  na.rm = T), 2)
## r414q06s r452q03s r452q06s r458q07s
##     0.32     0.03     0.28     0.39
round(colMeans(PISA09[PISA09$rtotal > 6, crsitems],
  na.rm = T), 2)
## r414q06s r452q03s r452q06s r458q07s
##     0.87     0.27     0.84     0.83
```

Although calculating p-values for different groups of individuals is still a useful approach to examining item discrimination, we lose information when we reduce scores on our construct to categories such as "high" and "low." Item discrimination is more often estimated using the correlation between item responses and construct scores. In the absence of scores on the construct, total scores are typically used as a proxy. The resulting correlation is referred to as an *item-total correlation* (ITC). When responses on the item are dichotomously scored, it is also sometimes called a point-biserial correlation.

Here, we take a subset of `PISA09` including CR item scores and the total reading score for German students. The correlation matrix for these five variables shows how scores on the items relate to one another, and to the total score. Relationships between items and the total are ITC estimates of item discrimination. The first item, with ITC of 0.7, is best able to discriminate between students of high and low ability.

```
# Create subset of data for German students, then reduce to complete
# data
pisadeu <- PISA09[PISA09$cnt == "DEU", c(crsitems, "rtotal")]
pisadeu <- pisadeu[complete.cases(pisadeu), ]
round(cor(pisadeu), 2)
##          r414q06s r452q03s r452q06s r458q07s rtotal
## r414q06s     1.00     0.27     0.42     0.39   0.70
## r452q03s     0.27     1.00     0.31     0.19   0.49
## r452q06s     0.42     0.31     1.00     0.31   0.65
## r458q07s     0.39     0.19     0.31     1.00   0.56
## rtotal       0.70     0.49     0.65     0.56   1.00
```

Note that when you correlate something with itself, the result should be a correlation of 1. When you correlate a component score, like an item, with a composite that includes that component, the correlation will increase simply because of the component in on both sides of the relationship. Correlations between item responses and total scores can be "corrected" for this spurious increase simply by excluding a given item when calculating the total. The result is referred to as a *corrected item-total correlation* (CITC). ITC and CITC are typically positively related with one another, and give relatively similar results. However, CITC is preferred, as it is considered more a conservative and more accurate estimate of discrimination.

Figure 6.1 contains scatter plots for two CR reading items from `PISA09`, items `r414q06s` and `r452q03s`. On the x-axis in each plot are total scores across all reading items, and on the y-axis are the scored item responses for each item. These plots help us visualize both item difficulty and discrimination. Difficulty is the amount of data horizontally aligned with 0, compared to 1, on the y-axis. More data points at 0 indicate more students getting the item wrong. Discrimination is then the bunching of data points at the low end of the x-axis for 0, and at the high end for 1.

```
# Scatter plots for visualizing item discrimination
ggplot(pisadeu, aes(rtotal, factor(r414q06s))) +
  geom_point(position = position_jitter(w = 0.2, h = 0.2))
ggplot(pisadeu, aes(rtotal, factor(r452q03s))) +
  geom_point(position = position_jitter(w = 0.2, h = 0.2))
```

Suppose you had to guess a student's reading ability based only on their score from a single item. Which of the two items shown in Figure 6.1 would best support your guess? Here's a hint: it's not item `r452q03`. Notice how students who got item `r452q03` wrong have `rtotal` scores that span almost the entire score scale? People with nearly perfect scores still got this item wrong. On the other hand, item `r414q06` shows a faster tapering off of students getting the item wrong as total scores increase, with a larger bunching of students of high ability scoring correct on the item. So, item `r414q06`, has a higher discrimination, and gives us more information about the construct than item `r452q03`.

Next, we calculate the ITC and CITC "by hand" for the first attitude toward school item, which was reverse coded as `st33q01r`. There is a sizable difference between the ITC and the CITC for this item, likely because the scale is so short to begin with. By removing the item from the total score, we reduce our scale length by 25%, and, presumably, our total score becomes that much less representative of the construct. Discrimination for the remaining attitude items will be examined later.

```
# Caculate ITC and CITC by hand for one of the attitude toward
# school items
```

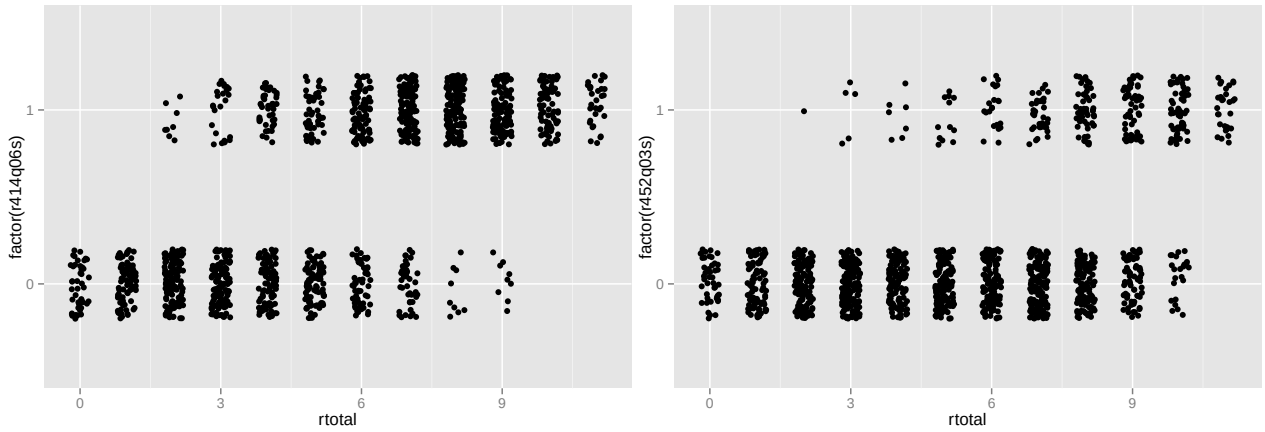


Figure 6.1: Scatter plots showing the relationship between total scores on the x-axis with dichotomous item response on two PISA items on the y-axis.

```
PISA09$atstotal <- rowSums(PISA09[, atsitems])
cor(PISA09$atstotal, PISA09$st33q01r, use = "c")
## [1] 0.7325649
cor(PISA09$atstotal - PISA09$st33q01r,
  PISA09$st33q01r, use = "c")
## [1] 0.4659522
```

Although there are no clear guidelines on acceptable or ideal levels of discrimination, 0.30 is sometimes used as a minimum. This minimum can decrease to 0.20 or 0.15 in lower-stakes settings where other constructs like motivation are expected to impact the quality of our data. Typically, the higher the discrimination, the better. However, when correlations between item scores and construct scores exceed 0.90 and approach 1.00, we should question how distinct our item really is from the construct. Items that correlate too strongly with the construct could be considered redundant and unnecessary.

6.2.3 Internal consistency

The last item analysis statistic we'll consider here indexes how individual items impact the overall internal consistency reliability of the scale. Internal consistency is estimated via coefficient alpha, introduced in Chapter 5.3.1. Alpha tells us how well our items work together as a set. "Working together" refers to how consistently the item responses change, overall, in similar ways. A high coefficient alpha tells us that people tend to respond in similar ways from one item to the next. If coefficient alpha were perfectly 1.00, we would know that each person responded in exactly the same rank-ordered way across all items. An item's contribution to internal consistency is measured by estimating alpha with that item removed from the set. The result is a statistic called alpha-if-item-deleted (AID).

AID answers the question, what happens to the internal consistency of a set of items with a given item is removed from the set? Because it involves the removal of an item, higher AID indicates a potential increase in internal consistency when an item is removed. Thus, when it is retained, the item is actually detracting from the internal consistency of the scale. Items that detract from the internal consistency should be considered for removal.

To clarify, it is bad news for an item if the AID is higher than the overall alpha for the full scale. It is good news for an item if AID is lower than alpha for the scale.

The `istudy()` function in the epmr package estimates AID, along with the other item statistics presented so far. AID is shown in the last column of the output. For the `PISA09` reading items, the overall alpha is 0.7596,

which is an acceptable level of internal consistency for a low-stakes measure like this (see Table 5.1). The AID results then tell us that alpha never increases beyond its original level after removing individual items, so, good news. Instead, alpha decreases to different degrees when items are removed. The lowest AID is 0.7251 for item `r414q06`. Removal of this item results in the largest decrease in internal consistency.

```
istudy(PISA09[, rsitems])
##
## Scored Item Study
##
## Alpha: 0.7596
##
## Item statistics:
##              m    sd     n   na   itc  citc   aid
## r414q02s 0.494 0.500 43958  920 0.551 0.411 0.741
## r414q11s 0.375 0.484 43821 1057 0.455 0.306 0.754
## r414q06s 0.547 0.498 43733 1145 0.652 0.533 0.725
## r414q09s 0.653 0.476 43628 1250 0.518 0.380 0.745
## r452q03s 0.128 0.334 44303  575 0.406 0.303 0.753
## r452q04s 0.647 0.478 44098  780 0.547 0.413 0.741
## r452q06s 0.509 0.500 44064  814 0.637 0.515 0.728
## r452q07s 0.482 0.500 43979  899 0.590 0.458 0.735
## r458q01s 0.556 0.497 44609  269 0.505 0.360 0.748
## r458q07s 0.570 0.495 44542  336 0.553 0.416 0.741
## r458q04s 0.590 0.492 44512  366 0.522 0.381 0.745
```

Note that discrimination and AID are typically positively related with one another. Discriminating items tend also to contribute to internal consistency. However, these two item statistics technically measure different things, and they need not correspond to one another. Thus, both should be considered when evaluating items in practice.

6.2.4 Item analysis applications

Now that we've covered the three major item analysis statistics, difficulty, discrimination, and contribution to internal consistency, we need to examine how they're used together to build a set of items. All of the scales in `PISA09` have already gone through extensive piloting and item analysis, so we'll work with a hypothetical scenario to make things more interesting.

Suppose we needed to identify a brief but effective subset of `PISA09` reading items for students in Hong Kong. The items will be used in a low-stakes setting where practical constraints limit us to only eight items, so long as those eight items maintain an internal consistency reliability at or above 0.60. First, let's use the Spearman-Brown formula to predict how low reliability would be expected to drop if we reduced our test length from eleven to eight items.

```
# Subset of data for Hong Kong, scored reading items
pisahkg <- PISA09[PISA09$cnt == "HKG", rsitems]
# Spearman-Brown based on original alpha and new test length of 8 items
sbr(r = alpha(pisahkg), k = 8/11)
## [1] 0.6039339
```

Our new reliability is estimated to be 0.604, which, thankfully, meets our hypothetical requirements. Now, we can explore our item statistics to find items for removal. The item analysis results for the full set of eleven items show two items with CITC below 0.20. These are items `r414q11` and `r452q03`. These items also have AID above alpha for the full set, indicating that both are detracting from the internal consistency of the

measure, though only to a small degree. Finally, notice that these are also the most difficult items in the set, with means of 0.36 and 0.04 respectively. Only 4% of students in Hong Kong got item `r452q03` right.

```
istudy(pisahkg)
##
## Scored Item Study
##
## Alpha: 0.6771
##
## Item statistics:
##               m    sd    n na   itc  citc   aid
## r414q02s 0.5109 0.500 1474 10 0.496 0.322 0.657
## r414q11s 0.3553 0.479 1469 15 0.349 0.165 0.684
## r414q06s 0.6667 0.472 1467 17 0.596 0.452 0.633
## r414q09s 0.7019 0.458 1466 18 0.439 0.273 0.666
## r452q03s 0.0405 0.197 1481  3 0.233 0.156 0.678
## r452q04s 0.6295 0.483 1479  5 0.541 0.382 0.645
## r452q06s 0.6303 0.483 1477  7 0.586 0.435 0.636
## r452q07s 0.4830 0.500 1474 10 0.510 0.338 0.654
## r458q01s 0.5037 0.500 1481  3 0.481 0.305 0.659
## r458q07s 0.5922 0.492 1481  3 0.515 0.349 0.652
## r458q04s 0.6914 0.462 1481  3 0.538 0.386 0.645
```

Let's remove these two lower-quality items and check the new results. The means and SD should stay the same for this new item analysis. However, the remaining statistics, ITC, CITC, and AID, all depend on the full set, so we would expect them to change. The results indicate that all of our AID are below alpha for the full set of nine items. The CITC are acceptable for a low-stakes test. However, item item `r414q09` has the weakest discrimination, making it the best candidate for removal, all else equal.

```
istudy(pisahkg[, rsitems[-c(2, 5)]])
##
## Scored Item Study
##
## Alpha: 0.6866
##
## Item statistics:
##              m    sd    n na   itc  citc   aid
## r414q02s 0.511 0.500 1474 10 0.506 0.320 0.670
## r414q06s 0.667 0.472 1467 17 0.605 0.451 0.642
## r414q09s 0.702 0.458 1466 18 0.454 0.277 0.678
## r452q04s 0.629 0.483 1479  5 0.542 0.370 0.659
## r452q06s 0.630 0.483 1477  7 0.601 0.442 0.644
## r452q07s 0.483 0.500 1474 10 0.514 0.330 0.667
## r458q01s 0.504 0.500 1481  3 0.500 0.314 0.670
## r458q07s 0.592 0.492 1481  3 0.537 0.361 0.660
## r458q04s 0.691 0.462 1481  3 0.550 0.388 0.655
```

Note that the reading items included in `PISA09` were not developed to function as a reading scale. Instead, these are merely a sample of items from the study, the items with content that was made publicly available. Also, in practice, an item analysis will typically involve other considerations besides the statistics we are covering here, most importantly, content coverage. Before removing an item from a set, we should consider how the test outline will be impacted, and whether or not the balance of content is still appropriate.

6.3 Additional analyses

Whereas item analysis is useful for evaluating the quality of items and their contribution to a test or scale, other analyses are available for digging deeper into the strengths and weaknesses of items. Two commonly used analyses are option analysis, also called distractor analysis, and differential item functioning analysis.

6.3.1 Option analysis

So far, our item analyses have focused on scored item responses. However, data from unscored SR items can also provide useful information about test takers. An *option analysis* involves the examination of unscored item responses by ability or trait groupings for each option in a selected-response item. Relationships between the construct and response patters over keyed and unkeyed options can give us insights into whether or not response options are functioning as intended.

To conduct an option analysis, we simply calculate bivariate (two variables) frequency distributions for ability or trait levels (one variable) and response choice (the other variable). Response choice is already a nominal variable, and the ability or trait is usually converted to ordinal categories to simplify interpretation.

The `ostudy()` function from epmr takes a matrix of unscored item responses, along with some information about the construct (a grouping variable, a vector of ability or trait scores, or a vector containing the keys for each item) and returns crosstabs for each item that break down response choices by construct groupings. By default, the construct is categorized into three equal-interval groups, with labels "low", "mid", and "high."

```
# Item option study on all SR reading items
pisao <- ostudy(PISA09[, sritems], scores = PISA09$rtotal)
# Print frequency results for one item
pisao$counts$r458q04
##    groups
##       low   mid  high
##   1  2117   890   297
##   2  5039  9191 11608
##   3  5288  2806   855
##   4  2944  1226   266
##   8   224    73    18
##   9   658   110    18
##   r     0     0     0
```

With large samples, a crosstab of counts can be difficult to interpret. Instead, we can also view percentages by option, in rows, or by construct group, in columns. Here, we check the percentages by column, where each column should sum to 100%. These results tell us the percentage of students within each ability group that chose each option.

