

# VALIDITY AND VALIDATION

U N D E R S T A N D I N G  
S T A T I S T I C S



CATHERINE S. TAYLOR

OXFORD

---

# VALIDITY AND VALIDATION

**SERIES IN UNDERSTANDING STATISTICS**

S. NATASHA BERETVAS      Series Editor

**SERIES IN UNDERSTANDING MEASUREMENT**

S. NATASHA BERETVAS      Series Editor

**SERIES IN UNDERSTANDING QUALITATIVE RESEARCH**

PATRICIA LEAVY      Series Editor

---

**Understanding Statistics**

*Exploratory Factor Analysis*  
Leandre R. Fabrigar and Duane  
T. Wegener

*Validity and Validation*  
Catherine S. Taylor

**Understanding Measurement**

*Item Response Theory*  
Christine DeMars

*Reliability*  
Patrick Meyer

**Understanding Qualitative  
Research**

*Oral History*  
Patricia Leavy

*Fundamentals of Qualitative Research*  
Johnny Saldaña

*The Internet*  
Christine Hine

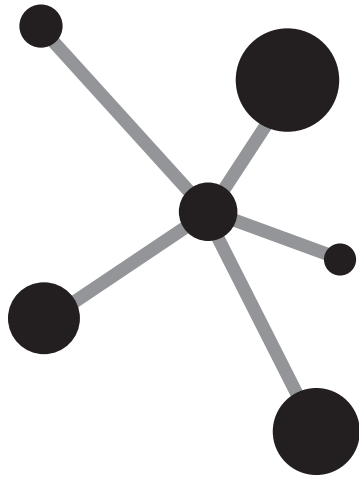
*Duoethnography*  
Richard D. Sawyer and Joe Norris

*Qualitative Interviewing*  
Svend Brinkmann

CATHERINE S. TAYLOR

---

# VALIDITY AND VALIDATION



OXFORD  
UNIVERSITY PRESS

# OXFORD

UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.  
It furthers the University's objective of excellence in research, scholarship,  
and education by publishing worldwide.

Oxford New York  
Auckland Cape Town Dar es Salaam Hong Kong Karachi  
Kuala Lumpur Madrid Melbourne Mexico City Nairobi  
New Delhi Shanghai Taipei Toronto

With offices in  
Argentina Austria Brazil Chile Czech Republic France Greece  
Guatemala Hungary Italy Japan Poland Portugal Singapore  
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trademark of Oxford University Press  
in the UK and certain other countries.

Published in the United States of America by  
Oxford University Press  
198 Madison Avenue, New York, NY 10016

© Oxford University Press 2013

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press, or as expressly permitted by law, by license, or under terms agreed with the appropriate reproduction rights organization. Inquiries concerning reproduction outside the scope of the above should be sent to the Rights Department, Oxford University Press, at the address above.

You must not circulate this work in any other form  
and you must impose this same condition on any acquirer

Library of Congress Cataloging-in-Publication Data  
Taylor, Catherine S.

Validity and validation / Catherine S. Taylor.

pages cm. — (Understanding statistics)

ISBN 978-0-19-979104-0

1. Social sciences—Statistical methods. 2. Social sciences—Methodology.

3. Research methods. I. Title.

HA29.T3237 2013

001.4'22—dc23 2013008389

9 8 7 6 5 4 3 2 1

Printed in the United States of America  
on acid-free paper

*For Laurie, Robin, and Courtney  
Thank you.*

*This page intentionally left blank*

---

# CONTENTS

	Acknowledgments . . . . .	viii
CHAPTER 1	Validity and Validation in Research and Assessment . . . . .	1
CHAPTER 2	Evidence for the Internal Validity of Research Results . . . . .	24
CHAPTER 3	External Threats to Validity . . . . .	55
CHAPTER 4	Validity of Statistical Conclusions . . . . .	65
CHAPTER 5	Construct-Related Evidence for Validity . . . . .	82
CHAPTER 6	Interpretation, Use, and Consequences of Scores from Assessments . . . . .	147
CHAPTER 7	Validity Theory and Validation Resources . . . . .	189
	Index . . . . .	201



---

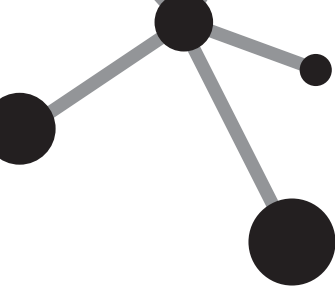
## ACKNOWLEDGMENTS

I would like to thank Robert Abbott for his willingness to think with me about how to use correlational models to control for threats to validity and Natasha Beretvas for her careful review of the chapters for their usefulness and accuracy. Thanks also to the Washington State Office of the Superintendent of Public Instruction for the use of state assessment data to generate examples. I am grateful to the individuals who have thought deeply about validity and the potential pitfalls that occur when its meaning is reduced to a handful of correlations. Thanks to Lee Cronbach, Michael Kane, Samuel Messick, Robert Linn, Robert Mislevy, Pamela Moss, and Lorrie Shepard for their elegant and thoughtful papers on assessment validity and validation. Their work has had a profound impact on my thinking about validity, my assessment development work, and the focuses of my research. They have deepened my understanding of validity and the ethics of assessment as well as the critical importance of assessment purpose in determining the range of research needed to validate assessment score interpretation and use.

---

# VALIDITY AND VALIDATION

*This page intentionally left blank*



# VALIDITY AND VALIDATION IN RESEARCH AND ASSESSMENT

THIS BOOK IS an effort to collect the thinking about validity from the past 50 years into a single volume. Most textbooks on measurement and research contain a chapter on validity. In the field of measurement, discussions about the definition of validity pepper conferences and journals every year. It is, perhaps, unfortunate that the term *validity* was ever coined. In common parlance, validity appears to be a thing with substance. In fact, validity is not a thing, nor is it a property of things. It is an adjective associated with claims. The term *validity* should always be phrased “the validity of...” You may ask, “The validity of what?” That is a fair question.

To understand validity, one must understand how humans make sense of their worlds—through inferences, interpretations, and conclusions. *Inferences* are statements about unseen connections between phenomena. For example, if plants grow better in full sunlight than in a dimly lit room, one can infer that plant growth is related to light. Generally, inferences are closely tied to observable evidence. *Interpretations*, like inferences, involve taking evidence and making sense of it; however, interpretations are

generally more value-laden. A psychologist might interpret a client's behaviors as friendly or hostile. *Conclusions* are summaries that take into account a range of available data. For example, suppose a scientist collects data from a range of individuals with and without skin cancer regarding their lifestyles and daily behaviors. If the scientist claims, based on the collection of evidence, that caffeine intake before sun exposure decreases the likelihood of skin cancer, the scientist is drawing a conclusion from a wide range of available data. Inferences, interpretations, and conclusions involve making sense of observable phenomena. Inferences, interpretations, and conclusions are not objects of substance. They are as substantial as air. They are claims made by researchers based on available evidence. As claims, they can be questioned, challenged, and tested. We can question the validity of inferences drawn from data; the validity of interpretations based on test scores; the validity of conclusions drawn from research results. In this book, I will use the word *claim* to refer to inferences, interpretations, or conclusions, unless use of a more specific term is appropriate.

"Sound," "well-founded," "justified," and "logical" are some of the words dictionaries use to define the term *valid*. When claims are sound, they are likely to be reasoned and logical. When claims are well-founded or justified, they are likely to be supported by evidence. They help to frame the strategies we use to question or support the validity of claims. What is the logical argument? What is the empirical evidence? How do we know if claims are warranted? The process of evaluating the logical arguments and scientific evidence that support claims is called *validation*.

Validation in research involves close scrutiny of logical arguments and the empirical evidence to determine whether they support theoretical claims. Similarly, validation in assessment involves evaluating logical arguments and empirical evidence to determine whether they support proposed inferences from, as well as interpretations and uses of, assessment results. Researchers make an effort to mitigate possible threats to the validity of their claims while they gather evidence to support their theories. Test developers gather evidence to support the interpretations to be made from scores and other measures<sup>1</sup> during and after the development of

---

1. A test score is typically a numerical value that results from some measurement procedure. However, measurement is not the only form of assessment,

an assessment tool. As consumers of research reports or users of assessment tools, scientists, educators, and psychologists have an obligation to examine both logical arguments and empirical evidence to determine whether the claims made by the researchers and the interpretations proposed by assessment developers can be trusted.

The purpose of this book is to further define validity and to explore the factors that should be considered when evaluating claims from research and assessment.<sup>2</sup> Those who write about validity generally focus on either assessment or research. However, research and assessment are inextricably related to one another. Research studies support the interpretation and uses of assessment results; assessment results support theory building and problem solving based on research. In this book, I will attempt to summarize current thinking about validity as it relates to both research and assessment.

This chapter is an overview of validity theory and its philosophical foundations, with connections between the philosophical foundations and specific ways that we consider validation in research and measurement. Chapter 2 presents strategies to address potential threats to the internal validity of research claims. Chapter 3 presents ways to address potential threats to the external validity of research claims. Chapter 4 discusses strategies for controlling potential threats to the validity of statistical conclusions. Chapters 5 and 6 focus on evidence for the validity of inferences and interpretations from test scores and other measures as well as evidence for the validity of uses of test scores and other measures. Chapter 5 addresses construct-related evidence for the validity of test scores, and Chapter 6 is focused on evidence for the validity of interpretations and uses of test scores, as well

---

and numerical test scores are not the only results of assessment procedures. Assessment results may include descriptive summaries, rubric levels, proficiency levels, and other summaries. As shorthand, I will use *test scores* or *scores* to describe any summary based on an assessment process. It is important to note that the validation issues that apply to numerical test scores apply to all summaries based on assessment procedures.

2. Throughout this book, I use *assessment* to refer to the systematic collection of information (numerical data, descriptions, etc.) and the interpretations made from that information. I use the term *assessment* rather than *measurement* because assessment encompasses traditional notions of measurement as well as more qualitative descriptions of phenomena.

as the consequences of test score interpretation and use. Finally, Chapter 7 provides references to other sources that deal with these subjects in more depth.

## **Validation in Theory Building and Assessment Development**

The primary purpose of research is to build theory and to develop causal explanations for phenomena. We also use research to solve human problems. However, even problem-solving requires theories (or, at a minimum, hypotheses) about the causes of problems and possible solutions. In what follows, I summarize perspectives on validity theory and describe how they influence our thinking about how to validate claims. I then present an overview of the key ideas that will be discussed more thoroughly in subsequent chapters.

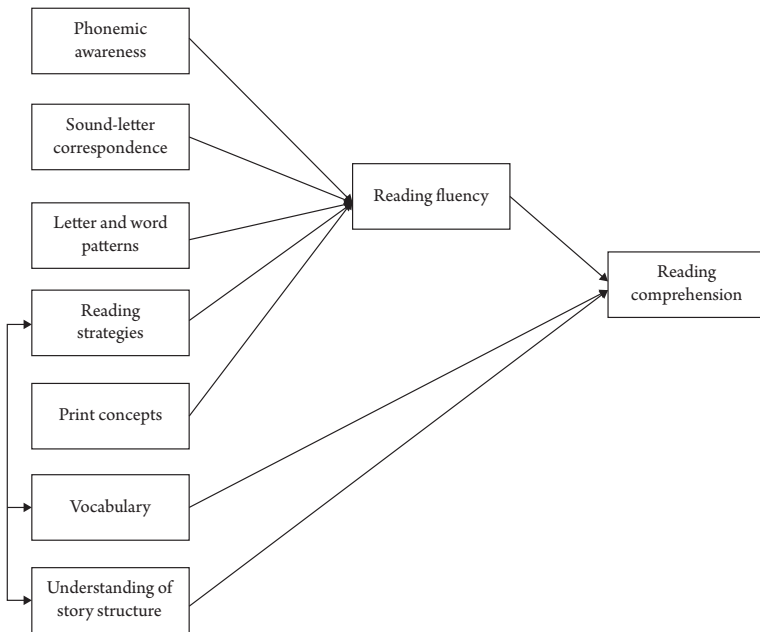
Figure 1–1 presents a hypothetical theory of the causes of students' ability to comprehend written text. These possible causes of reading comprehension ability are, by default, possible explanations for problems in reading comprehension. Each box represents a *construct*<sup>3</sup>—a set of related behaviors and/or cognitive processes that are grouped together and named by reading researchers. The task of reading theorists is to build and test such a theory through systematic research. Reading researchers observe and assess readers to see what they do as they develop reading comprehension abilities. They define constructs and generate causal statements. This theoretical system of relationships is called a “nomological network” and represents logical arguments within the theory.

To test these causal connections, researchers must have ways to assess each of the constructs. The development of assessments begins with a set of definitions—of the constructs, of the behaviors and tasks that will demonstrate each construct, of how those behaviors and tasks will be elicited from examinees, of how responses will be scored, and of how scores will be interpreted. These are the logical arguments underlying an assessment.

Logical arguments are necessary, but not sufficient, for validation. In the popular media, we often hear statements such as, “This drug has been clinically proven to treat symptoms of

---

3. The terms *ability*, *trait*, and *latent trait* are often used in place of *construct*. Regardless of the selected term, constructs are human labels for observed regularities in behaviors or phenomena.



**Figure 1–1** Nomological Network to Represent a Theory of Reading Comprehension

depression,” or “This achievement test is a valid and reliable measure of mathematical ability.” What evidence is brought to bear on these claims? The reader, viewer, or listener does not have access to the evidence and must trust the advertiser, testing company, or government agency to be telling the truth. However, in research and assessment, nothing is ever “proven”; no test scores are absolutely reliable and valid. Research is an ongoing process of refining and improving our understanding of cause-and-effect relationships such that our theories are generalizable across individuals in a population, over time, and in a wide range of settings. Testing, refining, and improving theories require empirical evidence. Developing, refining, and evaluating inferences from assessment scores also require empirical evidence.

## Validity Theory

Philosophers of science have long debated the ways in which to evaluate the validity of claims. Table 1–1 provides a brief, somewhat chronological summary of the philosophical antecedents to



Table 1–1

**Philosophical Foundations of Validity Theory**

<b>Philosophical Stance</b>	<b>Guiding Principles</b>
Positivism (Auguste Comte, 1848; Carl Hempel, 1967) and Instrumentalism (Ernst Mach, 1882)	<p>A theory is a well-defined set of statements that define a phenomenon</p> <p>Theory is defined through logical processes (nomological network)</p> <p>Axioms are established related to theory</p> <p>Axioms can be probabilistic as long as the probabilities are defined in advance</p> <p>Theoretical constructs and hypothesized relationships give the network deductive or predictive power</p> <p>Axioms are verified through observational data (i.e., obtain proof)</p> <p>Rules of interpretation determine how concrete observations are to be understood</p> <p>A statement is true if it fits within the logical system of other statements that explain reality</p> <p>All statements within an explanatory theory must form a coherent whole</p> <p>A statement is true if it is useful in directing inquiry or action</p>
Empirical Falsification (Karl Popper, 1959)	<p>Theories cannot be proven</p> <p>Theories can be falsified through empirical evidence</p> <p>Evidence is gathered over time to support and/or refine constructs</p>
Rationalism (Descartes, 1637, 1644; Stephen Toulmin, 1972)	<p>A real world exists independently of theories about it</p> <p>Theory is created through both deductive and inductive mechanisms</p> <p>A key tool of science is scientific falsification—using evidence to demonstrate that a theoretical claim is false</p>

	<p>Researchers build, reorganize, and refine theory through observations and systematic tests</p> <p>Researchers work within domains—bodies of information that form a coherent whole</p> <p>Variant explanations stemming from different theories about phenomena probably share a common body of knowledge</p> <p>Science is objective if it accepts into the domain only the knowledge claims that are supported by evidence</p>
<p>Relativism (Paul Feyerabend, 1975; Thomas Kuhn, 1962)</p>	<p>One can never ‘prove’ a theory; one can only gather evidence to support the theory or to falsify a theory</p> <p>Theories are not value neutral</p> <p>Observation and meanings are theory-laden; theories are value laden</p> <p>Theories, associated methodologies, and resulting observations are tied to the world view of the researcher; therefore, choices of methods for falsification are dependent on the world view of the researcher</p> <p>Researchers must generate and test rival hypotheses</p> <p>Naturalistic observations or qualitative research are the primary methodologies to support or falsify theory</p> <p>True falsification involves discordant observations and alternative explanations for results</p> <p>All theories and data are grounded in situation and time; validation requires cross validation of evidence</p> <p>In its most extreme form, all theory is literally in the mind of the theorist; theory and theorist cannot be separated</p>

(continued)

Table 1–1

**(Continued)**

<b>Philosophical Stance</b>	<b>Guiding Principles</b>
Realism (David Hume, 1739; Immanuel Kant, 1781, 1788; Dudley Shapere, 1988)	<p>Theoretical statements are conjectures about attributes of the observable world</p> <p>Our best theories yield knowledge of aspects of the world, including unobservable aspects</p> <p>The best scientific theories are at least partially true</p> <p>To say that a theory is approximately true is sufficient explanation of the degree of its power to predict</p> <p>The approximate truth of a theory is the only explanation of its predictive success</p> <p>Causal relationships exist outside the human mind but they are not perceived accurately because of our fallible sensory and intellectual capacities</p> <p><i>All theoretical statements should be subject to critical tests</i></p> <p>When observations are not consistent with theoretical statements, either theory or methodology for gathering evidence are questioned</p> <p>Theoretical statements are not about facts but about causal properties of phenomena</p> <p>Events always derive from a complex of multiple causes at many different levels</p> <p>Through systematic tests, we attempt to define causal relationships</p> <p>The validation process requires developing representations of the processes that could reasonably cause a phenomenon and test the proposed causal relationships as well as alternative causal models that could account for the phenomenon</p>

validity theory discussed by Messick (1989) in his landmark chapter on validity. In the table, one can see reference to ideas that were described above: theory, causal relationships, constructs, observation, evidence, alternate explanations, and probability. Each of the philosophical stances described in Table 1–1 has an impact on how we investigate validity. *Positivism* and *instrumentalism* press for testable, logical statements about our theories and the causal relationships within our theories. The idea of *empirical falsification* has become central to scientific work. Although we can never prove theories, we can falsify them through results that run counter to theoretical expectations. *Relativism* raises awareness of the potential for bias in theories, methodologies, and interpretations of results. Relativism also presses for testing of rival explanations for the causes of results using methodologies that are consonant with these rival explanations. *Rationalism* and *realism*, two stances developed in reaction to the notions of relativism, claim that phenomena are real, not the inventions of theorists. Therefore, our work is to refine theories about phenomena as more evidence is accumulated and to doubt both our theories and our methodologies when conflicting results are obtained. *Realism* presses for replication of investigations, given our fallible methods and thought processes. *Realism* adds another important idea—that some constructs cannot be directly observed and must be inferred from observable behaviors.

Implicit in the ideas presented in Table 1–1 is that the purpose of research is to build theory. Humans build theory to make sense of phenomena. As indicated in the statements derived from realism, all theoretical statements should be subject to critical tests. Validation is a process by which we test theoretical statements. Even the simplest investigation is situated within a theory and involves testing or generating a causal explanation in the larger theory. The results of investigations either provide support for or falsify causal explanations. For example, when headache researchers give a placebo to a treatment group and an analgesic to a control group, the experiment is situated within a theory shown in Figure 1–2 (i.e., Migraine headaches are caused by three main factors: stress, swelling in the outer brain cover, and hormonal changes<sup>4</sup>). Since analgesics reduce swelling, the hypothesis being

4. See the Mayo Clinic website—<http://www.mayoclinic.com/health/migraine-headache/DS00120/> DSECTION=causes.

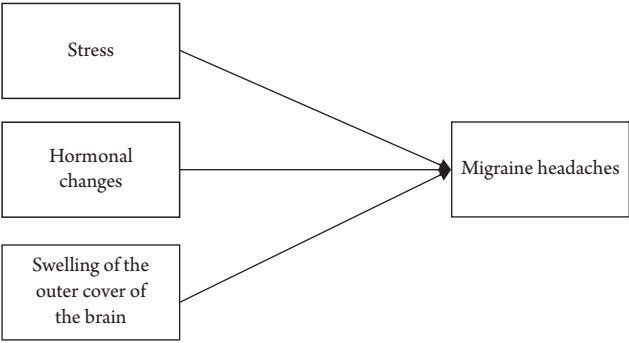


Figure 1–2 Theory of the Causes of Migraine Headaches

tested in this simple experiment is: If we reduce swelling, migraine headaches will lessen.

Threats to validity are those factors that lead us to doubt whether research and assessment claims can be trusted. These threats might derive from a myriad of sources. Generally, validation requires questioning the validity of claims in four areas: internal validity, external validity, statistical conclusion validity, and construct validity (Campbell & Stanley, 1966; Cook & Campbell, 1979, 1983; Cronbach & Meehl, 1955; Kane, 2006; Messick, 1989; Parker, 1993; Shadish, Cook, & Campbell, 2002). In research, internal validation asks whether the results of the investigation are truly due to expected causal relationships among variables. External validation asks whether the results of the investigation can be generalized beyond the situation in which the study was conducted. Statistical conclusion validity has to do with whether statistical conclusions can be trusted. Finally, in assessment we consider: (a) the connection between assessment results and the construct we intend to measure, (b) the usefulness of the results for the given purpose, and (c) the social consequences of inferences and actions based on test scores.

### Validation in Research

Quantitative research<sup>5</sup> is typically of two types: experimental or quasi-experimental research, and correlational research.

---

5. Many threats to validity in quantitative research also apply to qualitative research. The underlying principle is the same—to assess whether the results

Experimental or quasi-experimental research generally involves the control of variables and comparisons of groups in order to test causal relationships. Correlational research involves using statistical processes to look for explanatory patterns and trends in data. The goal of both types of research is to build models to explain phenomena. In this section, I briefly review potential threats to the validity of claims based on research results. Each of these potential threats is more fully discussed in Chapters 2 through 4.

### Internal Validity

Researchers must support their claims that the results of their investigations are attributable to the expected relationships among the identified variables in their investigations—using both logical arguments and empirical evidence. This is called *internal validity*. Common threats to internal validity of research claims can be grouped into four main categories: person factors (e.g., bias in selection, maturation, mortality, interactions with selection); measurement or statistical factors (e.g., pre-testing, instrumentation, statistical regression, ambiguity of results); situational factors (e.g., history, low reliability of treatment implementation, random irrelevancies in the treatment situation, diffusion or imitation of the treatment, and equalization of treatment); and alternate statistical models (e.g., alternative models that explain the relationships among the variables in the theory).

In terms of person factors, *bias in selection* occurs when individuals in the treatment and control groups differ from each other in a way that interacts with the construct being measured, or when the samples used for model-building do not adequately represent the target population. *Maturation* is a threat to internal validity when the natural changes in the participants affect the dependent variable, making it difficult to attribute change in the dependent variable to the causal variables described in a theory. *Mortality* is

---

of an investigation reflect something true beyond the biases of researchers or methodologies. Multiple lines of evidence are needed to support claims for both quantitative and qualitative research. In fact, the validation process for qualitative research requires much closer scrutiny of potential sources of bias. Qualitative researchers often rely on “critical friends” to examine their work or to cross-validate their observations. Textbooks on qualitative research generally include validation methodologies.

a threat to validity if subjects drop out of a study differentially for the treatment and control groups. *Interactions with selection* may be a threat to internal validity if participants are volunteers or if they can choose whether to be in the treatment or control condition in an experiment.

For statistical or measurement factors, *pre-testing* could be a threat to validity if participants altered their responses during post-testing because of familiarity with the items. *Instrumentation* is a threat to internal validity if the quality of the intervention deteriorates over time (e.g., if drugs are used after they exceed their shelf life) or if the scores from the measure are not reliable. *Statistical regression* could occur if some participants have extreme pre-test scores. Extreme scores tend to regress to the mean upon repeated assessment, even when no intervention or treatment has been provided. *Ambiguity of results* occurs when results do not clearly show a causal direction. In the migraine headache example described above, it may not be clear whether swelling causes migraine headaches or migraine headaches cause swelling.

In the case of situational factors that threaten internal validity, *history* threatens internal validity when events outside of the investigation influence the results. In the migraine headache example, suppose the researchers did not have a control group, the subjects were students, and final exams occurred during the treatment phase of the investigation. Final exams might increase the occurrence of migraine headaches; the end of the term could decrease the occurrence of migraine headaches. In this case, changes in the incidence of migraine headaches might have little to do with the use of an analgesic. *Unreliability of treatment implementation* could be a threat to internal validity if study participants do not complete research tasks as specified (e.g., if patients do not take prescribed medications consistently) or if the treatment providers differ in how they administer a treatment. *Random irrelevancies* could be a threat to internal validity if factors unrelated to the investigation impact the treatment or the post-test results (e.g., a fire drill during treatment or assessment could impact results). *Diffusion or imitation of treatment* could occur if participants in the control group are also exposed to a treatment during the period of the investigation. In the migraine headache example, participants in the control group might take a different analgesic to treat their headaches if a placebo does not decrease their pain.

Finally, *equalization of treatment* might occur if the treatment providers see the benefits of an intervention and provide the intervention to the control group members before the study is completed.

The idea of alternative statistical models is fairly self-explanatory. If an alternative statistical model explains the relationships among the variables better than or at least as well as the proposed theoretical model, this is a threat to the internal validity of theoretical claims.

The threats to internal validity reflect the ideas of *falsification*, *relativism*, and *realism*. The main stance of *falsification* is that researchers cannot prove a theoretical claim to be true; they can only attempt to falsify their claims. If a researcher can identify an alternative explanation for results (e.g., some person, situational, or measurement factor unrelated to the theory that may have caused the results) or if an alternate statistical model provides a better explanation for the relationships among variables in a model, the theoretical claims are weakened. If the theoretical claim is not falsified through close scrutiny of possible alternate explanations, there is more support for the theoretical claim.

One stance of *realism* is that, when observations are not consistent with theoretical claims, either the theory or the methodology is questioned. The first three categories of threats to internal validity focus on consistency of methodology. A stance of *realism* as well as *relativism* is that the validation process requires testing of alternative models that could account for the phenomenon. Examining research methodologies for flaws due to person, situational, and measurement factors, and examining alternative statistical models address this stance.

One of the most effective strategies for controlling for the first three categories of threats to the internal validity of research claims is through conducting experimental research in which individuals are randomly selected from a population and randomly assigned to treatment and control groups. However, true experimental design is rarely possible when conducting human research. A range of quasi-experimental and correlational designs can also be used to control for threats to internal validity (see Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). The most effective strategy for investigating alternate statistical explanations is through correlational research. Chapter 2 presents strategies researchers use for dealing with many of the internal



threats to the validity of claims using specific research designs and statistical methodologies.

### External Validity

The external validity of research claims is the degree to which results of investigations can be generalized beyond a specific investigation. Threats to the external validity of claims occur when research results are not generalizable across samples, times, and situations. We sample from known populations when conducting investigations. Generalization to the population as a whole requires that samples be representative of the targeted population. Generalization across times and situations requires that we make repeated tests during different times and in different situations or under different conditions. Specific external threats to validity that can be addressed by replication include: interactions among different treatments or conditions, interactions between treatments and methods of data collection, interactions of treatments with selection, interaction of situation with treatment, and interaction of history with treatment.

Using the migraine headache investigation as an example, *interactions among treatments* might occur if participants in the treatment condition also took a yoga class, making it difficult to generalize the results to individuals who did not participate in a yoga class. *Interactions between testing and treatment* could occur if pre-testing alerts participants to information that impacts their behaviors between pre-testing and post-testing. In this case, it would be difficult to generalize results to individuals who did not do a pre-test before the treatment. *Interaction of selection with treatment* might occur if those who volunteer for a study have dispositions that support or undermine the results of the investigation. *Interaction of treatment with setting* might occur if the setting of the study influences the investigative results. For example, if participants in the migraine headache study were inpatients in a hospital, the results could not be generalized to non-hospitalized individuals. *Interaction of history with treatment* could occur if the study took place in a location where a major crisis occurred (e.g., New Jersey during Hurricane Sandy). In this case, it would be difficult to generalize the results of the investigation to situations in which no crisis occurred. The threats to external validity described

here are related to *relativism* and *realism* in the same ways as threats to internal validity. Researchers must examine alternative explanations for research results in order to test research claims. Chapter 3 discusses threats to external validity in more detail and how researchers can address these threats.

## Validity of Statistical Conclusions

*Statistical conclusions* are conclusions that are made based on the strength of statistical results. Specific threats to statistical conclusion validity include: low statistical power, experiment-wise error, violating the assumptions of statistical tests, omitted variable bias, and over- or under-interpretation of statistical results.

*Statistical power* is a function of the relationship between probability of error, the number of participants in a sample, and the effect size (e.g., the size of differences between groups). Small sample sizes in an investigation generally have low statistical power. *Experiment-wise error* is a threat to validity when there are several statistical tests conducted in a single investigation. The potential for Type I (false positive) errors is accumulated over the statistical tests. *Omitted variable bias* in an investigation would be a threat to validity if an omitted but related variable impacted study results. For example, if there was a relationship between intensity of migraine headaches and blood pressure, statistical results of a study investigating the effects of analgesics on migraine headaches could be confounded by patients' blood pressure. *Violations of the assumptions of statistical tests* occur when the data to be analyzed are not consistent with basic assumptions for a specific statistical test. Parametric statistics generally assume that the scores from a test are equal-interval scores (i.e., the distance between adjacent scores is the same throughout the scale—as with inches or centimeters) and that the data are normally distributed in the population and in the samples. If the scores from a measure are ordinal and/or if the distributions of scores for samples are skewed, the statistical results are difficult to generalize beyond the samples in the study. Statistics texts (e.g., Garner, 2010; Urdan, 2010) generally provide detailed information about assumptions for various statistical tests. In addition, researchers have investigated whether parametric tests are robust to various violations to these assumptions (e.g., Boneau, 1960; Feir-Walsh & Toothaker, 1974; Hollingsworth,

1980; Keselman & Toothaker, 1974; Levy, 1980; Martin and Games, 1976; Ramsey, 1980; Wu, 1986; Zimmerman, 1998).

When researchers use experimental designs, statistical conclusions are generally focused on rejection of the *null hypothesis*—the hypothesis that there is no relationship between an independent variable and a dependent variable. In this case, we attempt to falsify the null hypothesis (see principles of *relativism* above). However, statistical conclusions do not always involve determining whether or not to reject a null hypothesis. We may want to find a statistical model that is the best explanation for the relationships among a set of measures. In this case, a threat to statistical conclusion validity would occur if we *over- or under-interpreted differences between alternative statistical models*. Threats to the validity of statistical conclusions in model testing are the same as those for null hypothesis testing. Chapter 3 presents strategies for addressing threats to the validity of statistical conclusions.

## **Validation and Assessment**

Validity theory, as it applies to assessment, has evolved a great deal over the past 50 years. Kane (2006) presents an excellent description of that evolution. Briefly, conceptions of validity began with the idea of prediction (Gulliksen, 1950). The validity question was, “Does the score from this assessment predict a criterion performance?” Even though validity theory has evolved over many years, this question is still an appropriate validity question when the criterion performance is well established (e.g., “Do the scores from this test predict a ship captain’s ability to pilot a ship?”). In the 1960s and 1970s, the increased use of achievement tests led to a new validity question, “Do the items on the test map onto the knowledge and skills in the achievement domain?” (Cronbach, 1971; Ebel, 1961). This is an appropriate validity question for all forms of testing. Once test developers outline the knowledge, skills, abilities, dispositions, and mental processes necessary for successful job performance or academic achievement, they must ensure that the tasks in the test represent this domain.

In 1955, Cronbach and Meehl published a seminal paper describing a third validity question—validation of the test as a measure of an underlying construct. The question—“What evidence do we have that the scores from this test reflect the

underlying trait?”—was considered appropriate for psychological testing. This question was proposed for situations in which there is no criterion performance per se, and when the domains involve internal psychological traits of examinees. By 1975, these three conceptions of validity were firmly established in professional literature. Criterion-related evidence for validity was seen as the *sine qua non* of assessment when the focus was on a known criterion such as job performance. Content-related evidence for validity was seen as a sufficient source of evidence for the validity of achievement test scores. Construct-related evidence for validity was considered the most essential type of evidence for scores from psychological tests.

Over time, more theorists and professional associations grappled with the concept of validity in assessment. Validity theorists recognized that, even for cases in which there is a known criterion performance, underlying traits (constructs) influence the likelihood of success on the criterion performance. Even in cases where an achievement domain is defined, test items tap into an array of underlying mental processes ranging from recall to complex reasoning skills. Current validity theorists agree that all three sources of evidence for validity are needed whenever test scores will be used to make inferences about examinee abilities, mental processes, or behaviors beyond responses to the test itself (Kane, 2006; Linn, 1997; Messick, 1989; Moss, 1998; Shepard, 1997).