```
# Print option analysis percentages by column
# This item discriminates relatively well
pisao$colpct$r458q04
##    groups
##     low mid high
##   1  13   6    2
##   2  31  64   89
##   3  33  20    7
##   4  18   9    2
##   8   1   1    0
##   9   4   1    0
##   r   0   0    0
```

Consider the distribution over response choices for high ability students, in the last column of the crosstab for item `r458q04`. The large majority of high ability students, 89%, chose the second option on this item, which is also the correct or keyed option. On the other hand, low ability students, in the left column of the crosstab, are spread out across the different options, with only 31% choosing the correct option. This is what we hope to see for an item that discriminates well.

```
# Print option analysis percentages by column
# This item discriminates relatively less well
pisao$colpct$r414q11
##    groups
##     low mid high
##   1  44  51   29
##   2  25   8    2
##   3  16  35   67
##   4   9   5    2
##   8   1   0    0
##   9   5   1    0
##   r   0   0    0
```

Item `r414q11` doesn't discriminate as well as item `r458q04`, and this is reflected in the option analysis results. The majority of high ability students are still choosing the correct option. However, some are also pulled toward another, incorrect option, option 1. The majority of low ability students are choosing option 1, with fewer choosing the correct option.

In addition to telling us more about the discrimination of an item by option, these results also help us identify dysfunctional options, ones which do not provide us with any useful information about test takers. Do you see any options in item `r414q11` that don't function well? The incorrect options are 1, 2, and 4.

A dysfunctional option is one that fails to attract the students we expect to choose it. We expect high ability students to choose the correct option. On more difficult questions, we also expect some high ability students to choose one or more incorrect options. We typically expect a somewhat uniform distribution of choices for low ability students. If our incorrect options were written to capture common misunderstandings of students, we may expect them to be chosen by low ability students more often than the correct option. A dysfunction option is then one that is chosen by few low or medium ability students, or is chosen by high ability students more often than the correct option.

When an incorrect option is infrequently or never chosen by lower ability groups, it is often because the option is obviously incorrect. In item `r414q11`, option 4 is chosen by only 9% of low ability students and 5% of medium ability students. Reviewing the content of the item in Appendix A.1, the option (D) doesn't stand out as being problematic. It may have been better worded as "It proves the No argument." Then, it would have conformed with the wording in option 2 (B). However, the option appears to follow the remaining item writing guidelines.

6.3.2 Differential item functioning

A second supplemental item analysis, frequently conducted with large-scale and commercially available tests, is an analysis of differential item functioning or DIF. Consider a test question where students of the same ability or trait level respond differently based on certain demographic or background features pertaining to the examinee, but *not* relating directly to the construct. In an option analysis, we examine categorical frequency distributions for each response option by *ability groups*. In DIF, we examine these same categorical frequency distributions, but by different demographic groups, where all test takers in the analysis have the same level on the construct. The presence of DIF in a test item is evidence of potential bias, as, after controlling for the construct, demographic variables should not produce significant differences in test taker responses.

A variety of statistics, with R packages to calculate them, are available for DIF analysis in educational and psychological testing. We will not review them all here. Instead, we focus on the concept of DIF and how it is displayed in item level performance in general. We'll return to this discussion in Chapter 7, within IRT.

DIF is most often based on a statistical model that compares item difficulty or the mean performance on an item for two groups of test takers, after *controlling for* their ability levels. Here, *controlling for* refers to either statistical or experimental control. The point is that we in some way remove the effects of ability on mean performance per group, so that we can then examine any leftover performance differences. Testing for the significance of DIF can be done, for example, using IRT, logistic regression, or chi-square statistics. Once DIF is identified for an item, the item itself is examined for potential sources of differential performance by subgroup. The item is either rewritten or deleted.

6.4 Summary

This chapter provided an introduction to item analysis in cognitive and noncognitive testing, with some guidelines on collecting and scoring pilot data, and an overview of five types of statistics used to examine item level performance, including item difficulty, item discrimination, internal consistency, distractor analysis, and differential item functioning. These statistics are used together to identify items that contribute or detract from the quality of a measure.

Item analysis, as described in this chapter, is based on a CTT model of test performance. We have assumed that a single construct is being measured, and that item analysis results are based on a representative sample from our population of test takers. Chapter 7 builds on the concepts introduced here by extending them to the more complex but also more popular IRT model of test performance.

6.4.1 Learning objectives

1. Explain how item bias and measurement error negatively impact the quality of an item, and how item analysis, in general, can be used to address these issues.
2. Describe general guidelines for collecting pilot data for item analysis, including how following these guidelines can improve item analysis results.
3. Identify items that may have been keyed or scored incorrectly.
4. Recode variables to reverse their scoring or keyed direction.
5. Use the appropriate terms to describe the process of item analysis with cognitive vs noncognitive constructs.
6. Calculate and interpret item difficulties and compare items in terms of difficulty.
7. Calculate and interpret item discrimination indices, and describe what they represent and how they are used in item analysis.
8. Describe the relationship between item difficulty and item discrimination and identify the practical implications of this relationship.
9. Calculate and interpret alpha-if-item-removed.
10. Utilize item analysis to distinguish between items that function well in a set and items that do not.
11. Remove items from an item set to achieve a target level of reliability.
12. Evaluate selected-response options using distractor analysis.

6.4.2 Exercises

1. Explain why we should be cautious about interpreting item analysis results based on pilot data.
2. For an item with high discrimination, how should p-values on the item compare for two groups known to differ in their true mean abilities?
3. Why is discrimination usually lower for CITC as compared with ITC for a given item?

4. What features of certain response options, in terms of the item content itself, would make them stand out as problematic within a distractor analysis?
5. Explain how AID is used to identify items contributing to internal consistency.
6. Conduct an item analysis on the `PISA09` reading items for students in Great Britain (`PISA09$cnt == "GBR"`). Examine and interpret results for item difficulty, discrimination, and AID.
7. Conduct a distractor analysis on SR reading item `r414q09`, with an interpretation of results.

Chapter 7

Item Response Theory

> One could make a case that item response theory is the most important statistical method about which most of us know little or nothing.
> — David Kenny

Item response theory (IRT) is arguably one of the most influential developments in the field of educational and psychological measurement. IRT provides a foundation for statistical methods that are utilized in contexts such as test development, item analysis, equating, item banking, and computerized adaptive testing. Its applications also extend to the measurement of a variety of latent constructs in a variety of disciplines.

Given its role and influence in educational and psychological measurement, the topic of IRT has accumulated an extensive literature. Rather than cover every detail, this chapter gives a broad overview of IRT, with the intention of helping you understand key concepts and common applications. For comprehensive treatments of IRT, see de Ayala (2009) and Embretson and Reise (2000). For a comparison of CTT and IRT, see Hambleton and Jones (1993). Harvey and Hammer (1999) describes IRT specifically in the context of psychological testing.

This chapter begins with a comparison of IRT with classical test theory (CTT), including a discussion of strengths and weaknesses and some typical uses of each. Next, the traditional dichotomous IRT models are introduced with definitions of key terms and a comparison based on assumptions, benefits, limitations, and uses. Finally, details are provided on applications of IRT in item analysis, test development, item banking, and computer adaptive testing.

```
# R setup for this chapter
# Required packages are assumed to be installed - see Chapter 1
library("epmr")
library("ggplot2")
# Functions we'll use in this chapter
```

7.1 IRT versus CTT

Since its development in the 1950s and 1960s (Lord 1952; Rasch 1960), IRT has become the preferred statistical methodology for item analysis and test development. The success of IRT over its predecessor CTT comes primarily from the focus on IRT on the individual components that make up a test, that is, the items themselves. By modeling outcomes at the item level, rather than at the test level as in CTT, IRT is more complex but also more comprehensive in terms of the information it provides about test performance.

7.1.1 CTT review

As presented in Chapter 5.3.1, CTT gives us a model for the observed total score X. This model decomposes X into two parts, truth T and error E:

$$X = T + E. \tag{7.1}$$

The true score T is the construct we're intending to measure, and we assume it plays some systematic role in causing people to obtain observed scores on X. The error E is everything randomly unrelated to the construct we're intending to measure. Error also has a direct impact on X. From Chapter 6, two item statistics that come from CTT are the mean performance on a given item, referred to as the p-value for dichotomous items, and the (corrected) item-total correlation for an item. Before moving on, you should be familiar with these two statistics, item difficulty and item discrimination, how they are related, and what they tell us about the items in a test.

It should be apparent that CTT is a relatively simple model of test performance. The simplicity of the model brings up its main limitation: the score scale is dependent on the items in the test and the people taking the test. The results of CTT are said to be *sample dependent* because 1) any X, T, or E that you obtain for a particular test taker only has meaning within the test she or he took, and 2) any item difficulty or discrimination statistics you estimate only have meaning within a particular sample of test takers. So, the person parameters are dependent on the test we use, and the item parameters are dependent on the test takers.

For example, suppose an instructor gives the same final exam to a new classroom of students each semester. At the first administration, the CITC discrimination for one item is 0.08. That's low, and it suggests that there's a problem with the item. However, in the second administration of the same exam to another group of students, the same item is found to have a CITC of 0.52. Which of these discriminations is correct? According to CTT, they're both correct, for the sample with which they are calculated. In CTT there is technically no absolute item difficulty or discrimination that generalizes across samples or populations of examinees. The same goes for ability estimates. If two students take different final exams for the same course, each with different items but the same number of items, ability estimates will depend on the difficulty and quality of the respective exams. There is no absolute ability estimate that generalizes across samples of items. This is the main limitation of CTT: parameters that come from the model are sample and test dependent.

A second major limitation of CTT results from the fact that the model is specified using total scores. Because we rely on total scores in CTT, a given test only produces one estimate of reliability and, thus, one estimate of SEM, and these are assumed to be unchanging for all people taking the test. The measurement error we expect to see in scores would be the same regardless of level on the construct. This limitation is especially problematic when test items do not match the ability level of a particular group of people. For example, consider a comprehensive vocabulary test covering all of the words taught in a fourth grade classroom. The test is given to a group of students just starting fourth grade, and another group who just completed fourth grade and is starting fifth. Students who have had the opportunity to learn the test content should respond more reliably than students who have not. Yet, the test itself has a single reliability and SEM that would be used to estimate measurement error for any score. Thus, the second major limitation of CTT is that reliability and SEM are constant and do not depend on the construct.

7.1.2 Comparing with IRT

IRT addresses the limitations of CTT, the limitations of sample and test dependence and a single constant SEM. As in CTT, IRT also provides a model of test performance. However, the model is defined at the item level, meaning there is, essentially, a separate model equation for each item in the test. So, IRT involves multiple item score models, as opposed to a single total score model. When the assumptions of the model are met, IRT parameters are, in theory, sample and item *independent*. This means that a person should have the same ability estimate no matter which set of items she or he takes, assuming the items pertain to the same

test. And in IRT, a given item should have the same difficulty and discrimination no matter who is taking the test.

IRT also takes into account the difficulty of the items that a person responds to when estimating the person's ability or trait level. Although the construct estimate itself, in theory, does not depend on the items, the precision with which we estimate it does depend on the items taken. Estimates of the ability or trait are more precise when they're based on items that are close to a person's construct level. Precision decreases when there are mismatches between person construct and item difficulty. Thus, SEM in IRT can vary by the ability of the person and the characteristics of the items given.

The main limitation of IRT is that it is a complex model requiring much larger samples of people than would be needed to utilize CTT. Whereas in CTT the recommended minimum is 100 examinees for conducting an item analysis (see Chapter 6), in IRT, as many as 500 or 1000 examinees may be needed to obtain stable results, depending on the complexity of the chosen model.

```
# Prepping data for examples
# Create subset of data for Great Britain, then reduce to complete data
ritems <- c("r414q02", "r414q11", "r414q06", "r414q09", "r452q03",
  "r452q04", "r452q06", "r452q07", "r458q01", "r458q07", "r458q04")
rsitems <- paste0(ritems, "s")
PISA09$rtotal <- rowSums(PISA09[, rsitems])
pisagbr <- PISA09[PISA09$cnt == "GBR", c(rsitems, "rtotal")]
pisagbr <- pisagbr[complete.cases(pisagbr), ]
```

Another key difference between IRT and CTT has to do with the shape of the relationship that we estimate between item score and construct score. The CTT discrimination models a simple linear relationship between the two, whereas IRT models a curvilinear relationship between them. Recall from Chapter 6 that the discrimination for an item can be visualized within a scatter plot, with the construct on the x-axis and item scores on the y-axis. A strong positive item discrimination would be shown by points for incorrect scores bunching up at the bottom of the scale, and points for correct scores bunching up at the top. A line passing through these points would then have a positive slope. Because they're based on correlations, ITC and CITC discriminations are always represented by a straight line. See Figure 7.1 for an example based on `PISA09` reading item `r452q06s`.

```
# Get p-values conditional on rtotal
# tapply() applies a function to the first argument over subsets of data
# defined by the second argument
pvalues <- data.frame(rtotal = 0:11, p = tapply(pisagbr$r452q06s,
  pisagbr$rtotal, mean))
# Plot CTT discrimination over scatter plot of scored item responses
ggplot(pisagbr, aes(rtotal, r414q06s)) +
  geom_point(position = position_jitter(w = 0.1, h = 0.1)) +
  geom_smooth(method = "lm", fill = NA) +
  geom_point(aes(rtotal, p), data = pvalues, col = "green", size = 3)
```

In IRT, the relationship between item and construct is similar to CTT, but the line follows what's called a logistic curve. To demonstrate the usefulness of a logistic curve, we calculate a set of p-values for item `r452q06s` *conditional on the construct*. In Figure 7.1, the green points are p-values calculated within each group of people having the same total reading score. For example, the p-value on this item for students with a total score of 3 is about 0.20. As expected, people with lower totals have more incorrect than correct responses. As total scores increase, the number of people getting the item correct steadily increases. At a certain total score, around 5.5, we see roughly half the people get the item correct. Then, as we continue up the theta scale, more and more people get the item correct. IRT is used to capture the trend shown by the conditional p-values in Figure 7.1.

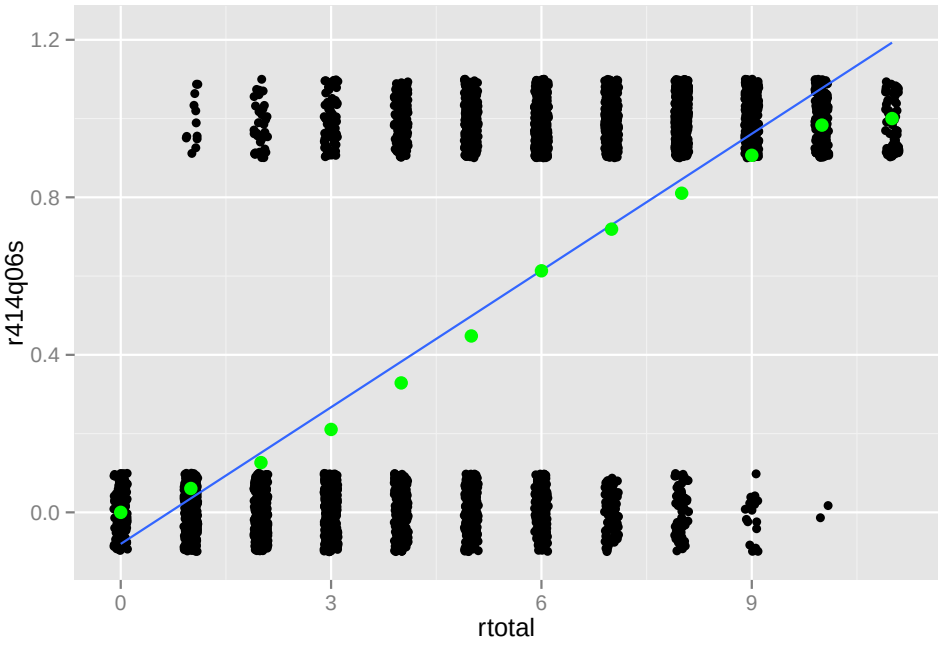


Figure 7.1: Scatter plots showing the relationship between total scores on the x-axis and scores from PISA item 'r452q06s' on the y-axis. Lines represent the relationship between the construct and item scores for CTT (straight) and IRT (curved).

7.2 Traditional IRT models

7.2.1 Terminology

Understanding IRT requires that we master some new terms. First, in IRT the underlying construct is referred to as theta or θ. Theta refers to the same construct we discussed in CTT, reliability, and our earlier measurement models. The underlying construct is the unobserved variable that we assume causes the behavior we observe in item responses. Different test takers are at different levels on the construct, and this results in them having different response patterns. In IRT we model these response patterns as a function of the construct theta. The theta scale is another name for the ability or trait scale.

Second, the dependent variable in IRT differs from CTT. The dependent variable is found on the left of the model, as in a regression or other statistical model. The dependent variable in the CTT model is the total observed score X. The IRT model instead has an item score as the dependent variable. The model then predicts the probability of a correct response for a given item, denoted $\Pr(X = 1)$.

Finally, in CTT we focus only on the person construct T within the model itself. The dependent variable X is modeled as a function of T, and whatever is left over via E. In IRT, we include the construct, now θ, along with parameters for the item that characterize its difficulty, discrimination, and lower-asymptote.

In the discussion that follows, we will frequently use the term *function*. A function is simply an equation that produces an output for each input we supply to it. The CTT model could be considered a function, as each T has a single corresponding X that is influenced, in part, by E. The IRT model also produces an output, but this output is now at the item level, and it is a probability of correct response for a given item. We can plug in θ, and get a prediction for the performance we'd expect on the item for that level on the construct. In IRT, this prediction of item performance depends on the item as well as the construct.

The IRT model for a given item has a special name in IRT. It's called the item response function (IRF), because it can be used to predict an item response. Each item has its own IRF. We can add up all the IRF in a test to obtain a test response function (TRF) that predicts not item scores but total scores on the test.

Two other functions, with corresponding equations, are also commonly used in IRT. One is the item information function (IIF). This gives us the predicted discriminatory power of an item across the theta scale. Remember that, as CTT discrimination can be visualized with a straight line, the discrimination in IRT can be visualized with a curve. This curve is flatter at some points than others, indicating the item is less discriminating at those points. In IRT, *information* tells us how the discriminating power of an item changes across the theta scale. IIF can also be added together to get an overall test information function (TIF).

The last IRT function we'll discuss here gives us the SEM for our test. This is called the test error function (TEF). As with all the other IRT functions, there is an equation that is used to estimate the function output, in this case, the standard error of measurement, for each point on the theta scale.

This overview of IRT terminology should help clarify the benefits of IRT over CTT. Recall that the main limitations of CTT are: 1) sample and test dependence, where our estimates of construct levels depend on the items in our test, and our estimates of item parameters depend on the construct levels for our sample of test takers; and 2) reliability and SEM that do not change depending on the construct. IRT addresses both of these limitations. The IRT model estimates the dependent variable of the model while accounting for both the construct and the properties of the item. As a result, estimates of ability or trait levels and item analysis statistics will be *sample and test independent*. This will be discussed further below. The IRT model also produces, via the TEF, measurement error estimates that vary by theta. Thus, the accuracy of a test depends on where the items are located on the construct.

Here's a recap of the key terms we'll be using throughout this chapter:

- Theta, θ, is our label for the construct measured for people.
- $\Pr(X = 1)$ is the probability of correct response, the outcome in the IRT model. Remember that X is now an item score, as opposed to a total score in CTT.
- The IRF is the visual representation of $\Pr(X = 1)$, showing us our predictions about how well people will do on an item based on theta.
- The logistic curve is the name for the shape we use to model performance via the IRF. It is a curve with certain properties, such as horizontal lower and upper asymptotes.
- The properties of the logistic curve are based on three item parameters, a, b, and c, which are the item discrimination, difficulty, and lower-asymptote, also known as the pseudo-guessing parameter.
- Functions are simply equations that produce a single output value for each value on the theta scale. IRT functions include the IRF, TRF, IIF, TIF, and TEF.
- Information refers to a summary of the discriminating power provided by an item or test.

7.2.2 The IRT models

We'll now use the terminology above to compare three traditional IRT models. Equation 7.2 contains what is referred to as the three-parameter IRT model, because it includes all three available item parameters. The model is usually labeled 3PL, for 3 parameter logistic. As noted above, in IRT we model the probably of correct response on a given item ($\Pr(X = 1)$) as a function of person ability (θ) and certain properties of the item itself, namely: a, how well the item discriminates between low and high ability examinees; b, how difficult the item is, or the construct level at which we'd expect people to have a probability $\Pr = 0.50$ of getting the keyed item response; and c, the lowest Pr that we'd expect to see on the item by chance alone.

$$\Pr(X = 1) = c + (1 - c)\frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}} \tag{7.2}$$

The a and b parameters should be familiar. They are the IRT versions of the item analysis statistics from CTT, presented in Chapter 5.3.1. The a parameter corresponds to ITC, where larger a indicate larger, better discrimination. The b parameter corresponds to the *opposite* of the p-value, where a low b indicates an easy item, and a high b indicates a difficult item; higher b require higher θ for a correct response. The c parameter should then be pretty intuitive if you think of its application to multiple-choice questions. When people low on the construct guess randomly on a multiple-choice item, the c parameter attempts to captures the chance of getting the item correct. In IRT, we acknowledge with the c parameter that the probability of correct response may never be zero. The smallest probability of correct response produced by Equation 7.2 will be c.

To better understand the IRF in Equation 7.2, focus first on the difference we take between theta and item difficulty, in $(\theta - b)$. Suppose we're using a cognitive test that measures some ability. If someone is high ability and taking a very easy item, with low b, we're going to get a large difference between theta and b. This large difference filters through the rest of the equation to give us a higher prediction of how well the person will do on the item. This difference is multiplied by the discrimination parameter, so that, if the item is highly discriminating, the difference between ability and difficulty is *magnified*. If the discrimination is low, for example, 0.50, the difference between ability and difficulty is cut in half before we use it to determine probability of correct response. The fractional part and the exponential term represented by e serve to make the straight line of ITC into a nice curve with a lower and upper asymptote at c and 1. Everything on the right side of Equation 7.2 is used to estimate the left side, that is, how well a person with a given ability would do on a given item.

Figure 7.2 contains IRF for five items with different item parameters. Let's examine the item with the IRF shown by the black line. This item would be considered the most difficult of this set, as it is located the furthest to the right. We only begin to predict that a person will get the item correct once we move past theta 0. The actual item difficulty, captured by the b parameter, is 3. This is found as the theta required to have a probability of keyed response of 0.50. This item also has the highest discrimination, as it is steeper than any other item. It is most useful for distinguishing between probabilities of correct response around theta 3, its b parameter; below and above this value, the item does not discriminate as well, as the IRF becomes more horizontal. Finally, this item appears to have a lower-asymptote of 0, suggesting test takers are likely not guessing on the item.

```
# Make up a, b, and c parameters for five items
# Get IRF using the rirf() function from epmr and plot
# rirf() will be demonstrated again later
ipar <- data.frame(a = c(2, 1, .5, 1, 1.5), b = c(3, 2, -.5, 0, -1),
  c = c(0, .2, .25, .1, .28), row.names = paste0("item", 1:5))
ggplot(rirf(ipar), aes(theta)) + scale_y_continuous("Pr(X)") +
  geom_line(aes(y = item1), col = 1) +
  geom_line(aes(y = item2), col = 2) +
  geom_line(aes(y = item3), col = 3) +
  geom_line(aes(y = item4), col = 4) +
  geom_line(aes(y = item5), col = 5)
```

Examine the remaining IRF in Figure 7.2. You should be able to compare the items in terms of easiness and difficulty, low and high discrimination, and low and high predicted probability of guessing correctly. Below are some specific questions and answers for comparing the items.

- Which item has the highest discrimination? Black, with the steepest slope.
- Which has the lowest discrimination? Green, with the shallowest slope.
- Which item is hardest, requiring the highest ability level, on average, to get it correct? Black, again, as it is furthest to the right.

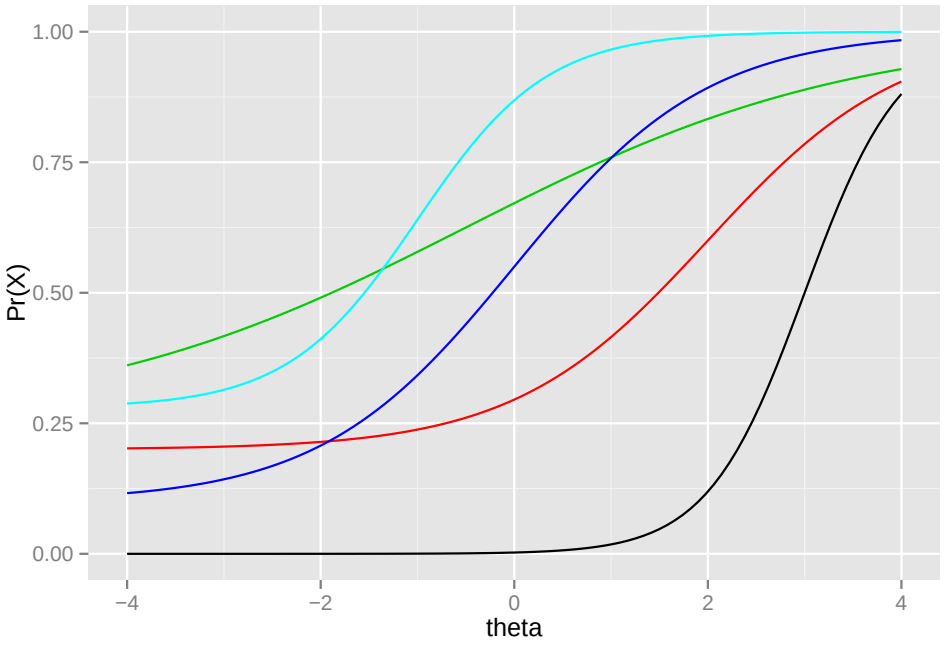


Figure 7.2: Comparison of five IRF having different item parameters.

- Which item is easiest? Cyan, followed by green, as they are furthest to the left.
- Which item are you most likely to guess correct? Green, cyan, and red appear to have the highest lower asymptotes.

Two other traditional, dichotomous IRT models are obtained as simplified versions of the three-parameter model in Equation 7.2. In the two-parameter IRT model or 2PL, we remove the c parameter and ignore the fact that guessing may impact our predictions regarding how well a person will do on an item. As a result,

$$\Pr(X = 1) = \frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}. \tag{7.3}$$

In the 2PL, we assume that the impact of guessing is negligible. Applying the model to the items shown in Figure 7.2, we would see all the lower asymptotes pulled down to zero.

In the one-parameter model or 1PL, and the Rasch model (Rasch 1960), we remove the c and a parameters and ignore the impact of guessing and of items having differing discriminations. We assume that guessing is again negligible, and that discrimination is the same for all items. In the Rasch model, we also assume that discrimination is one, that is, $a = 1$ for all items. As a result,

$$\Pr(X = 1) = \frac{e^{(\theta - b)}}{1 + e^{(\theta - b)}}. \tag{7.4}$$

Zero guessing and constant discrimination may seem like unrealistic assumptions, but the Rasch model is commonly used operationally. The PISA studies, for example, utilize a form of Rasch modeling. The popularity of the model is due to its simplicity. It requires smaller sample sizes (100 to 200 people per item may suffice) than the 2PL and 3PL (requiring 500 or more people). The theta scale produced by the Rasch model can also have certain desirable properties, such as consistent score intervals (see de Ayala 2009).

7.2.3 Assumptions

The three traditional IRT models discussed above all involve two main assumptions, both of them having to do with the overall requirement that the model we chose is "correct" for a given situation. This correctness is defined based on 1) the dimensionality of the construct, that is, how many constructs are causing people to respond in a certain way to the items, and 2) the shape of the IRF, that is, which of the three item parameters are necessary for modeling item performance.

In Equations 7.2, 7.3, and 7.4 we have a single θ parameter. Thus, in these IRT models we assume that a single person attribute or ability underlies the item responses. As mentioned above, this person parameter is similar to the true score parameter in CTT. The scale range and values for theta are arbitrary, so a z-score metric is typically used. The first IRT assumption then is that a single attribute underlies the item response process. The result is called a unidimensional IRT model.

The second assumption in IRT is that we've chosen the correct shape for our IRF. This implies that we have a choice regarding which item parameters to include, whether only b in the 1PL or Rasch model, b and a in the 2PL, or b, a, and c in the 3PL. So, in terms of shape, we assume that there is a nonlinear relationship between ability and probability of correct response, and this nonlinear relationship is captured completely by up to three item parameters.

Note that anytime we assume a given item parameter, for example, the c parameter, is unnecessary in a model, it is fixed to a certain value for all items. For example, in the Rasch and two-parameter IRT models, the c parameter is typically fixed to 0, which means we are assuming that guessing is not an issue. In the Rasch model we also assume that all items discriminate in the same way, and a is fixed to 1; then, the only item parameter we estimate is item difficulty.

7.3 Applications

7.3.1 In practice

Because of its simplicity and lower sample size requirements, the Rasch model is commonly used in small-scale achievement and aptitude testing, for example, with assessments developed and used at the district level, or instruments designed for use in research or lower-stakes decision making. The IGDI measures discussed in Chapter 2 are developed using the Rasch model. The MAP tests, published by Northwest Evaluation Association, are also based on a Rasch model. Some consider the Rasch model most appropriate for theoretical reasons. In this case, it is argued that we should seek to develop tests that have items that discriminate equally well; items that differ in discrimination should be replaced with ones that do not. Others utilize the Rasch model as a simplified IRT model, where the sample size needed to accurately estimate different item discriminations and lower asymptotes cannot be obtained. Either way, when using the Rasch model, we should be confident in our assumption that differences between items in discrimination and lower asymptote are negligible.

The 2PL and 3PL models are often used in larger-scale testing situations, for example, on high-stakes tests such as the GRE and ACT. The large samples available with these tests support the additional estimation required by these models. And proponents of the two-parameter and three-parameter models often argue that it is unreasonable to assume zero lower asymptote, or equal discriminations across items.

In terms of the properties of the model itself, as mentioned above, IRT overcomes the CTT limitation of sample and item dependence. As a result, ability estimates from an IRT model should not depend on the sample of items used to estimate ability, and item parameter estimates should not depend on the sample of people used to estimate them. An explanation of how this is possible is beyond the scope of this book. Instead, it is important to remember that, in theory, when IRT is correctly applied, the resulting parameters are sample and item independent. As a result, they can be generalized across samples for a given population of people and test items.

IRT is useful first in item analysis, where we pilot test a set of items and then examine item difficulty and discrimination, as discussed with CTT in Chapter 6. The benefit of IRT over CTT is that we can accumulate difficulty and discrimination statistics for items over multiple samples of people, and they are, in theory, always expressed on the same scale. So, our item analysis results are sample independent. This is especially useful for tests that are maintained across more than one administration. Many admissions tests, for example, have been in use for decades. State tests, as another example, must also maintain comparable item statistics from year to year, since new groups of students take the tests each year.

Item banking refers to the process of storing items for use in future, potentially undeveloped, forms of a test. Because IRT allows us to estimate sample independent item parameters, we can estimate parameters for certain items using pilot data, that is, before the items are used operationally. This is what happens in a computer adaptive test. For example, the difficulty of a bank of items is known, typically from pilot administrations. When you sit down to take the test, an item of known average difficulty can then be administered. If you get the item correct, you are given a more difficult item. The process continues, with the difficulty of the items adapting based on your performance, until the computer is confident it has identified your ability level. In this way, computer adaptive testing relies heavily on IRT.

7.3.2 Examples

The epmr package contains functions for estimating and manipulating results from the Rasch model. Other R packages are available for estimating the 2PL and 3PL (e.g., Rizopoulos 2006). Commercial software packages are also available, and are often used when IRT is applied operationally.

Here, we estimate the Rasch model for `PISA09` students in Great Britain. The data set was created at the beginning of the chapter. The `irtstudy()` function in epmr runs the Rasch model and prints some summary results, including model fit indices and variability estimates for random effect. Note that the model is being fit as a generalized linear model with random person and random item effects, using the lme4 package (Bates et al. 2015). The estimation process won't be discussed here. For details, see Doran et al. (2007) and De Boeck et al. (2011).