In the 1970s and 1980s, a unified validity theory emerged. Conceptions of validity shifted from the test itself to the interpretations derived from test scores and other summaries (Cronbach, 1971, 1982). In 1989, Messick wrote a seminal chapter for the third edition of *Educational Measurement* (Linn, 1989). He laid out the philosophical foundations of validity theory and, based on those philosophical stances, proposed a two dimensional framework for validation studies. According to Messick,

Validity is an integrative, evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores and modes of assessment. (p. 13; emphasis in the original)

Based on this definition, validation of inferences and actions requires grounding in theory and research as well as multiple

lines of empirical evidence. Messick (1989) also suggested that validation is an ongoing process. Even if test developers generate strong evidence that the assessment scores provide a measure of the targeted construct, studies are needed to investigate the relevance and utility of using test scores for each intended purpose. Messick also claimed that researchers should investigate alternate explanations for test scores and consider the social consequences of test score interpretation and use.

Figure 1–3 is Messick’s (1989) two-dimensional framework for thinking about validation of interpretations and uses of test scores and other measures. This framework takes into account the meaning of assessment results, their usefulness in a given context, and the consequences of their interpretation and use. The top row of the framework, the evidential basis, begins with construct validity (which incorporates both criterion-related and content-related evidence for validity) as the foundation of validity, and takes into account the relevance and utility of scores and other summaries for a given purpose and situation.

The bottom row of the framework is focused on the consequential basis of test score interpretation and use. Messick expanded the definition of validity by adding consequences to the validity framework—indicating that validation studies must consider the consequences of the value implications of score interpretations as well as intended and unintended consequences of test score use. According to Messick, unintended consequences were a threat to validity if the causes of unintended consequences were *an over- or under-representation* of aspects of a construct or *construct-irrelevant variance*. Although consequences of test interpretation and use had always been an important issue in testing, many validity researchers had not considered consequences as validity issues.

	Interpretation	Use
Evidential Basis	Construct Validity	Construct Validity + Relevance & Utility
Consequential Basis	Construct Validity + Value Implications	Construct Validity, Relevance, Utility, & Social Consequences

Figure 1–3 Messick’s (1989) Facets of Validity

Kane (2006) reflected on the history of validity theory and acknowledged that, although conceptions of validity are now more thoroughly articulated than in the past, test developers and researchers have little guidance about how to establish a scope for validation research. With such a sweeping mandate, test developers are likely to retreat to simplistic investigations and avoid more complex studies that are costly in time and money. Kane proposed an argument-based approach to establishing a plan for validation studies—*beginning* with a clear definition of the intended interpretation(s) and uses of test scores.

Validation in assessment requires a clearly defined construct (e.g., reading comprehension) or criterion performance (e.g., piloting an airplane) and a clear articulation of the intended score interpretations and uses. For example, suppose two test developers constructed tests of reading comprehension. One test developer might be concerned with identification and treatment of students who are at risk of failing in school due to reading comprehension problems. This developer might focus her definition of reading comprehension on literal comprehension. A second test developer might need a test to select students for gifted programs. This developer might include inferences, interpretations, and generalizations along with literal comprehension in his definition of reading comprehension.

Clearly, the intended interpretations of test scores for these two example tests are quite different (risk of school failure versus ability to engage in challenging intellectual work). The uses of the test scores are also quite different (selection for a reading intervention versus selection for a gifted program). Using an argument-based approach (Kane, 2006), strategies for the validation of scores from these two tests would align with the intended interpretation and use of test scores. In addition, since selection for either a reading intervention or a gifted program has significant educational consequences for the students, today's validity theorists would urge reading researchers to consider alternate interpretations of test scores (Cronbach, 1971; Kane, 2006; Messick, 1989) and the unintended consequences of test score interpretation and use (Kane 2006; Messick, 1989).

### Construct-Related Evidence for Validity

Construct-related evidence for the validity of inferences from test scores is largely focused on the inner workings of

an assessment tool and the meanings of scores. Threats to the validity of inferences from test scores arise from poor definitions of constructs, over- or under-representation of aspects of the construct, construct-irrelevant variance, problematic items or tasks, low correlations among items or between items or tasks and the construct or criterion performance the tool is intended to measure, problematic scoring models, and poor reliability. Chapter 5 presents detailed strategies for investigating construct-related evidence for the validity of test scores and other measures.

### Interpretation, Use, and Consequences of Assessment Scores

Validation of the uses of assessment scores requires consideration of whether assessment results serve their intended purpose given a larger context. Potential threats to the validity of interpretation and use of assessment scores include: inappropriate interpretations of results, lack of evidence supporting use of the results in a given context or situation, unintended or unsupported interpretations, and consequences due to construct-irrelevant variance or inappropriate test use. Chapter 6 presents more detailed explanation of the interpretations and uses of assessment scores as well as the consequences of score interpretation and use.

### Summary

Validation studies investigate both logical arguments and empirical evidence. In research, documentation of research procedures provides one avenue for validation of theoretical claims made from research investigations. Empirical evidence is also essential. Do the research results behave in ways predicted by theory? Can we show evidence that the results were *not* caused by variables other than what the theory suggests? Are there alternate explanations for results? When research results are inconsistent with what would be predicted by theory, we can question our methods, our instrumentation, and our theories.

Researchers use assessments as tools in their work of building theory, investigating social and psychological phenomena,

and implementing social policies.<sup>6</sup> The validity of the claims of researchers and policymakers depends on evidence for the validity of inferences from test scores and other measures. Documentation regarding the rationale behind the structure of assessments (e.g., test specifications) and expert judgments of the alignment of assessments to the targeted constructs provide logical sources of evidence for validity of assessment scores. However, empirical evidence to support inferences from scores is essential. Do item and test scores behave as expected? Are there alternative explanations for scores? Clearly, the responsibility for researchers and assessment developers is the same—to provide substantive evidence to support the validity of their claims.

The following chapters present detailed descriptions of methods researchers use to investigate the validity of claims for theories or to investigate the validity of assessment results. The methods described are illustrative, but not exhaustive. New statistical methodologies are developed all the time. Just as researchers test theories, statisticians test and refine methodologies. Similarly, *psychometrics*, a specialized area of statistics focused on assessment development, is constantly being refined. The statistics used by psychometricians are also tested and refined over time. Despite these caveats, this book will provide a good starting place for researchers who want to ensure that the results of their investigations are trustworthy, researchers who are concerned about the quality of the assessment tools they use in their research, and individuals who develop tests and other assessments.

## References

- Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin*, 57, 49–64.
- Campbell, D., & Stanley, J. (1966). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

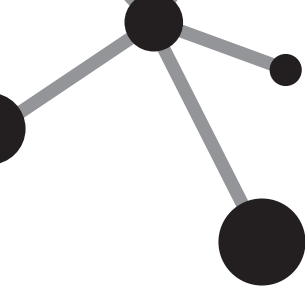
---

6. In education, achievement tests are increasingly used as instruments of social policy. Policymakers use test scores to drive educational reform, to evaluate the success of educational policies, and to reward or punish educational agencies for compliance with educational policies. In the world of work, policymakers use test scores insure that professionals are qualified to do their work, in order to protect the public.



- Comte, A. (1848). *A General View of Positivism*. London: Google Books.
- Cook, T., & Campbell, D. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.
- Cook, T., & Campbell, D. (1983). The design and conduct of quasi-experiments and true experiments in field settings. In M. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology* (pp. 223–326). Chicago: Rand McNally.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.; pp. 443–507). Washington, DC: American Council on Education.
- Cronbach, L. J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Descartes, R. (1637). Discourse on the method (J. Bennett, Trans.). Retrieved from [http://www.earlymoderntexts.com/f\\_descarte.html](http://www.earlymoderntexts.com/f_descarte.html), April 19, 2013.
- Descartes, R. (1644). Principles of philosophy (J. Bennett, Trans.). Retrieved from [http://www.earlymoderntexts.com/f\\_descarte.html](http://www.earlymoderntexts.com/f_descarte.html), April 19, 2013
- Ebel, R. (1961). Must all tests be valid? *American Psychologist*, 16, 640–647.
- Feir-Walsh, B. J., & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, Normal Scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement*, 34, 789–799.
- Feyerabend, P. (1975). *Against Method: Outline of an Anarchist Theory of Knowledge*. London: New Left Books.
- Garner, R. (2010). *Joy of Stats: A Short Guide to Introductory Statistics*. Toronto, CA: Toronto Press Incorporated.
- Gulliksen, H. (1950). *Theory of Mental Tests*. New York: Wiley.
- Hemple, C. G. (1967). Scientific explanation. In S. Morgenbesser, (Ed.), *Philosophy of Science Today* (pp. 79–88). New York: Basic Books.
- Hollingsworth, H. H. (1980). An analytical investigation of the effects of heterogeneous regression slopes in analysis of covariance. *Educational and Psychological Measurement*, 40, 611–618.
- Hume, D. (1739, 2000). In D. F. Norton and M. J. Norton, (Eds.), *A Treatise of Human Nature* (pp. 1–402). Oxford, England: Oxford University Press.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.; pp. 17–64). Washington, DC: American Council on Education.
- Kant, I. (1781, 1999). *Critique of Pure Reason*. Cambridge, England: Cambridge University Press.
- Kant, I. (1788, 1997). *Critique of Practical Reason*. Cambridge, England: Cambridge University Press.
- Keselman, H. J., & Toothaker, L. E. (1974). Comparison of Tukey's T-Method and Scheffe's S-Method for various numbers of all possible differences of averages contrasts under violation of assumptions. *Educational and Psychological Measurement*, 34, 511–519.
- Kuhn, T. (1962). *Structure of Scientific Revolution*. Chicago: University of Chicago Press.

- Levels of questioning: An alternative view. *Reading Research Quarterly*, 20, 586–602.
- Levy, K. (1980). A Monte Carlo study of analysis of covariance under violations of the assumptions of normality and equal regression slopes. *Educational and Psychological Measurement*, 40, 835–840.
- Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, 16(2), 14–16.
- Mach, E. (1882, 1910). The economical nature of physical inquiry. In T. J. McCormack (Trans.), *Mach, Popular Scientific Lectures* (pp. 186–213). Chicago: Open Court Publishers.
- Martin, C. G., & Games, P. A. (1976, April). ANOVA tests of homogeneity of variance when n's are unequal. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Messick, S. A. (1989). Validity. In Robert Linn (Ed.), *Educational Measurement* (3rd ed.; pp. 13–103). Washington, DC: American Council on Education.
- Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17(2), 6–12.
- Parker, R. M. (1993). Threats to validity of research. *Rehabilitation Counseling Bulletin*, 36, 130–138.
- Popper, K. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.
- Ramsey, P.H. (1980). Exact Type 1 error rates for robustness of student's t-test with unequal variances. *Journal of Educational Statistics*, 5, 337–349.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5–8, 24.
- Shapere, D. (1988). Modern physics and the philosophy of science. *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 201–210.
- Toulmin, S. (1972). *Human Understanding*. Princeton, NJ: Princeton University Press.
- Urdan, T. C. (2010). *Statistics in Plain English*. New York: Taylor and Francis Group.
- Wu, Y. B. (1984). Effects of heterogeneous regression slopes on the robustness of two test statistics in the analysis of covariance. *Educational and Psychological Measurement*, 44, 647–663.
- Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*, 67, 55–68.



## EVIDENCE FOR THE INTERNAL VALIDITY OF RESEARCH RESULTS

RESEARCH IS A systematic endeavor to build theory, to make sense of phenomena around us, and to solve problems. The most powerful research claims are causal claims. Theories involve a network of causal relationships. When explaining theories, researchers need to be confident that their research results support the causal relationships in the theory. Researchers may want to explain the causes of a phenomenon. For example, a researcher might want to explain the causes of bone deterioration in aging adults. In this case, the researcher needs to be confident that the causal relationships proposed in a theory are supported by evidence. When using research to solve problems, researchers need to understand the causes of the problems in order to intervene appropriately. For example, a reading researcher might be interested in the causes of reading failure in order to select the appropriate reading intervention for a child. Internal threats to validity are those that lead us to question the causal claims made by researchers.

In this chapter, I describe some of the strategies researchers use to control for or account for internal threats to the validity—factors that threaten causal claims. First, I describe experimental

and quasi-experimental research designs used to control internal threats to validity. I describe a limited number of research designs for illustrative purposes and focus on how those designs address internal threats to validity. Most research-method textbooks present a wider range and more detailed descriptions of research designs (e.g., Shadish, Cook, & Campbell, 2002). Next, I describe four correlational models that can account for factors that might threaten causal claims. Other books that explain these methods in more detail and provide examples of implementation include Cohen, Cohen, West, and Aiken (2002) and Hancock and Mueller (2006).

When investigating causal relationships, researchers may focus on a limited number of variables from within a larger nomological<sup>1</sup> network. Researchers might conduct a simple study focused on the relationship between two variables from within the overall theory in order to examine one causal relationship closely. Alternately, researchers might conduct studies in which they investigate several variables simultaneously.

Suppose two researchers were interested factors that affect reading comprehension. A nomological network for a theory of reading comprehension was posited in Chapter 1. An experimental researcher might focus a study on the causal relationship between two variables in the model: phonemic awareness and reading fluency. The researcher posits that students who receive instruction in phonemic awareness will read more fluently than students who do not receive this instruction. To test the proposed causal relationship, the researcher must: (a) define two key constructs (phonemic awareness and reading fluency), (b) have a clearly articulated intervention (instruction on phonemic awareness), (c) use trustworthy measures of both constructs, (d) set up a study in which some students receive instruction on phonemic awareness (treatment group) and others do not (control group), and (e) gather data to compare post treatment scores for students in the treatment and control groups.

A second reading researcher might focus on several causal variables in an attempt to test the theory as a whole. Rather than control variables through an experimental manipulation, the

---

1. See Chapter 1 for an explanation of nomological networks.

researcher might collect student performance data from tests measuring several different variables in the model and test the hypothesized relationships by examining the correlations among the variables.

For either of these cases, the researchers must consider internal threats to the validity of causal claims—threats that usually come from “nuisance variables” or variables that are not part of the theoretical model but that may affect the results. To set up complex investigations that control for internal threats to validity, researchers must clearly define *dependent* variables (those that are expected to be affected by changes in independent variables), *independent* variables (variables that cause changes), and *control* variables (variables that are controlled through features of the investigative design). The strategies for controlling or accounting for internal threats to validity differ depending on the research methodology used.

Four categories of internal threats to validity were introduced in Chapter 1:

1. Person factors (e.g., bias in selection, maturation, attrition, interactions with selection);
2. Measurement or statistical factors (e.g., pre-testing, instrumentation, statistical regression, ambiguity of results);
3. Situational factors (e.g., history, low reliability of treatment implementation, random irrelevancies in the treatment situation, diffusion or imitation of the treatment, and equalization of treatment); and
4. Alternate statistical models (e.g., alternate theoretical models that explain the patterns of relationships among the variables in the investigation).

In this chapter, I present several research designs and discuss how each design has attempted to manage potential internal threats to validity. I also discuss the internal threats to validity that have not been adequately addressed in the study design. At the end of the chapter, I summarize key ideas in the chapter and describe the responsibilities of researchers and consumers of research as they consider internal threats to the validity of causal claims.

## Controlling Internal Threats to Validity through Experimental Design: Random Selection and Random Assignment

One very powerful way to control for most internal threats to validity is to set up a tightly controlled experiment with random selection of participants and random assignment of participants to treatment and control conditions. This design is shown in Figure 2–1.

In many ways, the design is an elegant way to deal with person factors. If the participants are randomly selected from the target population, there is no bias in selection for the sample because it is likely to be a representative sample. If study participants are randomly assigned to treatment or control conditions, it is unlikely that the individuals in the treatment and control conditions are different in some systematic way (i.e., interaction with selection). Attrition (the likelihood of dropping out of the study) and maturation (the likelihood that growth will alter the behaviors of participants) are just as likely in either the treatment or the control conditions.

Several measurement and statistical factors are also managed through random sampling and random assignment. With random sampling, there is no need for pre-testing to address initial differences in the treatment and control groups; therefore, the influence of pre-testing (i.e., the likelihood that the pre-test gives participants a sense of what the study is about) is not a threat. Ambiguity of results (the likelihood that the causal direction may be reversed) is less likely because post-test differences between the two groups can probably be attributed to the treatment. However, researchers must still use care when selecting or developing instruments to ensure the reliability of results, and they must have sample sizes that are large enough to ensure statistical power.<sup>2</sup>

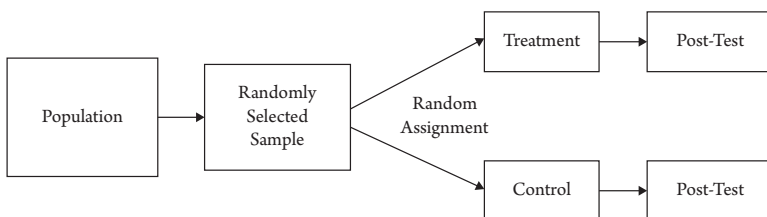


Figure 2–1 Simple Experimental Design

2. See Chapter 4 for a definition of statistical power.

In terms of situational factors, history and random irrelevancies are likely to be the same for both the treatment and control participants. Random assignment ensures that past experiences, or events taking place during the investigation, are equally likely to impact individuals in both groups.

### Potential Internal Threats to Validity Under Conditions of Random Selection and Random Assignment

The most likely internal threats to validity in a true experimental design are a result of situational factors. Researchers must implement the treatment and control conditions exactly as specified so that situational threats to validity (low reliability of treatment implementation, diffusion or imitation of the treatment, and equalization of treatment) do not occur. This means that researchers must control all aspects of the treatment. For research with humans, researchers must ensure that the treatment is administered in the same way to all participants and that aspects of the treatment are not inadvertently administered to individuals in the control condition.

Even with random assignment, humans are not mice or corn plants. We think about what we are doing and may change our behaviors based on our thinking. Three situational factors might affect the validity of claims based on experimental research results in human research: the Hawthorne effect, demand characteristics, and the experimenter effect.

In the case of *Hawthorne effect*, if participants in the control condition did not receive an intervention, there is a chance that those in the treatment condition may change simply because they are receiving some sort of treatment. For example, if a researcher wanted to know whether a given drug will alleviate depression, she might compare depression scores for patients who receive and who do not receive the drug. Changes in depression scores may occur simply because patients receive attention during the study. One way to address this potential threat is to give patients in the control group a placebo. In this way, the Hawthorne effect will affect participants in the treatment and control conditions equally.

*Demand characteristics* are the conditions of an investigation that give participants a sense of what is being studied and what outcomes are expected. Participants might adjust their behaviors

to be consistent with what they think the researcher is looking for. As with the Hawthorne effect, if individuals in both the treatment and control groups receive an intervention—only one of which is the treatment—demand characteristics are equally likely to influence results for the participants in both conditions.

Placebos help to control for both the Hawthorne effect and experimental demand characteristics; however, it may be difficult to provide a placebo in educational and psychological research. One strategy is to provide an alternate intervention. For example, in studies of terror management theory, Arndt, Greenberg, Pyszczynski, and Solomon (1997) investigated the degree to which being reminded of one's own death impacted social behaviors. Both the treatment and control groups completed a questionnaire. For the treatment condition, one item on the questionnaire referred to death. For the control condition, one item on the questionnaire referenced an unpleasant situation (e.g., pain). Once participants completed the questionnaire, they were asked to do tasks that required personal judgment. The researchers controlled for the demand characteristics of the investigation because participants in both treatment and control conditions completed the same questionnaire, and because participants in both groups responded to one item with a negative valence.

In the case of *experimenter effect*, a researcher, in her hopes for a positive effect, may act in ways that provide hints to participants about whether they are in the treatment or control condition. Again, this could lead participants in the treatment condition to behave in ways that support the desired results of the study. To control for this type of threat, researchers use double-blind studies. In a double-blind drug study, for example, the experimenter does not know whether individuals have taken a placebo or a treatment drug. Double-blind studies require a third party to maintain secure records of group assignment as well as similar treatment and control conditions so that the experimenter cannot determine which condition has been assigned (e.g., drug vs. placebo; questionnaire with death reference vs. questionnaire with pain reference).

## The Limits of Experimental Designs

The design in Figure 2–1 appears to be ideal for testing causal relationships between variables, and many internal threats to validity



can be controlled through this design. However, the design is difficult to implement, for several reasons.

### **Random Selection and Random Assignment**

In human research, it is nearly impossible to randomly select individuals from a larger population. Human experiments typically depend on available volunteers. In addition, if human research occurs in a laboratory, hospital, research center, or other controlled setting, and if the research involves an intervention that cannot be disguised, it is nearly impossible for researchers to be blind to the assigned conditions. For example, if the treatment is a particular non-pharmaceutical intervention for patients with depression, investigators and therapists would certainly know who is receiving the intervention. Also, since the treatment is provided by therapists, different therapists may implement the treatment differently (a threat to the reliability of treatment implementation). If treatment therapists are also seeing patients who have been assigned to the control condition, practice with the treatment could easily affect therapists' interactions with individuals who were assigned to a control condition (diffusion of treatment). In addition, if the treatment appears to be successful, therapists may feel obligated to give the treatment to all depressed patients (equalization of treatment).

In educational settings, if different teachers in the same school building are assigned to either the treatment or control conditions, they will know their assignment and may inadvertently influence students' performances (experimenter effects); teachers in the control condition might learn about the intervention during faculty meetings and emulate it in their own classrooms (diffusion of treatment). If the intervention has strong positive effects on students' learning, teachers may feel obligated to help other teachers learn how to use the intervention (equalization of treatment).

In human settings, not only are random selection and fidelity of intervention difficult to ensure, but participants may be clustered in some way (e.g., students within classrooms, clients with particular therapists, and citizens within communities). This increases the likelihood that nuisance variables will impact the results. For example, students are not randomly assigned to school districts or to school buildings, and rarely to teachers. The results of an educational research study will be impacted by factors such as the

socio-economic status (SES) of the community in which districts, schools, and classrooms are located. Communities with higher SES may put more money into their schools, have smaller class sizes, and provide more support for struggling students. Teachers may receive more professional development opportunities in more affluent school districts. Similar limitations occur in health care research when patients are clustered with doctors, clinics, hospitals, and so forth.

### **Feasibility**

Another barrier to pure experimental design in human research is that not all important theoretical questions can be answered using this strategy. For example, it would be immoral for a researcher to assign one group of participants to smoke a pack of cigarettes per day while the other group did not smoke cigarettes, in order to determine whether cigarette smoke causes lung diseases. Similarly, it would be inappropriate to require some teenagers to get pregnant and other teenagers to use birth control to determine the impact of teen pregnancy on high school graduation rates. Given the limitations of experimental design with humans, quasi-experimental designs and correlational designs have been developed.

### **Quasi-Experimental Strategies for Addressing Internal Threats to the Validity of Research Claims**

In what follows, I describe four quasi-experimental designs intended to address threats to the internal validity of causal claims. These designs are not exhaustive;<sup>3</sup> however, they illustrate a variety of strategies researchers use to control internal threats to the validity of causal claims when random selection and random assignment are not possible.

Suppose researchers are interested in whether cognitive therapy will decrease or eliminate symptoms of depression. Suppose their theory is that cognitive therapy will illuminate causes of patients' depressive thoughts and help patients address these underlying causes, thereby relieving their depression. The dependent variable

---

3. Most research-method textbooks provide comprehensive methods for conducting quasi-experimental research. See, for example, Cohen, Cohen, West, & Aiken (2002) or Shadish, Cook, & Campbell (2002).

is a measure of depression after treatment; the independent variable is the cognitive therapy treatment.

Controlling for Internal Threats to Validity Through Repeated Measures (Pre- and Post-Testing)

“Dr. Adams”<sup>4</sup> identifies two demographically comparable private hospitals in Arizona. Each hospital uses only antidepressant medications (drug therapy) as a treatment for severe depression. He designates one hospital as the treatment hospital and trains all of the therapists in the hospital to implement a particular cognitive therapy. He asks treatment therapists to meet with patients daily for ten days to provide cognitive therapy along with the drug therapy. Therapists in the control hospital continue to use only drug therapy. A depression scale is administered to patients as they enter each hospital and again after ten days. Dr. Adams selects a measure of depression that has well-documented evidence for the validity and reliability of scores in hospital settings. He uses a repeated-measures analysis of variance (ANOVA) to compare the levels of depression for patients from the two hospitals: the hospital that provides both cognitive therapy and drug therapy and the hospital that provides only antidepressants. Figure 2–2 shows Dr. Adam’s design.

In this design, Dr. Adams controls for sampling bias by comparing pre- and post-test differences in depression scores and through

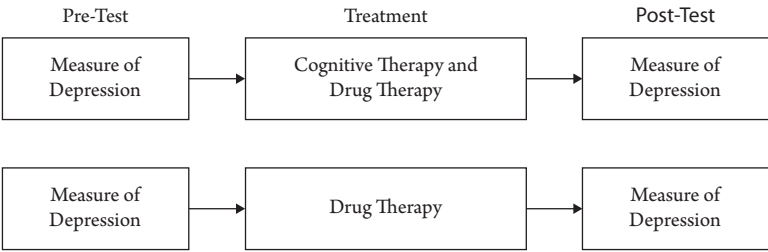


Figure 2–2 Pre-Test/Post-Test Quasi-Experimental Design

4. Throughout this book, I use pseudonyms for fictitious researchers. I have attempted to represent a variety of ethnicities in the names. Any pseudonyme that is shared by a researcher who is studying the described phenomenon is purely accidental. Name linked with citations, however, are authentic.

demographic matching of sites. Since he cannot randomly select patients and randomly assign them to treatment and control conditions, he uses pre-testing and repeated-measures ANOVA to control for differences in initial levels of depression for patients in the two hospitals. By selecting two demographically comparable hospitals, he controls for context differences that might influence the types of patients who enter the hospitals. By using two separate sites, he also controls for diffusion or imitation of treatment and equalization of treatment. Since therapists in both hospitals are likely to want patients to improve, experimenter effect will apply to patients in both hospitals. Since all patients are receiving some sort of treatment, the potential for Hawthorne effect applies to patients from both hospitals.

Dr. Adams uses a pre-test to control for initial differences in levels of depression because he cannot randomly assign patients to treatment or control conditions. Pre-testing is listed among the potential internal threats to validity. However, the potential for pre-testing to influence post-test responses will apply equally to the patients in both hospitals. Regression to the mean is also one of the potential threats to validity. Given that patients are hospitalized, it is entirely possible their depression scores will be very high at the beginning of treatment. Extreme scores have a tendency to regress toward the mean upon retesting. Therefore, it is possible that the post-treatment scores will be lower than pre-treatment scores without any treatment at all. However, as with pre-testing, this threat applies equally to patients from both hospitals.

Therefore, although changes in depression scores from pre-test to post-test may not be fully accounted for by the type of therapy (i.e., patients in both groups might systematically alter their responses due to experimenter effects or demand characteristics), the influence of pre-testing on scores is common to both groups. In terms of other measurement or statistical effects, regression to the mean is equally likely in both the treatment and the control groups. Finally, Dr. Adams uses a measure of depression that has strong evidence for the validity and reliability of scores, strengthening his argument that instrumentation is not an internal threat to validity.

One potential threat that Dr. Adams cannot control is the differences between hospitals. Although the two hospitals are demographically similar (as defined by the demographic

variables Dr. Adams uses), there will be differences that Dr. Adams cannot control. Another potential threat is reliability of treatment implementation. Dr. Adams may have to use some form of oversight to determine whether all therapists in the treatment hospital use the cognitive therapy consistently across patients and whether the therapists in the control hospital use only drug therapies. A third threat to validity in Dr. Adams’ study (an external threat that will be discussed in Chapter 3) is the possibility of treatment interactions, since patients who receive cognitive therapy also receive antidepressants.

Controlling for Internal Threats to Validity Through Block Design

“Dr. Bennett” conducts her study in four private hospitals—two in urban settings and two in suburban settings in the northeastern United States. She randomly designates one urban and one suburban hospital as treatment hospitals and the other two as comparison hospitals. She trains all therapists in the treatment hospitals to implement a specific cognitive therapy, which they will do in addition to drug therapy. She asks therapists to use the cognitive therapy with new patients for 10 consecutive days. She asks therapists in the control hospitals to use only drug therapy for the first 10 days of patients’ hospitalization. Dr. Bennett randomly selects four patients from each of the treatment and comparison therapists’ caseloads and measures their degree of depressive symptoms after 10 days of in-patient treatment for depression, using scores from a reliable measure of depression that has been validated in hospital settings. Figure 2–3 shows the design of Dr. Bennett’s study.

	Condition	
Setting	Treatment (Cognitive Therapy with Drug Therapy)	Control (Drug Therapy)
Urban	Post-Treatment Depression Measure	Post-Treatment Depression Measure
Suburban	Post-Treatment Depression Measure	Post-Treatment Depression Measure

Figure 2–3 Block Design to Control for Potential Differences in Hospital Settings

Dr. Bennett is using a quasi-random sampling process to attempt to control for bias in selection. Fully random assignment is not possible; however, Dr. Bennett knows that the location of a hospital can impact hospital conditions, and location may be related to the socio-economic status of entering patients. Therefore, she has selected two urban and two suburban hospitals. She controls for SES and hospital resources by assigning one of each type of hospital to the treatment and control conditions. By randomly assigning hospitals to treatment and control conditions and randomly selecting patients from each therapist within these hospitals, she has minimized (but not eliminated) selection bias as a potential internal threat to the validity of her research results.

Dr. Bennett attempts to control for diffusion or imitation of the treatment by designating different hospitals as treatment or comparison hospitals. She does not use a pre-test; therefore, pre-testing is not a potential internal threat to the validity of her research claims. History, maturation, and interactions with selection are equally likely to occur for patients in both the treatment and the control settings.

Dr. Bennett has not controlled for all potential internal threats to validity. For example, she may not be able to ensure that all treatment therapists are using the cognitive therapy consistently (reliability of treatment implementation) or that the therapists in the control hospital do not inadvertently provide cognitive therapy (equalization of treatment). Since patients in the treatment and control conditions all receive an intervention, and therapists want their patients to benefit from treatment, Hawthorne effects and experimenter effects are equally likely for both groups.

### Controlling Internal Threats to Validity Through Matching and Covariates

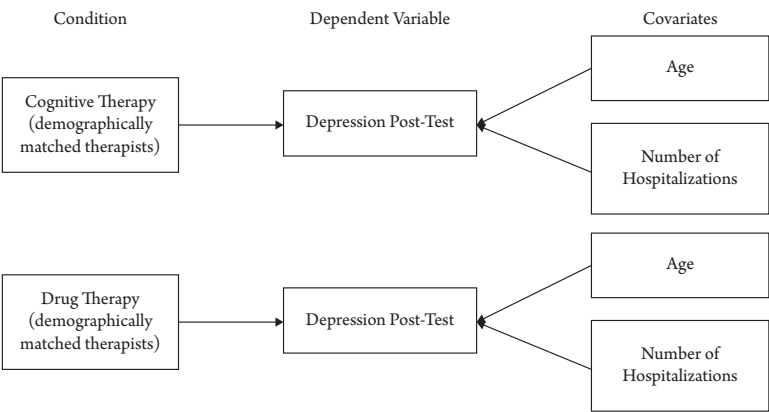
“Dr. Chang” contacts public and private hospitals in the Pacific Northwest and identifies therapists who already use cognitive therapy and therapists who use only drug therapy. He identifies variables that can be used to match treatment and comparison therapists (e.g., gender, age, and years of experience) and selects 24 matched pairs of therapists for each treatment condition.

He asks the therapists to test patients after two weeks of hospitalization, using a depression measure that has well-documented

evidence for the validity and reliability of scores in hospital settings. He asks the cognitive therapists to meet with their patients three times per week. In addition to post-treatment depression scores, Dr. Chang collects data on patients' ages and number of hospitalizations. He collects data from five randomly selected patients for each of the therapists. Dr. Chang uses analysis of covariance to compare the outcomes for patients who received either cognitive therapy or drug therapy. Figure 2–4 shows the design for his study.

Dr. Chang uses several strategies to control for internal threats to the validity of his research results. By carefully matching therapists based on gender and years of experience, he is attempting to control for therapist variables that might lead to random irrelevancies in the treatment implementation. By selecting therapists who already use either drug treatment or cognitive therapy, he controls for diffusion or imitation of treatment and equalization of treatment. He has statistically controlled for some potential biases in selection through covariates that may affect the effectiveness of treatment. Because all patients receive some sort of treatment, he is controlling for Hawthorne effect, demand characteristics, and experimenter effects. History, maturation, and attrition are equally likely to impact patients in both conditions. Finally, his post-test design means that pre-testing is not a threat, and his use of a reliable depression scale minimizes the threat from unreliability of scores.