```
# The irtstudy() function estimates theta and b parameters for a set of
# scored item responses
irtgbr <- irtstudy(pisagbr[, rsitems])
irtgbr
##
## Item Response Theory Study
##
## 3650 people, 11 items
##
## Model fit with lme4::glmer
##         AIC         BIC      logLik    deviance df.residual
##    46615.11    46632.31   -23305.55    38745.73    40148.00
##
## Random effects
##         Std.Dev      Var
## person 1.287442 1.657506
## item   1.011468 1.023067
```

The R object returned by `irtstudy()` contains a list of results. The first element in the list, `irtgbr$data`, contains the original data set with additional columns for the theta estimate and SEM for each person. The next element in the output list, `irtgbr$ip`, contains a matrix of item parameter estimates, with the first column fixed to 1 for a, the second column containing estimates of b, and the third column fixed to 0 for c.

```
head(irtgbr$ip)
##          a           b c
## r414q02s 1 -0.02535278 0
## r414q11s 1  0.61180610 0
## r414q06s 1 -0.11701587 0
## r414q09s 1 -0.96943042 0
## r452q03s 1  2.80576876 0
## r452q04s 1 -0.48851334 0
```

Figure 7.3 shows the IRF for a subset of the `PISA09` reading items based on data from Great Britain. These items pertain to the prompt "The play's the thing" in Appendix A.2. Item parameters are taken from `irtgbr$ip` and supplied to the `rirf()` function from epmr. This function is simply Equation 7.2 translated into R code. Thus, when provided with item parameters and a vector of thetas, `rirf()` returns the corresponding $\Pr(X = 1)$.

```
# Get IRF for the set of GBR reading item parameters and a vector of thetas
# Note the default thetas of seq(-4, 4, length = 100) could also be used
irfgbr <- rirf(irtgbr$ip, seq(-6, 6, length = 100))
# Plot IRF for items r452q03s, r452q04s, r452q06s, and r452q07s
ggplot(irfgbr, aes(theta)) + scale_y_continuous("Pr(X)") +
  geom_line(aes(y = irfgbr$r452q03s, col = "r452q03")) +
  geom_line(aes(y = irfgbr$r452q04s, col = "r452q04")) +
  geom_line(aes(y = irfgbr$r452q06s, col = "r452q06")) +
  geom_line(aes(y = irfgbr$r452q07s, col = "r452q07")) +
  scale_colour_discrete(name = "item")
```

Item `pisagbr$r452q03s` is estimated to be the most difficult for this set. The remaining three items in Figure 7.3 have locations or b parameters near theta 0. Notice that the IRF for these items, which come from the Rasch model, are all parallel, with the same slopes. They also have lower asymptotes of $\Pr = 0$.

We can also use the results in `irtgbr` to examine SEM over theta. The SEM is obtained using the `rtef()` function from epmr. Figure 7.4 compares the SEM over theta for the full set of items (the black line), with SEM for subsets of 4 (blue) and 8 items (green). Fewer items at a given point in the theta scale also results in higher SEM at that theta. Thus, the number and location of items on the scale impact the resulting SEM.

```
# Plot SEM curve conditional on theta for full items
# Then add SEM for the subset of items 1:8 and 1:4
ggplot(rtef(irtgbr$ip), aes(theta, se)) + geom_line() +
  geom_line(aes(theta, se), data = rtef(irtgbr$ip[1:8, ]),
    col = "blue") +
  geom_line(aes(theta, se), data = rtef(irtgbr$ip[1:4, ]),
    col = "green")
```

The `rtef()` function is used above with the default vector of theta `seq(-4, 4, length = 100)`. However, SEM can be obtained for any theta based on the item parameters provided. Suppose we want to estimate the SEM for a high ability student who only takes low difficulty items. This constitutes a mismatch in item and construct, which produces a higher SEM. These SEM are interpreted like SEM from CTT, as the average variability we'd expect to see in a score due to measurement error. They can be used to build confidence intervals around theta for an individual.

```
# SEM for theta 3 based on the four easiest and the four most
# difficult items
rtef(irtgbr$ip[c(4, 6, 9, 11), ], theta = 3)
```

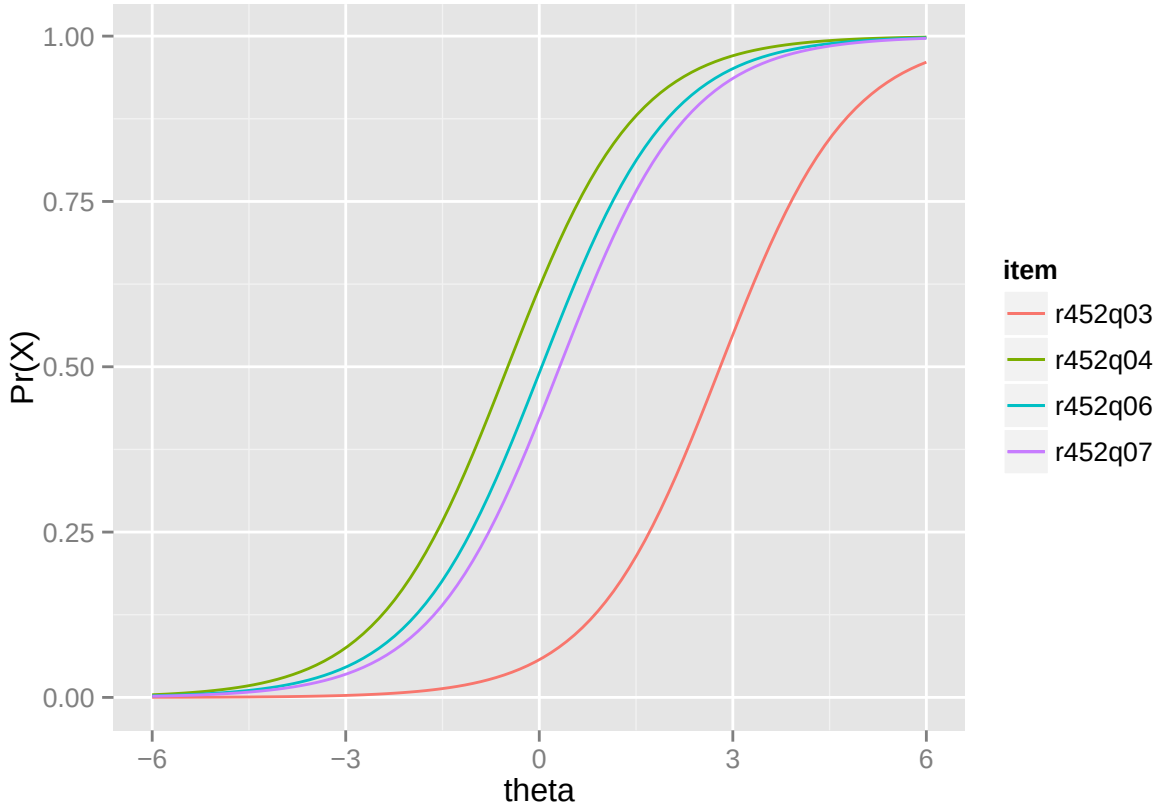


Figure 7.3: IRF for `PISA09` reading items from "The play's the thing" based on students in Great Britain.

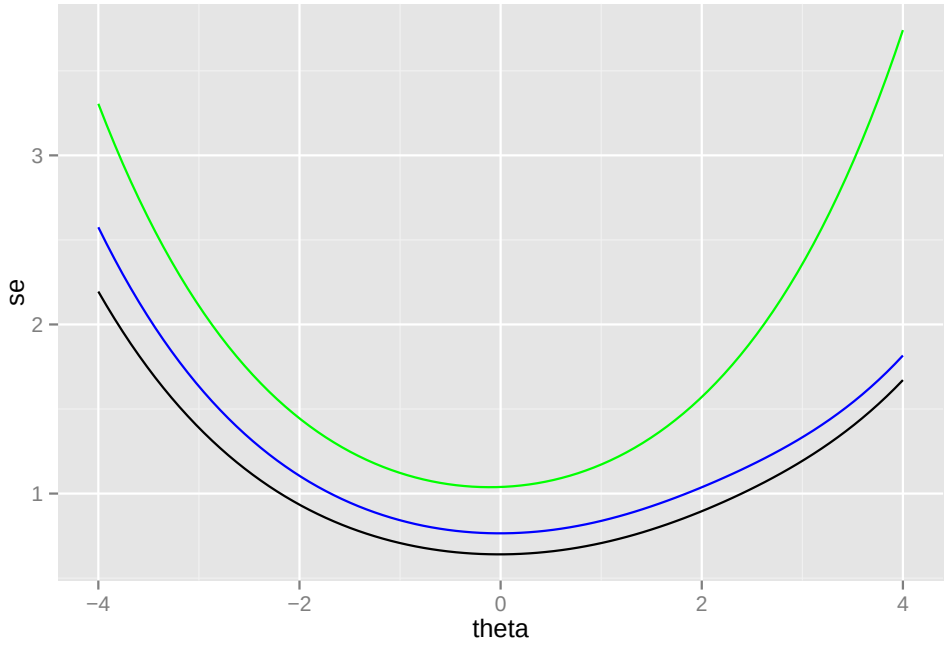


Figure 7.4: SEM for two subsets of 'PISA09' reading items based on students in Great Britain.

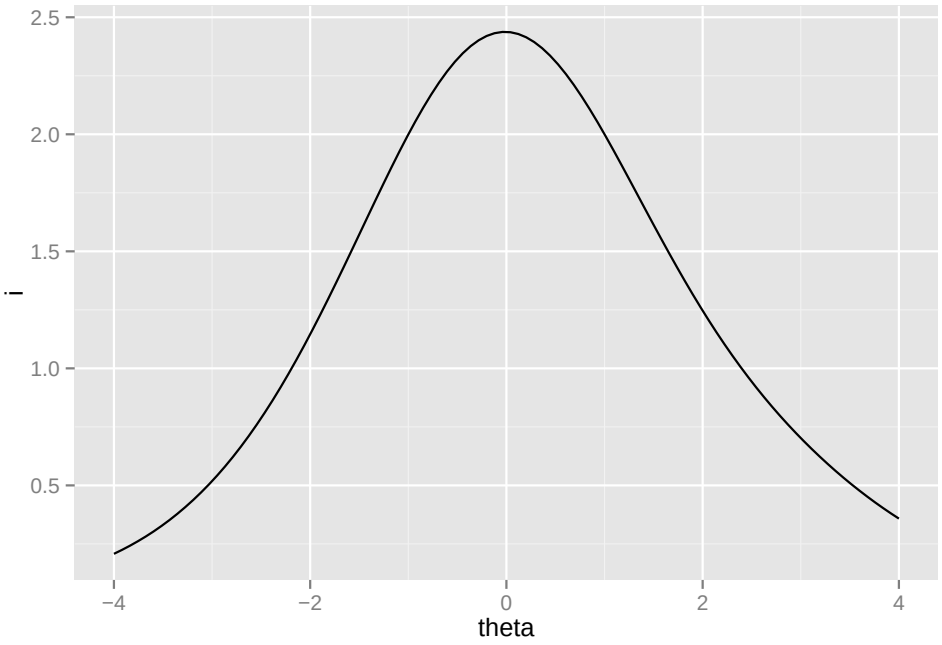


Figure 7.5: Test information for 'PISA09' reading items based on students in Great Britain.

```
##   theta       se
## 1     3 2.900782
rtef(irtgbr$ip[c(2, 7, 8, 10), ], theta = 3)
##   theta       se
## 1     3 1.996224
```

The reciprocal of the SEM function, via the TEF, is the test information function. This simply shows us where on the theta scale our test items are accumulating the most discrimination power, and, as a result, where measurement will be the most accurate. Higher information corresponds to higher accuracy and lower SEM. Figure 7.5 shows the test information for all the reading items, again based on students in Great Britain. Information is highest where SEM is lowest, at the center of the theta scale.

```
# Plot the test information function over theta
# Information is highest at the center of the theta scale
ggplot(rtif(irtgbr$ip), aes(theta, i)) + geom_line()
```

Finally, just as the IRF can be used to predict probability of correct response on an item, given theta and the item parameters, the TRF can be used to predict total score given theta and parameters for each item in the test. The TRF lets us convert person theta back into the raw score metric for the test. Similar to the IRF, the TRF is asymptotic at 0 and the number of dichotomous items in the test, in this case, 11. Thus, no matter how high or how low theta, our predicted total score can't exceed the bounds of the raw score scale. Figure 7.6 shows the test response function for the Great Britain results.