There are several internal threats to validity that Dr. Chang has not controlled in this design. Despite the fact that Dr. Chang



**Figure 2–4** Design for Studying the Effects of Cognitive Therapy on Depression

controlled for some patient variables through covariance analysis and some therapist variables through matching, the potential for situation effects based on the type of hospital (public or private) and selection biases based on uncontrolled patient variables (e.g., gender, ethnicity, and socio-economic status) is still a threat. In addition, he cannot control the quality of the treatment given to patients in either condition (low reliability of treatment). Finally, there may be an interaction between treatments and therapists: the types of therapists who choose cognitive therapy and the types of therapists who use only drug therapy may differentially affect the ways therapists interact with patients.

Controlling Internal Threats to Validity by Combining Block Design with Covariate Design

“Dr. Davis” invites therapists from a wide range of psychiatric hospitals to participate in her study. She identifies two therapist variables that could influence implementation of treatment: therapists’ training and therapists’ years of experience. She also identifies two patient variables that research suggests might influence the results of cognitive therapy: gender and age. Figure 2–5 presents Dr. Davis’s design.

	Cognitive Therapy (Treatment)							
Training	Clinical Psychology				Psychiatry			
Therapist Years of Experience	≤ 10		> 10		≤ 10		> 10	
Patient Age	Covariate							
Patient Gender	M	F	M	F	M	F	M	F
Patients	5	5	5	5	5	5	5	5

	Drug Therapy (Comparison)							
Training	Clinical Psychology				Psychiatry			
Therapist Years of Experience	≤ 10		> 10		≤ 10		> 10	
Patient Age	Covariate							
Patient Gender	M	F	M	F	M	F	M	F
Patients	5	5	5	5	5	5	5	5

**Figure 2–5** Block Design with Age Covariate for Investigation of the Impact of Treatment on Symptoms of Depression



From the therapists who agree to participate, Dr. Davis randomly selects 20 therapists who use only drug therapy to treat depression and 20 therapists who use only cognitive therapy to treat depression. Ten therapists in each condition have fewer than ten years of experience, and ten have more than ten years of experience. Within each experience block, she selects five therapists who have psychiatric training and five therapists who are trained psychologists. She sets up a block design to control for treatment, type of training, and therapists' years of experience. She randomly selects five female and five male patients from each therapist's caseload and asks the therapists to record the patients' ages.<sup>5</sup> She asks therapists to meet clients three times per week during a two-week period. Dr. Davis tests patients after two weeks of in-patient treatment using a measure of depression for which there is strong evidence for the validity and reliability of scores.

With this design, Dr. Davis attempts to control variables that might impact the treatment as well as variables that might affect the results of treatment. By randomly sampling therapists from among the volunteers, and then randomly sampling patients within therapists' rosters, Dr. Davis has attempted to control for bias in selection. Threats such as maturation, attrition, and interactions with selection are equally likely for patients in either condition. Since therapists are using their own preferred treatment (cognitive or drug therapy), Dr. Davis is not concerned about equalization of treatment or diffusion of treatment and is ensuring that both treatments are represented in the study. By selecting therapists with different experience levels, she can control for experience differences as a potential influence on treatment. By randomly selecting patients of each gender in the therapists' caseloads, she has ensured that one important patient variable is distributed evenly across the therapists and that its effect can be evaluated. By using age as a covariate, she is controlling for age as a factor that might influence the success of treatment. Maturation and attrition are equally likely in both conditions. Finally, by using a reliable post-test measure, she has minimized measurement as a potential threat to validity and can be more confident in the resulting scores.

---

5. Note that, although this is Dr. Davis's design, she may find that not all therapists have patients distributed across the age bands.

There may be internal threats to validity that Dr. Davis has not controlled. For example, situational factors may influence therapists' choices of therapeutic intervention. It may be that therapists from hospitals in higher-SES communities are more likely to use cognitive therapy, and therapists from hospitals in lower-SES communities are more likely to use drug therapy. In addition, Dr. Davis has not ensured the reliability of treatment across the therapists. She cannot be certain that patients in the drug therapy condition actually take the prescribed medicine.

### Summary of Quasi-Experimental Design Strategies for Controlling Internal Threats to Validity

In each of these quasi-experimental designs, the researchers identify person and situational factors that could cause bias in selection or unreliability of treatment and include these factors in the design of the investigations. A repeated-measures design, like the one used by Dr. Adams, allows him to control for initial differences between patient groups in their levels of depression upon entering treatment. Random assignment of hospitals and random sampling of patients within therapists, allows Dr. Bennett to control for some of the situational and person variables that might impact the results of her study. Matching, the strategy used by Dr. Chang, is an attempt to control for therapist variables that might differentially impact the reliability of treatment. Analysis of covariance allows researchers to statistically control person variables that would be automatically controlled through random selection and assignment.

For each of these designs, the researcher identifies the potential internal threats to validity before collecting data, so that these threats can be accounted for in the research design. Researchers often use previous studies to identify possible internal threats to validity. However, despite the care with which each researcher controls factors that may affect results, there will always be factors that the researcher has not planned for and that are omitted from a design. In addition, each of these designs depends on the cooperation of individuals in the settings in which the studies are conducted, as well as on the consistency in treatment across therapists and settings.

Not all identified internal threats to validity can be included in a research design. For example, suppose the cause of depression

was known to impact the success of cognitive therapy. Suppose further that some causes of depression have major impacts on degree of depression but are less frequent than other causes (for example, the death of a spouse or child may be less frequent but may have a greater impact on degree of depression and effectiveness of treatment than the loss of a job or failure in school). It may not be possible to ensure that all possible causes of depression are equally represented in the patients from different settings and therapists. In fact, when the focus of human research is on constructs like depression, achievement, motivation, and other complex and multidimensional phenomena, a vast number of person and situational variables may affect the results. Quasi-experimental research on humans requires permission from the individuals who participate. There may be differences between people who volunteer and people who do not volunteer that are not captured in the research designs.

In short, there are no perfect quasi-experimental research designs. The best that researchers can do is to use care in creating their designs so that they account for as many internal threats to the validity of their results as possible—considering person, situational, and measurement factors.

### **Correlational Designs for Addressing Internal Threats to the Validity of Research Claims**

The fundamental question being asked when we are concerned about internal threats to validity is whether we can trust causal claims based on the results of the research. Another way to frame the question is whether there are alternate explanations for the results. Correlational methods give researchers a way to determine what variables predict or explain outcomes. As such, correlational methods are often called *explanatory designs*. Correlational methods differ from experimental and quasi-experimental designs in that they are used when researchers have access to extant data sets, and experimental or quasi-experimental designs are not possible or appropriate.<sup>6</sup> Data may be available for a large number of demographic and outcome variables, including ones that offer alternative explanations for results and are counter to the explanations proposed in a theory.

---

6. For example, research on causes of cancer, teen pregnancy, school failure, etc.

Some might argue that correlation cannot be used to make causal claims. This raises one of the most important threats to internal validity in research—that of ambiguity of results. Does a migraine headache cause swelling in the cerebral cortex, or does swelling in the cerebral cortex cause migraine headaches? or are both caused by a third, unmeasured variable? However, there are some causal claims that can be made through correlational research without the use of experimentation. When the incidence of cancer is higher for smokers than for non-smokers, across many samples and contexts, a causal claim can be made. Ambiguity in the direction of the causal relationship is not an issue, since it would be ridiculous to claim that cancer causes smoking. However, a causal claim about the relationship between smoking and cancer requires a close look at alternate explanations (e.g., Are smokers more likely to live in urban environments with high levels of smog? Are smokers more likely to be from communities where toxic waste is prevalent in the air and water [e.g., due to mining or agribusiness]?).

Correlational methods are used to explain variability in scores (i.e., differences among research participants) on one or more dependent variables (e.g., level of depression). This makes them ideal for investigating possible threats to the validity of causal claims. In this section, I briefly introduce four correlational methods (multiple-regression, path analysis, hierarchical linear modeling, and structural equation modeling) and how researchers might use them to investigate potential internal threats to the validity of causal claims.

### Using Multiple-Regression to Control for Internal Threats to Validity

“Dr. Fahd” uses multiple-regression to investigate which variables are the best predictors of post-treatment levels of depression after ten cognitive therapy sessions. He identifies 40 therapists who use cognitive therapy and 40 therapists who use antidepressants to treat depression. The therapists are from community out-patient clinics and private practices. He gathers post-therapy measures of depression (using a depression scale for which there is strong evidence for validity and reliability of scores) from clients who have

worked with the cognitive therapists for at least ten sessions and clients who have taken antidepressants for two months. Dr. Fahd identifies four other variables he wants to account for in his analysis: therapists' training (psychiatry or psychology), therapists' years of experience, client's number of episodes of depression, and age of client. In his design, Dr. Fahd attempts to control for selection bias and situational factors that could affect results. As with other researchers, since the therapists use their preferred mode of treatment, Dr. Fahd controls for two of the situational factors that are internal threats to the validity of claims about the efficacy of cognitive therapy—diffusion of treatment and equalization of treatment. By accounting for therapists' training and years of experience, he is controlling for factors that might impact the reliability of treatment. Capturing information about clients' past episodes of depression and age allows him to control for two person factors that could result in biased selection of cases for his study. Finally, by using a measure of depression with reliable scores that have been validated for out-patient settings, Dr. Fahd controls for some of the measurement threats to the validity of the results.

The use of regression allows Dr. Fahd to determine the relative strength of each variable in explaining client's post-treatment depression scores. In this way, he is testing whether cognitive therapy has a stronger impact on the post-treatment depression scores than the other variables. Table 2-1 presents the results of Dr. Fahd's regression analysis. The results suggest that the age of the patient is the strongest predictor of post-treatment depression scores. The beta weight for age is negative, suggesting that age is inversely related to post-treatment depression scores (i.e., the younger the patient, the higher the post-treatment depression scores). The results suggest that age could impact the effectiveness of treatment. The next strongest predictor is treatment, followed by type of training. The fact that training is such a strong predictor suggests that treatment implementation might depend on the therapists' training.

### Using Path Analysis to Account for Potential Internal Threats to Validity of Claims

"Dr. Garcia" uses path analysis to explain the possible sources of variability in changes between pre- and post-treatment depression

Table 2–1

**Results of Regression Analysis Examining Treatment, Training, Years of Experience, Therapist Gender, and Age of Patient as Possible Predictors of Post–Treatment Depression**

	Unstandardized Coefficients		Standardized Coefficients			Partial Correlation
	B	Std. Error	Beta	t	Sig.	
(Constant)	58.487	2.417		24.201	0.000	
Treatment	4.026	0.585	0.244	6.887	0.000	0.306
Training	2.793	0.552	0.169	5.055	0.000	0.229
Years of Experience	–0.479	0.325	–0.047	–1.471	0.142	–0.068
Gender of Therapist	0.651	0.531	0.039	1.224	0.221	0.057
Age of Patient	–1.454	0.102	–0.527	–14.267	0.000	–0.554

scores. Path analysis allows the researcher to develop a simple causal model, and to include variables that may present threats to the validity of causal claims about the efficacy of cognitive therapy. A path coefficient shows the strength of the relationship between two variables after correlations with other variables in the model have been accounted for.<sup>7</sup> Dr. Garcia wants to assess the relationship between cognitive therapy treatment and changes in depression levels after accounting for other sources of variability from therapists (years of experience) and patients (number of depression episodes, attitude toward therapy, and age<sup>8</sup>).

In selecting the therapists, Dr. Garcia invites therapists from community and private clinics. He gathers data for a sample of 10 clients from each of 20 therapists who use only cognitive therapies to treat depression, and a sample of 10 clients from each of 20 therapists who use drug therapies to treat depression. Recognizing that even therapists who use drug therapies may also employ cognitive therapies, he asks each therapist to complete a checklist of characteristics of their therapeutic strategies. He uses this checklist to formulate a rating scale that ranges from zero (“uses none of the cognitive therapy strategies”) to ten (“uses all identified cognitive therapy strategies”).

Dr. Garcia asks therapists to administer a depression scale and an attitude toward therapy scale to their clients before their first session and to administer the depression scale again after two months of therapy. Scores from the depression and attitude scales have been validated for patients with depression as a presenting problem, and both measures provide reliable scores. Figure 2–6 shows the path diagram for Dr. Garcia’s study. The path analysis allows Dr. Garcia to posit a set of causal relationships rather

- 
7. It is important to note that, in path analysis, causality is not assured. Two variables may be correlated; however, variations in one may not cause variations in the other. Researchers who use path analysis may posit causal relationships and test them through the statistical analyses, but in order to support causal claims, researchers would have to evaluate whether changes in one variable lead to changes in another (e.g., changes in attitude toward therapy lead to changes in the outcomes of therapy).
  8. Note that, although number of depressive episodes, attitudes towards therapy, and age may be correlated with changes in depression scores over time, one cannot assume that these variables “cause” changes in levels of depression, despite the suggestion of causation indicated in the directional arrows.

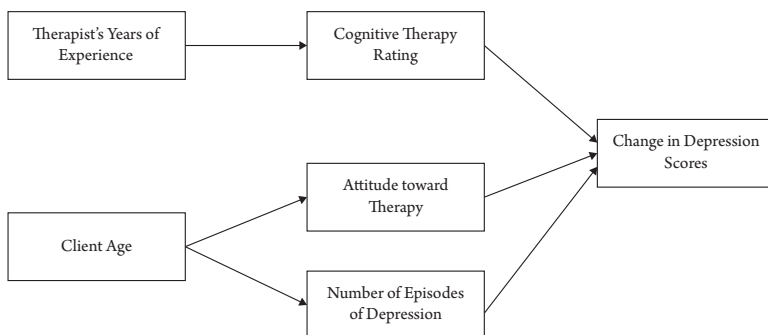


Figure 2-6 Path Model to Investigate Changes in Depression Levels

a single regression equation. For example, he has structured his proposed model such that client age predicts both the number of episodes of depression and client attitudes toward therapy, which in turn explain some of the variance in post-treatment depression scores. He posits that years of experience will influence the number of cognitive therapy strategies used by therapists.

Dr. Garcia's design can help him investigate several variables that could threaten the validity of claims about the efficacy of cognitive therapy in treating depression: differential characteristics of therapists (situational threats) and differential characteristics of clients (selection bias). He attempts to control for some of the selection bias by sampling clients from a range of community and private clinics. Dr. Garcia cannot randomly assign clients to a treatment or to a therapist. He attempts to control for the reliability of treatment by asking therapists to rate their own therapeutic practices. Since therapists are using their preferred methods of treatment, diffusion of treatment and equalization of treatment are unlikely.

Given that this research takes place in community and private clinics, it is likely that the data will be collected over an extended period of time. Therefore, Dr. Garcia cannot control the impact of history on his results. In addition, attrition may be a problem since the length of time between entry into therapy and post-test is two months. In using a self-report rating for cognitive therapy, Dr. Garcia may have added measurement threats to the validity of his results in that therapists' self-reports may not reflect their actual practices. Therapists may feel pressure to identify strategies



that they do not, in fact, use, or may identify strategies that they use infrequently. Another threat to validity is in the use of change scores as a post-treatment measure. Even when a scale produces fairly reliable scores, there will be some level of unreliability (measurement error). With change scores, measurement error is compounded across the two assessment events. Finally, regression to the mean is a threat when using pre-test and post-test measurements.

### Using Multilevel Modeling to Account for Potential Internal Threats to Validity

Throughout the examples of quasi-experimental and correlational designs described above, one threat to statistical conclusion validity<sup>9</sup> has not been adequately addressed—that of *nesting*. In each example, clients are “nested” within therapists and therapists within clinics or hospitals. Each researcher has attempted to compensate for issues of sampling through some form of random selection, the use of multiple therapists or multiple sites; however, the fact remains, all clients and patients have been nested in a context that influence the results of the investigation. It may be that variability within a setting is smaller than variability between settings for the same treatment or intervention. This can undermine causal claims that are based on statistical analyses of individual differences in response to a treatment or intervention.

“Dr. Hamma” uses hierarchical linear modeling to investigate factors that affect post-treatment depression and that address potential internal threats to validity. In a multilevel model, client variables, therapist variables, and context variables can be accounted for in hierarchical regression equations. Dr. Hamma identifies therapists from community and private clinics who use cognitive therapy to treat depression. Dr. Hamma’s model takes into account four variables that could cause sample bias: age of clients, gender of clients, number of previous episodes of depression, and patients’ attitude toward therapy. Her model takes into account three therapist variables (self-report rating of cognitive therapy,<sup>10</sup> years of experience, and gender). Finally, her model takes into

---

9. See Chapter 4 for a more thorough discussion of this issue.

10. See description in the previous section.

account variables related to the contexts in which the therapists work: type of provider (public or private) and client-to-therapist ratio. The multilevel model has three levels: clients, therapists, and settings. The equation that represents the post-treatment depression of client  $i$  within therapist  $j$  within setting  $k$  ( $Y_{ijk}$ ) is represented as a function of the client's background characteristics,  $X_{ijkl}$ , and a random error,  $R_{ijk}$ :

$$Y_{ijk} = \beta_{jk0} + \beta_{jk1}X_{jk1} + \beta_{jk2}X_{jk2} + \beta_{jk3}X_{jk3} + \beta_{jk4}X_{jk4} + R_{ijk}$$

In this equation,  $\beta_{jk0}$  represents the mean post-treatment depression score associated with each therapist, and  $\beta_{jk0}$ ,  $\beta_{jk2}$ ,  $\beta_{jk3}$ , and  $\beta_{jk4}$  represent the regression coefficients for client age, client gender, number of previous episodes of depression, and patient's attitude toward therapy, respectively. In hierarchical linear modeling, regression coefficients can be assumed to vary across therapists; therefore, each regression coefficient is a function of therapist variables (years of experience, gender, and rating of cognitive therapy) and random error associated with therapist  $j$  within setting  $k$  ( $U_{jkl}$ ).

$$\beta_{jkl} = \gamma_{0kl} + \gamma_{1kl}W_{1kl} + \gamma_{2kl}W_{2kl} + \gamma_{3kl}W_{3kl} + U_{jkl}$$

In this equation,  $\gamma_{0kl}$  represents the mean post-treatment depression score associated with the setting, and  $\gamma_{1kl}$ ,  $\gamma_{2kl}$ , and  $\gamma_{3kl}$  are the between-therapist regression coefficients associated with therapist years of experience, therapist gender, and cognitive therapy rating, respectively. The therapist regression coefficients are also assumed to vary across settings; therefore, each therapist regression coefficient is a function of setting variables (type of provider and client-to-therapist ratio) and random error associated with setting ( $E_{jkl}$ ).

$$\gamma_{jkl} = \theta_{j0l} + \theta_{j1l}Z_{j1l} + \theta_{j2l}Z_{j2l} + E_{jkl}$$

In this equation,  $\theta_{j0l}$  represents the grand mean post-treatment depression score;  $\theta_{j1l}$  and  $\theta_{j2l}$  represent the between-setting regression coefficients for type of setting (public or private) and client-to-therapist ratio, respectively.

Through a hierarchical model, Dr. Hamma attempts to control for sample bias by controlling for patient variables, for possible variations in the treatment and therapist characteristics, and

for possible situational variables that could impact treatment by considering the therapists' contexts. However, she does so in a way that takes into account the nesting effects of these variables. The results of her analysis will provide information regarding whether level of cognitive therapy, as reported by the therapists, is a strong predictor of post-treatment depression measures as well as the strength of other client, therapist, and setting variables in predicting post-treatment depression scores.

### Using Structural Equation Modeling to Investigate Threats to Internal Validity and to Consider Alternate Explanations

For each of the three correlational models described above, observed variables are used in the analyses. Structural equation modeling (SEM) is a correlational method that includes both observed variables and possible latent variables. The benefit of SEM is that it can be used to test alternate or competing explanations. Sometimes, SEM is only used to explain relationships among variables that are theoretically linked—in which case, causal claims are not made. At other times, SEM is used to test alternate causal explanations. As with any correlational model, a major threat to the validity of causal claims for SEM is that correlation does not equate with causation.

Two components are needed to propose causal models in SEM. First, theory based on solid empirical research should suggest causal relationships among the variables. Second, to support causal claims, causal variables must occur before dependent variables.<sup>11</sup>

In an SEM diagram, the *measured* variables (e.g., scores on a measure of depression), are indicated by rectangles or squares; *latent* variables (also called *factors* or *constructs*) are indicated by ellipses or circles. Error terms are included in the SEM diagram, represented by “e” for measured variables and “d” (disturbances) for latent variables. The error terms represent residual variances within variables not accounted for by pathways hypothesized in the model. Variables are considered either *endogenous* (i.e., measures of these variables are caused by other variables) or *exogenous* (i.e., no causal pathways lead to the given variable).

---

11. More detailed discussions of causality can be found in Pearl (2009) and Freedman, Collier, Sekhon, & Stark (2009).

For SEM, researchers develop models to explain the results of an investigation and test those models through model-fit statistics—testing to see whether the model fits the data. If more than one model is theoretically possible, SEM researchers set up competing models to determine which best fits the data.

With SEM, fit between the SEM model-implied and the observed sample covariance matrices can be evaluated using the  $\chi^2$  model-fit statistic. However, because  $\chi^2$  statistics are sensitive to sample size, and estimation of SEM models generally requires large sample sizes, a number of additional fit statistics have been developed. For example, the comparative fit index (CFI) is useful for comparing models; the standardized root mean square residual (SRMR; Bentler, 1995), and the root mean square error of approximation (RMSEA; Steiger & Lind, 1980) are the commonly used fit indices.

In most correlational methods, observed data are used to develop the models—regardless of the measurement errors that are likely to be found in the data. SEM can be used to control for measurement error by partitioning score variability into two components: the component attributable to a factor or latent variable (construct) measured via commonality among observed variable indicators, and the component relevant to variable-specific error (which includes measurement error).<sup>12</sup>

“Dr. Iocomo” uses SEM to describe the influence of a number of situational and client variables in his investigation of the influence of cognitive therapy on post-treatment depression scores. He selects therapists from both private and public clinics. He gathers post-treatment depression scores from three depression measures (Beck Depression Inventory—BDI; Hamilton Rating Scale for Depression—HRSD; and the Center for Epidemiology Studies Depression Scale—CES-D). He uses three scales in order to generate more a more valid and reliable estimate of post-treatment depression.

Dr. Iocomo also collects data for three patient variables that could influence post-treatment outcomes (age, number of depressive episodes, and attitude toward therapy). He collects data for two therapist variables that could influence the reliability of treatment (years of experience and level of training in cognitive therapy

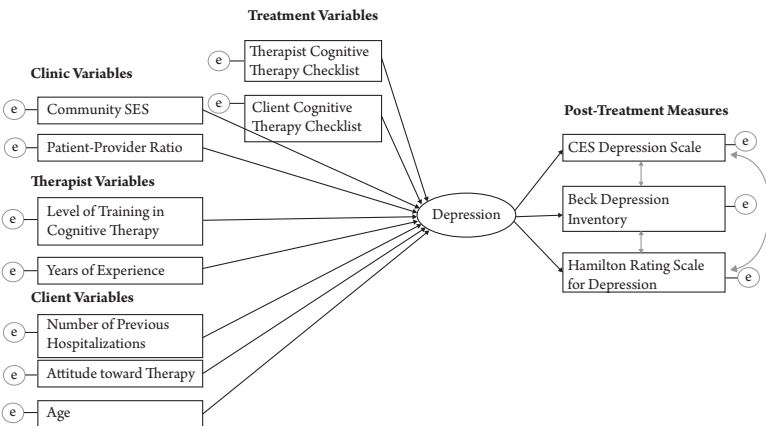
---

12. A more complete introduction to structural equation modeling can be found in Kline (2011).

techniques) and two variables related to the therapists' contexts (mean socio-economic status of clients and client-to-therapist ratio). Dr. Iocomo also collects data for two measures to assess the level of cognitive therapy used by the therapist (therapist self-report checklist of the cognitive therapy strategies used, and a client checklist of cognitive therapy strategies used). Dr. Iocomo proposes three competing models to explain the relationships among the variables (see Figures 2-7 through 2-9).

The model in Figure 2-7 includes each variable as a predictor in the same way as a typical regression analysis, except that the predictor variables are predicting a factor score based on the three measures of depression. This is called a "base model" against which the other models can be compared. Figure 2-8 presents a hierarchical model wherein four latent variables are proposed that represent level of cognitive therapy, context conditions, therapist expertise, and one patient factor (severity factor). Figure 2-9 presents a second hierarchical model wherein quality of treatment is caused by context conditions and therapist expertise.

If the models being compared are found to fit the data, then Dr. Iocomo can test each model to identify the best-fitting model. The tests will provide path coefficients that allow him to determine the strength of each factor in predicting post-treatment depression. In this way, he will be able to tell whether patient or



**Figure 2-7** Base Model to Explain Variables Related to Post-Treatment Depression

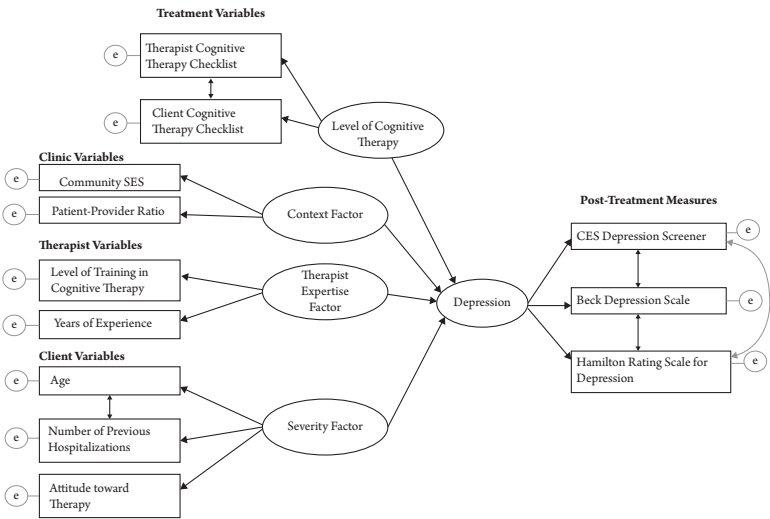


Figure 2-8 Structural Model 1 to Explain Variables Related to Post-Treatment Depression

situational factors (threats to validity of claims about the efficacy of cognitive therapy) are stronger predictors than the treatment.

Dr. Iocomo’s design addresses many threats to the validity of claims about the relationship between cognitive therapy and depression. He has controlled for several person variables that

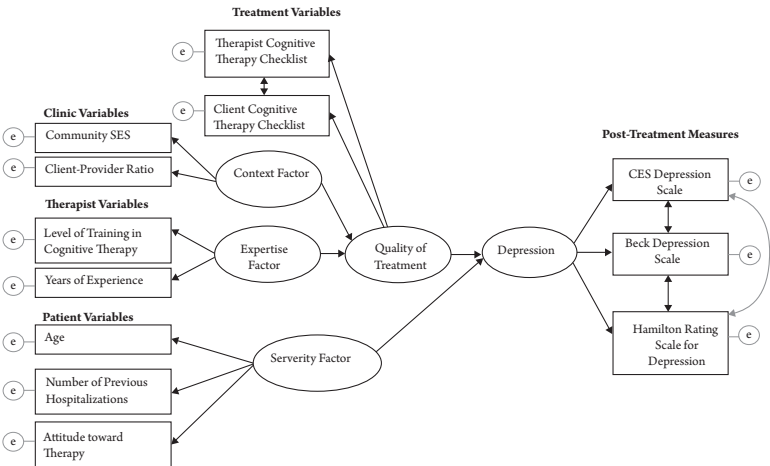


Figure 2-9 Structural Model 2 to Explain Variables Related to Post-Treatment Depression

could bias selection; he has addressed many of the measurement threats through the use of latent traits rather than direct measurement. By having therapists use their preferred type of treatment, he has minimized the threats of diffusion of treatment and equalization of treatment. Dr. Iocomo's use of self-report to assess the level of cognitive therapy introduces a measurement threat; however, some of this threat is offset by the fact that he asks clients to identify the same behaviors.

SEM is very useful for theory testing; however, it requires large sample sizes to implement. Given the large data-collection process, there may be threats to internal validity that cannot be controlled. Dr. Iocomo will need the cooperation of a wide range of private and public clinics. Clinics that volunteer and therapists who volunteer may be different from those that do not volunteer. Large-scale data collection takes time; therefore, unaccounted-for events (history) may influence some locations more than others and at different points in time.

### Summary of Correlational Strategies for Controlling Internal Threats to Validity

Correlational designs are used when researchers wish to investigate relationships among several variables in a single investigation and when experimental or quasi-experimental studies are not feasible or appropriate. A quick perusal of the correlational designs described here demonstrates that they can be very complex—a far cry from the simple experimental design described at the beginning of this chapter. Experimental designs can employ smaller sample sizes because they rely on random selection and random assignment. However, as I will discuss in the next chapter, it is difficult to generalize from experimental studies to the population as a whole because of very real confounding interrelationships among variables when phenomena occur in situ.

When experimental designs are not possible and quasi-experimental designs are not feasible or ethical, correlational studies allow for simultaneous investigation of interactions among several variables. SEM takes this further to allow for investigations of relationships among latent traits. Larger sample sizes help mitigate the threats to validity that arise when random selection and random assignment are not possible. Yet, even with large

samples, correlational research results are sample-dependent. Correlational researchers should cross-validate results using additional samples before making any causal claims.

Correlational studies are subject to the same potential internal threats to validity as are experimental and quasi-experimental studies. Researchers should consider the potential internal threats to validity in correlational research designs, using past research as a guide, and then work to minimize their influence.

## Summary

Whether researchers use experimental, quasi-experimental, or correlational designs to control for internal threats to the validity of claims, their research designs will address some but not all of the internal threats. Even if researchers match therapists and hospitals based on identified demographic similarities, no two psychiatric hospitals and no two therapists will be exactly the same. Therefore, unaccounted-for situational variables may still be internal threats to validity. Random sampling and assignment help with some internal threats to validity; however, research on human behaviors requires human beings to implement treatments. Inevitably, there will be differences in how a treatment is implemented from one therapist to the next, from one teacher to the next, from one graduate assistant to the next, and so on.

Demographic matching and block designs may control some variables, but there will be other variables related to implementation of a treatment that are not captured in the design. Giving a pre-test may help account for initial differences in groups, but pre-testing introduces another internal threat to validity. Researchers may identify some covariates, but there will always be covariates that are not accounted for.

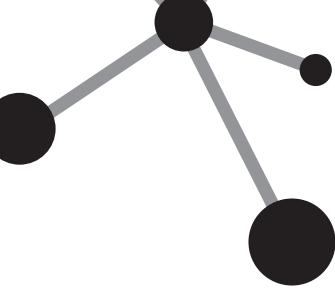
In addition to these challenges, each identified variable is, in fact, a construct. As such, these constructs require definition and a mechanism for their assessment. Some variables are relatively easy to assess (e.g., gender, age, number of hospitalizations, public or private hospital, years of experience). Covariates that involve internal characteristics of participants (e.g., attitude toward therapy), measures to evaluate implementation of treatment, and measures of dependent variables such as depression require significant efforts in assessment design, development, and validation research.



Given that no research design, no matter how elaborate, can control all possible internal threats to the validity of results, what can researchers do to address the realities of human research? First, researchers are responsible for identifying the potential threats to validity for the research situation. Next, they must identify ways to mitigate these threats through research designs and statistical controls. Finally, they must acknowledge likely threats when reporting the results of an investigation. Consumers of others' research should consider internal threats to validity and assess whether the investigators considered such threats in their designs, in their analyses, and in their any claims they make based on their results.

## References

- Arndt, J., Greenberg, J., Pyszczynski, T., & Solomon, S. (1997). Subliminal exposure to death-related stimuli increases defense of the cultural worldview. *Psychological Science*, 8, 379–385.
- Bentler, P. M. (1995). *EQS Structural Equations Program Manual*. Encino, Multivariate Software.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). London: Routledge Academic Press.
- Freedman, D. A., Collier, D., Sekhon, J. S., & Stark, P. B. (2009). *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. Cambridge, UK: Cambridge University Press.
- Hancock, G. R., & Mueller, R. O. (Eds.) (2006). *Structural Equation Modeling: A Second Course*. Greenwich, CT: Information Age Publishing.
- Kline, R. (2011). *Principles and Practice of Structural Equation Modeling* (3rd ed.). New York: The Guilford Press.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2nd ed.). Florence, KY: Cengage Learning.
- Steiger, J. H., & Lind, J. C. (1980, May). Statistically based tests for the number of common factors. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.



## EXTERNAL THREATS TO VALIDITY

AS STATED IN Chapter 1, the external validity of causal claims is the degree to which results of an investigation are generalizable beyond the specific investigation. In this chapter, I describe strategies researchers use to address threats to the external validity (i.e., generalizability) of research results. I describe threats to external validity in terms of populations, times, and situations. Within these factors, I describe specific threats to external validity, such as:

1. Interactions among different treatments or conditions
2. Interactions between treatments and methods of data collection
3. Interactions of treatments with selection
4. Interactions of situation with treatment, and
5. Interactions of history with treatment

Each of these factors can limit the generalizability of research results and, therefore, the validity of claims based on those results.

### **Populations, Samples, and Generalizability**

Generalization to a population as a whole requires that samples be representative of the targeted population. Therefore, one of the

central tasks for a researcher is to define the population to which results are to be generalized. Populations must be defined in terms of the specific variables that are of interest to the researcher (e.g., age, demographic characteristics [race/ethnicity, cultural background, primary language, gender, etc.], location [region, nation, etc.], and so forth).

Suppose a researcher is investigating the effectiveness of a new reading program for low-income children. What is the population to which she should be able to generalize? All elementary school children from low-income families? All low-income children at primary grade levels? All low-income children in urban settings? The definition of the population circumscribes one critical feature of an investigator's research design.

Once the population is defined, the researcher must find a group of individuals to include in the study. Sampling is a very challenging task. As mentioned in Chapter 2, the best way to be sure that the results of a study are due to the treatment or intervention, rather than some characteristic(s) of the study participants, is to use random selection and random assignment. Similarly, the best way to ensure that the results of a study can be generalized to all members of the population is to use random selection and random assignment. In the case of the reading program, the ideal study would involve randomly selecting students in the targeted grade levels and randomly assigning half of them the new reading program, and randomly assigning the other half to an existing reading program; then determining which program was most successful in teaching students how to read.

## **Interactions and Generalizability**

### **Interaction of Treatment with Selection**

Chapter 2 described two studies in which the researchers used quasi-random sampling processes. "Dr. Bennett" identifies four private hospitals in the northeastern United States. She randomly assigns one urban and one suburban hospital as treatment hospitals; one urban and one suburban hospital as comparison hospitals. After training all therapists in the treatment hospitals, she randomly selects four patients for each of her trained and comparison therapists. Through her sampling design, Dr. Bennett attempts to ensure that the patients in her study represent the population of individuals who suffer from depression. Through

random sampling of patients within settings, she is likely to have patients who represent a broad range of ages, males and females, and a variety of ethnicities. The patients in both the treatment and the control hospitals are likely to represent individuals from both urban and suburban settings in the northeastern United States.

“Dr. Chang” conducted his study with therapists who already use cognitive therapy to treat depression, and demographically matched them—based on therapists’ gender, ethnicity, age, and years of experience—to therapists who use only drug therapy. He collected pre- and post-test data from ten randomly selected patients for each of the therapists. As did Dr. Bennett, Dr. Chang used a quasi-random sampling process for patients. By randomly sampling patients within therapists’ client lists, Dr. Bennett can be more certain that the results of his investigation can be generalized to a wide range of patients.

Suppose all the patients in Dr. Bennett’s and Dr. Chang’s studies agreed to participate in the research. One possible limitation to their results would be that results could be generalized only to patients who volunteer to participate in research. There may be characteristics of volunteers that influence their receptiveness to treatment. In addition, individuals for whom treatment (cognitive or drug therapy) was successful might be more willing to volunteer than individuals for whom treatment was unsuccessful. Either of these factors would result in an *interaction between selection and treatment*. On the other hand, if post-treatment testing were a routine aspect of hospital policy and if permission to use patient data was not a requirement for these studies, it would be easier to assert that there is no interaction between selection and treatment based on volunteer status.

Another potential source of interaction between selection and treatment is the fact that the patients in Dr. Bennett’s and Dr. Chang’s studies are in hospitals. Patients in hospitals may have more acute or more severe forms of depression than patients in out-patient settings. This would limit generalizability of results to in-patients treated for depression.

### Interaction of Treatment and Situation

As was discussed in Chapter 2, random assignment is very difficult to do in human research. Participants are generally grouped in some way (e.g., children in classrooms; patients within hospitals).

An intervention or treatment depends on the teachers or therapists who implement it. In addition, teachers, therapists, researchers, and professionals involved in studies are in specific contexts (e.g., schools and districts for teachers; private practices, clinics, and hospitals for therapists; universities and research centers for researchers). Researchers must account for grouping when they sample from populations.

Dr. Bennett conducts her study using therapists who work in private hospitals. Suppose she invites both public and private hospitals to participate in her study, and only private hospitals agree to participate? This fact would limit the generalizability of her results to patients in hospitals that are willing to participate in research. Both Dr. Bennett and Dr. Chang use therapists who have agreed to be part of their studies. This will limit the generalizability of their results to patients who work with therapists who are willing to contribute to research.

Dr. Bennett intentionally selects urban and suburban hospitals as sites for her research, and she randomly assigns one hospital from each setting to be a treatment hospital. This allows Dr. Bennett to generalize her results to patients in both urban and suburban settings.

Dr. Bennett, Dr. Chang, and Dr. Adams (all from Chapter 2) conduct their studies in hospitals. A hospital setting may create a context in which cognitive therapy is more or less successful than in an out-patient setting. Therefore, these researchers can generalize their results only to patients who are treated in hospital settings.

All of these studies present examples of the potential for *interaction between situation and treatment*. There may also be interactions between situation and treatment that are more difficult to recognize. Suppose, for example, that one of the treatment hospitals in Dr. Bennett's study provides ongoing support for implementation of new procedures. This could lead to better implementation of treatment and, therefore, more successful outcomes for patients. In such a case, the results of Dr. Bennett's study would be affected by internal hospital policies. When conducting studies in defined settings, researchers are obligated to find out about potential sources of interaction and to describe them as possible limitations to the generalizability of the results of their studies.

Dr. Bennett conducts her study in the northeastern United States. Dr. Chang conducts his study in the Pacific Northwest. Dr. Adams conducts his study in two private hospitals in Arizona. These also represent an interaction of treatment and situation. Therefore, each researcher can only generalize her or his results to patients from particular regions of the United States.

Drs. “Fahd,” “Garcia,” “Hamma,” and “Iocomo” attempt to control for interactions between treatment and selection by obtaining data from a large number of therapists from multiple public and private clinics. Any interactions between treatment and setting are likely to be distributed randomly across the settings.

### Interaction Between Treatments

Another situational factor that could have an impact on the external validity of research is *interaction between different treatments*. For example, in the studies conducted by Drs. Bennett, Chang, and Adams, the patients are administered drug therapies along with cognitive therapies. It is possible that drug therapy and cognitive therapy together have a different impact on depression than cognitive therapy alone. Therefore, Drs. Bennett, Chang, and Adams would have to acknowledge this limitation to the generalizability of their studies. In the study conducted by Dr. Davis and Dr. Fahd, the researchers recruited therapists who used either cognitive therapy or drug therapy. These researchers may have minimized interactions among treatments as a potential threat to external validity. However, it is possible that therapists who claim use of drug therapy alone also provide some cognitive therapy—which would lead to the potential for interaction between treatments.

Dr. Garcia also selects therapists who use either cognitive therapy or drug therapy; however, he anticipates that even therapists who claim to use only drug therapy may also use some cognitive therapy strategies; therefore, he has the therapists complete a checklist of cognitive therapy strategies and gives each therapist a cognitive therapy rating. Rather than use a dichotomous variable (drug therapy *or* cognitive therapy), he uses this cognitive therapy rating in the path analysis. While this acknowledges that cognitive therapy may be present in any therapy session, it does not remove the potential of interaction between drug therapy and cognitive therapy.

### Interaction Between Data Collection and Treatment

All of the studies described in Chapter 2 require data collection. It is possible for there to be an *interaction between data collection and a treatment or intervention*. In Dr. Adams's study, he administers a pre-test to all patients before the study begins so that he can control for initial differences in the samples from the two hospitals. This can alert patients to the purpose of a study. Even though this awareness is likely to have the same effect on patients in both conditions, Dr. Adams would have to acknowledge that his results are only generalizable to patients who take a pre-test. Research designs that involve only post-testing are more generalizable.

As another example of an interaction between data collection and treatment, suppose all of the studies described in Chapter 2 required patients' permission to use their data. If permission is obtained before the study begins, patients will be alerted to the purpose of the study and may alter their behaviors due to this knowledge. Although the interaction between data collection and treatment or intervention is equally likely to affect all patients and clients, in terms of generalizability, the researchers will have to limit their generalization to individuals who are alerted to the purpose of the study. If testing is completed before researchers ask patients for permission to use their test results, researchers can better control for the interaction between treatment and data collection.

### Interaction of History and Treatment

Events unrelated to the purpose of a study can significantly impact the generalizability of results. Any number of significant events could affect the results of studies, thereby threatening the external validity of any causal claims. For example, the results of the studies described in Chapter 2 might be different if they were conducted before or after a significant economic downturn, such as the one that began in 2008, which caused high unemployment; before or after major floods in the Midwest damaged farms and homes in 1993; before or after wildfires burned homes and forests throughout the southwestern United States in 2009 through 2011; before or after the terrorist attacks in 2001.

Researchers must be alert to the *interaction of historical events and treatment or intervention* and acknowledge them in the

discussion of their results. For the examples given above, these historical events had widespread impact of a type that could affect psychological variables such as depression.

In studies of educational interventions, the historical events may be less catastrophic and more local. For example, suppose that a control school for a study of project-based learning is in a district that requires senior portfolios for graduation. Students' focus on their senior portfolios might depress post-test scores if post-testing is timed to occur before the portfolios are due—making an instructional intervention appear to be less effective than it is. When possible, researchers should time their studies so that the results are not influenced by events that can be avoided. Researchers must address historical events when discussing the generalizability of their results.

### Summary of Potential Threats to External Validity of Causal Claims

From these examples, it is easy to see that researchers must be very circumspect when drawing conclusions from their research. Researchers must acknowledge the limitations of their results in terms of the populations, the situations and contexts relevant to the study, and the timing of the studies. Generalization across times and situations requires that we make repeated tests during different times and in different situations or under different conditions. Researchers may refer to previous studies in their discussions so that, as with “many hands on the elephant,” the true picture can be known.

### **Controlling for External Threats to Validity Through Research Designs**

Clearly, it is impossible to ensure that the results of any single study are generalizable to a population across settings and at different times. However, some research methodologies are better for generalization than others. For example, a perfect experimental design is tightly defined in terms of random selection, random assignment, treatment condition, control condition, and post-testing. It is highly effective in controlling internal threats to the validity of claims. However, a single study using this design is the least likely to be generalizable. As it is a tightly constructed study,



any interactions between the treatment and other variables are not addressed. Random selection and random assignment are rarely possible in such a case; however, even if they were possible, making causal claims that could be generalized from a single study to a population is unwise. Even with random selection and random assignment, generalization of results to natural situations involving a complex of uncontrolled variables is not possible.

Human research generally prevents both random selection and random assignment. In addition, human research generally involves administration of an intervention by humans. This increases the likelihood of unreliability of treatment (a threat to internal validity). Even when a tightly controlled study is possible, many variables are likely to intervene. For example, with a tightly controlled reading study, teacher variables, school variables, interactions between the reading intervention and other school programs, interactions between the intervention and parent support at home, and so forth, will all influence the outcome of the study. A single study using an experimental or quasi-experimental design does not address these complexities.

Ensuring external validity (generalizability) requires either more complex research designs or multiple replications of tightly controlled experimental or quasi-experimental studies. In Chapter 2, each subsequent research design was more complex than the last. The more complex designs take into account more of the situational and person variables that might influence the effectiveness of cognitive therapy in treating depression. Results from studies that involve multiple settings can be generalized to more settings. Results from designs that account for multiple situational factors (e.g., characteristics of therapists, characteristics of settings) can be generalized to more contexts. Results from studies with large, representative samples can be generalized to more of the population.

### Using Replication to Support External Validity of Causal Claims

The most effective strategy for ensuring external validity is replication. When studies are replicated in multiple settings with different samples of participants and at different times, the results are more trustworthy. An excellent example of multiple replications in human-subject research can be found in the work of Greenberg and colleagues (see Greenberg, Solomon, & Pyszczynski, 1997,

for a review). Using experimental designs with volunteer samples, these researchers tested terror-management theory in a range of contexts and with a wide range of participants (from college students to judges) and with varying stimuli and tasks. They found remarkably similar results across the different conditions. Other examples can be found in research on factors that affect reading comprehension. Since no single study can be generalized to an entire population, in multiple settings, and over time, ensuring the external validity of research results requires a focused research program—preferably with multiple researchers considering the same research questions.

### Sample Sizes

A second way to strengthen the generalizability of causal claims is through the use of large sample sizes involving cases from multiple sites. For the correlational designs described in Chapter 2, Drs. Hamma and Iocomo use hierarchical linear modeling (HLM) and structural equation modeling (SEM), respectively—designs that require large samples. Large samples can compensate, to a certain extent, for the lack of experimental controls in correlation research. Rather than accounting for all possible confounding variables in the research design, unaccounted-for variables are allowed to function randomly across individuals and settings. If targeted causal relationships are strong enough to emerge from the overall complexity of the study, large samples strengthen the generalizability causal claims. Large samples derived from multiple contexts randomly distribute many of the confounding variables that could threaten generalizability, thereby making the causal claims more generalizable.

In addition, Dr. Hamma's use of HLM not only involves multiple sites and large samples, her design results in hierarchical regression equations for clients nested within therapists and for therapists nested within settings. In this way, Dr. Hamma acknowledges the uniqueness of different contexts and can reveal common patterns in diverse contexts.

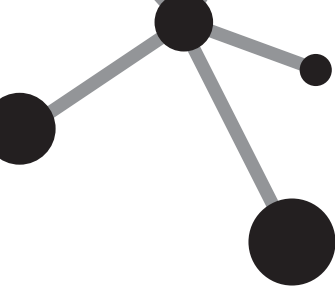
### Summary

External threats to validity are those that make it difficult to generalize results from one study to the target population as a whole.

To make generalization possible, researchers must be very clear about the nature of the population and provide information that documents the degree to which samples are representative of the population. Researchers must acknowledge factors in the investigation that could have affected results—factors that are intentionally or unintentionally incorporated in the study and that could influence results. Intentional factors might include pre-tests (Will the results be the same if individuals taking this drug don't take a pre-test?), interactions among treatments (Will the results of this study be the same if the participants are not simultaneously involved in a different treatment?), interactions between treatment and selection (Will the results of this study be the same for people who are not volunteering in a study?), interactions of treatment with situation (Will the results of this study be the same in a different setting, with different providers?), and interactions of treatment with history (Will the results of this study be the same two years hence; would they have been the same two years ago?). Researchers are obligated to present the limitations to the generalization of results in their reports. Consumers of research have a responsibility to look for and evaluate the significance of these limitations to generalization as they consider the implications of the causal claims in their own work.

## Reference

- Greenberg, J., Solomon, S., & Pyszczynski, T. (1997). Terror management theory and research: Empirical assessments and conceptual refinements. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology*, 30, 61–139. New York: Academic Press.



## VALIDITY OF STATISTICAL CONCLUSIONS

AS STATED IN previous chapters, the purposes of educational and psychological research are to build theories about human behavior, to make sense of the phenomena around us, and to solve problems. For all of these purposes, researchers attempt to develop causal explanations. Chapters 2 and 3 in this volume described internal and external threats to the validity of research claims. Although it is not possible to prove a theory, researchers can gather evidence to test claims about expected theoretical relationships. Quantitative researchers use statistical tests to evaluate the strength of the relationships among the variables in their data. Their goal is to determine whether the data behave in ways that are consistent with theory.

Statistical conclusions are claims made based on the strength of statistical results. When thinking about the validity of statistical conclusions, we are applying the principles of *relativism* described in Chapter 1. The goal is to falsify the null hypothesis or test competing explanations for phenomena. Threats to the validity of statistical conclusions include: *low statistical power*, *experiment-wise error*, *violating the assumptions of statistical tests*, *omitted-variable bias*, and *over- or under-interpretation of results*.

In this chapter, I briefly address how to examine each of these factors when considering the validity of statistical conclusions. More detailed treatment of these issues can be found in introductory statistics textbooks (e.g., Garner, 2010; Field, 2005; Howell, 2011; Urdan, 2010).

## Statistics Fundamentals

Several key ideas are important to any discussion of statistics. These include: *null hypothesis*, *alternative hypothesis*, *Type I error*, *Type II error*, and *alpha (probability) level*. These ideas derive from validity theory—in particular the notion of *falsification* (Popper, 1959). Evaluation of the validity of theoretical claims requires an understanding of these concepts.

The job of the researcher is to test theoretical claims and determine whether the data support those claims. Theories cannot be proven; however, if research results do not falsify the claims, researchers are more confident in their theoretical claims. When building theory, we posit expected causal relationships among theoretical constructs and gather data to test those relationships. The *null hypothesis* is the hypothesis that no relationships exist among the targeted variables:

$H_0$  = There is no relationship between the targeted variables.<sup>1</sup>

Needless to say, *alternative hypotheses* suggest that expected theoretical relationships will be evident in the data. Statistical tests help us determine whether or not statistical results provide support for theoretical relationships. If statistical results do not provide support for expected relationships, the researcher *fails to reject* the null hypothesis. If statistical results suggest that the expected relationship *is* present in the data, the researcher has the *option* of *rejecting* the null hypothesis. Even with statistical significance, the researcher may not reject the null hypothesis because statistical significance does not always reflect a meaningful relationship.

When conducting statistical tests, there are two possible types of error. *Type I error* occurs when a statistically significant test result suggests a relationship that does not actually exist (false positive).

---

1. Measured variables represent the theoretical constructs.

*Type II error* occurs when a statistical test suggests no relationship, but a relationship *does* exist (false negative). The idea of error is central to all research and measurement. Researchers can never be absolutely certain about or prove a theory. Therefore, they must decide how much uncertainty they are willing to tolerate.

The *alpha level* is the probability of Type I error. When researchers set an alpha level for their statistical tests, they are setting a level of tolerated error. The most commonly used alpha levels are 0.05 and 0.01. For example, when a researcher sets an alpha level to 0.05, the researcher is willing to tolerate less than 5 percent likelihood of a false positive.

Error is always possible. There is nothing magical about a statistical test or a particular alpha level. In fact, alpha levels for statistical tests are *probability estimates* of error based on randomly generated data in ideal conditions. Rarely does a research study result in ideal data. Therefore, statisticians conduct studies to determine whether various statistical tests are *robust* (function well) under less than ideal conditions (e.g., Boneau, 1960; Feir-Walsh & Toothaker, 1974; Hollingsworth, 1980; Keselman & Toothaker, 1974; Levy, 1980; Martin & Games, 1976; Ramsey, 1980; Wu, 1984; Zimmerman, 1998).

In the following, I discuss each of the potential threats to the validity of conclusions based on statistical results. Each of the concepts described above (Type I and Type II error, probability, alpha level, null hypothesis, and alternative hypothesis) are referenced in this discussion.

## **Factors to Consider Regarding the Validity of Statistical Conclusions**

### Statistical Significance

The term *statistical significance* is used to describe a situation in which a statistical test suggests nontrivial relationships among the variables in the data. The researcher hopes that this nontrivial difference supports his or her theory. For example, suppose Dr. Adams (from Chapter 2) conducts a statistical test to determine whether patients who receive both drug therapy and cognitive therapy have lower depression scores than patients who receive only drug therapy. If a t-test shows a statistically significant difference in the means at some established alpha level (e.g.,  $p < 0.05$ ),

Dr. Adams has more confidence in his theory about the effect of cognitive therapy on depression. Dr. Adams's statistical results appear to provide support for his theory. However, he must determine whether his statistical conclusions are trustworthy.

Although statistical significance may be a wonderful outcome for research (assuming that the patterns in the data are consistent with theory), statistically significant differences do not automatically ensure meaningful differences. A factor that has a great influence on the likelihood of detecting a true relationship is the number of cases in a sample. For example, it is possible that a very small difference between two groups will be statistically significant if the sample sizes are fairly large. On the other hand, a large difference may not be statistically significant when the sample sizes are small.

## Effect Size

One way that researchers deal with the meaning of a statistical result is to examine *effect size*. Effect size is a measure of the strength of a statistical relationship; the greater the effect size, the greater the support for the validity of causal claims. In experimental research, a standardized difference between two means ( $d$ ) is a measure of effect size. The standardized difference measures the difference between means in standard deviation units. Cohen (1992) proposed three levels of effect size. His levels were  $d \cong 0.20$  (small effect),  $d \cong 0.50$  (medium effect), and  $d \cong 0.80$  (large effect). These correspond to differences between means (in standard deviation units) of approximately one-fifth of a standard deviation, half of a standard deviation, and four-fifths of a standard deviation.

In correlational research, a *Pearson correlation* ( $r$ ) is considered a measure of effect size. For correlations ( $r$ ), Cohen's (1992) levels<sup>2</sup> were  $r \cong 0.10$  (small effect),  $r \cong 0.30$  (medium effect), and  $r \cong 0.50$  (large effect). In a regression analysis,  $R^2$  (the squared correlation) is a measure of effect size.  $R^2$  is an estimate of the variance of one variable explained by or shared with another variable. Using Cohen's criteria, it follows that 1 percent of variance ( $R^2 = 0.01$ ) is a small effect, 9 percent of variance ( $R^2 = 0.09$ ) is a medium effect, and 25 percent of variance ( $R^2 = 0.25$ ) is a large effect.

---

2 Note that Cohen's effect sizes were based on bodies of experimental and correlational research at the time of his paper.

One of the benefits of examining effect size is that it helps researchers make sense of the results of statistical tests. If a researcher has a very large sample, a small difference between two means may be statistically significant; effect size will indicate whether the difference is meaningful. For example, suppose a researcher wants to determine whether gender predicts mathematics ability. She compares test performance for males and females on a mathematical aptitude test for which scores range from 1 to 100. She finds that the average score for males is 49.7, and the average score for females is 48.5. The standard deviation for both groups is about 12. If her sample size is 400 (200 for each gender), the *t*-test ( $t = 10.00$ ) is statistically significant at the 0.05 level. In contrast, if her sample size is 40 (20 males and 20 females), the *t*-test ( $t = 1.00$ ) suggests a negligible relationship between gender and mathematical ability. How does the researcher make sense of her results? In this hypothetical study, the effect size,  $d = (\bar{X}_1 - \bar{X}_2)/s$ , where  $s$  is the pooled standard deviation of the outcome (here, mathematics ability) across the two independent groups, would be 0.10 for both sample sizes. Clearly the effect size is small, suggesting that the difference between the means is not meaningful, even when it is statistically significant.

The effect-size levels established by Cohen have been criticized by other researchers. Haase, Ellis, and Ladany (1989) suggest that researchers can evaluate effect-size estimates in three ways. They can compare the effect sizes with those found in similar studies. They can compare their effect sizes with Cohen's categories for small, medium, and large effects. They can compare resulting effect sizes with effect sizes established *a priori* (e.g., a researcher who understands the data and knows what a meaningful difference in scores would be, based on his or her knowledge of the scale and prior research, could establish effect-size ranges prior to conducting the study).

One of the most common uses of effect-size estimates is in meta-analysis research. For example, Hattie (2009) synthesized results from a large number of meta-analysis studies on student achievement. In his report, he compared the relative effect sizes of different factors that affect student achievement and ranked the different factors from the smallest to the largest effect. A second common use of effect-size is in estimating statistical power.



## Statistical Power

The validity of a statistical argument is strengthened when the results have statistical power. *Statistical power* is a function of the relationship between probability of error, the number of participants in a sample, and the effect size. Statistical power is an estimate of the probability that the test will *not* lead to a *Type II error* or generate a false negative result. Alternately stated, it is the probability of finding a difference that *does* exist, as opposed to the likelihood of declaring a difference that does not exist (*Type I error*). As statistical power increases, the chance of making a Type II error decreases. The probability of a Type II error occurring is referred to as the *false negative rate* ( $\beta$ ). Therefore power is equal to  $1 - \beta$ , which is also known as the *sensitivity* of a statistical test.

Suppose, for example, that a causal relationship does exist between two variables. In such a case, the smaller the alpha level (e.g.,  $p < 0.01$ ), the higher the chance of failing to reject the null hypothesis and the lower the statistical power. Low statistical power occurs when the likelihood of Type II error is high, which can result from the use of a small sample size and/or when the true effect size is small. For example, suppose the alpha level is set to an exceedingly low level (e.g.,  $p < 0.001$ ) and the power of a statistical test is 0.10. This means that the probability of a false negative ( $\beta$ ) outcome is  $1 - 0.10 = 0.90$ . In deciding whether this is an acceptable risk, the researcher can compare the seriousness of Type I error with the seriousness of Type II error. In this case, the researcher compares a 90 percent likelihood of making a Type II error with a 0.1 percent likelihood of making a Type I error. Put differently, setting alpha at 0.001 results in 900 to 1 odds of making a false negative decision for the given effect size. In some fields, it is less detrimental to make a false negative decision than a false positive decision (e.g., when using multiple regression to help decide who will be admitted to a prestigious college). In other fields, it is better to make a false positive decision than a false negative one (e.g., using a regression formula to identify teens likely to attempt suicide).

When assessing power, researchers must take into account whether or not their statistical tests are *one-tailed* or *two-tailed*, as well as the direction of expected differences if they choose a one-tailed test. A one-tailed statistical test is testing, not only whether

two groups are different, but is focusing the test on whether the difference is in the expected direction; a two-tailed statistical test is testing for differences between groups in either direction. For example, suppose a researcher sets the alpha level at  $p < 0.05$  and her expectation is that group A will have a higher mean than group B. By using a one-tailed test, the researcher focuses the t-test on whether the mean for group A is significantly higher than the mean for group B but does not test whether the mean for group B is higher than the mean of group A. In contrast, if the researcher uses a two-tailed test, half of the alpha level of  $p < 0.05$  is allotted to testing the significance of differences in one direction and half is allotted to testing the significance of differences in the other direction. This means that 0.025 of the alpha level is in each tail of the distribution of the test statistic. When using a two-tailed test, the researcher hypothesizes that differences between groups could be in either direction.

If a researcher uses a one-tailed test and the results of the study are in the expected direction, then the researcher has increased the power of the test. If, however, the results of the study are in the direction opposite to that assumed under the alternative hypothesis (e.g., participants in the treatment condition perform more poorly on a post-test measure than participants in the control condition), then use of the one-tailed over a two-tailed test will have drastically reduced the power of the test.

Because of the interrelationship between error, effect size, sample size, and statistical power, if any three are known, the fourth can be computed. For example, if a researcher knows the sample size, the effect size, and the tolerable level of Type I error, he can compute a power estimate. Alternately, if a researcher knows the typical effect size in studies of a given phenomenon and she is willing to tolerate a certain level of Type I error, she can compute the sample size necessary to obtain a desired level of statistical power (i.e., a sample size that will help her minimize Type II error).

Table 4-1 presents an excerpt from statistical power table for two-sample t-tests with an alpha level of  $p < 0.05$  for samples of 2 to 20 (Bissonette, 2011). Needless to say, a more extensive table is needed to investigate power for larger samples and many such tables would be needed to investigate power for different alpha levels, even with such a simple statistical test. However, the table demonstrates the interrelationship between alpha level, effect size,

Table 4-1

**Power of a Two-Sample, Two-Tailed Test at 0.05 Level**

	<b>Cohen's d</b>																			
Sample n	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.90	1.00	1.10	1.20	1.30	1.40	1.50
3	.05	.05	.05	.06	.06	.06	.06	.06	.07	.07	.07	.08	.08	.09	.11	.12	.14	.16	.18	.21
4	.05	.06	.06	.06	.07	.07	.08	.08	.09	.10	.10	.11	.12	.15	.17	.21	.24	.29	.33	.38
5	.06	.06	.06	.07	.08	.08	.09	.10	.11	.12	.14	.15	.17	.20	.25	.30	.35	.41	.47	.53
6	.06	.06	.07	.08	.09	.10	.11	.12	.14	.15	.17	.19	.21	.26	.32	.38	.44	.51	.58	.64
7	.06	.07	.07	.08	.10	.11	.12	.14	.16	.18	.20	.23	.26	.32	.38	.45	.53	.60	.67	.73
8	.06	.07	.08	.09	.11	.12	.14	.16	.18	.21	.24	.27	.30	.37	.44	.52	.60	.67	.74	.80
9	.07	.07	.09	.10	.12	.13	.16	.18	.21	.24	.27	.30	.34	.42	.50	.58	.66	.73	.80	.85
10	.07	.08	.09	.11	.13	.15	.17	.20	.23	.26	.30	.34	.38	.47	.55	.64	.72	.78	.84	.89
11	.07	.08	.10	.11	.14	.16	.19	.22	.25	.29	.33	.37	.42	.51	.60	.69	.76	.83	.88	.92
12	.07	.09	.10	.12	.15	.17	.20	.24	.28	.32	.36	.41	.46	.55	.64	.73	.80	.86	.91	.94
13	.07	.09	.11	.13	.16	.19	.22	.26	.30	.34	.39	.44	.49	.59	.68	.77	.84	.89	.93	.95
14	.08	.09	.11	.14	.17	.20	.24	.28	.32	.37	.42	.47	.52	.63	.72	.80	.86	.91	.94	.97
15	.08	.10	.12	.15	.18	.21	.25	.30	.34	.40	.45	.50	.56	.66	.75	.83	.89	.93	.96	.98
16	.08	.10	.12	.15	.19	.23	.27	.32	.37	.42	.48	.53	.59	.69	.78	.85	.91	.94	.97	.98
17	.08	.10	.13	.16	.20	.24	.28	.33	.39	.44	.50	.56	.62	.72	.81	.87	.92	.96	.98	.99
18	.09	.11	.14	.17	.21	.25	.30	.35	.41	.47	.53	.59	.64	.75	.83	.89	.94	.96	.98	.99
19	.09	.11	.14	.18	.22	.26	.32	.37	.43	.49	.55	.61	.67	.77	.85	.91	.95	.97	.99	.99
20	.09	.12	.15	.18	.23	.28	.33	.39	.45	.51	.57	.63	.69	.79	.87	.92	.96	.98	.99	.99

sample size, and power. Reading across the row for a sample size of 3, it is evident that, even with an effect size of  $d = 1.5$ , the statistical test has very little power. The power for  $d = 1.5$  and a sample size of 3 is 0.21, suggesting that likelihood of a false negative is 79 percent. For an effect size of 1.5, an increase of sample size to 4 immediately improves power from 0.21 to 0.38.

There are no formal standards for an acceptable level of power. A common recommendation is to use a power estimate of 0.80. If power is 0.80, the risk of Type II error is  $1 - 0.80 = 0.20$ . Using the  $p < 0.05$  as the alpha level, this results in a four to one ratio of the likelihood of Type II error to Type I error. Using 0.80 as a desired goal for power, one can see that, as sample sizes increase, the necessary effect size decreases. However, even with a sample size of 20, an effect size would have to be 1.0 (one standard deviation of difference between groups) to reach a power level of 0.80. Clearly, sample size is an essential factor in ensuring statistical power.

Similar tables have been computed for a wide range of applications of power analysis. Power analysis is a routine part of simulation studies when researchers test new statistical models. As statistical analyses become more complex (e.g., multi-factor analysis of variance, multiple-regression analysis, hierarchical linear modeling, structural equation modeling), the computations for power become more complex. Statistical software packages offer power analysis routines to assist researchers in conducting power analyses and in using power analysis to estimate ideal sample sizes for proposed research.

Examination of statistical power is one tool to help researchers (and consumers of others' research) make sense of their statistical results and evaluate the strength of their claims. In particular, analysis of statistical power helps researchers balance considerations of Type I and Type II error.

### Experiment-Wise Error

*Experiment-wise error* is a threat to the validity of claims when there are several statistical tests conducted in a single investigation. The potential for error is accumulated over the statistical tests.

Suppose that researchers are interested in differences between males and females within cognitive therapy and drug therapy conditions in a study of the effectiveness of cognitive therapy for treatment of depression. When the researchers conduct three statistical

tests—one comparing the two treatment groups, one comparing males and females, and one testing the interaction between treatment and gender—the potential for a Type I error occurring in at least one of the three tests can be up to three times the rate for a single statistical test. If a researcher sets the alpha level at  $p < 0.05$  for each test, then the experiment-wise error could be as high as  $p < 0.15$ . In other words, the researcher might have a chance of making a false positive decision that is as high as 15 percent. The simplest strategy for controlling experiment-wise error is to divide the desired alpha level *for the investigation as a whole* by the number of statistical tests to be conducted in the investigation (Abdi, 2007). In this example, if the total level of error allowable across all of the statistical tests is set at  $p < 0.05$ , each test would have an allowable alpha level (likelihood of Type I error) of 0.017.