```
# Plot the test response function over theta
ggplot(rtrf(irtgbr$ip), aes(theta, p)) + geom_line()
```

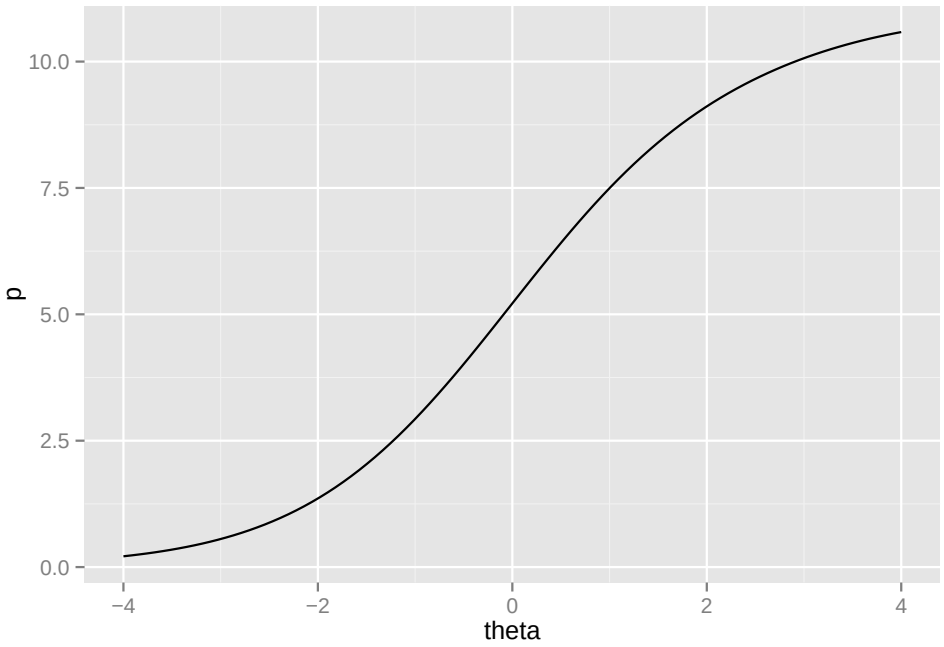


Figure 7.6: Test response function for 'PISA09' reading items based on students in Great Britain.

7.4 Summary

This chapter provides an introduction to IRT, with a comparison to CTT, and details regarding the three traditional, dichotomous, unidimensional IRT models. Assumptions and applications of the models are discussed. The Rasch model is demonstrated using data from `PISA09`, with examples of the IRF, TEF, TIF, and TRF.

7.4.1 Learning objectives

1. Compare and contrast IRT and CTT in terms of their strengths and weaknesses.
2. Identify the two main assumptions that are made when using a traditional IRT model, regarding dimensionality and functional form or the number of model parameters.
3. Identify key terms in IRT, including probability of correct response, logistic curve, theta, IRF, TRF, SEM, and information functions.
4. Define the three item parameters and one ability parameter in the traditional IRT models, and describe the role of each in modeling performance with the IIF.
5. Distinguish between the 1PL, 2PL, and 3PL IRT models in terms of assumptions made, benefits and limitations, and applications of each.
6. Describe how IRT is utilized in item analysis, test development, item banking, and computer adaptive testing.

7.4.2 Exercises

1. Sketch out a plot of IRF for the following two items: a difficult item 1 with a high discrimination and negligible lower asymptote, and an easier item 2 with low discrimination and high lower asymptote. Be sure to label the axes of your plot.
2. Sketch another plot of IRF for two items having the same difficulties, but different discriminations and lower asymptotes.

3. Examine the IRF for the remaining `PISA09` reading items for Great Britain. Check the content of the items in Appendix A to see what features of an item or prompt seemed to make it relatively easier or more difficult.
4. Using the `PISA09` reading test results for Great Britain, find the predicted total scores associated with thetas of -1, 0, and 1.
5. Estimate the Rasch model for the `PISA09` memorization strategies scale. First, dichotomize responses by recoding 1 to 0, and the remaining valid responses to 1. After fitting the model, plot the IRF for each item.
6. Based on the distribution of item difficulties for the memorization scale, where should SEM be lowest? Check the SEM by plotting the TEF for the full scale.

Chapter 8

Dimensionality

8.1 Exploratory factor analysis

8.2 Confirmatory factor analysis

8.3 Summary

8.3.1 Learning objectives

8.3.2 Exercises

Chapter 9

Validity

> Validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few
> — Robert Ebel

As noted by Ebel (1961), validity is universally considered the most important feature of a testing program. Validity encompasses everything relating to the testing process that makes score inferences useful and meaningful. All of the topics covered in Chapters 1 through 7, including measurement, test construction, reliability, and item analysis, provide evidence supporting the validity of scores. Scores that are consistent and based on items written according to specified content standards with appropriate levels of difficulty and discrimination are more useful and meaningful than scores that do not have these qualities. Correct measurement, sound test construction, reliability, and certain item properties are thus all prerequisites for validity.

This chapter begins with a definition of validity and some related terms. After defining validity, three common sources of validity evidence are discussed: test content, via what's referred to as a test blueprint or test outline; relationships with criterion variables; and theoretical models of the construct being measured. These three sources of validity evidence are then discussed within a unified view of validity. Finally, threats to validity are addressed.

9.1 Overview of validity

9.1.1 Definitions

Suppose you are conducting a research study on the efficacy of a reading intervention. Scores on a reading test will be compared for a treatment group who participated in the intervention and a control group who did not. A statistically significant difference in mean reading scores for the two groups will be taken as evidence of an effective intervention. This is an inferential use of statistics, as discussed in Chapter 1.

In measurement, we step back and evaluate the extent to which our mean scores for each group accurately measure what they are intended to measure. On the surface, the means themselves may differ. But if neither mean actually captures average reading ability, our results are misleading, as our intervention may not actually be effective.

Reliability, from Chapter 5.3.1, focuses on the consistency of measurement. With reliability, we estimate the amount of variability in scores that can be attributed to a reliable source, and, conversely, the variability that can be attributed to an unreliable source, that is, random error. While reliability is useful, it does not tell us whether that reliable source of variability is the source we hope it is. This is the job of validity. With validity,

we additionally examine the quality of our items as individual components of the target construct. We examine other sources of variability in our scores, such as item and test bias. We also examine relationships between scores on our items and other measures of the target construct.

The Standards for Educational and Psychological Testing (AERA, APA, and NCME 1999) define validity as "the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of a test." This definition is simple, but very broad, encompassing a wide range of evidence and theory. We'll focus on three specific types of validity evidence, evidence based on test content, other measures, and theoretical models.

Recent literature on validity theory has clarified that tests and even test scores themselves are not valid or invalid; only score *inferences* and *interpretations* are valid or invalid (e.g., Kane 2013). Tests are then described as being valid only for a particular use. This is a simple distinction in the definition of validity, but some authors continue to highlight it. Referring to a test or test score as valid implies that it is valid for any use, even though this is not the case. Shorthand is sometimes used to refer to tests themselves as valid, because it is simpler than distinguishing between tests, uses, and interpretations. However, the assumption is always that validity only applies to a specific test use and not broadly to the test itself.

Finally, Kane (2013) also clarifies that validity is a matter of degree. It is establish incrementally through an accumulation of supporting evidence. Validity is not inherent in a test, and it is not simply declared to exist by a test developer. Instead, data are collected and research is conducted to establish evidence supporting a test for a particular use. As this evidence builds, so does our confidence that test scores can be used for their intended purpose.

9.1.2 Validity examples

To evaluate the proposed score interpretations and uses for a test, and the extent to which they are valid, we should first examine the purpose of the test itself. As discussed in Chapters 2 and 3, a good test purpose articulates key information about the test, including what it measures (the construct), for whom (the intended population), and why (for what purpose). The question then becomes, given the quality of its contents, how they were constructed, and how they are implemented, is the test valid for this purpose?

As a first example, lets return to the test of early literacy introduced in Chapter 2. Documentation for the test (www.myigdis.com) claims that,

> myIGDIs are a comprehensive set of assessments for monitoring the growth and development of young children. myIGDIs are easy to collect, sensitive to small changes in children's achievement, and mark progress toward a long-term desired outcome. For these reasons, myIGDIs are an excellent choice for monitoring English Language Learners and making more informed Special Education evaluations.

Different types of validity evidence would be needed to support the claims made for the IGDI measures. The comprehensiveness of the measures could be documented via test outlines that are based on a broad but well-defined content domain, and that are vetted by content experts, including teachers. Multiple test forms would be needed to monitor growth, and the quality and equivalence of these forms could be established using appropriate reliability estimates and measurement scaling techniques, such as Rasch modeling. Ease of data collection could be documented by the simplicity and clarity of the test manual and administration instructions, which could be evaluated by users, and the length and complexity of the measures. The sensitivity of the measures to small changes in achievement and their relevance to long-term desired outcomes could be documented using statistical relationships between IGDI scores and other measures of growth and achievement within a longitudinal study. Finally, all of these sources of validity evidence would need to be gathered both for English Language Learners and other target groups in special education. These various forms of information fit into the *sources of validity evidence* discussed below.

As a second example, consider a test construct that interests you. What construct are you interested in measuring? Perhaps it is one construct measured within a larger research study? How could you measure this

construct? What type of test are you going to use? And what types of score(s) from the test will be used to support decision making? Next, consider who is going to take this test. Be as specific as possible when identifying your target population, the individuals that your work or research focuses on. Finally, consider why these people are taking your test. What are you going to do with the test scores? What are your proposed score interpretations and uses? Having defined your test purpose, consider what type of evidence would prove that the test is doing what you intend it to do, or that the score interpretations and uses are what you intend them to be. What information would support your test purpose?

9.1.3 Sources of validity evidence

The information gathered to support a test purpose, and establish validity evidence for the intended uses of a test, is often categorized into three main areas of validity evidence. These are content, criterion, and construct validity. Nowadays, these are referred to as *sources of validity evidence*, where content focuses on the test content and procedures for developing the test, criterion focuses on external measures of the same target construct, and construct focuses on the theory underlying the construct and includes relationships with other measures. In certain testing situations, one source of validity evidence may be more relevant than another. However, all three are often used together to argue that the evidence supporting a test is "adequate."

We will review each source of validity evidence in detail, and go over some practical examples of when one is more relevant than another. In this discussion, consider your own example, and examples of other tests you've encountered, and what type of validity evidence could be used to support their use.

9.2 Content validity

According to Haynes, Richard, and Kubany (1995), content validity is "the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose." Note that this definition of content validity is very similar to our original definition of validity. The difference is that content validity focuses on *elements* of the construct and how well they are represented in our test. Thus, content validity assumes the target construct can be broken down into elements, and that we can obtain a representative sample of these elements.

There are four main steps to establishing content validity evidence. First, as always, we define the purpose of our test and the construct we are measuring. Next, we define the content domain based on relevant standards, skills, tasks, behaviors, facets, factors, etc. that represent the construct. The idea here is that our construct can be represented in terms of specific identifiable dimensions or components, some of which may be more relevant to the construct than others. Next, we use this definition of the content domain to create a blueprint or outline for our test. The blueprint organizes the test based on the relevant components of the content domain, and describes how each of these components will be represented within the test. Finally, subject matter experts evaluate the extent to which our test blueprint adequately captures the content domain, and the extent to which our test items will adequately sample from the content domain.

Here is a brief overview of how content validity could be established for the IGDI measures of early literacy. Again, the purpose of the test is to identify preschoolers in need of additional support in developing early literacy skills. The early literacy content domain is broken down into a variety of content areas, including alphabet principles (e.g., knowledge of the names and sounds of letters), phonemic awareness (e.g., awareness of the sounds that make up words), and oral language (e.g., definitional vocabulary). The literature on early literacy has identified other important skills, but we'll focus here on these three. Note that the content domain for a construct should be established both by research and practice. Next, we map the portions of our test that will address each area of the content domain. The test outline can include information about the type of items used, the cognitive skills required, and the difficulty levels that are targeted, among other things. Review Chapter 4 for additional details on test outlines or blueprints.

Table 9.1 contains an example of a test outline for the IGDI measures. The three content areas listed above are shown in the first column. These are then broken down further into cognitive processes or skills. Theory

Table 9.1: Example Test Outline for a Measure of Early Literacy

Content Area	Cognitive process	Items	Weight
Alphabet principles	Letter naming	20	13%
	Sound identification	20	13%
Phonological awareness	Rhyming	15	10%
	Alliteration	15	10%
	Sound blending	10	7%
Oral language	Picture naming	30	20%
	Which one doesn't belong	20	13%
	Sentence completion	20	13%

and practical constraints determine reasonable numbers and types of test items or tasks devoted to each cognitive process in the test itself. The final column shows the percentage of the total test that is devoted to each area.

Validity evidence requires that the test outline be appropriate, given the construct and test purpose. The appropriateness of an outline is typically evaluated by content experts. In the case of the IGDI measures, these experts could be researchers in the area of early literacy, and teachers who work directly with students from the target population.

Content validity is important for non-cognitive psychological tests as well. Suppose the purpose of a test is to measure client experience with panic attacks so as to determine the efficacy of treatment. The domain for this test could be defined using criteria listed in the DSM-V (www.dsm5.org), reports about panic attack frequency, and secondary effects of panic attacks. The test outline would organize the number and types of items written to address all relevant criteria from the DSM-V. Finally, experts who work directly in clinical settings would evaluate the test outline to determine its quality, and their evaluation would provide evidence supporting the content validity of the test for this purpose.

When considering content validity, we must also be aware of how content validity can be compromised. Think about this issue for tests in general, and then for this specific example. If our panic attack scores were not valid for a particular use, how would this lack of validity manifest itself in the process of establishing content validity?

Here are two possible sources of content invalidity. First, if items reflecting criteria that are important to the construct are omitted, the construct will be *underrepresented* in the test. For example, if the test does not include items addressing "nausea or abdominal distress," other criteria, such as "fear of dying," may have too much sway in determining an individual's score. Second, if items measuring irrelevant or tangential material are included, the construct will be *misrepresented* in the test. For example, if items measuring depression are included in the scoring process, the score itself is less valid as a measure of the target construct.

Here is a third example of content validity from the area of licensure/certification testing. I have experience working with tests of medical imaging, including knowledge assessments taken by candidates for certification in radiography. This area provides a unique example of content validity, because the test itself measures a construct that is directly tied to professional practice. If practicing radiographers utilize a certain procedure, that procedure, or the knowledge required to perform it, should be included in the test.

The domain for a licensure/certification test such as this is defined using what is referred to as a job analysis or practice analysis (Raymond 2001). A job analysis is a research study, the central feature of which is a survey sent to practitioners that lists a wide range of procedures and skills potentially used in the field. Respondents indicate how often they perform each procedure or use each skill on the survey. Procedures and skills performed by a high percentage of professionals are then included in the test outline. As in the previous examples, the final step in establishing content validity is having a select group of experts review the procedures and skills and their distribution across the test, as organized in the test outline.

9.3 Criterion validity

9.3.1 Definition

Criterion validity is the degree to which test scores correlate with, predict, or inform decisions regarding another measure or outcome. If you think of content validity as the extent to which a test *correlates with* or corresponds to the content domain, criterion validity is similar in that it is the extent to which a test correlates with or corresponds to another test. So, in content validity we compare our test to the content domain, and hope for a strong relationship, and in criterion validity we compare our test to a criterion variable, and again hope for a strong relationship.

The keyword in this definition of criterion validity is *correlate*, which is synonymous with relate or predict. The assumption here is that the construct we are hoping to measure with our test is known to be measured by another test or variable. This other test or variable is often referred to as a "gold standard," a label presumably given to it because it is based on strong validity evidence. So, in a way, criterion validity is a form of validity by association. If our test correlates with a known measure of the construct, we can be more confident that our test measures the same construct.

Criterion validity is sometimes distinguished further as concurrent validity, where our test and the criterion are administered concurrently, or predictive validity, where our test is measured first and can then be used to predict the future criterion.

Criterion validity is limited because it does not actually require that our test be a reasonable measure of the construct, only that it relate strongly with another measure of the construct. Nunnally and Bernstein (1994) clarify this point with a hypothetical example:

> If it were found that accuracy in horseshoe pitching correlated highly with success in college, horseshoe pitching would be a valid measure for predicting success in college.

The example is silly, but it highlights the fact that, on it's own, criterion validity is limited. The take-home message is that you should never use or trust a criterion relationship as your sole source of validity evidence.

There are two other challenges associated with criterion validity. First, finding a suitable criterion can be difficult, especially if your test measures a new or not well defined construct. Second, a correlation coefficient is attenuated, or reduced in strength, by any unreliability present in the two measures being correlated. So, if your test and the criterion test are unreliable, a low validity coefficient (the correlation between the two tests) may not necessarily represent a lack of relationship between the two tests. It may instead represent a lack of reliable information with which to estimate the criterion validity coefficient.

The steps for establishing criterion validity evidence are relatively simple. After defining the purpose of the test, a suitable criterion is identified. The two tests are administered to the same sample of individuals from the target population, and a correlation coefficient is obtained. Note that criterion validity can be maximized by writing items that are predictive of or correlate with the criterion.

9.3.2 Examples

A popular example of criterion validity is the GRE, which has come up numerous times in this book. The GRE is designed to predict performance in graduate school. Admissions programs use it as one indicator of how well you are likely to do as a graduate student. Given this purpose, what is a suitable criterion variable that the GRE should predict? And how strong of a correlation would you expect to see between the GRE and this graduate performance variable?

The simplest criterion for establishing criterion-related validity evidence for the GRE would be some measure of performance in graduate school. First-year graduate GPA is commonly chosen. The GRE has been shown to correlate around 0.30 with first-year graduate GPA. A correlation of 0.30 is evidence of a small positive

relationship, suggesting that most of the variability in GPA, our criterion, is not predicted by the GRE. In other words, many students score high or low on the GRE and do not have a similarly high or low graduate GPA.

Although this modest correlation may at first seem disappointing, a few different considerations suggest that it is actually pretty impressive. First, GPA is likely not a reliable measure of graduate performance. It's hardly a "gold standard." Instead, it's the best we have. It's one of the few quantitative measures available for all graduate students. Second, there is likely some restriction of range happening in the relationship between GRE and GPA. People who score low on the GRE are less likely to get into graduate school, so their data are not represented. Restriction of range tends to reduce correlation coefficients. Third, what other measure of pre-graduate school performance correlates at 0.30 with graduate GPA? More importantly, what other measure of pre-grad school performance that only takes a few hours to obtain correlates at 0.30 with graduate GPA? In conclusion, the GRE isn't perfect, but it's the best we've got. Admissions programs just need to make sure they don't rely too much on it in admissions decisions, as discussed in Chapter 3.

A substantial amount of research has been conducted documenting predictive validity evidence for the GRE. See Kuncel, Hezlett, and Ones (2001) for a meta-analysis of results from this literature.

9.4 Construct validity

9.4.1 Definition

As noted above, validity focuses on the extent to which our construct is in fact what we think it is. If our construct is what we think it is, it should relate in known ways with other measures of the same or different constructs. On the other hand, if it is not what we think it is, relationships that should exist with other constructs will not be found.

Construct validity is established when relationships between our test and other variables confirm what is predicted by theory. For example, theory might indicate that the personality traits of conscientiousness and neuroticism should be negatively related. If we develop a test of conscientiousness and then demonstrate that scores on our test correlate negatively with scores on a test of neuroticism, we've established construct validity evidence for our test. Furthermore, theory might indicate that conscientiousness contains three specific dimensions. Statistical analysis of the items within our test could show that the items tend to cluster, or perform similarly, in three specific groups. This too would establish construct validity evidence for our test.

9.4.2 Examples

The entire set of relationships between our construct and other available constructs is sometimes referred to as a *nomological network*. This network outlines what the construct is, based on what it relates positively with, and what it is not, based on what it relates negatively with. For example, what variables would you expect to relate positively with depression? As a person gets more depressed, what else tends to increase? What variables would you not expect to correlate with depression? Finally, what variables would you expect to relate negatively with depression?

Table 9.2 contains an example of a correlation matrix that describes a nomological network for a hypothetical new depression scale. The BDI would be considered a well known criterion measure of depression. The remaining labels in this table refer to other related or unrelated variables. "Fake bad" is a measure of a person's tendency to pretend to be "bad" or associate themselves with negative behaviors or characteristics. Positive correlations in this table represent what is referred to as *convergence*. Our hypothetical new scale converges with the BDI, a measure of anxiety, and a measure of faking bad. Negative correlations represent *divergence.* Our new scale diverges with measures of happiness and health. Both types of correlation should be predicted by a theory of depression.

Table 9.2: Nomological Network for a Hypothetical Depression Inventory

	New Scale	BDI	Anxiety	Happy	Health	Fake Bad
New Scale	1.00					
BDI	0.80	1.00				
Anxiety	0.65	0.50	1.00			
Happy	-0.59	-0.61	-0.40	1.00		
Health	-0.35	-0.10	-0.35	0.32	1.00	
Fake Bad	0.10	0.14	0.07	-0.05	0.07	1.00

9.5 Unified validity and threats

In the early 1980s, the three types of validity were reconceptualized as a single construct validity (e.g., Messick 1980). This reconceptualization clarifies how content and criterion evidence do not, on their own, establish validity. Instead, both contribute to an overarching evaluation of construct validity. The literature has also clarified that validation is an ongoing process, where evidence supporting test use is accumulated over time from multiple sources. As a result, validity is a matter of degree instead of being evaluated as a simple and absolute yes or no.

As should be clear, scores are valid measures of a construct when they accurately represent the construct. When they do not, they are not valid. Two types of threats to content validity were mentioned previously. These are content underrepresentation and content misrepresentation. These can both be extended to more broadly refer to construct underrepresentation and construct misrepresentation. In the first, we fail to include all aspects of the construct in our test. In the second, our test is impacted by variables or constructs other than our target construct, including systematic and random error. And in both, we introduce construct irrelevant variance into our scores.

Construct underrepresentation and misrepresentation can both be identified using a test outline. If the content domain is missing an important aspect of the construct, or the test is missing an important aspect of the content domain, the outline should make it apparent. Subject matter experts provide an external evaluation of these issues. Unfortunately, the construct is often underrepresented or misrepresented by individual items, and item-level content information is not provided in the test outline. As a result, the test development process also involves item-level reviews by subject matter experts and others who provide input on potential bias and unreliability at the item level.

Underrepresenting or misrepresenting the construct in a test can have a negative impact on testing outcomes, both at the item level and the test level. Item bias refers to differential performance for subgroups of individuals, where the performance difference is not related to true differences in ability or trait. An item may address content that is relevant to the content domain, but it may do so in a way that is less easily understood by one group than another. For example, in educational tests, questions often involve word problems that provide context to an application. This context may address material, for example, a vacation to the beach, that is more familiar to students from a particular region, for example, coastal areas. This item might be biased against students who less familiar with the context because they don't live near the beach. Given that we aren't interested in measuring proximity to coastline, this constitutes test bias that reduces the validity of our test scores.

Other specific results of invalid test scores include misplacement of students, misallocation of funding, and implementation of programs that should not be implemented. Can you think of anything else to add to the list? What are some practical consequences of using test scores that do not measure what they purport to measure?

9.6 Summary

This chapter provides an overview of validity, with examples of content, criterion, and construct validity, and details on how these three sources of validity evidence come together to support the intended interpretations and uses of test scores.

9.6.1 Learning objectives

1. Define validity in terms of test score interpretation and use, and describe and identify examples of this definition in context.
2. Compare and contrast three main types of validity evidence (content, criterion, and construct) and identify examples of how each type is established, including the validation process involved with each.
3. Explain the structure and function of a test blueprint, and how it is used to provide evidence of content validity.
4. Identify appropriate sources of validity evidence for given testing applications and describe how certain sources are more appropriate than others for certain applications.
5. Describe the unified view of validity and how it differs from and improves upon the traditional view of validity.
6. Identify threats to validity, including features of a test, testing process, or score interpretation or use, that impact validity. Consider, for example, the issues of content underrepresentation and misrepresentation, and construct irrelevant variance.

9.6.2 Exercises

1. Consider your own testing application and how you would define a content domain. What is this definition of the content domain based on? In education, for example, end-of-year testing, it's typically based on a curriculum. In psychology, it's typically based on research and practice. How would you confirm that this content domain is adequate or representative of the construct? And how could content validity be compromised for your test?
2. Consider your own testing application and a potential criterion measure for it. How do you go about choosing the criterion? How would you confirm that a relationship exists between your test and the criterion? How could criterion validity be compromised in this case?
3. Construct underrepresentation and misrepresentation are reviewed briefly for the hypothetical test of panic attacks. Consider how underrepresentation and misrepresentation could each impact content validity for the early literacy measures.
4. Consider what threats might impact the validity of your own testing application.

Chapter 10

Test Evaluation

Test evaluation summarizes many of the topics that precede it in this course, including test purpose, reliability and validity study design and results, scoring and reporting guidelines, and recommendations for test use. The new material for this chapter includes the process of evaluating this information within a test review or technical manual when considering one or more tests for a particular use. Our perspective will be that of a test consumer, for example, a researcher or practitioner in the market for a test to inform some application, for example, a research question or decision making process.

This chapter utilizes the construct of creative problem solving to demonstrate some of the critical considerations in the test evaluation process. Within this context, recommendations are provided on test purpose, study design, reliability, validity, scoring, and test use.

10.1 Test purpose

We'll begin, as usual, with a discussion of test purpose. The question is, why does a test needs to have a clear purpose? Suppose you have a research question that requires measurement of some kind, and you don't have the time or resources to develop your own test. So, you start looking for an existing measure that will meet your needs. How do you find such a measure?

It is possible that the gold standard measures for your field or area of work will be well known to you. But, if they aren't, you will need to compare tests based on their test purposes. So, what construct(s) are you hoping to measure, for what population, and for what reason? And what tests have purposes that match your application?

Maybe you are interested in helping children develop their creative problem solving. You have received a grant to fund a four-week summer camp, but the funding agency requires that you conduct an evaluation of your creativity camp. Aside from qualitative measures, for example, based on interviews, you also need a quantitative measure of effectiveness. So, you decide to do a pre/post study using a test of creative problem solving. The problem you can't solve is where to find the ideal test.

We will use information from the Buros test database buros.org/ in our test evaluations. The Buros Center for Testing includes a test review library. The library is physically located at the University of Nebraska-Lincoln. Electronic access is available online. Each year, Buros solicits and publishes reviews of newly published tests in what is called the *Mental Measurements Yearbook* (MMY). These reviews are written by testing professionals and they summarize and evaluate the information that test users need to know about a test before using it. This information is essentially what we are covering in this chapter.

If you have access to the electronic version of MMY, for example, through a university library, you should go search for a test of creative problem solving. If you use the search terms "creative problem solving," you probably won't get many results. I only had ten at the time of writing this. Judging by their titles, some

Table 10.1: Comparison of Two Tests of Creativity

	CAPSAT	CAP
Publication year	2011	1980
Format	Self-administered	Parent and teacher ratings
Number of scales	4	8
Number of items	36	48 SR, 4 CR
Score scale range	0 to 100	Three-point frequency scale
Referencing	Criterion	Norm
Population	Adults 17 to 40	Children 6 to 18

of the results sounded like they might be appropriate, but I decided to try again with just the search term "creativity." This returned 182 results.

As you browse through these tests, think about the construct of creative problem solving, and creativity in general. How could we test a construct like this? What types of tasks would we expect children to respond to, and how would we score their responses? It turns out, creativity is not easily measured, primarily because a standardized test with a structured scoring guide does not leave room for responses that come from "outside the box," or that represent divergent thinking, or that we would consider novel or... "creative." In a way, the term *standardized* suggests the opposite of creativity. Still, quite a few commercial measures of creativity exist.

Table 10.1 contains information for two tests of creativity, the Creativity and Problem-Solving Aptitude Test (CAPSAT) and the Creativity Assessment Packet (CAP). Only one test (the CAP) is available in the electronic version of MMY. When you perform your search on MMY with search term "creativity" you should see the CAP near the top of your results. Neither test purpose is stated clearly in the technical documentation, so we'll have to assume that they're both intended to describe creativity and problem-solving for the stated populations. The intended uses of these tests are also not clear.

Given the limited information in Table 10.1, which test would be more appropriate for your summer camp evaluation? The CAP seems right, given that it's designed for children 6 to 18. Unfortunately, as we dig deeper into the technical information for the test, we discover that there is essentially no reliability or validity evidence for it. One MMY review states that test-retest reliability and criterion validity were established for the CAP in the 1960s, but no coefficients are actually reported in the CAP documentation.

No matter how well a test purpose matches your own, a lack of reliability and validity evidence makes a test unusable. The only solution for the CAP would be to conduct your own reliability and validity studies. Technical documentation for the CAPSAT indicates that internal consistency (alpha) for the entire test is 0.90, with subscales having coefficient alphas between 0.72 and 0.87. Those are all acceptable. However, as noted in the MMY reviews, validity information for the CAPSAT is missing.

We clearly need a better test. In MMY, search for information on the Torrance Tests of Creative Thinking (TTCT). Read through the two reviews and try to find any weaknesses or limitations that put this test in the unusable category, along with the CAPSAT and the CAP. First, you need to make sure that the test purpose is acceptable, since there's no point in using a test that has been validated for the wrong purpose.

For the uninitiated, the MMY provides a brief statement of purpose for the TTCT: "to identify and evaluate creative potential through words (verbal forms) and pictures (figural forms)." And here's a short summary of the TTCT from one reviewer:

> In the complex and still-evolving domain of creativity assessment, the TTCT can be recommended as sound examples of instruments useful for research, evaluation, and general instructional planning decisions.

Given this background information, the TTCT sounds like a viable solution. Next, we need to look into the reliability and validity information for the test.

10.2 Study design