If not all tests are important to the purpose of a study, researchers can control experiment-wise error through planned comparisons using procedures proposed by Tukey (1977) or Scheffé (1959). When multiple dependent variables are involved, researchers can consider all tests in a single omnibus test. Most advanced statistics textbooks (see, for example, Agresti & Finlay, 2009; Sirkin, 2006) present a range of strategies for managing experiment-wise error.

Cook and Campbell (1979) called the use of multiple statistical tests “fishing for results.” Fishing occurs when researchers analyze the relationships among many variables using many statistical tests, hoping to find a few statistically significant results. Usually, fishing happens when a researcher has not focused his or her research on a particular set of research questions or is not grounded in a theory with an existing history of studies. Fishing is not theory-building, nor does it solve or explain educational or psychological problems. Researchers must be circumspect, not only in the conclusions they draw from their data, but also in the research choices they make.

Imagine a situation in which a researcher calculates over 100 different inter-correlations in a single study (for example, there are 105 unique correlations that can be estimated between each possible pair taken from a set of 15 variables) and focuses the discussion section of the paper on the handful of correlations that are statistically significant. Even if the true value of each of the correlation is zero, under the laws of probability, the researcher is likely to find at least five correlations to be statistically significant.

## Omitted Variable Bias

*Omitted variable bias* in an investigation would be a threat to the validity of causal claims if a third, correlated variable impacted study results. For example, most cognitive ability tests require the use of language to respond to test items. In addition, mathematical word problems typically require reading in order for students to extract the mathematical information and solve a given problem. Suppose “Dr. Jacobs” is investigating the relationship between gender and mathematical ability. He uses a mathematics problem-solving test to measure students’ ability to solve non-routine mathematical problems. He finds that females score higher than males on the test. The results are inconsistent with previous research on problem-solving that indicates that males do better than females in solving non-routine problems. “Dr. Kim” is also researching the relationship between gender and the ability to solve non-routine mathematics problems. She considers the possibility that the ability to read will influence students’ ability to solve non-routine mathematical word problems. She includes scores from a reading achievement test as a covariate in her study and finds no differences between males and females.

When conducting research, investigators must consider alternate explanations for their results. This requires consideration of variables that may not be the focus of the theoretical relationships but that covary with independent and/or dependent variables. Several of the studies described in Chapter 2 included potential covariates in the research design. The rationale for selecting covariates should be grounded in research.

## Violating Assumptions of Statistical Tests

Many of the statistical procedures used today are based on certain assumptions. One critical assumption made when using most conventional statistical tests is that that scores on measured variables follow a normal distribution. A second critical assumption underlying group comparison tests is that even if the groups’ means differ, the groups’ scores have the same variability. A third assumption is that errors are random.

It is common to find situations wherein the data violate these assumptions. For example, “Dr. Litzenberg” investigates the relationship between post-traumatic stress disorder (PTSD) and

depression. He administers the Beck Depression Scale and compares it with therapists' judgments about patients' level of PTSD. He finds that the scores on the Beck Depression Scale demonstrate a ceiling effect; the hospital patients have such high scores on the depression measure that there is very little score variability. This violates a basic assumption of the correlation coefficient. The resulting correlation between the test scores and the therapists' rating of PTSD will be near zero.

"Dr. Malatea" compares reading achievement scores for middle-school-aged English language learners and native English speakers. The English language learners students have much lower scores than native English speakers' students. In addition, the variance of scores for English language learners' scores is much smaller than the variance of scores for native English speakers. This violates a basic assumption of t-tests and F-tests—that of the homogeneity of the variances across groups, which would make it difficult to interpret the test results.

Despite the best-laid plans, researchers will face situations in which their data violate assumptions of the statistical tests they use. Many simulation studies have been conducted to determine whether statistical tests are robust in the face of violations of assumptions. Simulation studies can be conducted that introduce systematic violations of statistical assumptions and assess how and whether the probability of Type I errors is altered due to these violations. Researchers should review studies that examine the robustness of statistical procedures to determine whether the statistical tests function well in the face of violated assumptions.

### Over- and Under-Interpretation of Results

Two common situations could result in the over- or under-interpretation of statistical results. One situation was mentioned in Chapter 2—that of nesting. The second occurs when researchers are comparing competing statistical models.

As mentioned in Chapter 2, random selection and random assignment are difficult to accomplish in much human research. When researchers investigate a phenomenon or test a theory, they may be forced to use samples that are grouped in some way (e.g., classrooms, schools, clinics, hospitals). When statistical analyses are based on individual scores without taking into account

the grouping context, results may be biased and lead to over-interpretation of results. For example, teachers may have systematic effects on their students; therapists may have systematic effects on their clients. Whenever possible, researchers should account for nesting factors when conducting statistical analyses. Multilevel statistical models attempt to control for nesting effects by accounting for ways in which study participants are grouped.

Statistical conclusions do not always involve rejecting or failing to reject a null hypothesis. We may want to find a statistical model that is the best explanation for the relationships among a set of constructs. In this case, a threat to the validity of statistical conclusions would occur if we over- or under-interpreted differences between alternate models. Power, level of error in the models, and covariates must also be taken into account. For example, confirmatory factor analysis might be used to test two competing models for effective reading comprehension. In one model, each of the constructs shown in Figure 4-1 is an independent predictor of reading comprehension. In the second model (Figure 4-2),

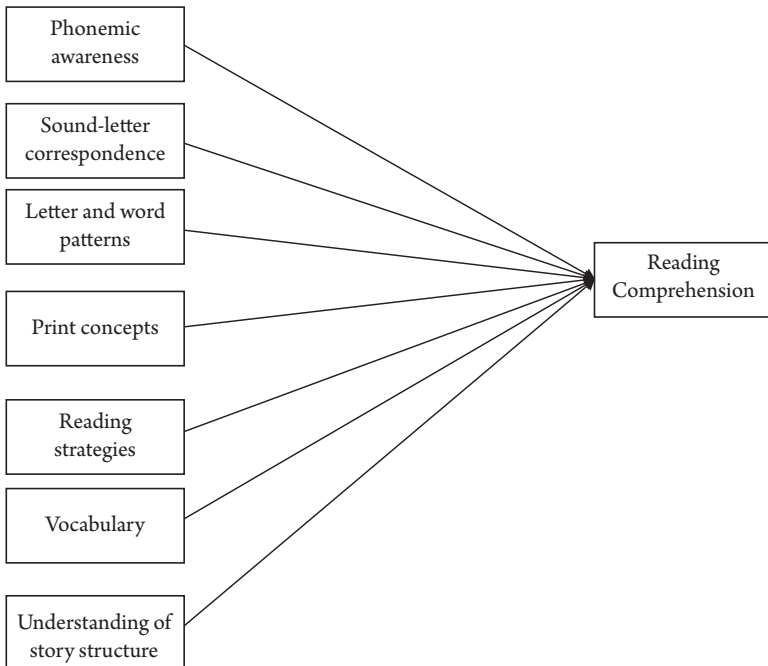


Figure 4-1 Theoretical Model 1 for Reading Comprehension



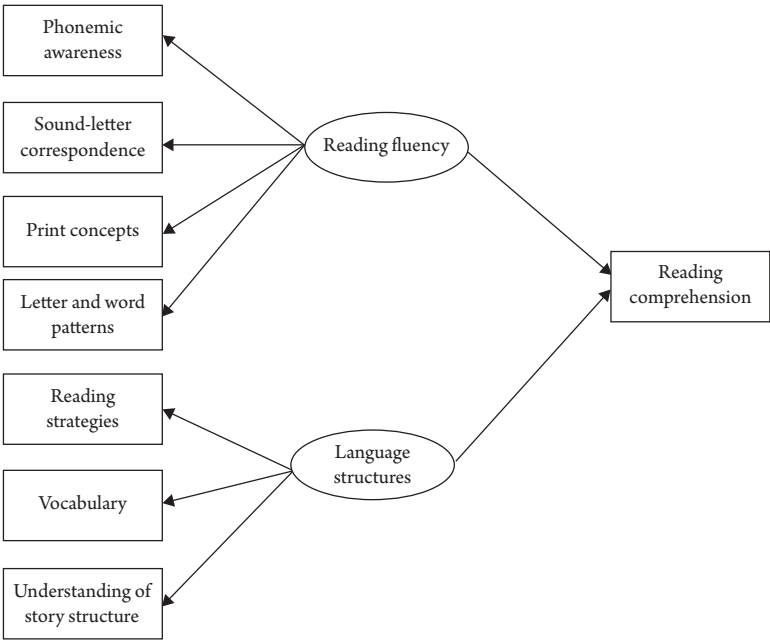


Figure 4-2 Theoretical Model 2 for Effective Reading Comprehension

understanding of story structure, vocabulary knowledge, and reading strategies are grouped together in one higher order factor called *language structures*. Phonemic awareness, sound-letter correspondence, letter and word patterns, and print concepts are combined to form another higher order factor called *reading fluency*. A researcher could compare these two models to find the model supported by better fit statistics.

Threats to statistical conclusion validity in model testing are the same as those for null hypothesis testing. Sometimes statistically significant differences in model fit might be associated with differences that lack practical significance. For example, the fit indices for two models might differ by a small amount (see Table 4-2). Although Model 2 has a slightly higher fit statistic and slightly lower root mean square error of approximation, the values are considered evidence of a good fit for both of the models. In such a case, it is important for researchers to be modest in their claims and to use reason and substantiated theory rather than the statistical significance of the results to choose the best model. The choice of model should be grounded in a strong theoretical foundation or

Table 4-2

**Model Fit Results from Structural Equation Model Analyses of Two Competing Models**

<b>Model</b>	<b>Chi Square</b>	<b>Degrees of Freedom</b>	<b>Incremental Fit Index</b>	<b>Root Mean Square Error of Approximation</b>
Model 1	348.24	28	0.95	0.05
Model 2	242.38	15	0.97	0.04

a practical purpose (e.g., a unidimensional model of reading may fit the data well; however, a multi-dimensional model may provide more diagnostic information).

### **Summary of Threats to the Validity of Statistical Conclusions**

No single statistical test can “prove” a theory or answer a research question. Validation of theoretical claims is an ongoing process of gathering evidence to test claims that are derived from theories. When making theoretical claims based on the results of research, researchers must determine whether statistical evidence provides adequate support for their claims. The validity of statistical conclusions depends on the strength of statistical results. Statistical significance may be an artifact of large sample sizes; therefore, researchers must evaluate effect size (the magnitude of an effect) and power (probability of a false negative) when making sense of statistical results. Statistical conclusions are stronger, and provide stronger support for theoretical claims, when results are statistically significant, when effect sizes are moderate to large, and when statistical power is strong.

Each of the potential threats to the validity of statistical conclusions can be managed if researchers consider these threats before beginning research. Prior research should provide some guidance regarding covariates that may influence the relationships among variables. Sample sizes can be planned to help ensure statistical power. Statistical analyses can be selected that minimize the number of statistical tests within a single investigation.

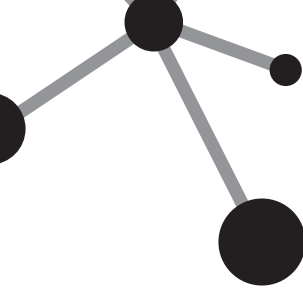
Steps can also be taken to address threats after a study has been conducted. If the results of an investigation yield data that violate assumptions of statistical tests, researchers can apply non-parametric statistics that are less sensitive to these violations. Planned comparisons can minimize the number of statistical tests. Effect size can be used to determine whether statistically significant results are meaningful.

One of the best ways to minimize threats to the validity of statistical conclusions is to include statistical analysis in the overall plans for research. Selecting the right statistical tool helps ensure that design, sample sizes, and data collection serve the researcher's purposes.

## References

- Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In N. J. Salkind (ed.), *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage.
- Agresti, A., & Finlay, B. (2009). *Statistical Methods for the Social Sciences* (4th ed.). Upper Saddle River, NJ: Pearson.
- Bissonnette, V. (2011). Statistical power of the t-test for two independent samples. Statistics links: Some useful statistical tables. Retrieved July 14, 2012, from <http://facultyweb.berry.edu/vbissonnette/tables/pwr2samp.pdf>.
- Boneau, C. A. (1960). The effects of violations of assumptions underlying the test. *Psychological Bulletin*, 57, 49–64.
- Cohen, J. (1992). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin Company.
- Feir-Walsh, B. J., & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, Normal Scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement*, 34, 789–799.
- Field, A. (2005). *Discovering Statistics Using SPSS*. London: Sage Publications.
- Garner, R. (2010). *Joy of Stats: A Short Guide to Introductory Statistics*. Toronto, CA: Toronto Press.
- Haase, R. F., Ellis, M. V., & Ladany, H. (1989). Multiple criteria for evaluating the magnitude of experimental effects. *Journal of Counseling Psychology*, 36, 511–516.
- Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Related to Achievement*. New York: Routledge.
- Hollingsworth, H. H. (1980). An analytical investigation of the effects of heterogeneous regression slopes in analysis of covariance. *Educational and Psychological Measurement*, 40, 611–618.

- Howell, D. C. (2011). *Fundamental Statistics for the Behavioral Sciences* (7th ed.). Belmont, CA: Wadsworth-Cengage Learning.
- Keselman, H. J., & Toothaker, L. E. (1974). Comparison of Tukey's T-Method and Scheffé's S-Method for various numbers of all possible differences of averages contrasts under violation of assumptions. *Educational and Psychological Measurement*, 34, 511–519.
- Levy, K. (1980). A Monte Carlo Study of analysis of covariance under violations of the assumptions of normality and equal regression slopes. *Educational and Psychological Measurement*, 40, 835–840.
- Martin, C. G., & Games, P. A. (1976, April). ANOVA tests of homogeneity of variance when n's are unequal. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Popper, K. (1959). *The Logic of Scientific Discovery* London: Hutchison & Company.
- Ramsey, P. H. (1980). Exact Type 1 error rates for robustness of student's t test with unequal variances. *Journal of Educational Statistics*, 5, 337–349.
- Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.
- Sirkin, R. M. (2006). *Statistics for the Social Sciences*. Thousand Oaks, CA: Sage Publications.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Boston: Addison-Wesley.
- Urdan, T. C. (2010). *Statistics in Plain English*. New York: Taylor and Francis Group.
- Wu, Y. B. (1984). Effects of heterogeneous regression slopes on the robustness of two test statistics in the analysis of covariance. *Educational and Psychological Measurement*, 44, 647–663.
- Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*, 67, 55–68.



## CONSTRUCT - RELATED EVIDENCE FOR VALIDITY

CHAPTER 1 INTRODUCED the idea that questions of validity apply to research and assessment; that research and assessment are interdependent activities. One cannot conduct high-quality research without assessment tools and strategies; one cannot build high-quality assessment tools without using a wide variety of research methodologies during development and in validation of the inferences from scores.

Chapters 5 and 6 are focused on the validity of *inferences* and *actions* from assessment scores and other measures.<sup>1</sup> This chapter describes issues related to construct validation. Chapter 6 describes issues related to validation of the interpretations and uses of assessment scores. Both chapters present strategies used to investigate questions of validity. Many of the ideas in these

- 
1. As stated in Chapter 1, a test score is usually a numerical value that results from some measurement procedure; however, measurement is not the only form of assessment and numerical test scores are not the only results of assessment procedures. As shorthand, I will use *assessment scores* or *scores* to describe any descriptive or numerical summary based on an assessment process. It is important to note that the validation issues that apply to numerical test scores apply to all summaries based on assessment procedures.

chapters come from *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, & NCME, 1999)) as well as guidance provided by validity theorists (e.g., Cronbach & Meehl, 1955; Cronbach, 1971; Kane, 2006; Messick, 1989).

Evaluating construct-related evidence for validity involves evaluating the logical arguments and empirical evidence supporting the central claim (implied or explicit) in assessment—that the scores from an assessment can be interpreted and used in a particular way. Score interpretations may include predictions about examinees' most likely performance on a criterion (e.g., piloting an airplane or success in college) or to make inferences about examinees' measures on internal constructs (e.g., depression, reading comprehension, attitudes toward science). Evidence is needed to determine whether these inferences can be made. Over the past 50 years, most validity theorists have reached a consensus that construct-related evidence for validity is the cornerstone of validation research.

## Constructs

Understanding construct-related evidence for validity requires understanding the idea of a construct. The first definition of the term *construct* in most dictionaries includes words such as “build,” “make,” or “form”—which is fitting when we speak of the constructs we assess. Constructs are human inventions. We create constructs to explain consistencies in complex human behaviors. In much educational and psychological research, constructs are latent. In other words, they are unseen and must be inferred from observable behaviors—either naturally occurring behaviors or responses to standardized stimuli. Researchers work together to define constructs so that shared definitions guide their work. Sometimes there are competing definitions of constructs (e.g., achievement motivation, mathematical ability, and historical literacy). Constructs may also be called *traits*, *proficiencies*, *personalities*, *dispositions*, *abilities* and so forth.

## Criterion Performances

Criterion performances differ from constructs in some ways. A criterion performance might involve on-the-job behaviors

such as flying an airplane or something more abstract such as college freshman grade point average (GPA). The main difference between criterion performances and constructs is the degree to which a criterion performance can be observed. This does not mean that constructs are not in play when we attempt to predict criterion performances. For example, a department store might give applicants a test to assess how well they are likely to handle money or to figure discounts for shoppers. The test might include items asking about the value of a discount, the discounted price, or the amount of change given to customers. Even in a case where the criterion performance can be observed, test developers select a limited number of items from which to infer unseen future performances. Sometimes, the assessment tasks will be direct samples from the domain of criterion behaviors (e.g., a writing sample, on-the-job performance); sometimes they will be indirect measures of the domain of criterion behaviors (e.g., multiple-choice test questions). Regardless of the form of assessment, inferences must be made from performance on the assessment to the criterion performance.

### Claims and Construct Validation

A critical aspect of validity theory is that assessment tools themselves are not “valid” or “invalid.” Assessment tools produce scores from which inferences are made—inferences about the examinee in relation to a construct or criterion performance. It is these inferences (claims) that must be validated. Although assessment tools are not valid or invalid, the first steps in validation research involve examination of the scientific rationale for the substantive and structural components of any assessment tool. For illustrative purposes, the following description highlights substantive and structural components of a physical measurement.

Suppose a scientist creates a measurement tool. The scientist claims that the tool measures linear dimensions. The first validity question to be asked is whether the tool can measure linear dimensions. Any rigid straight-edge with interval markers on it could be used to measure linear dimensions. If the tool meets these three simple criteria (rigid, straight edge, and interval markers), we have a structural source of evidence to support the claim that the tool can measure linear dimensions.

However, the tool has no inherent worth until it is used to measure an object. Suppose that the scientist uses the tool to measure the height of plants. She claims that the plants in her laboratory have an average height of 8 units. The second validity question is whether we can trust this claim. To trust this claim, we need to know the following: Can measurements from the tool be used to compute averages? Has the researcher used the same (standardized) procedure to measure each and every plant in the lab? Was the tool long enough to measure the tallest plants, and were the units small enough to measure the shortest plants and to distinguish between heights of plants? If we are satisfied that the units are adequate for computing averages, the scientist used the same procedure for all plants, and the size of the intervals and length of the tool are adequate, we have structural evidence to support her second claim regarding average height. Note that the trustworthiness of the measurements depends on features of the tool as well as the implementation of the tool by the scientist. In other words, trustworthy measures require consistent behaviors by the individuals who do the measurement.

The researcher has an average measure in units, but the number has no meaning. Is a height of 8 units typical of this species of plant? Is it taller or shorter than the typical height of this species of plant? The scientist claims that her plants are shorter than average. The third validity question is whether this claim is warranted based on evidence *beyond* her lab. Suppose that most plants of this species grow to be about 25 centimeters tall. We compare the units on the new tool with a centimeter ruler and determine that each unit interval on the new tool is about 2.5 centimeters. From this information, we can infer that the units are large compared with centimeters. We can extrapolate to determine that the scientist's plants have an average height of about 20 centimeters and, therefore, are shorter than average. We now have evidence to support the third claim made using the measurement tool. Whereas earlier claims were descriptive of the tool itself, this claim is inferential, and required corroboration between the measurements from the new tool and measurements from a known tool.

Suppose further that the scientist has two groups of plants: those that received a fertilizer and those that did not. The average height of the non-fertilized plants is 6 units, and the average height of the fertilized plants is 12 units. The scientist claims that the fertilized



plants are twice the height of the non-fertilized plants and that the fertilized plants are taller than average. We now have two more inferences drawn from the measurements that must be validated. If the scientist has used a standardized measurement procedure, we have support for her claim about the comparisons between the heights of fertilized and non-fertilized plants. Based on the average height of this plant species, we have support for her claim about the relative height of the fertilized plants.

The scientist claims that the fertilized plants are healthier than the non-fertilized plants and fertilized plants are healthier than average plants of this species. The word *healthier* places a completely new burden of proof on the validation process. Up to this point, claims were straightforward and could be easily validated. However, once a value judgment is placed on the measurement, the validation process becomes much more complex. What does “healthier” mean? Is height a sufficient characteristic to determine plant health? What other evidence is needed to support this claim?

In this example, the scientist has created a new measurement tool of a familiar phenomenon that already has many established measurement tools. It is a simple phenomenon (linear measurement), and measurement is straightforward. The first steps of the validation process are also straightforward. The fact that other standardized and widely accepted linear measurement tools exist makes validation much easier.<sup>2</sup> When measuring latent constructs (e.g., adaptive behaviors, achievement motivation, depression, reading comprehension), assessment development and score validation processes are much more complex. Not all researchers agree on the definitions of the constructs. Measurement units are not necessarily equal intervals at all locations on a scale—which makes both absolute and relative measurement very difficult. The meanings of scores are more difficult to define. Too often, assessment users move quickly from numbers to inferences to interpretations—which has significant implications for the validation process.

- 
2. One might ask why anyone would create a new measurement tool for linear measurement when such tools already exist. The same question could be asked for measures of depression, reading comprehension, and pain. Generally, new tools are developed because research suggests the need for new tools or because current tools are considered to be inadequate or invalid.

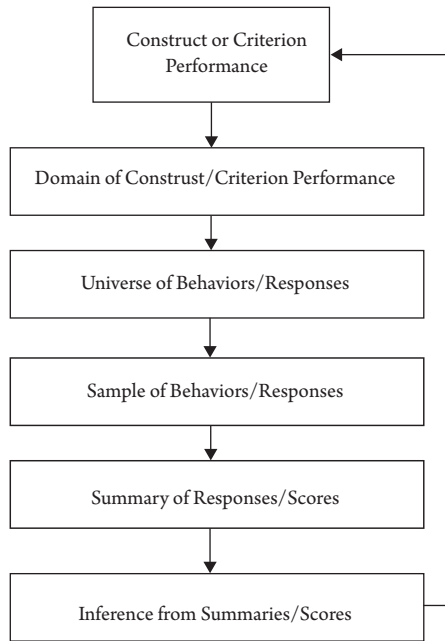
## Scores as Signs

A critical difference between validation of scores from measures of physical phenomena and scores from measures of latent constructs (or scores that predict criterion performances) is the directness of the measurement. Measurement of many physical constructs (e.g., weight, height, distance) involves measuring the actual phenomenon. Not only is the measurement direct, if the units of the measurement tool are sufficiently small, it is possible to keep measurement error to a minimum.

To assess a psychological construct or predict a criterion performance, we must make inferences from what is seen (e.g., behaviors and responses to assessment items/tasks) to what is unseen—the construct or criterion performance. We cannot observe indefinitely, nor can we present a sufficient number of items or tasks to represent the construct or criterion performance in all situations and contexts. Assessment development processes involve defining the domain of the construct, identifying how the domain will be assessed (specifying the “universe” of behaviors or responses), sampling from the universe of behaviors or responses, summarizing observations or generating scores, and drawing inferences from the scores to the domain of the construct or criterion performance (Kane, 2006). Figure 5–1 shows the process used to develop tools that will yield scores from which inferences about the strength of a construct or criterion performance can be made.

Beyond inferences are the interpretations and actions based on those inferences. The validation process involves questioning every step, from our definition of the construct or criterion performance, to the intended interpretations we make using the results of assessments, to the consequences of score interpretation and use.

Measurement of latent constructs involves observing *representations* (signs) of the phenomenon of interest. We cannot measure depression or reading comprehension directly, and the strength of depression or reading comprehension for an individual may differ from situation to situation. Therefore, we infer from a given assessment event the likely strength of a construct within the individual, and we infer that this measure applies to a broad range of situations. When the focus is a criterion, assessment performance is used to predict likely criterion performance over time and in different



**Figure 5–1** From Construct or Criterion Performance to Inferences from Descriptive Summaries or Scores (Adapted from Kane, 2005)

circumstances. The challenge in assessment validation research is to “ascertain the degree to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported” (Messick, 1989, p. 13).

### Gathering Evidence to Support Construct-Related Validity Claims

Since the 1950s, construct validity has emerged as the heart of the validity question. What evidence supports the inferences to be made from test scores? In addition, testing standards clearly indicate that test developers should provide evidence supporting the use of test scores in a given context. Messick (1989) expanded our understanding of validity when he indicated that validation studies must consider the consequences caused by the value implications of score interpretations, as well as intended and unintended consequences of test score use. Figure 5–2 is a two-dimensional framework for thinking about validation of assessment results

	Interpretation	Use
Evidential Basis	Construct Validity	Construct Validity + Relevance & Utility
Consequential Basis	Construct Validity + Value Implications	Construct Validity, Relevance, Utility, & Social Consequences

Figure 5–2 Facets of Validity (Adapted from Messick, 1989)

(adapted from Messick, 1989). This framework takes into account the meaning of assessment results, their usefulness, and the consequences of their interpretation and use.

To investigate these facets of validity, Messick suggests “a half dozen or so” sources of evidence:

We can look at the content of the test in relation to the content of the domain of reference. We can probe the ways in which individuals respond to items or tasks. We can examine relationships among responses to the tasks, items, or parts of the test, that is, the internal structure of test responses. We can survey relationships of the test scores with other measures and background variables, that is, the test’s external structure. We can investigate differences in these test processes and structures over time, across groups and settings, and in response to experimental interventions—such as instructional or therapeutic treatment and manipulation of content, task requirements, or motivational conditions. Finally, we can trace the social consequences of interpreting and using test scores in particular ways, scrutinizing not only the intended outcomes but also unintended side effects. (Messick, 1989, p. 16)

The first four sources provide construct-related evidence for validity. The last two sources can provide evidence related to score interpretation and use, including the consequences of score interpretation and use.

Kane (2006) suggests that Messick’s (1989) model does not provide sufficient guidance for validation research. He proposes that validation research begin with a clear statement of the intended interpretation and uses of assessment scores and that all validation studies should be framed in terms of claims

and arguments in reference to the intended interpretation and use of scores. Kane's approach can be helpful when assessment developers must establish a research agenda for validation studies and when assessment users must determine whether there is sufficient evidence to warrant the use of assessment scores for a given purpose.

Kane (2006) describes four claims that should guide validation research (see Claims 1 through 4 in Table 5-1). The first claim is that the scores can be used to make inferences about examinees in terms of the strength of a construct for the individual or their likely criterion performance. In order to support this claim, evidence must be provided for two arguments: the scoring rules used for the items are appropriate, and the scoring rules are applied consistently. The second claim is that one can generalize from the score on the assessment to the universe of behaviors or responses that the items and/or tasks represent. Two arguments must be made to support this claim: the items and/or tasks on the assessment represent the domain of the construct or criterion performance, and the sample of items and/or tasks is large enough to minimize random errors. The third claim is that it is possible to extrapolate from the score on the assessment to the domain of the construct or criterion performance. Two arguments must be made to support this claim: the items and tasks require the same abilities as are required in the domain of the construct or criterion performance, and responses to items/tasks require no abilities that are irrelevant to the construct or criterion performance. The fourth claim is that interpretations of assessment scores and the decisions made based on assessment scores are appropriate. Two arguments must be made to support this claim: the implications of assessment score use are appropriate, and the properties of the observed scores support the of interpretations. Validation research involves obtaining evidence to support each of the arguments.

Although Kane (2006) argues that consequences are relevant to validity, he omits consequences from his list of testable claims. The fifth claim in Table 5-1 adds "consequences" to the validity claims and arguments. Messick's (1989) writing would suggest two arguments related to consequences: the value implications of assessment score interpretations are appropriate, and the consequences of assessment score use are appropriate.

Table 5–1 <b>Combined Framework for Validation Research and Evaluation</b>		
	<b>Claim/Argument</b>	<b>Source of Evidence</b>
<b>Claim 1</b>	<b>Scores can be used to make inferences.</b>	
Argument 1	Scoring rules are appropriate	Check answer keys Check fit of rating scale with behavior or statement Check rubric alignment with domain behaviors Evaluate evidence for scoring model (e.g., fit to IRT model for scaling and item calibration; factor analysis to investigate dimensionality of internal structure) Check that scoring model fits purpose (e.g., norm-referenced vs. criterion-referenced)
Argument 2	Scoring rules are applied consistently	Check application of scoring key Check conversion of ratings to scores Compute and evaluate item analysis statistics (classical and IRT) Compute and evaluate inter-rater agreement statistics Check for rater drift

(continued)

Table 5–1

**(Continued)**

	<b>Claim/Argument</b>	<b>Source of Evidence</b>
<b>Claim 2</b>	<b>It is possible to generalize from scores to a universe of behaviors related to the construct or criterion performance.</b>	
Argument 1	Items and tasks represent the universe of behaviors or responses defined in specifications	Review domain definition for fit with theory and research Review alignment of test specifications with domain definition Review alignment of item specifications with domain definition Review assessment content (items and tasks) in relation to universe of behaviors or responses (domain definition)
Argument 2	Sample of items/tasks is large enough to minimize errors in assessment scores	Examine generalizability analysis to identify sources of error Evaluate the reliability coefficients Review inter-rater agreement resultsEvaluate decision consistency Evaluate standard error of measurement

<b>Claim 3</b>	<b>It is possible to extrapolate from the score to the domain of the construct or criterion performance.</b>	
Argument 1	The same knowledge, skills, abilities, and dispositions are required on the assessment as in the domain of the construct or criterion performance	<p>Evaluate alignment of items with the construct or criterion performance</p> <p>Observe examinees during assessment event (e.g., think-aloud study) to verify use of expected knowledge, skills, abilities, etc.</p> <p>Conduct factor analyses to examine internal structure of the test scores</p> <p>Obtain correlations between assessment scores and other measures of the same construct</p> <p>Obtain correlations between assessment scores and measure of the criterion performance</p> <p>Conduct experimental studies to determine whether manipulations of independent variables have expected effects on assessment scores</p>
Argument 2	No knowledge, skills, abilities, or dispositions irrelevant to the domain of the construct or criterion performance are required	<p>Conduct bias and sensitivity reviews to identify factors that may depress or enhance scores</p> <p>Observe examinees during assessment event to verify no irrelevant knowledge, skills, or abilities are required</p> <p>Conduct differential item functioning studies to look for secondary dimensions or sources of bias</p> <p>Conduct experimental and correlational (e.g., multi-trait/multi-method; factor analysis) studies to examine alternate explanations for scores</p>

(continued)



Table 5–1

**(Continued)**

	<b>Claim/Argument</b>	<b>Source of Evidence</b>
<b>Claim 4</b>	<b>Interpretations of assessment scores or decisions made from scores are appropriate.</b>	
Argument 1	Properties of the observed scores support interpretations	Conduct studies to determine whether scores behave as expected in response to interventions Conduct studies to determine whether scores behave as expected over time Conduct studies to determine whether scores behave as expected across different groups Conduct studies to determine whether items and scores behave consistently across linguistic and cultural groups
Argument 2	Implications of assessment score use are appropriate	Conduct studies to determine whether assessment scores serve intended purpose(s) Conduct studies to determine whether assessment is needed for the purpose(s) Conduct new studies when scores are to be used for new purposes

<b>Claim 5</b>	<b>Consequences of score interpretation and use are appropriate.</b>	
Argument 1	Value implications of assessment scores are relevant to construct or criterion performance	Identify potential unintended consequences of score interpretations Conduct studies to examine intended and unintended consequences of score interpretations
Argument 2	Social consequences of assessment score interpretation and use are appropriate	Identify potential unintended consequences of score use Conduct studies to examine intended and unintended consequences of score use

The first three validity claims can guide research to obtain construct-related evidence for validity of inferences from scores. Trustworthiness of scores, generalizability of scores, and extrapolation of scores to the domain go hand in hand. If scoring rules are not appropriate, it is difficult to generalize from assessment scores to a larger universe of items or tasks that represent the domain. If the universe of items or tasks does not truly represent the construct, appropriate scoring rules and generalizability are nice but not very helpful. The last two claims in Table 5–1 can guide research to obtain evidence for the validity of score interpretation and use. These claims will be discussed in Chapter 6.