Because the CAPSAT and the CAP both lacked reliability and validity evidence, there really wasn't anything to evaluate. Most tests will include some form of evidence supporting reliability and validity, and we will need to evaluate this evidence both in terms of its strength and its relevance to the test purpose.

In Chapters 5.3.1 and 9, we discussed how to evaluate the reliability and validity evidence for a test. Whether or not reliability and validity coefficients are acceptable depends on guidelines established in a particular field, based on previous work. This also depends on the test purpose. For example, when making high-stakes decisions, coefficients should be in the 0.90s. When making low-stakes decisions, lower values are acceptable. A minimum of 0.70 reliability is sometimes discussed, but values below it are common and may not be cause for alarm, depending on the context, such as with shorter tests. Validity coefficients, that is, correlations with a criterion variable, can vary widely and must be interpreted in-context.

To interpret and evaluate reliability and validity, we should first consider the strength of the reliability or validity *study designs*. Keep in mind that reliability and validity are based on actual test data (except for a theoretical foundation in construct validity). The quality of data depends on the quality of the test administration design and data collection procedures. Here are some basic questions to ask when evaluating a test.

- Is the study sample representative?
- Is the sample randomly or intentionally selected?
- Are appropriate age/gender/ethnic other groups included?
- Are administration conditions standardized?
- Are accommodations made when necessary, and do they impact results?

Regardless of strength or magnitude, reliability and validity coefficients may be irrelevant if they are based on a weak (e.g., non-random or biased) study design or the wrong population. These questions help us determine how relevant and appropriate the reliability and validity evidence is. Note that answers to these questions will typically not be found in test reviews. Instead, you may have to go to published articles or the technical documentation published with the test. A number of reliability and validity studies have been conducted with the TTCT (e.g., Torrance 1981a; Torrance 1981b).

10.3 Reliability

Here are some of the questions that you need to ask when evaluating the relevance of reliability evidence for your test purpose. What types of reliability are estimated? Are these types appropriate given the test purpose? Are the study designs appropriate given the chosen types? Are the resulting reliability estimates strong or supportive? These, along with the questions presented below, simply review what we have covered in previous chapters. The point here is to highlight how these features of a test (reliability, validity, scoring, and test use) are important when evaluating a test for a particular use.

So, what type of reliability is reported for the TTCT? The test reviews note that a "collection of reliability and validity data" are available. Tests of creativity, like the TTCT, require that individuals create things, that is, perform tasks. Given that the TTCT involves performance assessment, we need to know about interrater consistency. One review highlights interrater reliabilities ranging from 0.66 to 0.99 for trained scorers and classroom teachers. The 0.66 is low, but acceptable, given the low-stakes nature of our test use. The 0.99 is optimal, and is something we would expect when correlating scores given by two well trained test administrators. Are we concerned that correlations were reported, and not generalizability coefficients?

Given the ordinal/interval nature of the scale (discussed further below), percentage agreement and kappa are not appropriate, but generalizability coefficients, taking into account systematic scoring differences, would be more informative than correlations.

According to the test reviews, we also have test-retest reliabilities from the 0.50s in one study to 0.93 in another. A reviewer notes that these different results could be due to differences in the variability in the samples of students used in each study; a wider age range was used to obtain the test-retest of 0.93. Other studies reported test-retest reliabilities in the 0.60s and 0.70s, which the reviewer states are acceptable, given the number of weeks that pass between administrations.

10.4 Validity

Here are some questions that you need to ask when evaluating the relevance of validity evidence for your test purpose. These are essentially the same as the questions for reliability. What types of validity evidence are examined? Are these types appropriate given the test purpose? Are the study designs appropriate given the chosen types of validity evidence? Does the resulting evidence support the intended uses and inferences of the test?

Remember, factors impacting construct validity fall into two categories. Construct underrepresentation is failure to represent what the construct contains or consists of, and construct misrepresentation happens when we measure other constructs or factors including measurement error.

One reviewer of the TTCT notes that "validity data are provided relating TTCT scores to various measures of personality and intelligence without structuring a theory of creativity to describe what the relationship of these variables to creativity should be." However, the same reviewer then makes this comment, which is especially relevant to our purpose:

> A second type of construct validation deals with the impact of treatment to alter creative performance and its relation to changing scores on the TTCT. If the theory says that treatment X should increase scores on tests of creativity, then pre and posttest differences should be found on the TTCT, with treatment intervening. The manual reports a broad summary of these studies, and the conclusion is that when the treatment deals with tasks somewhat like those on the test, posttest scores do indeed show significant gains on the TTCT.

Although we don't have specific details, I think we can be confident, based on this MMY review, that the TTCT is suitable for a test-retest study. In reality, we would go to the specific studies in the references that provide this information.

10.5 Scoring

Methods for creating and reporting scores for a test can vary widely from one test to the next. However, there are a few key questions to ask when evaluating the scoring that is implemented with a test. What types of scores are produced? That is, what type of measurement scale is used? What is the score scale range? How is meaning given to scores? And what type of score referencing is used, and does this seem reasonable? Finally, what kinds of score reporting guidelines are provided, and do they seem appropriate? As with reliability and validity, these questions should be considered in reference to the purpose of the test.

As with the majority of educational and psychological tests, the score scale for the TTCT is based on a sum of individual item/task scores. In this case, scores come from a trained rater. We assume this is an interval scale of measurement. The score scale range isn't important for our test purpose, as long as it can capture growth, which we assume it can, given the information presented above. However, note that a small scale range, for example, 1 to 5 points, might be problematic in a pre/post test administration, depending on how much growth is expected to take place.

Regarding scoring, one MMY reviewer notes:

> The scoring system for the tests also has some problems. For example, the assignment of points to responses on the Originality scale was based on the frequency of appearance of these responses among 500 unspecified test takers. For the Asking subtest, for example, responses that appeared in less than 2% of the protocols receive a weight of 2. What is the basis for setting this criterion? How would it alter scores if 2s were assigned to 10%, for example? Or 7%? How does one decide? No empirical basis for this scoring decision is given. Further, not all tests use the same frequency criteria. This further complicates the rationale. Also, 2s may be awarded for unlisted items on the basis of their "creative strength." A feel for this "variable" is hard to get.

This statement clarifies how meaning is given to scores. If an individual's response only matches with 2% or less of the sample of 500 test takers, it is considered creative enough to merit 2 points, instead of, presumably, 1. Although the appropriateness of this process is questioned by this reviewer, it should be evident what type of score referencing this is.

In terms of evaluating the scoring process, the reference to "500 unspecified test takers" and the use of different percentages without empirical basis is concerning. This weighting in the scoring process could introduce bias that leads to construct misrepresentation. But this is an issue we might be willing to overlook, as it only seems to influence the relative weighting of scores for each task, in comparison to one another, as opposed to the overall score given. In other words, we could assume this would introduce a systematic error that could impact how creative we determine a child to be at a given test administration, but not how much change in creativity we measure over time.

10.6 Test use

As a final general consideration we need to examine the recommended uses for the test we are evaluating. Here are the questions we should ask. What are the recommended uses or test score inferences or interpretations? Do these uses match the test purpose? Does the test development process support the intended use? And are there appropriate cautions against unsupported test uses?

Regarding the TTCT, the test purpose according to the MMY summary is broad and vague: "To identify and evaluate creative potential through words (verbal forms) and pictures (figural forms)." However, research and technical information cited in the reviews demonstrates test use in pre/post studies similar to the hypothetical one we have devised here. Thus, the TTCT does appear to match our test purpose.

Note that we did not get into the development process of the TTCT. This would require a literature review. Established tests like the TTCT typically have published research documenting their development and use in practice. For example, after using the test, we could publish our own reliability and validity evidence, and other important results from our study. This would contribute to the body of work supporting (or not) its further use.

10.7 Summary

This chapter provides a brief overview of the test evaluation process, using tests of creative problem solving as a context. Some key questions are discussed, including questions about test purpose, study design, reliability, validity, scoring, and test use.

10.7.1 Learning objectives

1. Review and critique the documentation contained in a test review, test manual, or technical report, including:

a. Data collection and test administration designs,

b. Reliability analysis results,

c. Validity analysis results,

d. Scoring and reporting guidelines,

e. Recommendations for test use.

2. Compare and contrast tests using reported information.

3. Use information reported for a test to determine the appropriateness of the test for a given application.

10.7.2 Exercises

1. Why is it important to consider first the purpose of a test when evaluating it for use in your own work or research?

2. Conduct a literature review on the TTCT.

3. Search the MMY reviews for a test on a topic that interests you. Use the information in the reviews for a given test to evaluate the test for its intended purpose.

Appendix A

PISA Reading Items

The items and scoring information below are excerpted from the Reading Literacy Items and Scoring Guides document available at nces.ed.gov/surveys/pisa/educators.asp (Organization for Economic Cooperation and Development 2009). Note that the format and content of the items and scoring information have been modified slightly for presentation here. In the PISA 2009 study, these items were administered in the context of other reading items and some general instructions not shown here were also given to students.

Three sets of reading items are included here under the headings Cell phone safety, The play's the thing, and Telecommuting. Each set contains multiple items, some of them selected-response and some constructed-response. Scoring rubrics are provided for the constructed-response items. These item sets correspond to `PISA09` items: `r414q02`, `r414q11`, `r414q06`, and `r414q09` for cell phone safety; `r452q03`, `r452q04`, `r452q06`, and `r452q07` for the play's the thing; and `r458q01`, `r458q07`, and `r458q04` for telecommuting.

A.1 Cell phone safety

Cell Phone Safety

Are cell phones dangerous?

	Yes	No
1.	Radio waves given off by cell phones can heat up body tissue, having damaging effects.	Radio waves are not powerful enough to cause heat damage to the body.
2.	Magnetic fields created by cell phones can affect the way that your body cells work.	The magnetic fields are incredibly weak, and so unlikely to affect cells in our body.
3.	People who make long cell phone calls sometimes complain of fatigue, headaches, and loss of concentration.	These effects have never been observed under laboratory conditions and may be due to other factors in modern lifestyles.
4.	Cell phone users are 2.5 times more likely to develop cancer in areas of the brain adjacent to their phone ears.	Researchers admit it's unclear this increase is linked to using cell phones.
5.	The International Agency for Research on Cancer found a link between childhood cancer and power lines. Like cell phones, power lines also emit radiation.	The radiation produced by power lines is a different kind of radiation, with much more energy than that coming from cell phones.
6.	Radio frequency waves similar to those in cell phones altered the gene expression in nematode worms.	Worms are not humans, so there is no guarantee that our brain cells will react in the same way.

Key Point
Conflicting reports about the health risks of cell phones appeared in the late 1990s.

Key Point
Millions of dollars have now been invested in scientific research to investigate the effects of cell phones.

If you use a cell phone …

Do	Don't
Keep the calls short.	Don't use your cell phone when the reception is weak, as the phone needs more power to communicate with the base station, and so the radio-wave emissions are higher.
Carry the cell phone away from your body when it is on standby.	Don't buy a cell phone with a high "SAR" value[1]. This means that it emits more radiation.
Buy a cell phone with a long "talk time". It is more efficient, and has less powerful emissions.	Don't buy protective gadgets unless they have been independently tested.

Key Point
Given the immense numbers of cell phone users, even small adverse effects on health could have major public health implications.

Key Point
In 2000, the Stewart Report (a British report) found no known health problems caused by cell phones, but advised caution, especially among the young, until more research was carried out. A further report in 2004 backed this up.

"Cell Phone Safety" above is from a website. Use "Cell Phone Safety" to answer the questions that follow.

Question 1 (R414Q02)

What is the purpose of the **Key points**?

A. To describe the dangers of using cell phones. B. To suggest that debate about cell phone safety is ongoing.

C. To describe the precautions that people who use cell phones should take. D. To suggest that there are no known health problems caused by cell phones.

Scoring

Correct: Answer B. To suggest that debate about mobile phone safety is ongoing.

Incorrect: Other responses.

Question 2 (R414Q11)

"It is difficult to prove that one thing has definitely caused another."

What is the relationship of this piece of information to the Point 4 **Yes** and **No** statements in the table **Are cell phones dangerous**?

A. It supports the Yes argument but does not prove it.
B. It proves the Yes argument.
C. It supports the No argument but does not prove it.
D. It shows that the No argument is wrong.

Scoring

Correct: Answer C. It supports the No argument but does not prove it.

Incorrect: Other responses.

Question 3 (R414Q06)

Look at Point 3 in the **No** column of the table. In this context, what might one of these "other factors" be? Give a reason for your answer.

Scoring

Correct: Answers which identify a factor in modern lifestyles that could be related to fatigue, headaches, or loss of concentration. The explanation may be self-evident, or explicitly stated.

Incorrect: Answers which give an insufficient or vague response.

Fatigue. [Repeats information in the text.]

Tiredness. [Repeats information in the text.]

Answers which show inaccurate comprehension of the material or are implausible or irrelevant.

Question 4 (R414Q09)

Look at the table with the heading **If you use a cell phone...**

Which of these ideas is the table based on?

A. There is no danger involved in using cell phones. B. There is a proven risk involved in using cell phones. C. There may or may not be danger involved in using cell phones, but it is worth taking precautions. D. There may or may not be danger involved in using cell phones, but they should not be used until we know for sure. E. The **Do** instructions are for those who take the threat seriously, and the **Don't** instructions are for everyone else.

Scoring

Correct: Answer C. There may or may not be danger involved in using cell phones, but it is worth taking precautions.

Incorrect: Other responses.

A.2 The play's the thing

Takes place in a castle by the beach in Italy.

FIRST ACT *Ornate guest room in a very nice beachside castle. Doors on the right and left. Sitting room set in the middle of the stage: couch, table, and two armchairs. Large windows at the back. Starry night. It is dark on the stage. When the curtain goes up we hear men conversing loudly behind the door on the left. The door opens and three tuxedoed gentlemen enter. One turns the light on immediately. They walk to the center in silence and stand around the table. They sit down together, Gál in the armchair to the left, Turai in the one on the right, Ádám on the couch in the middle. Very long, almost awkward silence. Comfortable stretches. Silence. Then:*

GAL: Why are you so deep in thought?

TURAI: I'm thinking about how difficult it is to begin a play. To introduce all the principal characters in the beginning, when it all starts.

ADAM: I suppose it must be hard.

TURAI: It is - devilishly hard. The play starts. The audience goes quiet. The actors enter the stage and the torment begins. It's an eternity, sometimes as much as a quarter of an hour before the audience finds out who's who and what they are all up to.

GAL: Quite a peculiar brain you've got. Can't you forget your profession for a single minute?

TURAI: That cannot be done.

GAL: Not half an hour passes without you discussing theatre, actors, plays. There are other things in this world.

TURAI: There aren't. I am a dramatist. That is my curse.

GAL: You shouldn't become such a slave to your profession.