### **Validity Claim 1—Scores Can Be Used to Make Inferences**

*We can examine relationships among responses to the tasks, items, or parts of the test; that is, the internal structure of test responses.* (Messick, 1989, p. 16)

It may seem premature to discuss scores when considering validity arguments. Scores come after the construct has been defined, item and test specifications have been developed, and items have been written. Scores follow content reviews and bias and sensitivity reviews. Scores follow pilot or field testing. Yet scores are the basis of all inferences made about examinees—even when the scores are descriptive summaries of examinee behaviors. Therefore, Kane’s (2006) claims and arguments must be viewed as mutually supportive rather than sequential. The purpose of assessment scores influences how we write incorrect answers to a multiple-choice item or rubrics for constructed-response items and performance tasks. The purpose of assessment scores influences how we generate scales and create score reports. Therefore, it is critical that assessment developers consider the scores derived from an assessment at the same time as they consider how to assess the targeted construct or criterion performance.

#### **Evidence for Appropriateness of Scoring Rules**

Messick identifies the internal structure of an assessment as one of his “half a dozen or so” sources of evidence for validation of test

scores. Verification of item scores through systematic reviews and statistical item analyses is a critical aspect of the internal structure of assessments. Table 5–1 suggests two arguments to support the claims regarding the trustworthiness of scores. The first argument has to do with whether the scoring rules are appropriate.

### Scoring Rules

The appropriateness of scoring rules includes a wide range of issues. Scoring rules include answer keys for multiple-choice items, rating scales for items on questionnaires and psychological assessments, rubrics for performance items and tasks; and conversions from observed scores to derived scores. For example, suppose “Dr. Li” creates an assessment tool to measure achievement motivation. Figure 5–3 shows four items that might be included in her assessment. The first version of the rating scale is a consistency rating, whereas the second version focuses on the degree to which an examinee agrees with the statement.

These different versions of ratings result in different meanings from assessment scores, even though the item statements are the

Rating Version One

I try to get higher grades than my classmates.	Consistently	Occasionally	Rarely	Never
I like it when teachers grade ‘on the curve’.	Consistently	Occasionally	Rarely	Never
I always try to do my best work in school.	Consistently	Occasionally	Rarely	Never
I like to learn while I do projects in school.	Consistently	Occasionally	Rarely	Never

Rating Version Two

I try to get higher grades than my classmates.	Strongly agree	Agree	Disagree	Strongly disagree
I like it when teachers grade ‘on the curve’.	Strongly agree	Agree	Disagree	Strongly disagree
I always try to do my best work in school.	Strongly agree	Agree	Disagree	Strongly disagree
I like to learn while I do projects in school.	Strongly agree	Agree	Disagree	Strongly disagree

**Figure 5–3** Example of the Role of Ratings in Score Meaning

same. The choice of rating scale will partly depend on the purpose of the assessment. If the purpose is to measure the consistency of students’ achievement motivations, rating version one is more appropriate. If the purpose is to compare the strength of achievement goals versus performance goals, rating version two is more appropriate.

Similar issues arise with scoring rubrics for performance tasks. Suppose “Dr. Olin” is creating a test of mathematical problem-solving. Figure 5–4 shows an item with two possible rubrics. Although the rubrics are similar, they focus on different ideas. Rubric version one gives full credit if the student considers sales numbers in the solution; however, other factors (e.g., pleasing the largest number of customers, providing a range of options) are

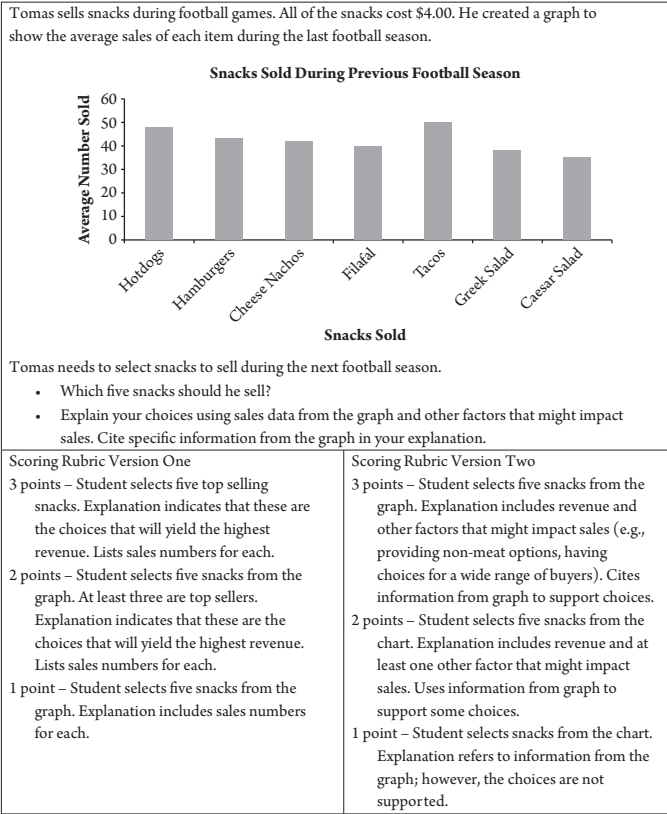


Figure 5–4 Example of the Role of Scoring Rubrics in Score Meaning

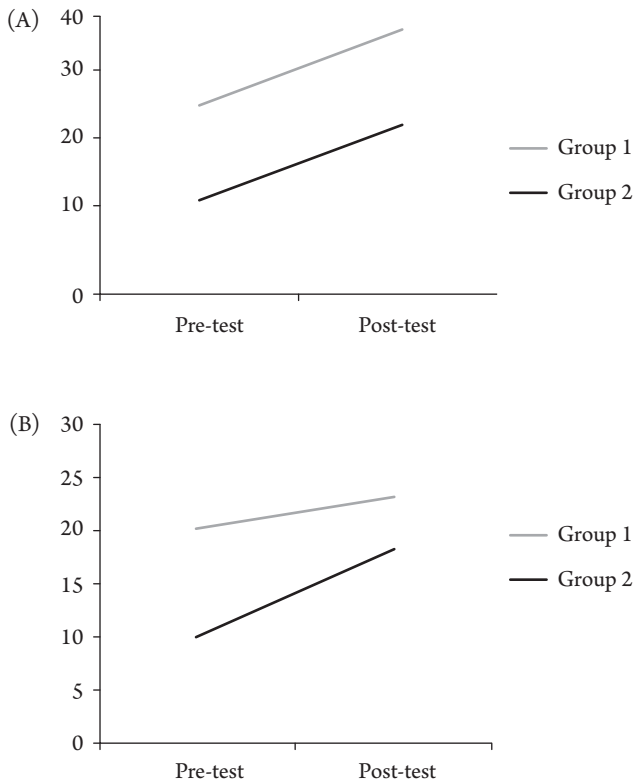
not valued as part of mathematical problem-solving. The rubric assumes a single correct answer. Reference to information in the graph is limited to a list of the highest five numbers. In contrast, rubric version two focuses on a broader range of factors that might influence choices and also attends to the degree to which students support their choices with information from the graph, which might include numerical comparisons, sums across several categories, and so on.

These differences in the rubrics have bearing on the meaning of scores. One rubric is focused on number order. The other is focused on reasoning with data. These purposes must be evident in the construct definition in order to evaluate the appropriateness of item or task scores. Clearly, review of scoring rules has a significant role in validation of inferences from test scores.

### Level of Measurement

In addition to scoring rules, assessment developers must decide whether they plan to classify examinees into groups, rank-order examinees, or measure on an interval scale. A six-level rubric used to assess writing performance is unlikely to be an interval scale. If the purpose of the assessment is to determine the relative quality of written essays or whether individual examinees have improved their writing skill over time, ordinal scores are acceptable. In contrast, if the purpose of the assessment is to determine whether changes over time are consistent, or to compare examinees who receive a particular intervention with those who do not receive the intervention, an interval scale will be more appropriate.

An example of how interval and ordinal scales can affect interpretations is shown in Figures 5–5A and 5–5B. The graph in Figure 5–5A shows score changes for two groups. The lines on the graph suggest that score changes were the same for both groups. However, a close look at the vertical axis shows that the scale is not an equal-interval scale. Score intervals at the top of the scale indicate less change than score intervals near the bottom of the scale. Figure 5–5B shows what the score changes might look like on an interval scale. Clearly, the meaning of score changes differs depending on the type of scale used. To ensure interval scales, assessment developers should use a scaling model



**Figure 5-5** (A) Score Changes on an Ordinal Scale. (B) Score Changes on an Interval Scale

designed to produce interval scales (e.g., item response theory [IRT] modeling).<sup>3</sup>

### Total Score Models

Another issue related to the appropriateness of scoring rules has to do with the scoring model used when generating total scores. Assessment developers have many options when deciding how to report scores. If the purpose of an assessment is to compare examinees with each other (e.g., determining whether a score from a

3. Several IRT scaling models are possible. The 1-parameter (Rasch) model generates a scale by parameterizing item difficulty (location) on the underlying trait scale. The 2-parameter model generates a scale by parameterizing

depression scale is “normal”), a *norm-referenced* scoring model is appropriate (e.g., percentile ranks). For a norm-referenced score model, the assessment must include items or tasks that examinees who are lowest on the underlying scale are likely to respond to correctly or favorably, and the assessment must include items or tasks that examinees who are highest on the scale are likely to respond to incorrectly or unfavorably. For the ideal norm-referenced score model, all examinees would get at least one point on the scale, and no examinees would get all of the points.<sup>4</sup> This has implications for item selection.

On the other hand, if the purpose of an assessment is to identify individuals who are at risk, to select individuals for a special program, or to determine whether students have mastered a corpus of knowledge and skills, a *criterion-referenced* scoring model is more appropriate. To identify individuals who are at risk of attempting suicide, a criterion-referenced model for a depression scale might involve setting a cut-score based on a comparison of scores for individuals who have attempted suicide with individuals who have not attempted suicide.

Whether norm-referenced or criterion-referenced, assessment developers should select the scoring model that is most appropriate for their assessment purpose before they begin developing the assessment and provide a rationale for the selected scoring model anchored in the stated purpose of the assessment.

Selection of a scoring model also has implications for the selection of items and tasks. For example, if an achievement test is to

---

item difficulty and discrimination (how well the item discriminates between individuals above and below the item's difficulty). The 3-parameter model generates a scale by parameterizing item difficulty, discrimination, and the likelihood that examinees will guess a correct response on a multiple-choice item. Partial-credit IRT models are used with items that have ratings (e.g., *strongly agree* to *strongly disagree*) or rubric scores. The choice of model will impact how well the items fit the scale and how well interval scaling works.

4. The rationale for this ideal is that the scale must be useful for measuring everyone in the population. For example, a 7-foot-long tape measure would be adequate for measuring the height of nearly everyone in the world, including Michael Jordan (66) and Kobe Bryant (66). However, it would not be long enough to measure the basketball players Kareem Abdul-Jabbar (72) or Yao Ming (76).



be used to assess whether students are meeting grade-level standards, it is not necessary to include items and tasks that measure knowledge or skills more appropriate for students at higher or lower grade levels. However, if an achievement test is to be used to provide normative comparisons, it may be necessary to include above-grade-level and below-grade-level content in order to adequately assess the highest and lowest performing students.

Another scoring model issue has to do with dimensionality in scores. Assessment developers generally provide a total score for assessments. A single total score suggests that a construct is unidimensional. However, many constructs and criterion performances are composed of multiple dimensions. In such a case, a single total score could obfuscate score meaning. Individuals with high scores are probably strong on all dimensions, and individuals with low scores may be weak on all dimensions.<sup>5</sup> However, individuals with middle scores could have many different profiles within the score. A middle score might reflect a high score on one dimension and a low score on a second dimension, or a medium score on two dimensions. Multidimensional scores are difficult to interpret. For example, suppose a reading comprehension test includes items that measure both text comprehension and vocabulary knowledge. These may be correlated but distinct dimensions of reading and, as such, separation of scores into two subscores would be important to ensuring that examinee performance can be interpreted. As assessment developers plan their scoring models, they should conduct studies (e.g., factor analyses, multidimensionality analyses) that provide support for the appropriateness of all summary scores. If more than one dimension is a better fit to the data, assessment developers should provide sub-scores for each of the dimensions.

### Evidence for Scoring Consistency

The second argument related to the trustworthiness of scores has to do with scoring consistency (See Table 5–1). When students write essays for a class, most teachers will admit that they change

- 
5. One of the major challenges in measurement is that low scores are the most difficult ones from which to make inferences; low scores generally have the most measurement error. When an examinee earns a low score, it may be caused by low level of the construct, poor motivation, confusion about the intent of items, or a myriad of other possible factors.

their evaluation criteria as they work through the papers. Teachers might be lenient at the beginning of the evaluation process and become more stringent once they have scored some highly proficient papers. Alternately, teachers might have very stringent expectations when they begin, but loosen their expectations as they discover that most students failed to meet these expectations. Teachers may have high expectations early in the day and become more lenient as the scoring process extends to the evening. Without a tightly defined scoring rule (rubric), teachers may give different scores to two papers that are essentially of the same quality. Scorer consistency is only one of many score consistency issues relevant to the validity of scores.

Scoring inconsistencies lead to scores that cannot be trusted to mean the same thing from one examinee to the next. Assessment developers and users must document the strategies they use to ensure scoring consistencies. Several issues could arise to threaten score consistency, including faulty technology, incorrect answer keys, incorrect translation of item responses to item scores, more than one correct answer for an achievement test item, and rater inconsistencies for performance items and tasks.

### **Technology Documentation**

When scanners are used to capture and score examinees' responses on scannable score sheets, assessment developers must have documentation that the scoring keys for selected-response items are accurate. Technical materials for scanning operations should describe the strategies used to ensure correct scoring keys. For example, scanning operators should document that they test all answer spaces on the documents, that the marked locations translate into appropriate score records, and that the score records translate into correct item scores and total scores. Similarly, when computers are used to administer and score items and tasks, accuracy of examinees' scores depends on whether the scoring program correctly captures examinee responses and correctly applies scoring algorithms to translate responses to item and task scores. Technical reports should describe the procedures used to ensure accuracy of electronic scoring programs.

### **Item Analyses**

Item analyses are a routine part of test development. Generally, two types of item analyses are conducted: classical item analyses

and item response theory (IRT) analyses. Classical item analyses include item means, item score to total score correlations, inter-correlations among item scores, options analyses (statistical performance of answer choices for a multiple-choice item), rubric analyses (statistical performance of each of the score levels on a rubric), and rating scale analyses (e.g., statistical performance of each response option for a Likert-type scale). The two most popular IRT analyses include item difficulty/popularity (i.e., the item’s location on an underlying scale) and item fit.

Item analysis statistics should provide support for item quality, correctness of answer keys, effectiveness of distractors for multiple-choice items, effectiveness of scoring rubrics, and appropriateness of rating scales. An item’s data should support expected item difficulty or popularity<sup>6</sup> and reveal no technical flaws. If statistics reveal that an item that was expected to be easy or popular is quite difficult or unpopular, this *could* suggest a scoring error. Item statistics for multiple-choice items should also show that all distractors function—that there is a positive item-to-total correlation for correct answers and a zero or negative response-to-total correlation for incorrect answers. For items that are scored with a rubric, rating scale, or Likert-type scale, data should demonstrate that all ratings or rubric levels are functional and that there is a positive correlation between item scores and total scores.

Sometimes survey and psychological test items are worded in a way that requires reverse scoring. For example, suppose responses to the items in Figure 5–3 were scored as follows: Strongly Agree = 4; Agree = 3; Disagree = 2; Strongly Disagree = 1. If the item in Figure 5–6 were included in the same test as the four items in Figure 5–3, the negative wording would suggest reverse

I hide my grades from others when I get an A on a test.	Strongly Agree	Agree	Disagree	Strongly Disagree
---------------------------------------------------------	----------------	-------	----------	-------------------

**Figure 5–6** An Example of a Negatively Worded Item that Requires Reverse Scoring

6. In achievement testing, the term “item difficulty” is appropriate; however, in survey research or psychological testing, “popularity” is a more appropriate label.

scoring (Strongly Agree = 1; Agree = 2; Disagree = 3; Strongly Disagree = 4). A lower-than-expected mean item score and a negative correlation between the item score and the total score for Ego Involvement would suggest reverse scoring was not implemented.

Table 5–2 presents some item-analysis data for 20 items from a mathematics test. Data for several items are questionable. Items 1, 2, 10, 12, 14, and 16 appear to be extremely difficult, with item means of less than one-fourth of the possible points. In addition, items 1, 10, 14, and 16 have very low correlations between item scores and total test scores. In fact, item score to total test score correlations for items 1 and 16 are near zero.

Inter-item correlations can also flag problematic item scores. If all items are expected to contribute to total scores, assessment developers might expect that inter-item correlations will be positive. If the assessment represents a heterogeneous set of items measuring a complex trait, these correlations may not be very strong. However, the item scores should at least be positively related if they are to contribute to the same total score. A negative correlation between one item and several other items on the assessment could suggest a mis-key, inappropriate score conversions, a faulty item (e.g., an item with confusing wording, multiple correct answers, a faulty rubric), or an item that is measuring something unrelated to the other items in the assessment. All of these factors can present threats to score consistency. An item that is measuring something unrelated to the other items in the assessment (e.g., construct-irrelevant content) threatens the validity of the meaning of total scores. Assessment users should examine technical reports to evaluate the degree to which item statistics support the validity argument of scoring consistency.

### **Item Fit**

If an assessment is developed using item response theory (IRT), item/task analysis data should provide information about whether items fit the selected IRT model. Poor fit may suggest that the item or task is unrelated to the construct or that the item or task is problematic. If items have poor fit, they threaten the validity of claims about the meaning of total scores and should be eliminated from the assessment.

Table 5-2

**Classical Test Theory and Item Response Theory Item-Analysis Statistics**

<b>Item</b>	<b>Item Type</b>	<b>Points Possible</b>	<b>Item Mean</b>	<b>Item-Test Correlation</b>	<b>Rasch Item Difficulty</b>	<b>Standard Error</b>	<b>INFIT</b>	<b>OUTFIT</b>
1	MC	1	0.11	0.02	-0.64	0.03	1.25	1.33
2	MC	1	0.23	0.35	1.00	0.04	0.97	1.03
3	SA	2	1.23	0.51	-1.04	0.02	0.88	0.88
4	MC	1	0.58	0.34	-0.73	0.03	0.98	0.96
5	ER	4	1.81	0.56	0.80	0.02	1.01	0.94
6	MC	1	0.39	0.34	0.10	0.03	1.00	1.01
7	MC	1	0.51	0.28	-0.39	0.03	1.08	1.10
8	SA	2	1.00	0.58	-0.42	0.02	0.84	0.79
9	MC	1	0.44	0.29	-0.08	0.03	1.02	1.03
10	MC	1	0.20	0.13	1.18	0.04	1.06	1.35
11	SA	2	0.75	0.50	0.06	0.02	1.08	1.14
12	SA	2	0.49	0.52	0.94	0.02	0.88	0.82
13	MC	1	0.78	0.25	-1.83	0.04	1.05	1.15
14	MC	1	0.19	0.11	1.29	0.04	1.11	1.50
15	SA	2	1.11	0.38	-0.53	0.02	1.18	1.30
16	MC	1	0.22	-0.03	0.74	0.03	1.24	1.45
17	SA	2	0.57	0.50	0.53	0.02	0.94	0.88
18	MC	1	0.25	0.33	1.07	0.04	0.97	1.03
19	MC	1	0.29	0.18	0.68	0.03	1.07	1.19
20	SA	2	0.76	0.45	0.27	0.02	1.03	1.00

MC = Multiple-Choice; SA = Short Answer

A look at the item data in Table 5–2 shows two IRT item-fit statistics<sup>7</sup>: INFIT is sensitive to odd patterns of responses for examinees whose location on the underlying scale is near an item's IRT location. OUTFIT is sensitive to odd patterns of responses from examinees whose location on the underlying scale is far from the item's IRT location. OUTFIT statistics between 0.70 and 1.30 are considered fairly good fit. Note that the OUTFIT statistics for items 1, 10, 14, and 16 are greater than 1.30. These data, combined with the classical item analysis means and item-to-test correlations, suggest that items 1, 10, 14, and 16 are very problematic and are likely to detract from the validity of the total scores.

### Rater Agreement

A final source of evidence for score consistency should come from the rater agreement data. Assessment developers must document the consistency with which raters apply scoring rules (rubrics and checklists) to examinee performances. For example, Table 5–3 shows rater agreement data from a mathematics achievement test. The table shows counts for exact agreement, counts for different levels of discrepancy, and percent of exact agreement. These data suggest that the scorers were in strong agreement with each other as they applied the scoring rubrics—with exact agreement ranging from 79.4–98.5 percent.

Raters may agree with each other but, as a group, lose touch with the intent of a rubric. This is called *rater drift*. To check for rater drift, one can have scorers apply rubrics to pre-scored responses (also called criterion scores). Table 5–4 presents data showing how well raters agreed with the criterion scores for responses to the mathematics items referenced in Table 5–3. These data add to the support for the consistency with which raters applied scoring rubrics, with exact agreement ranging from 78.1–96.7 percent.

If rater disagreement is biased in some way, discrepant scores could result in different total scores for examinees, depending on the rater—even when the inter-rater agreement data appear to be strong. Table 5–5 shows the correlation between total scores resulting from different raters for the 15 mathematics items referenced in Tables 5–3 and 5–4. These data suggest that rater discrepancies

---

7. These are fit statistics from WINSTEPS, a 1-parameter item response theory analysis program.

Table 5–3

**Inter-Rater Agreement Data for a Mathematics Achievement Test**

<b>Item</b>	<b>Points Possible</b>	<b>Exact Score Match</b>	<b>Adjacent Scores</b>	<b>Discrepant by Two Points</b>	<b>Discrepant by Three Points</b>	<b>Discrepant by Four Points</b>	<b>Percent Exact Agreement</b>
1	2	3789	231	24	0	0	93.7%
2	2	3811	216	17	0	0	94.2%
3	4	3209	676	132	25	2	79.4%
4	2	3840	197	7	0	0	95.0%
5	2	3917	123	4	0	0	96.9%
6	4	3448	556	37	3	0	85.3%
7	2	3985	56	3	0	0	98.5%
8	4	3574	452	17	1	0	88.4%
9	2	3867	161	16	0	0	95.6%
10	2	3948	94	2	0	0	97.6%
11	2	3919	119	6	0	0	96.9%
12	4	3689	332	20	1	2	91.2%
13	2	3857	184	3	0	0	95.4%
14	2	3718	320	6	0	0	91.9%
15	2	3926	99	19	0	0	97.1%

Table 5–4

**Rater Criterion Score Agreement Data for a Mathematics Achievement Test**

<b>Item</b>	<b>Points Possible</b>	<b>Exact Score Match</b>	<b>Adjacent Scores</b>	<b>Discrepant by Two Points</b>	<b>Discrepant by Three Points</b>	<b>Discrepant by Four Points</b>	<b>Percent Exact Agreement</b>
1	2	34	2	0	0	0	93.9%
2	2	33	4	0	0	0	89.6%
3	4	29	6	2	0	0	78.1%
4	2	34	2	0	0	0	93.6%
5	2	34	3	0	0	0	92.1%
6	4	30	6	0	0	0	82.6%
7	2	35	1	0	0	0	96.7%
8	4	30	3	2	1	0	82.9%
9	2	33	3	0	0	0	90.6%
10	2	35	2	0	0	0	95.2%
11	2	33	4	0	0	0	90.2%
12	4	34	2	1	0	0	92.6%
13	2	35	1	0	0	0	96.4%
14	2	32	4	0	0	0	87.6%
15	2	35	1	0	0	0	96.2%



Table 5–5 Correlation Between Mathematics Total Scores from First and Second Scorers				
Correlation	Mean First Reading	Standard Deviation First Reading	Mean Second Reading	Standard Deviation Second Reading
0.99	17.22	6.21	17.23	6.19

are random and are unlikely to impact examinees’ total scores. Not only is the correlation between total scores from first and second raters extremely high ( $r = 0.99$ ), but the means and standard deviations of total scores for first and second raters are nearly identical. Tables 5–3 through 5–5 present the type of data that support an argument for scoring consistency. Technical manuals for assessments should provide evidence that scoring rubrics can be applied consistently across different scorers.

Summary of Validity Claim 1

Validity Claim 1 is that scores from the assessment can be used to make intended inferences about examinees. Two arguments support this claim: (1) scoring rules are appropriate, and (2) scoring rules are applied consistently. Historically, validation studies have focused on external evidence for the validity of assessment scores (e.g., correlations among measures of the same construct). Novices may be surprised that close attention is paid to item-analysis data in the validation process. However, if item scores are faulty, either because of scoring errors or scoring inconsistencies, total scores are faulty. Faulty total scores result in faulty inferences from scores, which invalidates the first claim in Table 5–1 (“Scores can be used to make inferences”).

Assessment developers have an obligation to provide technical reports that describe the strategies used to ensure the appropriateness of scoring models, the procedures used to ensure the accuracy of scores, and the consistency with which scoring processes are implemented. Item-analysis data and rater-agreement data should be available to assessment users so that they can determine whether scores can be trusted to make inferences. Scoring

procedures and score models should be explicitly described along with a rationale for the selected score model. Together, these sources of evidence provide a foundation of support for the validity of inferences from assessment scores.

Assessment developers should be able to provide a plethora of information to support the two arguments for Claim 1. Some of the evidence is mundane and need only be described (e.g., the accuracy of scoring keys and scanners); some evidence is central to the assessment development efforts and to claims about score meaning (e.g., statistical item-analyses data; how item data are used to evaluate the quality of items, scoring rubrics, rating scales, and distractors for multiple-choice items and to select items for the final form of the assessment; rationales for scaling and scoring models; evidence that scoring rules can be applied consistently across scorers). The strength of this claim depends a great deal on the decisions made during assessment development stages; however, assessment developers should continue to evaluate the strength of these claims when assessments are used for their intended purposes. Most of this evidence should be provided in technical reports or be readily available to users if requested, along with narratives that explain the data and rationales. Documentation of procedures and processes should also be available in technical reports or journal articles (e.g., descriptions of how item/task data were evaluated, and the qualifications of the reviewers; description of the methods used to verify unidimensionality in test scores).

Assessment users should review the available evidence and make sure that sufficient evidence is provided before selecting assessment tools for their own purposes. In addition, assessment users should make certain that they understand the scoring model used for the assessment. For example, interpretation of derived scores is a common problem among assessment users (e.g.: Educators and parents often misinterpret derived scores such as percentile ranks and grade equivalent scores; percentile ranks may be confused with percent correct scores; students may be placed in a higher or lower grade in school, based on inappropriate interpretations of grade-equivalent scores). Researchers often use parametric statistics (which are designed for interval data) on data from ordinal scales (e.g., rubric scores). Technical reports should explain and justify the score model used and the limits of the chosen model.

## **Validity Claim 2—It Is Possible to Generalize from Scores to a Universe of Behaviors or Responses Related to the Construct**

Validation studies related to Validity Claim 2 relate to the internal structure of an assessment. Two arguments are relevant to Validity Claim 2: items and tasks represent the universe of behaviors or responses defined in the specifications, and the sample of items/tasks is large enough to minimize errors. Typical validation studies related to Claim 2 involve content reviews; however, Kane (2006) also addressed the degree to which the number of items and/or tasks is sufficient to provide a reliable score as an aspect of the generalization claim.

### Representativeness of Items and Tasks

*We can look at the content of the test in relation to the content of the domain of reference.* (Messick, 1989, p. 16)

The first argument to support Validity Claim 2 has to do with the alignment of the items/tasks on the assessment to a defined universe of behaviors or responses. When assessment developers design an assessment tool, they begin by defining the construct or criterion performance and then write item and test specifications to describe how the construct will be assessed.

Kane (2006) distinguishes between the construct itself and the universe of items and tasks that could be developed for an assessment. This is an important distinction. He notes that it is impossible to accurately represent the entire domain of a construct or criterion performance. Domains may be too complex to assess in a systematic way. Standardization of the measurement will limit some of the tasks that can be presented to examinees; yet standardization enhances the reliability of measurement. Therefore, the assessment developer must carefully balance the breadth and depth of measurement with the need for reliable scores.

For example, the construct “Ability to play basketball” is quite complex. It is composed of knowledge of the rules of the game, use of offensive and defensive strategies, communication among players, application of skills (e.g., dribbling, free throws, layups, running, passing balls), and players’ physical characteristics (e.g., agility, stamina, speed, hand–eye coordination). The best way to

assess the construct is to observe a game of basketball. However, basketball games introduce variables that are difficult to control—making it difficult to ensure assessment of all dimensions of the construct for all players. A standardized basketball test designed to measure each player’s capacity related to all the targeted knowledge and skills would, of necessity, limit the complexity of assessment and, thereby, generalizability to the domain. On the other hand, if the basketball test were based on observations of the game, the complexity of the game would result in uneven assessment of all the players, limiting generalizability to the domain for each player.

This example illustrates the challenges in creating any assessment. The construct or criterion performance may be well defined; however, assessment may require proxies for the behaviors and responses characteristic of the construct or criterion performance. A standardized, end-of-year mathematics test composed of 40 or 50 items cannot represent all the knowledge and skills a student has learned in mathematics class during a single year. In addition, individual mathematics items generally ask students to demonstrate an isolated skill or concept rather than the real work of mathematicians or the work of individuals who use mathematics in their daily work.

The universe of behaviors or responses defined in test and item or task specifications narrows the focus of the test and allows for control over what is demonstrated in any given item. Therefore, the specifications are a critical aspect of the representativeness of items and tasks. Beyond the specifications, test developers consider whether each item or task is a representative of the universe of items defined through the specifications.

### **Evaluations of the Definitions of Constructs and Criterion Performances**

Before developing an assessment, developers should provide a clear definition of the construct or criterion performance and submit that definition to expert review. This explanation should be grounded in theory and/or research. Expert-reviews of construct definitions or definitions of criterion performances (e.g., job analyses) for their fit with current thinking about the knowledge, skills, abilities, behaviors, and dispositions in each domain are part of the validity argument regarding the representativeness of items and tasks.

Definitions of criterion performances are often based on research rather than theory. For example, to predict job performance, assessment developers must do a job analysis to identify all of the knowledge, skills, and abilities demonstrated by individuals who are minimally proficient. Qualified judges should review the job analysis and verify that the listed knowledge, skills, and abilities are indeed required for the job. Judges might be personnel directors, supervisors, employees who are successful on the job, and those who prepare individuals for the job.

Constructs for psychological assessments are grounded in both theory and research. For example, research on achievement motivation suggests two primary ways in which students engage with schoolwork—ego involvement (Nicholls, 1989; Nicholls, Cobb, Yackel, Wood, & Wheatley, 1990; Nicholls, Patashnick, & Nolen, 1985; Nolen, 1988; Thorkildsen & Nicholls, 1998) and task involvement (Nolen, 1988, 1995). Students who are task-involved view school as an opportunity for learning; therefore, their goals are learning goals. In contrast, students who are ego-involved view school as a competitive environment in which their abilities are compared with the abilities of others. Their goals are to perform better than others or to avoid demonstrating lower ability than others (Elliot & Harackiewicz, 1996). To design an assessment, a developer would have to review theoretical arguments on achievement motivation, examine the past and current research, identify behaviors that demonstrate motivational orientations, and determine what types of stimuli would elicit these behaviors. Expert reviews of the construct definitions for psychological assessments often occur through peer-reviewed journal articles.

In educational testing, the domain of content is generally determined by examination of the content in textbooks or in state and/or national curriculum standards. However, these may be very limiting sources. To what extent is the content of a textbook validated in terms of the disciplines represented by those textbooks? To what extent do state or national curriculum standards represent current research on reading, writing, and mathematics? Recently, educational researchers have begun to consider the cognitive processes as well as the content within educational domains. For example, for most of the nineteenth and twentieth centuries, geographical literacy was limited to knowledge of locations (e.g., states, countries, and their capitols) and the terms for land forms

and bodies of water; knowledge of history was defined as a body of historical knowledge (Beadie, 1999). In the late twentieth century, educators began to consider dimensions of geographical and historical thinking as they relate to the work of geographers and historians (National Council for the Social Studies, 2010). Notions of geographical literacy now include understanding of place and location, regions, human movement, human–environment interactions, and human culture. Notions of historical literacy now include understanding of time and chronology, continuity and change, cause and effect, and perspective. Wineburg (2001) has generated a model for “reading like a historian.” This model is derived from his research on how historians approach primary and secondary documents to ascertain their validity. Based on these ideas, a twenty-first-century social studies test might assess students’ ability to use primary and secondary documents to develop interpretations about the causes and effects of historical events, their understanding of chronology, their ability to critically read primary documents, and their ability to evaluate the causes and effects of human movement. Content experts, such as educational researchers, teachers, and subject-matter experts, should review construct definitions for achievement tests for the fidelity of the construct definitions with what is known about the targeted domains. In educational test development, expert judgment about the alignment of the test with the construct(s) is usually done as part of the test development process.

### **Evaluation of Test and Item/Task Specifications**

The assessment developer should demonstrate that the specifications for the assessment represent the knowledge, skills, abilities, behaviors, and/or dispositions that are indicative of the construct or that are required to successfully complete the criterion performance. With a clear definition of the construct or criterion performance, validity researchers examine test specifications to evaluate whether they clearly describe how the construct or criterion performance will be elicited and demonstrated. They judge the fidelity of the specifications to the construct or criterion performance definition—to see whether the proposed content of the test represents the breadth and depth of the construct or criterion performance. If there is a poor match, or if the slice of the domain is too narrow, the assessment developer will lack sufficient

evidence to support the claim that scores from the assessment can be generalized to the construct or criterion performance.

Item and task specifications provide detailed information about how items and tasks will tap into the targeted knowledge, skills, abilities, behaviors, or disposition. Details might include rules for stimulus materials, frames for item stems, response formats, allowable terms and vocabulary, appropriate rating scales for survey items, rules for wrong answer choices in multiple-choice items, etc. Test specifications indicate the number of items or tasks for each component of the assessment, features of the final assessment (e.g., font sizes, number of items on a page, formatting of electronic presentation), and any requirements for the test as a whole (e.g., total number of reading passages, total number of graphic elements). Figure 5–7 presents an abbreviated set of test and item specifications for a fictitious academic self-concept scale. A content review of the specifications would involve expert review of the degree to which the specifications represent the defined construct.

Experts who review test specifications are assessing whether or not the specifications for the test will result in an appropriate representation of the defined construct or criterion performance. Reviews of item specifications provide evidence of whether or not judges believe the items are designed to tap into the knowledge, skills, abilities, behaviors, and/or dispositions relevant to the construct definition or criterion performance. When documenting specification reviews for validity evidence, assessment developers should describe the procedures used in the review as well as the qualifications of the judges.

### **Item/Task Content Reviews**

Item/task content reviews are a routine part of the test development process. Experts review items and tasks to make a judgment about whether the items or tasks will elicit the knowledge, skills, abilities, behaviors, or dispositions that are the target of the assessment. Sometimes, the review is guided by item/task specifications. Judges evaluate whether the items/tasks actually align with requirements in the specifications. In other studies, judges evaluate the fit between items and the construct or criterion performance without the use of item/task specifications. Item/task reviews also require evaluation of the scoring rules to ensure that

**Construct Definition<sup>1</sup>**

The construct is called "academic self-concept". Academic self-concept is defined as the faith an individual has in her or his ability to learn and succeed in traditional and formal schooling. There are two dimensions in academic self-concept: "scientific literacy self-concept" and "cultural literacy self-concept". Scientific self-concept is the sense of ability to learn in mathematical and scientific areas. These subjects are generally taught in highly abstract ways, so the student's scientific self-concept defines her/his belief that she can learn easily in highly technical and abstract subjects. Cultural literacy self-concept is the sense of ability to learn in sociological and literary areas. These subjects are generally taught in terms of how humans are involved in social and political relationships, so the student's cultural literacy self-concept defines his/her belief that s/he can learn easily in subjects of the human domain.

**Test Specifications for the Academic Self-concept (ASC) Test**

**Construct:** The items will measure cultural literacy academic self-concept and scientific literacy academic self-concept. Given a situation involving a scientific or cultural literacy activity, the examinee will indicate preference for or against the situation.

**Purpose:** There are two purposes for this test. First, the test will determine whether students have faith in their own ability to learn in different subject areas. Teachers can work to strengthen students' academic self-concept based on the results. Second, based on examinees' profiles, teachers can individualize instructional activities to build on students' strengths. If students have low scientific literacy self-concept but high cultural literacy self-concept, teachers can connect scientific and mathematical ideas and skills to real world social and political situations. Alternately, if students have low cultural literacy academic self-concept and high scientific literacy academic self-concept, teachers can help students connect literature and social science issues to science and mathematics. If students are low in both areas, teachers must work to strengthen students' academic self-concept in general, helping them have successful learning experiences in both areas.

The test will be composed of 40 Likert-type items.

**Test Structure:** Each item will stand alone. Responses to each item will be independent of responses to all other items. Items will NOT be grouped by the content of activities. The content focus of activities will be randomly distributed throughout the test. Approximately 25% of items will be phrased in terms of negative affect. No activity will be repeated in the test as a whole.

<sup>1</sup>. This is entirely invented.

Test Map for the Academic Self-concept (ASC) Test		
Dimension	Types of Activities	Number of Items
Scientific Literacy Academic Self-concept	Mathematical Activities	10
	Scientific Activities	10
Cultural Literacy Academic Self-concept	Social Science Activities	10
	Literacy Activities	10

Figure 5–7 Example Test Specifications for Test of Academic Self-concept



Item Specifications for Academic Self-concept (ASC) Test				
<p><u>Item Format:</u> All items will be statements of preferences followed by Likert-type response choices. Likert-responses will be Strongly Agree (SD), Agree (A), Disagree (D), and Strongly Disagree (SD).</p>				
<p><u>Grammatical Form of Items:</u> All items will be 'I statements' in subject / affective verb / circumstance order. Circumstances will be participation in activities or events. No subordinate clauses will be used.</p>				
<p><u>Length of Item Stems:</u> Statements will be brief and fit on a single line (when presented along with the answer choices) with 11 point type.</p>				
<p><u>Font:</u> Items will be sans serif font.</p>				
<p><u>Item Content:</u> All items will describe scientific, mathematical, literary, or social science activities that are familiar to middle and high school children. Contexts will include both in school and out of school events. Affective verbs can be: like, enjoy, dislike, feel anxious when, avoid, or similar terms.</p>				

Example Items					
1.	I feel anxious when I have to do a large number of math computations.	SD	D	A	SA
2.	I enjoy reading biographies of people in the past.	SD	D	A	SA
3.	I like going to science museums.	SD	D	A	SA
4.	I feel annoyed when I have to write about mathematical ideas in class.	SD	D	A	SA
5.	I get bored when people debate political issues.	SD	D	A	SD
6.	I avoid doing my science homework.	SD	D	A	SD

Figure 5–7 (Continued)

they are (1) tied to the construct or criterion performance and (2) aligned with the expectations of the items/tasks.

Crocker and Algina (1986) describe methods for conducting content reviews and issues that the assessment developer must consider before, during, and after content reviews (e.g.: Do judges make yes/no decisions or rate items in terms of their alignment? What aspects of the items or tasks should be reviewed? How should reviews be captured and summarized? What information should be provided to judges before the review?).

Assessment developers often revise items and tasks immediately after reviews to take advantage of the judges’ expertise. Documentation of expert reviews of test specifications, item/task specifications, and the items or tasks should be part of the

validity documentation provided for any assessment. Assessment developers can describe the procedures used in the review, the qualifications of the judges, statistics regarding expert agreement about the alignment between specifications and the construct or criterion performance, and the statistics on the percent of items/tasks judged to be aligned.

### Minimizing Errors in Assessment Scores

The second argument for Validity Claim 2 is that the sample of items or tasks in the assessment is sufficient to minimize errors in assessment scores. The beginning of this chapter described the relative simplicity of creating a linear measurement tool when compared with more complex constructs (e.g., reading comprehension, depression, achievement motivation). If a person measures his height once a week, there will be slight differences from one measurement to the next; this is measurement error. Similarly, if an examinee completes a reading comprehension test at two different times, there will be differences in the earned score. If nothing else has changed (e.g., if reading skills have not improved), the score differences between time 1 and time 2 represent measurement error. Unfortunately, it is not possible to know which score is accurate. One of the fundamental assumptions of any assessment is that there is some measurement error.

When measuring height, it is possible to measure a great many times. The median or mean measure is likely to be the height of the object. In the case of psychological, educational, or employment assessments, it is not possible to assess someone a tremendous number of times. Examinees are affected by the assessment process. They might remember items and respond from recall rather than from their current state. They might learn from the assessment process. They may become fatigued from all of the testing and respond haphazardly. In short, their actual skill level, psychological state, or conceptual understanding may change from one time to the next for reasons unrelated to the measured construct. Therefore, it is not possible to be absolutely certain we have *true scores* for examinees on psychological and educational assessments.

The focus of this argument for Claim 2 is that the assessment is sufficient to minimize measurement error and, therefore, to

strengthen our confidence that scores represent “true scores.” How do we know whether we have sufficient measurement? Generally, evaluation of error is through estimates of reliability and measurement error.

### Reliability Estimates

A *reliability coefficient* is an estimate of the likelihood that an examinee’s score from an assessment is her or his true score on that assessment. In test theory, this estimate is an extrapolation from groups to individuals. Since we cannot effectively assess an individual hundreds of times, we use the performance of large samples to estimate reliability for individuals. Using group data, the reliability coefficient provides an estimate of the proportion of observed score variance that is true score variance.

The reliability coefficient allows us to estimate the likelihood that a score from a given assessment is the examinee’s true score *on the given assessment* or on an assessment that is strictly parallel to it. The reliability coefficient does not tell us whether the earned score, regardless of how reliable, is *the* true score on the construct or criterion performance. This is an important distinction. Just because the reliability coefficient for a test is high does not mean we have a true measure of a person in relation to the construct or criterion performance. In the basketball example described above, a teacher may have a reliable assessment of students’ ability to dribble, pass a ball, and do free throws, and of their knowledge of basketball rules; however, her test is unlikely to tell her how well students can play basketball.

Another way to conceptualize reliability is in terms of replicability. Chapter 3 introduced the idea of *replication* as a way to determine whether the results of an investigation can be generalized across times, settings, and samples. Similarly, assessors want to know whether examinees’ test scores are replicable. The reliability coefficient provides an estimate of the likelihood that examinees will obtain the same score upon repeated testing (assuming no change in the examinee). This is an essential validity issue. If scores are not repeatable, they cannot be trusted.

Assessment developers routinely conduct studies to evaluate the reliability of scores. The most commonly reported estimate of reliability is the alpha coefficient or, for dichotomously scored items, the KR20. This is a measure of the internal consistency of

examinees' responses to items. If an assessment measures a complex construct or criterion performance composed of a heterogeneous collection of items and tasks, internal consistency is likely to yield an underestimate of reliability.

Other reliability estimation processes involve retesting with a parallel test (alternate forms reliability estimates), retesting with the same test at a different time (test-retest reliability estimate), or splitting the test in half and, based on the correlation between scores on each half of the test, estimating reliability for the total test score.

More items or tasks generally result in fewer errors. Tests composed of items or tasks with objectively scorable responses (e.g., multiple-choice items, rating scales) can be administered efficiently and cost-effectively. Therefore, it is possible to sample a broader array of the domain in a relatively short period of time. However, it takes many objectively scorable items to obtain reliable total scores. Performance tasks may be more like criterion performances or provide better demonstrations of the targeted construct; however, they are expensive to score and time-consuming to complete. Therefore, far fewer performance tasks are possible in a relatively short period of time.

### **Estimating Error**

The ultimate value of a reliability coefficient is that it allows test users to estimate the amount of error in scores. Since the reliability coefficient is an estimate of the proportion of observed score variance that is true score variance, it is possible to use the reliability coefficient to estimate the proportion of observed score variance that is error variance. This estimate allows a test user to obtain an estimate of the standard error of measurement. The standard error of measurement can be used to estimate the degree of confidence we have that an examinee's score will fall within a given range of scores.

Problematic items and tasks can detract from the reliability of a test score and increase measurement error. Other possible sources of measurement error have already been discussed. For example, if inter-rater agreement is low, measurement error will increase. If items are negatively correlated with total scores or with scores from other items on the assessment, reliability estimates will decrease.

### **Generalizability**

Generalizability studies (G studies) are another strategy used to identify potential sources of error in scores. Generalizability

analysis is founded on the principles of analysis of variance using aspects of testing as the factors in the analysis. G studies can provide estimates of variance due to raters, tasks, methods, and other facets of an assessment. During the assessment development process, developers can use generalizability studies to determine the optimum number of tasks, raters, and methods needed to minimize error, given the amount of testing time available.

Various factors can lessen the generalizability of scores. If items and tasks are tied to stimulus materials, it may be difficult to generalize examinee performance to other stimuli. If tasks are complex, it may be difficult to generalize from one task to the next. For example, Shavelson, Baxter, and Gao (1993) found that a large number of performance tasks are necessary to obtain reliable estimates of examinees' abilities in science and mathematics. They also found that performances did not generalize between different methods (i.e., hands-on science investigations versus paper-and-pencil science tasks).

### Decision Consistency

One issue that is often overlooked in discussions of reliability is that of decision consistency. When assessments are used to make criterion-referenced decisions (e.g., placement decisions, selection for special programs, identification of patients who are at risk, selection for jobs, and determining whether examinees have met proficiency standards), *cut-scores* are established. Cut-scores are points on the total score scale where decisions are made. For example, scores above a cut-score may suggest a "proficient" reading performance whereas scores below the cut-score suggest that the student's reading performance is at a "basic" level.

Criterion-referenced decisions are often high-stakes decisions, and assessment users want to have confidence that decisions made based on the cut-scores are reliable. For example, if an assessment will be used to identify patients who are likely to attempt suicide, the therapist wants some assurance that the assessment will identify the same patients regardless of when the test is administered. Figure 5–8 provides an image of the type of reliability that is at issue.

The cells in the chart represent percentages. Cell A in Figure 5–8 is the percent of examinees who are identified as "at risk" on the first and second administration of the test; Cell B is the percent of examinees who are identified as "at-risk" at Testing Time 1 and

		Testing Time 2	
Testing Time 1		At Risk (At or Above Cut Score)	Not at Risk
	At Risk (At or Above Cut Score)	Cell A	Cell B
	Not At Risk (Below Cut Score)	Cell C	Cell D

Figure 5–8 Decision-Consistency Reliability Hit Rate Table

“not at risk” at Testing Time 2, and so forth. The sum of percents in Cell A and Cell D indicates the percentage of time that the decisions are consistent across testing events. The percents in Cells B and C indicate the percentage of times that decisions are inconsistent across testing events.

The consistency of decisions depends greatly on the degree of error at the cut-score.<sup>8</sup> If the standard error of measurement is very small at the cut-score, the assessment developer can be more confident in the consistency of the decisions based on scores above and below the cut-score. Gathering evidence of decision consistency generally involves administering an assessment twice and determining the sum of the percent of examinees in Cells A and D from Figure 5–8. If it is not possible to administer the test twice, several researchers have proposed methods to estimate decision consistency using classical test theory and item response theory methods (e.g., Hanson & Brennan, 1990; Kolen, Zeng, & Hanson, 1996; Livingston & Lewis, 1993; Subkoviak, 1976; Wang, Kolen, & Harris, 1996).

When assessments are developed in order to make placement, selection, proficiency, or risk-based decisions, assessment developers should provide information about decision consistency reliability for published cut-scores. Assessment users should scrutinize this reliability data to determine whether the published cut-scores will yield reliable decisions about examinees. If assessment users plan to set their own cut-scores for assessments, they

8. In classical test theory, the standard error of measurement is assumed to be the same for all assessment scores. Item response theory methodologies provide standard errors of measurement at each assessment score. Generally, the greatest amount of error is found for high and low scores. Therefore, cut scores closer to the middle of the score distribution are likely to be the most reliable.

should conduct studies that support the use of the cut-scores and to demonstrate that error in classification has been minimized.

In summary, assessment developers must provide evidence to support the reliability and generalizability of all summary scores and, in doing so, consider potential sources of error. Assessment users should consider whether the standard error of measurement is small enough for them to have confidence in the observed scores for examinees. Users should also examine these data to determine whether the evidence warrants use of the assessment for the intended purpose.

### Summary of Validity Claim 2

Validity Claim 2 is that it is possible to generalize from scores to a universe of behaviors or responses related to the construct. Assessment developers have an obligation to provide evidence related to two major arguments to support this claim: (1) evidence that the composition of the assessment is representative of the universe of behaviors or responses defined for the construct or criterion performance, and (2) evidence that the scores from the assessment are generalizable. Evidence for the first argument should include expert reviews of construct definitions or expert reviews of the definitions of criterion performances, the alignment between test specifications and definitions for constructs and criterion performances, and the alignment of items to the defined domain. Evidence for the second argument should include evidence for the reliability of scores—including measures of consistency, inter-judge agreement, generalizability, and error. Assessment users should be able to find documentation of these sources of evidence in readily available forms, along with explanations of the data.

### **Validity Claim 3: It Is Possible to Extrapolate from the Score to the Domain of the Construct or Criterion Performance**

*We can probe the ways in which individuals respond to items or tasks. . . . We can survey relationships of the test scores with other measures and background variables, that is, the test's external structure. We can investigate differences in these test processes and structures over time, across groups and*

*settings, and in response to experimental interventions—such as instructional or therapeutic treatment and manipulation of content, task requirements, or motivational conditions.* (Messick, 1989, p. 16)

Most published validation research is focused on the third claim defined in Table 5–1. In common parlance, the validity question is, “Can the scores be used to infer an examinees’ status in relation to the intended construct or criterion performance?” Put differently, the question is also whether assessment scores may be caused by something other than what is intended. In a way, this is the most central construct validity issue. Assuming that scores are trustworthy and measurement error is minimized, score inferences and interpretations depend upon the degree to which the assessment taps the same knowledge, skills, abilities, behaviors, and dispositions as the construct or criterion performance.

### Alignment with the Construct or Criterion Performance

The first argument required to support Claim 3 is that the knowledge, skills, abilities, behaviors, and/or dispositions required when responding to the items and tasks in the assessment are the same as those in the domain of the construct or criterion performance. Strategies for substantiating this argument are quite varied and range from observations of examinees to complex empirical studies. In this section, I describe six strategies: observation of examinees, correlational studies, profile analysis, experimental or quasi-experimental studies, and factor analysis. They are generally considered studies designed to obtain *convergent evidence* in support of validity claims.

#### Observation of Examinees

Expert reviews of items and tasks are necessary but not sufficient sources of evidence for whether items and tasks elicit the knowledge, skills, abilities, or dispositions of the domain of the construct or criterion performance. Evidence must show that examinees actually *use* the targeted knowledge, skills, abilities and behaviors, or that examinees truly have the dispositions that are the focus of the assessment. A set of items and/or tasks may appear to map to a construct or criterion performance, yet, without empirical evidence, one cannot know whether the items or tasks truly



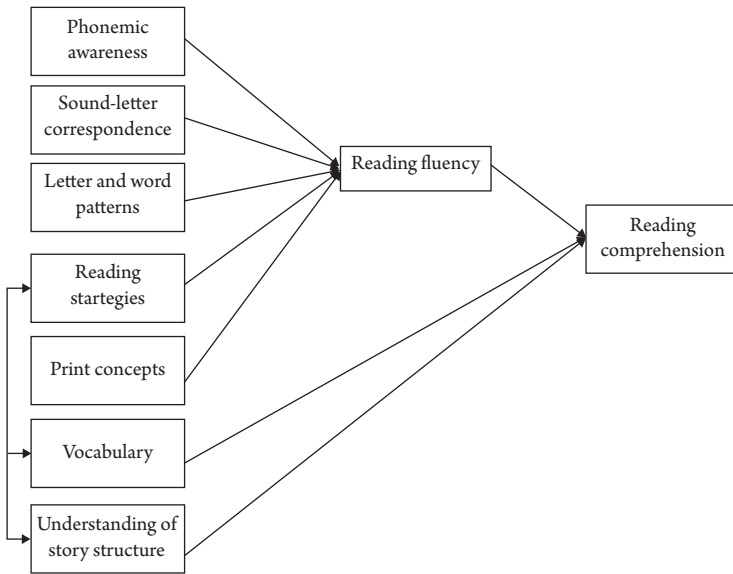
elicit what is described in the construct definition or criterion performance description.

Direct observations of examinees during an assessment event can provide evidence about what actually occurs for examinees. Think-aloud studies are a popular strategy for eliciting examinees' cognitive or psychological processes during an assessment. The researcher asks examinees to describe their thinking and reasoning processes while they complete each item or task. The researcher documents examinees' comments and uses qualitative strategies to analyze and describe examinees' internal processes. For example, Kazemi (2002) asked fourth-grade students to think aloud while completing multiple-choice and open-ended items measuring the same mathematical content. She found that students focused on the answer choices for multiple-choice items and made their choices based on (often incorrect) generalizations rather than thinking through the problem first. She found that students' thinking was focused on the problems themselves when responding to open-ended items asking for the same mathematical processes. Studies such as this think-aloud study may call into question the claims made about the meaning of scores from a test.

### **Correlational Studies**

Another typical construct validation study involves correlations between two assessments of the same construct. If the two assessments are measuring the same construct, the correlations are expected to be strong. Suppose "Dr. Rodriguez" wanted to investigate the validity of scores for a new assessment of post-traumatic stress disorder (PTSD). He asks 200 clients to complete the new assessment, and he has therapists judge the degree of PTSD for these 200 clients. He obtains a correlation of  $r = 0.65$  between scores on his new assessment and judgments by therapists. After correcting for the unreliability of each measure, he obtains a correlation between scores from the two assessments of  $r = 0.74$ . This gives Dr. Rodriguez some confidence that the two assessments are measuring the same construct.

Theories and research not only help define the domain of the construct and the mental processes examinees are expected to apply, but they also help frame a wide range of other validity studies. Figure 5–9 presents the theoretical network of



**Figure 5–9** A Nomological Network to Represent a Theory of Reading Comprehension

relationships among variables in a reading theory shown in Chapter 1. Suppose a new test of reading fluency is developed. The assessment developers would probably obtain correlations between scores on the new test and an existing test of reading fluency. However, given the theoretical model shown in Figure 5–9, the assessment developers would also conduct studies to examine correlations between the scores from the new test of reading fluency and the scores from the reading-comprehension assessment.

Correlational studies are also useful in construct validation when evaluating whether scores from an assessment are adequate predictors of criterion performances. However, prediction studies can present several challenges. First, assessment of the actual criterion performance can suffer from any of the threats to validity that arise in assessment. For example, suppose the measure of the criterion job performance is a performance evaluation from a supervisor. Assessment of job performance may be unreliable due to unreliability of, or bias in, supervisor judgments. Supervisors may focus on only part of a job performance rather than all of the knowledge, skills, and abilities required for the job. Validation researchers must consider potential validation issues with criterion

performances and attempt to address them as they conduct their research.

Suppose that validation of the criterion performance has been addressed and a new assessment is to be used to select individuals for an entry-level position. “Dr. Jagatheesan” administers the new assessment and obtains supervisor judgments for current employees to compute the correlation between supervisor judgment and performance on the new assessment. She obtains a correlation of  $r = 0.50$  between supervisor judgments and assessment scores. This is a moderate correlation and suggests that supervisor ratings are not strongly related to performance on the new assessment. However, correlations between supervisor judgments and assessment performance may be depressed. Individuals on the job are usually those who have been successful; therefore, restriction in the range of score is likely to impact correlations. The researcher will not have access to job performance for unsuccessful employees or individuals who were not hired. Despite these limitations, correlations between job performance and predictor assessments are routinely used to support the use of assessment scores in selection of employees and in job placements. Assessment users should consider factors that might impact correlations when making decisions about the validity of score inferences. Assessment developers should provide explanations of the evidence in their technical reports to help users make sense of the correlational data in building the validity arguments.

### **Profile Analysis**

One way to strengthen the argument that scores from the assessment can be extrapolated to the domain of knowledge, skills, abilities, behaviors, and/or dispositions within the domain of the construct or criterion performance is to look at patterns of scores across measures of different constructs. For example, suppose a researcher is developing an assessment for post-traumatic stress disorder. The researcher might administer the PTSD assessment, along with assessments for depression, bipolar disorder, and obsessive-compulsive disorder, to a group of clients who already have diagnoses of these four disorders. A profile analysis would allow the researcher to determine whether the patterns of scores differ for the different groups and whether the patterns make sense in terms of the specific disorder. For example, if she can

extrapolate from the measure of PTSD to the domain, then, in the profile of scores, clients with a PTSD diagnosis would have elevated scores on the PTSD measure and lower scores on the other measures; clients with depression would have elevated scores on the depression measure and lower scores on the other measures; and so forth.

### Experimental and Quasi-Experimental Studies

Experimental and quasi-experimental research can also be used to investigate the argument that the same traits are required on the assessment as in the construct or criterion-performance domain. In experimental and quasi-experimental validation studies, rather than investigating whether a particular causal relationship exists, the focus is on whether scores from assessments *function as expected* based on particular interventions or factors.

“Dr. Cook” develops a test to assess middle-school students’ ability to solve non-routine mathematical problems. It is composed of ten open-ended tasks that require knowledge and skills in computation, data, measurement, geometry, and algebra. He plans to use the assessment to evaluate the effectiveness of different mathematics instructional programs. However, he needs to know whether his test will actually assesses students’ ability to solve non-routine problems. He designs an experimental study to investigate whether scores on the test will differ in response to instruction on problem-solving. He selects two schools with similar demographics. At the beginning of summer school, Dr. Cook administers a standardized mathematics achievement test. Throughout the summer term, he provides instruction to students in one school on how to solve non-routine problems. At the end of summer school, Dr. Cook administers his problem-solving test to students in both schools. He uses analysis of covariance to analyze the scores. Figure 5–10 shows the design of his study.

All other things being equal (e.g., scoring of responses is reliable; initial differences in mathematics achievement have been controlled), if scores from his test are higher for the students who receive problem-solving instruction, the results provide initial support for his claim that the test scores reflect students’ ability to solve problems.

### Factor Analysis

Factor analysis is another type of correlational study that can be used to provide support for the claim that scores are related to the

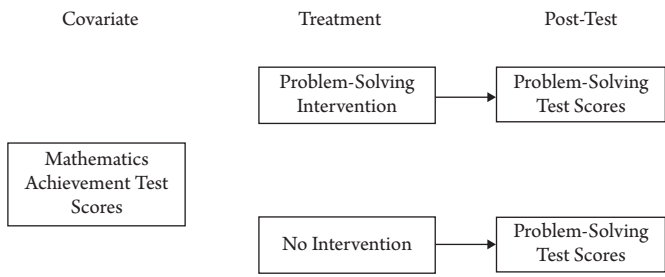


Figure 5–10 Design of Experimental Study for Construct Validation

targeted construct or criterion performance. Factor analysis is useful for testing whether a collection of items and/or tasks measures a single construct. Or, when an assessment is designed to measure multiple dimensions of a construct or criterion performance, factor analysis can provide evidence to support distinct scores for the expected dimensions. Generally, when the focus is on providing evidence to support scores, validity researchers use confirmatory factor analyses (CFA) and evaluate whether the expected factor patterns fit the data.

In summary, to evaluate the argument that the scores on the test reflect the same knowledge, skills, abilities, behaviors, and dispositions in the construct domain or criterion performance, users should expect to find a range of evidence to support this argument. Multiple lines of evidence should be available to support the alignment of the assessment scores with the construct or the criterion performance (Messick, 1989). Users should be able to examine construct definitions or criterion-performance descriptions, the results of think-aloud studies, correlational studies, and experimental/quasi-experimental studies in order to judge whether sufficient evidence supports the first argument of Validity Claim 3—that scores can be extrapolated to the domain of the construct or criterion performance.

### Construct-Irrelevant Variance

*[T]wo different kinds of evidence are needed in construct validation, one to assess the degree to which the construct’s implications are realized in empirical score relationships and the other to argue that these relationships are not attributable instead to distinct alternative constructs. (Messick, 1989, p. 34)*

The second argument for Validity Claim 3 is that no knowledge, skills, or abilities *irrelevant* to the domain of the construct or criterion performance are required to complete the assessment. Chapter 1 introduced the idea of alternate explanations as an aspect of the validation process. In assessment, alternate explanations for scores are sometimes referred to as *construct-irrelevant variance*. In other words, the variability in scores can, at least in part, be explained by factors unrelated to the construct or criterion performance. Construct-irrelevant variance is systematic error; therefore, it cannot be accounted for through the standard error of measurement, which is based on the assumption that error is random. If research supports the claim that construct-irrelevant variance is *not* a factor in assessment scores, this is often called *discriminant evidence*. Construct-irrelevant variance can come from a wide range of factors (e.g., method bias, response bias from examinees, assessment demands that require use of traits other than the one being targeted by the assessment tool). Validation research focused on construct-irrelevant variance is designed to unearth these factors.

The range of strategies for investigating construct-irrelevant variance is quite large. In this section, I will present five commonly used strategies: bias and sensitivity reviews, differential item functioning, factor analysis, multi-trait/multi-method correlations, and experimental manipulations.

### **Bias and Sensitivity Reviews**

Bias and sensitivity reviews are a routine part of achievement test development and should be a routine part of the development for all assessments. The purpose of bias and sensitivity reviews is to evaluate items and tasks, stimulus materials, and other features of assessments to determine whether there are features unrelated to the targeted construct or criterion performance that could have a negative effect on examinees. These features may not show up in a statistical analysis of item bias; however, they could have an overall effect on the performance of some examinees.

Bias and sensitivity review panels are composed of individuals who represent and are knowledgeable about the sensitivity issues for particular demographic groups. Bias and sensitivity review panels might represent different ethnic and cultural groups; individuals from different socio-economic strata; individuals who

represent different types of families (e.g., single parents, foster parents), different genders and sexual preferences, different age groups, and so on. For bias and sensitivity review to be useful, these individuals should be knowledgeable about issues that might negatively affect members of their demographic group.

Once reviewers are impaneled, they review stimulus materials, items, tasks, and tests for a wide range of possible issues. Some examples of review foci are: over- or under-representation of individuals from any group; negative or patronizing representations of individuals from any group; idiomatic phrases or vocabulary that might be differentially familiar based on cultural or socioeconomic background, region, or type of community; stereotypical representations; unfamiliar contexts for items/tasks; and presentation of content that has negative valence for any group. Assessment developers use the results of the reviews to revise or eliminate items, tasks, and stimulus materials before they are used in an assessment.

As with content reviews, there may be a real difference between examinee performance and what a reviewer perceives as features of items, tasks, or stimulus materials that could negatively impact examinees. For example, Engelhard, Hansche, and Rutledge (1990) found little agreement between bias and sensitivity reviews and statistical analyses in terms of which items were likely to be biased against African-American students. Hambleton and Jones (1992), asked Native American reviewers to identify items with specific features likely to result in differential item functioning (DIF) and found higher rates of agreement than were shown in the Engelhard et al. (1990) study.

Regardless of whether there is agreement between judgmental and statistical indices of bias, assessment developers should conduct these reviews, and document the qualifications of the reviewers, the procedures used, the results of the reviews, and how the data were used in assessment-development decisions, as part of the evidence to support their argument that no knowledge, skills, abilities, and/or dispositions irrelevant to the domain of the construct or criterion performance are required in response to items and tasks. Technical reports should present these data so that assessment users can judge whether adequate attention has been given to these potential sources of construct-irrelevant variance.

### Differential Item Functioning

*Differential item functioning* or DIF refers to a phenomenon wherein two examinees with the same knowledge, skill, ability, and/or disposition have different probabilities of responding correctly or favorably to items or tasks. In a perfect world, items and tasks on an assessment would create interval “rulers”—each item at its own location on the ruler. If an examinee responds correctly or positively to an item, the examinee would be expected to respond correctly or positively to all items that are lower on the scale. In educational, psychological, and employment testing, this is unlikely, if not impossible.

Different patterns of item responses are expected, even when examinees have the same total test score. However, these differences should be random. When differences are systematic, it is called DIF. DIF studies generally focus at the item or task level but can also focus on bundles of items or tasks (differential bundle functioning). DIF statistics are conditioned on examinees’ total test scores; therefore, to conduct DIF studies, one needs both item/task scores and total scores for all examinees.

DIF studies are a routine step in large-scale achievement test development. Most DIF studies involve comparisons of demographic groups when there is a concern about inequities in the assessment process. Males are compared with females (e.g., Taylor & Lee, 2012); whites are compared with non-white groups (e.g., Taylor & Lee, 2011). DIF studies have also been conducted that compare native English speakers with English-language learners (Smith, 2010; Ilich, 2010; Brown, 2010). Many researchers recommend that DIF studies be conducted comparing typically developing students with students who are served by special-education programs (e.g., Almond, Browder, Crawford, Ferrara, Haladyna, Huynh, Tindal, & Zigmond, 2005).