TURAI: If you do not master it, you are its slave. There is no middle ground. Trust me, it's no joke starting a play well. It is one of the toughest problems of stage mechanics. Introducing your characters promptly. Let's look at this scene here, the three of us. Three gentlemen in tuxedoes. Say they enter not this room in this lordly castle, but rather a stage, just when a play begins. They would have to chat about a whole lot of uninteresting topics until it came out who we are. Wouldn't it be much easier to start all this by standing up and introducing ourselves? *Stands up.* Good evening. The three of us are guests in this castle. We have just arrived from the dining room where we had an excellent dinner and drank two bottles of champagne. My name is Sandor Turai, I'm a playwright, I've been writing plays for thirty years, that's my profession. Full stop. Your turn.

GAL: Stands up. My name is Gal, I'm also a playwright. I write plays as well, all of them in the company of this gentleman here. We are a famous playwright duo. All playbills of good comedies and operettas read: written by Gal and Turai. Naturally, this is my profession as well.

GAL and TURAI: *Together.* And this young man...

ADAM: *Stands up.* This young man is, if you allow me, Albert Adam, twenty-five years old, composer. I wrote the music for these kind gentlemen for their latest operetta. This is my first work for the stage. These two elderly angels have discovered me and now, with their help, I'd like to become famous. They got me invited to this castle. They got my dress-coat and tuxedo made. In other words, I am poor and unknown, for

now. Other than that I'm an orphan and my grandmother raised me. My grandmother has passed away. I am all alone in this world. I have no name, I have no money.

TURAI: But you are young.

GAL: And gifted.

ADAM: And I am in love with the soloist.

TURAI: You shouldn't have added that. Everyone in the audience would figure that out anyway.

They all sit down.

TURAI: Now wouldn't this be the easiest way to start a play?

GAL: If we were allowed to do this, it would be easy to write plays.

TURAI: Trust me, it's not that hard. Just think of this whole thing as...

GAL: All right, all right, all right, just don't start talking about the theatre again. I'm fed up with it. We'll talk tomorrow, if you wish.

"The Play's the Thing" is the beginning of a play by the Hungarian dramatist Ferenc Molnar.

Use "The Play's the Thing" to answer the questions that follow. (Note that line numbers are given in the margin of the script to help you find parts that are referred to in the questions.)

Question 1 (R452Q03)

What were the characters in the play doing just before the curtain went up?

Scoring

Correct: Answer which refer to dinner or drinking champagne. May paraphrase or quote the text directly.

They have just had dinner and champagne.

"We have just arrived from the dining room where we had an excellent dinner." [direct quotation]

"An excellent dinner and drank two bottles of champagne." [direct quotation]

Incorrect: Answers which give an insufficient or vague response, show inaccurate comprehension of the material, or are implausible or irrelevant.

Question 2 (R452Q04)

"It's an eternity, sometimes as much as a quarter of an hour..." (lines 29-30)

According to Turai, why is a quarter of an hour "an eternity"?

A. It is a long time to expect an audience to sit still in a crowded theatre. B. It seems to take forever for the situation to be clarified at the beginning of a play. C. It always seems to take a long time for a dramatist to write the beginning of a play. D. It seems that time moves slowly when a significant event is happening in a play.

Scoring

Correct: Answer B. It seems to take forever for the situation to be clarified at the beginning of a play.

Incorrect: Other responses.

Question 3 (R452Q06)

A reader said, "Adam is probably the most excited of the three characters about staying at the castle."

What could the reader say to support this opinion? Use the text to give a reason for your answer.

Scoring

Correct: Indicates a contrast between Adam and the other two characters by referring to one or more of the following: Adam's status as the poorest or youngest of the three characters; his inexperience (as a celebrity).

- Adam is poor, he must be excited to stay at a fancy castle.
- He must be happy to be with the two guys who can make him famous.
- He is writing music with two really famous people.
- He is young, and young people just get more excited about things, it's a fact!
- He's young to stay at the castle. [minimal]
- He has the least experience. [minimal]

Incorrect: Answers which give an insufficient or vague response.

- He is excited. [Repeats stem.]

Answers which show inaccurate comprehension of the material or give an implausible or irrelevant response.

- He is an artist.
- He has fallen in love. [not an explanation of why he is excited to be staying at the castle]
- Adam must be excited; surely the soloist will show up. [no support in the text]
- He has been given a tuxedo. [an explanatory detail, not the reason itself]

Question 4 (R452Q07)

Overall, what is the dramatist Molnar doing in this extract?

A. He is showing the way that each character will solve his own problems. B. He is making his characters demonstrate what an eternity in a play is like. C. He is giving an example of a typical and traditional opening scene for a play. D. He is using the characters to act out one of his own creative problems.

Scoring

Correct: Answer D. He is using the characters to act out one of his own creative problems.

Incorrect: Other responses.

A.3 Telecommuting

The way of the future, by Molly

Just imagine how wonderful it would be to "telecommute"^1 to work on the electronic highway, with all your work done on a computer or by phone! No longer would you have to jam your body into crowded buses or trains or waste hours and hours travelling to and from work. You could work wherever you want to – just think of all the job opportunities this would open up!

Disaster in the making, by Richard

Cutting down on commuting hours and reducing the energy consumption involved is obviously a good idea. But such a goal should be accomplished by improving public transportation or by ensuring that workplaces are located near where people live. The ambitious idea that telecommuting should be part of everyone's way of life will only lead people to become more and more self-absorbed. Do we really want our sense of being part of a community to deteriorate even further?

ˆ1"Telecommuting" is a term coined by Jack Nilles in the early 1970s to describe a situation in which workers work on a computer away from a central office (for example, at home) and transmit data and documents to the central office via telephone lines.

Use "Telecommuting" above to answer the questions that follow.

Question 1 (R458Q01)

What is the relationship between "The way of the future" and "Disaster in the making"?

A. They use different arguments to reach the same general conclusion. B. They are written in the same style but they are about completely different topics. C. They express the same general point of view, but arrive at different conclusions. D. They express opposing points of view on the same topic.

Scoring

Correct: Answer D. They express opposing points of view on the same topic.

Incorrect: Other responses.

Question 2 (R458Q07)

What is one kind of work for which it would be difficult to telecommute? Give a reason for your answer.

Scoring

Correct: Answers which identify a kind of work and give a plausible explanation as to why a person who does that kind of work could not telecommute. Responses MUST indicate (explicitly or implicitly) that it is necessary to be physically present for the specific work.

- Building. It's hard to work with the wood and bricks from just anywhere.
- Sportsperson. You need to really be there to play the sport.
- Plumber. You can't fix someone else's sink from your home!
- Digging ditches because you need to be there.
- Nursing – it's hard to check if patients are ok over the Internet.

Incorrect: Answers which identify a kind of work but include no explanation OR provide an explanation that does not relate to telecommuting.

- Digging ditches.
- Fire fighter.
- Student.
- Digging ditches because it would be hard work. [Explanation does not show why this would make it difficult to telecommute.]
- Gives an insufficient or vague response.
- You need to be there.

- Shows inaccurate comprehension of the material or gives an implausible or irrelevant response.
- Manager. No-one takes any notice of you anyway. [irrelevant explanation]

Question 3 (R458Q04)

Which statement would **both** Molly and Richard agree with?

A. People should be allowed to work for as many hours as they want to. B. It is not a good idea for people to spend too much time getting to work. C. Telecommuting would not work for everyone. D. Forming social relationships is the most important part of work.

Scoring

Correct: Answer B. It is not a good idea for people to spend too much time getting to work.

Incorrect: Other responses.

Abedi, Jamal. 2004. "The No Child Left behind Act and English Language Learners: Assessment and Accountability Issues." *Educational Researcher* 33: 4–14.

AERA, APA, and NCME. 1999. *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. doi:10.18637/jss.v067.i01.

Beck, A T, C H Ward, M Mendelson, J Mock, and J Erbaugh. 1961. "An Inventory for Measuring Depression." *Archives of General Psychiatry* 4: 53–63.

Bennett, Randy Elliot. 2011. "Formative Assessment: A Critical Review." *Assessment in Education: Principles, Policy & Practice* 18 (1). Taylor & Francis: 5–25.

Black, P., and D. Wiliam. 1998. "Inside the Black Box: Raising Standards Through Classroom Assessment." *Phi Delta Kappan* 80: 139–48.

Bloom, Benjamin Samuel, and David R Krathwohl. 1956. "Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain." Longmans.

Bond, L. A. 1996. "Norm- and Criterion-Referenced Testing." *Practical Assessment Research & Evaluation* 5 (2).

Bradfield, T. A., A. C. Besner, A. K. Wackerle-Hollman, A. D. Albano, M. C. Rodriguez, and S. R. McConnell. 2014. "Redefining Individual Growth and Development Indicators: Oral Language." *Assessment for Effective Intervention* 39: 233–44.

Brennan, R. L. 1992. "Generalizability Theory." *Educational Measurement: Issues and Practice* 11: 27–34.

———. 2001. *Generalizability Theory*. New York, NY: Springer.

———. 2013. "Commentary on 'Validating the Interpretations and Uses of Test Scores'." *Journal of Educational Measurement* 50: 74–83.

Briggs, D. C. 2009. "Preparation for College Admission Exams." Arlington, VA: National Association for College Admission Counseling.

Carter, S D. 2002. "Matching Training Methods and Factors of Cognitive Ability: A Means to Improve Training Outcomes." *Human Resource Development Quarterly* 13: 71–88.

Cizek, G. J. 2010. "An Introduction to Formative Assessment." In *Handbook of Formative Assessment*, edited by H. L. Andrade and G. J. Cizek, 3–17. New York, NY: Routledge.

College Board. 2012. "The SAT Report on College and Career Readiness: 2012." New York, NY: College

Board.

Cronbach, L J, and R. J. Shavelson. 2004. "My Current Thoughts on Coefficient Alpha and Successor Procedures." *Educational and Psychological Measurement* 64: 391–418.

de Ayala, R. J. 2009. *The Theory and Practice of Item Response Theory.* New York, NY: The Guilford Press.

De Boeck, Paul, Marjan Bakker, Robert Zwitser, Michel Nivard, Abe Hofman, Francis Tuerlinckx, and Ivailo Partchev. 2011. "The Estimation of Item Response Models with the Lmer Function from the Lme4 Package in R." *Journal of Statistical Software* 39 (12). American Statistical Association: 1–28.

Deno, S. L. 1985. "Curriculum-based measurement: The emerging alternative." *Exceptional Children* 52: 219–32.

Deno, S. L., L. S. Fuchs, D. Marston, and J. Shin. 2001. "Using curriculum-based measurement to establish growth standards for students with learning disabilities." *School Psychology Review* 30: 507–24.

Doran, H., D. Bates, P. Bliese, and M. Dowling. 2007. "Estimating the Multilevel Rasch Model: With the Lme4 Package." *Journal of Statistical Software* 20 (2): 1–18.

Ebel, R. 1961. "Must All Tests Be Valid?" *American Psychologist* 16 (640–647).

Embretson, S. E., and S. P. Reise. 2000. *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Ferketich, S. 1991. "Focus on Psychometrics: Aspects of Item Analysis." *Research in Nursing & Health* 14: 165–68.

Fuchs, L. S., and D. Fuchs. 1999. "Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment." *School Psychology Review* 28: 659–71.

Gardner, William L, and Mark J Martinko. 1996. "Using the Myers-Briggs Type Indicator to Study Managers: A Literature Review and Research Agenda." *Journal of Management* 22: 45–83.

Goodwin, Laura D. 2001. "Interrater Agreement and Reliability." *Measurement in Physical Education and Exercise Science* 5: 13–34.

Haladyna, T. M., and M. C. Rodriguez. 2013. *Developing and Validating Test Items.* New York, NY: Routledge.

Haladyna, T. M., S. M. Downing, and M. C. Rodriguez. 2002. "A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment." *Applied Measurement in Education* 15: 309–34.

Hambleton, R. K., and R. W. Jones. 1993. "Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development." *Educational Measurement: Issues and Practice*, 38–47.

Harvey, R. J., and A. L. Hammer. 1999. "Item Response Theory." *The Counseling Psychologist* 27: 353–83.

Haynes, S. N., D. C. S. Richard, and E. S. Kubany. 1995. "Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods." *Psychological Assessment* 7: 238–47.

Hockenberry, Marilyn J, and D Wilson. 2012. *Wong's Essentials of Pediatric Nursing.* St Louis, MO: Mosby.

Hursh, D. 2005. "The Growth of High-Stakes Testing in the USA: Accountability, Markets, and the Decline in Educational Equality." *British Educational Research Journal* 31: 605–22.

Kane, M. T. 2013. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50: 1–73.

Kelley, T. L. 1927. *Interpretation of Educational Measurements.* Yonkers, NY: World Book Co.

Kline, Paul. 1986. *A Handbook of Test Construction: Introduction to Psychometric Design.* New York, NY: Methuen.

Kuncel, N. R., S. A. Hezlett, and D. S. Ones. 2001. "A Comprehensive Meta-Analysis of the Predictive Validity of the Graduate Record Examinations: Implications for Graduate Student Selection and Performance."

Psychological Bulletin 127: 162–81.

Likert, R. 1932. "A Technique for the Measurement of Attitudes." *Archives of Psychology* 22: 5–55.

Linn, R. L., E. L. Baker, and D. W. Betebenner. 2002. "Accountability Systems: Implications of Requirements of the No Child Left Behind Act of 2001." *Educational Researcher* 31 (3–16).

Lord, F. M. 1952. "A theory of test scores." *Psychometric Monographs. No. 7.*

Mehrens, W. A. 1992. "Using performance assessment for accountability purposes." *Educational Measurement: Issues and Practice* 11: 3–9.

Messick, S. 1980. "Test validity and the ethics of assessment." *American Psychologist* 35: 1012–27.

Meyer, A. N. D., and J. M. Logan. 2013. "Taking the Testing Effect Beyond the College Freshman: Benefits for Lifelong Learning." *Psychology and Aging* 28: 142–47.

Militello, Matthew, Jason Schweid, and Stephen G Sireci. 2010. "Formative Assessment Systems: Evaluating the Fit Between School Districts' Needs and Assessment Systems' Characteristics." *Educational Assessment, Evaluation and Accountability* 22 (1). Springer: 29–52.

Miller, C., and K. Stassun. 2014. "A Test That Fails." *Nature* 510: 303–4.

Myers, I. B., M. H. McCaulley, N. L. Quenk, and A. L. Hammer. 1998. "Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator." Palo Alto, CA: Consulting Psychologist Press.

Nelson, D. A., C. C. Robinson, C. H. Hart, A. D. Albano, and S. J. Marshall. 2010. "Italian Preschoolers' Peer-Status Linkages with Sociability and Subtypes of Aggression and Victimization." *Social Development* 19: 698–720.

Nelson, H. 2013. "Testing More, Teaching Less: What America's Obsession with Student Testing Costs in Money and Lost Instructional Time." American Federation of Teachers.

Nunnally, J. C, and I. H. Bernstein. 1994. *Psychometric Theory.* New York, NY: McGraw-Hill.

Organization for Economic Cooperation and Development. 2009. "PISA 2009 Reading Literacy Items and Scoring Guides." Retrieved on April 20, 2016 from https://www.oecd.org/pisa/pisaproducts/pisa-test-questions.htm.

Pittenger, D. J. 2005. "Cautionary Comments Regarding the Myers-Brigg Type Indicator." *Consulting Psychology Journal: Practive and Research* 57: 210–21.

Pope, K S, J N Butcher, and J Seelen. 2006. *The MMPI, MMPI-2, & MMPI-A in Court: A Practical Guide for Expert Witnesses and Attorneys (3rd).* Washington, DC: American Psychological Association.

Popham, W. J., and T. R. Husek. 1969. "Implications of Criterion-Referenced Measurement." *Journal of Educational Measurement* 6: 1–9.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago, IL: University of Chicago Press.

Raymond, M. 2001. "Job Analysis and the Specification of Content for Licensure and Certification Examinations." *Applied Measurement in Education* 14: 369–415.

Rizopoulos, Dimitris. 2006. "ltm: An R Package for Latent Variable Modelling and Item Response Theory Analyses." *Journal of Statistical Software* 17 (5): 1–25. http://www.jstatsoft.org/v17/i05/.

Rodriguez, M. C. 2005. "Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research." *Educational Measurement: Issues and Practice* 24: 3–13.

Roediger, H. L., P. K. Agarwal, M. A. McDaniel, and K. B. McDermott. 2011. "Test-Enhanced Learning in the Classroom: Long-Term Improvements from Quizzing." *Journal of Experimental Psychology: Applied* 17:

382–95.

Santelices, M. V., and M. Wilson. 2010. "Unfair Treatment? The Case of Freedle, the SAT, and the Standardization Approach to Differential Item Functioning." *Harvard Educational Review* 80: 106–34.

Shavelson, R. J., and Norman L. Webb. 1991. *Generalizability Theory: A Primer*. SAGE Publications: Thousand Oaks, CA.

Spector, Paul E. 1992. *Summated Rating Scale Construction: An Introduction*. SAGE Publications.

Stevens, S. S. 1946. "On the Theory of Scales of Measurment." *Science* 103: 677–80.

Stiggins, R. J. 1987. "The Design and Development of Performance Assessments." *Educational Measurement: Issues and Practice* 6: 33–42.

———. 1991. "Assessment Literacy." *Phi Delta Kappan* 72 (534–539).

Torrance, E. P. 1981a. "Empirical Validation of Criterion-Referenced Indicators of Creative Ability Through a Longitudinal Study." *Creative Child and Adult Quarterly* 6: 136–40.

———. 1981b. "Predicting the Creativity of Elementary School Children." *Gifted Child Quarterly* 25: 55–62.

US Department of Education. 2002. "A New Era: Revitalizing Special Education for Children and Their Families." Washington, DC: US Department of Education.

Webb, Norman L. 2002. "Depth-of-Knowledge Levels for Four Content Areas." *Language Arts*.

Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York. http://had.co.nz/ggplot2/book.

Wickham, Hadley, and Winston Chang. 2016. *devtools: Tools to Make Developing R Packages Easier*. https://CRAN.R-project.org/package=devtools.

Wiliam, D., and P. Black. 1996. "Meanings and consequences: A basis for distinguishing formative and summative functions of assessment?" *British Educational Research Journal* 22: 537–48.

Xie, Y. 2016. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.12.3.

Xie, Yihui. n.d. *Bookdown: Authoring Books with R Markdown*. https://github.com/rstudio/bookdown.