The purpose of demographic DIF studies is to look for items that might be biased against historically under-served groups. However, many theorists now consider DIF as evidence for multidimensionality in assessment (e.g., Ackerman, 1992, 1994, 1996; Shealy & Stout, 1993). Multidimensionality occurs when examinees are the same on a primary dimension but differ on a secondary dimension. For example, suppose a mathematics test requires both problem-solving skills and recall of memorized content. If examinees have the same problem-solving ability but different



memorization ability, their performance on items could differ due to the combined demands of items. Similarly, if one mathematics program focuses on problem-solving and another focuses on memorization of algorithms, item/task responses could show DIF based on the students' educational program.

In the typical case, DIF researchers identify two groups (e.g., males and females) and evaluate item performance for each group, controlling for overall ability or level of the latent trait. Suppose “Dr. Vokos” is interested in finding out whether the items on his anxiety scale demonstrate gender DIF. Dr. Vokos can compare item performance for male and female examinees at each total score level. Figure 5–11 shows item responses for males and females to one item on Dr. Vokos’ test (“I feel like hurting myself”). The curves represent the probability of a favorable response to the item based on the examinee’s underlying level of depression. As the level of depression increases, the item characteristic curves move to the right and rise. The curves show little change in probability at first, then a marked increase in probability, and then probability tapers off. The main difference between the two curves is their location on the x-axis. The location of the curve shows what level of depression is likely to yield a positive response to the item. According to the data, males and females with the same level of depression (depression total score) have different probabilities

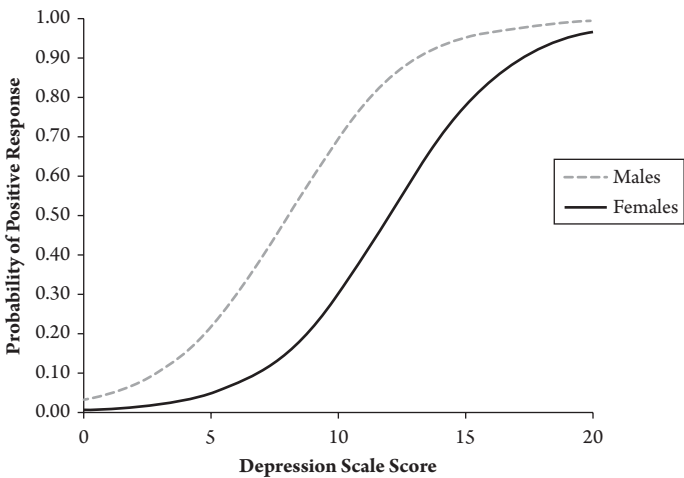


Figure 5–11 Item Characteristic Curves for Males and Females on an Item from a Depression Scale

of responding favorably to the item. Males will respond positively with lower levels of depression than females. Suppose that an unacknowledged secondary dimension of Dr. Vokos's scale is "tendency toward violent behavior in response to depression." If females are less prone to violent behavior than males, DIF is likely to emerge for items that suggest violent behaviors.

Validity Claim 1 references the appropriateness of scores from tests. If DIF studies suggest that assessment scores are multidimensional, assessment developers should consider separating items into different subscales for the distinct dimensions. Alternately, a multidimensional scoring model could be used to generate scores for each dimension using the same items.

Deciding how to address DIF in items may depend on the source of the secondary dimension. If a second dimension is not intended or desired, then items flagged for DIF could be removed from a scale. However, DIF results may suggest that items must be balanced to achieve valid scores. For example, in a study of reading and mathematics tests, Taylor and Lee (2012) found that nearly all short-answer and extended-response items flagged for DIF favored females, while nearly all multiple-choice items flagged for DIF favored males. This suggests an interaction between gender and assessment method. Elimination of items would distort the measurement of the underlying constructs; therefore, using a balance of item types is more appropriate in this context.

### Factor Analysis

Factor analysis is another strategy assessment developers use to investigate construct-irrelevant variance and dimensionality in scores. Validity researchers have the option of conducting exploratory factor analyses (EFA) or confirmatory factor analyses (CFA). Exploratory factor analysis is useful if the researcher does not have an *a priori* notion of how items and subscale scores will cluster. Confirmatory factor analysis is useful for testing alternate models when the likely sources of construct-irrelevant variance are assumed to be known.

Table 5–6 shows the results of an EFA. The purpose of the study was to investigate whether reading was a factor in mathematics test performance. For this study, the researcher obtained subtest scores from two achievement tests—a norm-referenced test and a state criterion-referenced test. This is an ideal case for CFA. The

Table 5–6			
<b>Factor Analysis of Scores from a Norm-Referenced Test and a Criterion-Referenced Test</b>			
<b>Test</b>	<b>Factor 1: Mathematics</b>	<b>Factor 2: Reading</b>	<b>Factor 3: ?</b>
NRT Vocabulary	.161	.449	.731
NRT Reading Comprehension	.185	.438	.749
NRT Math Concepts	.429	.168	.764
NRT Math Problem-Solving	.412	.235	.765
NRT Math Computation	.430	.057	.661
CRT Reading Main Ideas & Details–Fiction	.178	.764	.162
CRT Reading Analysis & Interpretation–Fiction	.210	.745	.172
CRT Reading Main Ideas & Details–Nonfiction	.260	.650	.340
CRT Reading Analysis & Interpretation–Nonfiction	.299	.653	.285
CRT Number Sense	.662	.208	.250
CRT Measurement	.518	.310	.249
CRT Geometric Sense	.589	.134	.362
CRT Statistics & Probability	.581	.299	.229
CRT Algebraic Sense	.653	.273	.109
CRT Solves Problems & Reasons Logically	.668	.311	.272
CRT Communicates Understanding	.673	.247	.292
CRT Makes Connections	.670	.152	.235
NRT = Norm-Referenced Test CRT = Criterion-Referenced Test			

researcher could test two models: one with a single factor combining reading and mathematics, and a second with two factors—one for reading and a second for math. However, this study demonstrates the value of EFA.

The factor loadings (correlations between scores for each variable in the analysis and the latent score for each factor) from this analysis suggest two fairly strong factors—one for reading (including the norm-referenced and criterion-referenced reading subscale scores) and one for mathematics (including scores from the norm-referenced and criterion-referenced mathematics subscales). The highest factor loadings for the first factor include all of the scores from the mathematics subscales in both tests; the highest factor loadings for the second factor include all of the scores from the reading subscales in both tests. It is noteworthy that the factor loadings for the norm-referenced subtests are not terribly strong on either of the first two factors, particularly in comparison with the factor loadings for the criterion-referenced subscale scores. The value of an EFA is apparent in the third factor. The third factor shows strong factor loadings for all of the norm-referenced subscale scores. Although the purpose of the study was to investigate whether reading is a factor in the criterion-referenced mathematics test scores, the results call into question the validity of scores from the norm-referenced test. These data suggest that there may be a method effect that generates construct-irrelevant variance.

### **Multi-Trait/Multi-Method**

Multi-trait/multi-method studies can also be used to investigate construct-irrelevant variance. For example, if a method effect exists, correlations among scores from tests of different constructs using a similar method will be higher than correlations between scores from tests of same construct using different methods. The potential for a method effect was suggested by the factor loadings in Table 5–6. Table 5–7 also suggests a method effect.

For the data shown in Table 5–7, the norm-referenced test (NRT) is entirely composed of multiple-choice items, whereas the criterion-referenced test is composed of multiple-choice, short-answer, and extended-response items. Correlations in the upper left-hand quadrant are within-method correlations for the NRT. Correlations in the lower right-hand quadrant are the

Table 5–7

**Correlations Among Fourth-Grade CRT Scores and NRT Test Scores**

<b>Tests</b>	<b>NRT Reading</b>	<b>NRT Language</b>	<b>NRT Mathematics</b>	<b>CRT Reading</b>	<b>CRT Writing</b>	<b>CRT Mathematics</b>
NRT Reading	(.91)	.75	.74	.70	.55	.62
NRT Language		(.90)	.77	.62	.66	.60
NRT Mathematics			(.92)	.60	.55	.74
CRT Reading				(.91)	.63	.68
CRT Writing					(.79)	.61
CRT Mathematics						(.92)

NRT = Norm-Referenced Test

CRT = Criterion-Referenced Test

within-method correlations for the CRT. Correlations in the upper right-hand quadrant are the between method correlations. With the exception of the correlation between the NRT and CRT mathematics scores, the within-method NRT correlations are the highest correlations in the entire matrix ( $r = 0.75, 0.74$ , and  $0.77$ ). The within-method correlations among the CRT test scores are not as strong as the within-method correlations for the NRT. The strongest within-method correlation for the CRT is between reading and mathematics ( $r = 0.69$ ), which is not as high as any of the intra-method correlations for the NRT but is still a strong correlation.

The correlations in the upper right-hand quadrant also provide evidence that can be evaluated against a construct-irrelevant variance argument. The highest correlations in the upper right-hand quadrant are between the CRT and NRT reading scores ( $r = 0.70$ ) and between the CRT and NRT mathematics scores ( $r = 0.74$ ). However, the correlation between the CRT and NRT reading scores is nearly the same as the correlation between the CRT reading and CRT mathematics scores ( $r = 0.70$  and  $0.69$  respectively). When correlations within method are as strong as between-method correlations for the same construct, this also suggests a method effect.

Not only does this matrix of correlations suggest method effects, it also prompts validation questions about the role of reading and language in all of the tests. The correlation between the CRT reading test scores and the NRT mathematics test scores is moderately strong ( $r = 0.60$ ). Similarly the correlation between CRT mathematics test scores and the NRT reading and language scores is moderately strong ( $r = 0.62$  and  $0.60$ , respectively). Finally, the correlations between CRT writing scores and the other CRT scores suggest a moderately strong relationship ( $r = 0.63$  and  $0.61$ ). The data in Table 5–7 present a challenge to the second argument of Validity Claim 3. A validation researcher would have to consider further study of method effects for both tests.

A question that must be asked when considering construct-irrelevant variance is whether the additional variance is truly construct-irrelevant. The within-method correlations among the criterion-referenced test scores in Table 5–7 suggest a strong relationship between reading and mathematics. Is reading skill a source of construct-irrelevant variance, or is it a secondary dimension of mathematics? Students must read equations, graphs, tables, charts, and diagrams, as well as text, to successfully perform on a

mathematics test. If research suggests that reading mathematically is a secondary dimension in mathematics learning and performance, then assessment developers could consider scoring models for mathematics tests that extract the reading dimension. Item-development activities could focus on the most effective ways to tap into mathematical reading in a systematic way.

Experimental Research

Experiments can be used to test for construct-irrelevant variance when researchers have previous evidence to suggest a potential source. In an experimental design, validation researchers would intentionally manipulate items based on the possible source of construct-irrelevant variance.

“Dr. Potter” is interested in whether item format (multiple-choice versus short-answer) is a source of construct-irrelevant variance in reading tests composed of both literal comprehension and higher order reading skills. She creates two tests—one composed of multiple-choice items and one composed of short-answer items. The reading passages and item stems are the same for both tests. Half of the items assess higher-order reading skills.

Dr. Potter gives a published standardized reading test composed of a mixture of multiple-choice and short-answer items to all examinees as a pre-test. Then she randomly assigns students to each test format. She conducts an analysis of covariance using the pre-test as a covariate (to control for initial differences between the two groups). Table 5–8 shows the means and standard deviations

Table 5–8 Means and Standard Deviations for Covariate and Post-Test for Two Groups				
Test Score	Multiple-Choice Condition		Short-Answer Condition	
	Mean	Standard Deviation	Mean	Standard Deviation
Covariate	15.93	7.48	15.73	7.81
Total Reading Score	12.40	3.65	13.83	3.63
Literal Comprehension	7.57	1.96	7.50	1.94
Higher-Order Reading	4.83	1.78	6.33	1.73

Table 5–9 <b>F-Test Results on Differences Between Reading Test Scores from Multiple-Choice and Construct Response Items</b>		
Score	F	Sig.
Total Reading Score	13.733	.000
Literal Comprehension Score	.307	.582
Higher-Order Reading Skills Score	53.014	.000

for the covariate, the reading total score, the literal comprehension score, and the higher-order reading score for students in the two groups.

These means suggest that there may be a method effect when assessing higher-order reading skills. The data suggest that the differences between the students in the two assessment conditions are negligible. Table 5–9 present the F-test results of the ANCOVAs for each of the three scores (total reading, literal comprehension, higher-order reading) after controlling for initial differences in the two groups. Dr. Potter’s results show significant differences in the total test score and the higher-order reading score between students who took the multiple-choice version of the test when compared with students who took the version with short-answer items, after controlling for initial differences among students. There is no difference between groups for the literal-comprehension items. Since item type should not influence scores, these data would suggest that item type is a source of construct-irrelevant variance for higher-order reading items.

Summary of Validity Claim 3

Validity Claim 3 states that one can extrapolate from the score on a test to the domain of the construct or criterion performance. Support for this claim is based on evidence for two arguments: (1) evidence that the same behaviors and responses are required on the assessment as in the domain of the construct or criterion performance, and (2) evidence that no knowledge, skills, abilities, or traits irrelevant to the domain of the construct or



criterion performance are required. Although it is never possible to *prove* this claim or the arguments, the job of test developers is to consider the intended inferences from an assessment as they relate to the constructs or criterion performance and to examine multiple sources of evidence that both support and potentially refute this claim.

Messick (1989) states that “multiple lines of evidence” should be brought to bear on inferences based on assessment scores. Messick cautions assessment developers to build strong validation programs and to avoid studies that are guaranteed to provide support for Validity Claim 3. He notes, “This very variety of methodological approaches in the validation armamentarium, in the absence of specific criteria for choosing among them, makes it possible to select evidence opportunistically and to ignore negative findings” (p. 33). As Cronbach (1989) indicated, technical manuals “rarely report... checks into rival hypotheses, followed by an integrative argument. Rather, they rake together miscellaneous correlations” (p. 155). Shepard (1993) states, “Explanations for this type of practice are possibly that the integrative nature of construct validation is not understood or that its demands are perceived to be too complex... to be implemented” (p. 407).

Kane (2006) indicates that planned interpretations and uses of assessment scores should guide selection of the studies that are important to investigate. I would add that assessment developers may want to consider likely public criticisms when selecting their validation studies. For example, a common criticism of tests of mathematical problem-solving is that they demand strong reading and writing skills. Assessment developers should anticipate these competing claims and conduct studies to investigate them.

Technical reports for assessments should provide results from studies providing multiple lines of evidence. As Messick states, “The varieties of evidence are not alternatives but rather supplements to one another” (1989, p. 16). Documentation of content reviews is necessary but not sufficient in the argument that the same abilities are required on the assessment as in the domain of the construct or criterion performance. Empirical studies should be conducted to show relationships between scores of different assessments that measure the same constructs, or between

assessment scores and the criterion performances. Empirical studies are needed to show whether patterns of relationships among tests of different constructs behave as expected. Studies are also needed to investigate possible sources of construct-irrelevant variance. Technical reports should provide a narrative that explains the rationales for the studies and an interpretation of the results. “What serves as evidence is the results of a process of interpretation—facts do not speak for themselves” (Kaplan, 1964, p. 375). Assessment users should closely examine the lines of evidence and the arguments to determine whether there is sufficient convergent and discriminant evidence to warrant use of the scores for their purposes.

## **Summary of Construct Validation**

Construct validation is the core of validity research. Construct validity research is focused on whether inferences from test scores are appropriate. Assessment developers should begin with a clear definition of the construct or criterion performance and a clear idea about the intended interpretations and uses of assessment scores when planning the structure of their assessments and the studies investigating the construct validity claims.

This chapter elaborates on Kane’s (2006) argument-based approach to validation research. Construct validation involves three major claims: scores can be trusted, scores are generalizable to a universe of items or tasks, and one can extrapolate from scores to the domain or criterion performance. The first two claims center on the technical qualities of assessments and are certainly the responsibility of assessment developers as they design and build their assessments. The third claim is one that can be jointly shared by assessment developers and validity researchers. References provided in this chapter demonstrate that validation research does not cease when an assessment is published and widely used (e.g., Kazemi, 2002; Taylor & Lee, 2011, 2012).

However, assessment developers have an obligation to conduct research related to each of these claims in terms of the intended inferences from test scores. Developers also have an obligation to publish technical reports that document and explain their validation work. One benefit of conducting validation research

is that the results can provide information to guide future assessment development work. Studies that support the arguments for these validity claims increase our trust in the inferences to be made from test scores and increase our confidence in the current methods of assessment. However, studies that refute these arguments may expand our understanding of the constructs and criterion performances, which may lead to improved test development.

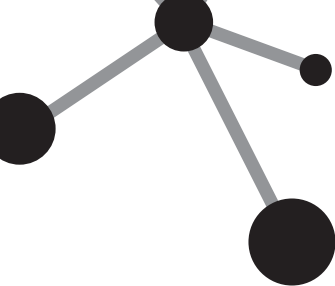
Assessment users also have responsibilities. Before selecting a tool for their assessment purposes, they should review technical reports to evaluate whether developers have adequately tested each of the claims described here. Users should also examine the evidence to assess whether the evidence warrants use of scores for their intended interpretations and uses. If no evidence exists, it is the responsibility of test users to investigate new interpretations and uses. Chapter 6 elaborates on studies focused on interpretations and uses of test scores.

## References

- Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67–91.
- Ackerman, T. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, 7, 255–278.
- Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, 20, 311–329.
- Almond, P., Browder, D., Crawford, L., Ferrara, S., Haladyna, T., Huynh, H., et al. (2005). *Including Students with Disabilities in Large-Scale Assessment Systems*. Retrieved January, 14, 2007. From <http://www2.ed.gov/programs/specedassessment>.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: AERA, APA & NCME.
- American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed). Arlington, VA: APA.
- Beadie, N. (1999). From student markets to credential markets: The creation of the Regents Examination system in New York State, 1864–1890. *History of Education Quarterly*, 39, 1–30.
- Brown, R. (2010, Dec.). English language learners and differential item functioning on the Easy CBM mathematics assessment Paper presented at the annual meeting of the Washington Educational Research Association, Seattle, WA.

- Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory* (pp. 218–222). New York: Holt, Rinehart and Winston.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.; pp. 443–507). Washington, DC: American Council on Education.
- Cronbach, L. J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.
- Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement Theory and Public Policy* (Proceedings of a symposium in honor of Lloyd G. Humphreys; pp. 147–171). Urbana: University of Illinois Press.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Elliot, A., & Harackiewicz, J. M. (1996). Approach and avoidance achievement goals and intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology*, 70(3), 461–475.
- Engelhard, G., Jr., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning. *Applied Measurement in Education*, 3, 347–360.
- Hambleton, R. K., & Jones, R. W. (1992, April). Comparison of empirical and judgmental methods for detecting differential item functioning. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification indices estimated under alternative strong true score models. *Journal of Educational Measurement*, 27, 345–359.
- Ilich, M. (2010, Dec.). English language learners and differential item functioning on NAEP mathematics items. Paper presented at the annual meeting of the Washington Educational Research Association, Seattle, WA.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.; pp. 17–64). Washington, DC: American Council on Education.
- Kaplan, A. (1964). *The Conduct of Inquiry: Methodology for Behavioral Science*. San Francisco: Chandler & Sharp.
- Kazemi, E. (2002). Exploring test performance in mathematics: The questions children's answers raise. *Journal of Mathematical Behavior*, 21, 203–224.
- Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, 33, 129–140.
- Livingston, S. A., & Lewis, C. (1993). *Estimating the Consistency and Accuracy of Classifications Based on Test Scores*. ETS Research Report 93–48. Princeton, NJ: Educational Testing Service.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed.; pp. 13–103). Washington DC: American Council on Education.
- National Council for the Social Studies (2010). *National Curriculum Standards for Social Studies*. Silver Springs, MD: NCSS.
- Nicholls, J. G. (1989). *The Competitive Ethos and Democratic Education*. Cambridge, MA: Harvard University Press.

- Nicholls, J. G., Cobb, P., Yackel, E., Wood, T., & Wheatley, G. (1990). Students' theories about mathematics and their mathematical knowledge: Multiple dimensions of assessment. In G. Kulm (Ed.), *Assessing Higher-Order Thinking in Mathematics* (pp. 138–154). Washington: American Association for the Advancement of Science.
- Nicholls, J. G., Patashnick, M., & Nolen, S. B. (1985). Adolescents' theories of education. *Journal of Educational Psychology*, 77, 683–692.
- Nolen, S. B. (1988). Reasons for studying: Motivational orientations and study strategies. *Cognition and Instruction*, 5(4), 269–287.
- Nolen, S. B. (1995). Teaching for autonomous learning. In C. Desforges (Ed.), *An Introduction to Teaching: Psychological Perspectives* (pp. 197–215). Oxford, England: Blackwell.
- Shavelson, R., Baxter, G., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30, 215–232.
- Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58, 159–194.
- Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammon (Ed.), *Review of Research in Education*, Vol. 19 (pp. 405–450). Washington, DC: AERA.
- Smith, W. (2010, Dec.). Language-related DIF in the WASL Mathematics Test. Paper presented at the annual meeting of the Washington Educational Research Association, Seattle, WA.
- Subkoviak, M. J. (1976). Estimating reliability from a single administration of a mastery test. *Journal of Educational Measurement*, 13, 265–276.
- Taylor, C. S., & Lee, Y. (2011). Ethnic DIF and DBF in reading tests with mixed item formats. *Educational Assessment*, 16, 1–34.
- Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education*, 25, 246–280.
- Thorkildsen, T. A., & Nicholls, J. G. (1998). Fifth-graders' achievement orientations and beliefs: Individual and classroom differences. *Journal of Educational Psychology*, 90(2), 179–201.
- Wang, T., Kolen, M. J., & Harris, D. J. (1996, April). Conditional standard errors, reliability, and decision consistency of performance levels using polytomous IRT. Paper presented at the Annual Meeting of the American Educational Research Association, New York City.
- Wineburg, S. (2001). *Historical Thinking and Other Unnatural Acts: Charting the Future of Teaching the Past. Critical Perspectives on the Past*. Philadelphia: Temple University Press.



## INTERPRETATION, USE, AND CONSEQUENCES OF SCORES FROM ASSESSMENTS

IN HIS GROUNDBREAKING treatise on validity, Messick (1989) states:

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness of inferences* and *actions* based on test score or other modes of assessment. (p. 13; emphasis in the original)

Actions based on test scores begin with inferences (What do the scores represent?) and interpretations (What do these representations mean?). Recall from Chapter 1 that an inference is generally closely tied to data. A student earns 90 percent of the points on a classroom reading test; the inference is that the student reads and comprehends well. A client has a very high score on a measure of depression; the inference is that the client is very depressed. A typist types 40 words per minute; the inference is that he has average typing skills. The inferences are made from the observed behaviors and responses to the larger domain of the construct or criterion performance.

The first three validity claims in Table 5–1 centered on inferences from scores: whether scores can be trusted, whether scores represent a universe of behaviors, and whether one can extrapolate (infer) from scores to the construct or criterion performance. Evidence to support each of these claims is tightly tied to the definition of a construct or criterion performance and the degree to which we can trust scores enough to make inferences about examinees. The studies, processes, and documentation described in Chapter 5 fit into the upper-left-hand quadrant of Figure 5–1 showing Messick's (1989) four facets of validity.

This chapter is focused on the last two claims in Table 5–1: Validity Claim 4—interpretations and decisions based on scores are appropriate; Validity Claim 5—the consequences of score interpretation and use are appropriate. These two claims are made—explicitly or implicitly—as soon as someone uses assessment scores for any purpose. In many ways, validation of claims about the interpretations, uses, and consequences of assessment scores are the most important validation studies to be done. Interpretations and uses of assessment scores have the greatest impact on examinees and assessment users. Yet, investigations of score interpretations, uses, and consequences are least likely to be conducted by assessment developers or included in technical reports for new or existing assessments.

Kane (2006) claimed that we must begin any plans for development of an assessment and for validation research related to the assessment with a clear idea of the intended interpretations of scores and the assessment purposes—the actions to be made based on assessment scores. Examples in Chapter 5 demonstrated the critical role of purpose in assessment development. The rating associated with the items in the achievement motivation test differed depending on the purpose of the assessment (research on consistency of goal versus strength of goal). The scoring rubric for the mathematics item differed depending on whether the purpose was to measure content knowledge or problem-solving ability. Purpose influences the scoring model (e.g., criterion-referenced versus norm-referenced) and the type of scale (e.g., interval versus ordinal). Each decision internal to the assessment development process is guided by the ultimate purpose of the assessment.

This chapter is focused on the types of research needed to support (or refute) interpretations and uses of test scores. It

begins with a discussion of studies to investigate whether there is support for score interpretations. Next, it presents examples of research for investigating score uses. Finally, it presents considerations of the social consequences of test score interpretations and uses.

#### **Claim 4: Interpretations of Scores and Decisions Made from Test Scores Are Appropriate**

*We can investigate differences in these test processes and structures over time, across groups and settings, and in response to experimental interventions—such as instructional or therapeutic treatment and manipulation of content, task requirements, or motivational conditions.* (Messick, 1989, p. 16)

Kane (2006) proposed two arguments related to Validity Claim 4: (1) the properties of observed scores support interpretations and (2) the implications of score use are appropriate. The distinction between interpretations of scores and the uses of scores is a critical distinction. For example, in their research on terror management theory, Arndt, Greenberg, Pyszczynski, & Solomon (1997) manipulated participants' fears in order to examine the influence of fear of death on expressions of nationalism, racism, and political conservatism. Fear of death was important to the research process; however, it was a manipulated variable. The researchers did not use the participants' level of fear to label them. However, if a psychologist uses a measure of fear to determine whether patients are paranoid, the scores are used to label patients based on an interpretation of the scores. Assessment developers and users have an obligation to investigate the validity of the proposed interpretations of assessment scores.

Similarly, when individuals use assessment scores to make decisions, they should provide evidence to support the validity of these decisions. Unfortunately, test users may not consider the meanings of scores. Interpretations may be implicit. For example, throughout the United States, scores from state tests are used to identify struggling schools. Laws have established the criteria to be used to identify these schools. However, the meaning of the test scores has little bearing on the decision-making process. Test score use



is procedural rather than meaningful. This occurs when assessment policy overrides validation of score meaning. What meanings might educators and policymakers assign to low-performing schools? A quick survey might show very different ideas about score meaning. Policymakers might use phrases such as “failing schools” or “dysfunctional schools.” Educators might interpret the scores to mean that teachers provide poor reading and/or mathematics instruction. Social scientists might interpret the scores to mean that students are from low income communities. Each of these interpretations moves well beyond the actual scores on tests. When assessment users make interpretations that go beyond the planned interpretations, or use the assessment scores for purposes that were not part of the original intent, they have a responsibility to conduct research to investigate the appropriateness of their interpretations and uses of scores.

### Properties of Observed Scores Support Interpretations

The first argument for Claim 4 in Table 5–1 is that the properties of the observed scores support *interpretations* of scores. Properties of observed scores are the ways in which scores function. A validation researcher would investigate whether scores function in a way that supports (or refutes) the validity of score interpretations. As mentioned above, score interpretations have value implications. If a student, a client, a program, an agency, or a professional is to be labeled in some way based on assessment scores, assessment users should be confident that the labels are warranted. Any interpretation of assessment scores merits attention—especially in situations where test scores are used to assign labels in ways that have social impacts. Do low scores on an achievement measure truly reflect low cognitive function? Do scores from an interest inventory truly indicate which career might be a good fit for an individual? Do thoughts about inkblots truly indicate that a patient has problems with authority?

### Investigating Score Interpretations

Chapter 1 introduced the idea that interpretations differ from inferences; they extend beyond the data and incorporate values. For example, low scores on a mathematics test might be interpreted to mean that students have poor mathematical reasoning

skills or that they are missing some prerequisite knowledge and skills. Both of these interpretations extend beyond the scores as a measure of mathematical achievement and have implications for action. To label a student as having poor mathematical reasoning skills is a value judgment; to say that a student lacks prerequisite skills is also a value judgment. The validation question must be whether the assessment interpretation is supported by evidence.

In the example of high-stakes achievement testing provided above, different interpretations of achievement test scores are made by policymakers, educators, and sociologists. Suppose a researcher claims that mathematics assessment scores provide evidence of the quality of mathematics instruction in schools. Typically, mathematics tests are designed to assess mathematics achievement. Developers generally don't do studies to evaluate the efficacy of scores in evaluating instructional quality. To test the claim that low mathematics scores reflect poor mathematics instruction, researchers would observe and evaluate the quality of mathematics instruction in a range of schools/classrooms and compare assessment scores with instructional quality evaluations. If scores from the assessment are to be interpreted to reflect the quality of instruction, researchers should find a positive relationship between evaluations of instructional quality and achievement test scores.

### **Investigating Alternative Interpretations**

It is possible that more than one interpretation is credible for assessment responses. A variety of strategies can be used to investigate alternate interpretations. Three strategies discussed here are experimental or quasi-experimental research, multidimensionality analyses, and structural equation modeling analyses.

*Experimental or Quasi-Experimental Research* Suppose a validity researcher believes that scores from a mathematics achievement test are a reflection of mathematical reasoning skills. To investigate this claim, the researcher might randomly assign students to either a treatment or a control condition. In the treatment condition, students receive supplemental instruction focused on mathematical reasoning skills (e.g., comparison based on attributes, sorting, ranking, inferring, drawing conclusions, extracting information). The other group serves as a control group and receives no supplemental instruction. If the mathematics scores can be interpreted as measures of mathematical reasoning, post-treatment scores for

the treatment group should be much better than post-treatment scores for the control group.

On the other hand, suppose the researcher believes that the mathematics scores are a reflection of whether students have (or lack) prerequisite mathematics skills. The researcher might conduct an experimental study wherein the treatment group has an intervention in which students review and practice with prerequisite skills and the control group receives no additional instruction. If low scores can be interpreted as indicators that students lack prerequisite knowledge and skills, students in the treatment condition should have higher post-treatment scores than students in the control condition.

*Multidimensionality Research* Alternate interpretations are possible when item and task responses may be caused by factors other than the targeted construct. In construct-validation research, multidimensionality and differential item functioning (DIF) analyses were proposed as ways to examine whether item and task scores are caused by the expected construct or whether secondary dimensions influence scores. Researchers can also use these analyses to examine the appropriateness of score interpretations.

For example, Walker and Beretvas (2000) conducted a study to examine whether examinees with strong writing scores did better on open-ended mathematics items than did examinees with weak writing scores. Their findings were surprising. All of the open-ended mathematics items showed DIF in favor of strong writers except the item that required the most writing. Even items that required no writing at all (e.g., “draw a figure with two lines of symmetry,” “write numbers in order from greatest to least”) demonstrated DIF in favor of strong writers. In such a case, interpretation of mathematics scores as indicative of mathematical proficiency alone is problematic. It is possible that the open-ended mathematics items tapped into a secondary dimension—one that also influences writing performance (e.g., the ability to organize and represent information). In this case, interpretations of total scores limited to mathematical proficiency, without taking into account this secondary dimension, could lead to inappropriate and ineffective interventions for examinees.

In another study, Taylor and Lee (2004) used DIF analyses to investigate whether reading was a dimension of mathematics

performance on achievement tests composed of word problems in multiple-choice and constructed-response formats. They found that the majority of DIF items favored struggling readers. The majority of the items favoring struggling readers involved visual models (graphs, charts, and figures) except at the high-school level. For high school, the items and tasks favoring struggling readers were those that involved applications of mathematics in contexts that would be familiar to teenagers (e.g., using a system of equations to select the best cell phone plan). Their conclusion was that, although there may be multidimensionality issues within the scores, reading was not a secondary dimension of mathematics scores from the studied test. Their research supported claims that scores could be interpreted in terms of mathematical proficiency.

*Structural Equation Modeling (SEM) Research* Suppose “Dr. Nguyen” has developed a new self-confidence scale and wants to provide subscale scores so that she can conduct research factors that influence dimensions of self-concept. Prior researchers have proposed that self-concept can be divided into five dimensions: aspiration, anxiety, academic confidence, initiative, and social identification. However, some theories suggest that aspiration and academic confidence actually compose a single dimension. To investigate the validity of score interpretations from the new scale, Dr. Nguyen conducts SEM research. First, she investigates a base model, which posits a unidimensional scale. Then she tests the two competing models and compares their fit with the base model. Figures 6–1 and 6–2 demonstrate the two competing models for interpretation of scores.

By testing these models, Dr. Nguyen can select the model that best fits the data from her assessment so that score interpretations will be appropriate.

### **Investigating Interpretations over Time**

When investigating score meaning, validation researchers should examine the properties of scores over time. Table 6–1 shows the factor structure for scores from a hypothetical reading test given to samples of students from the seventh grade in two different years. As can be seen, the factor structures differ between Time 1 and Time 2. For Time 1, the first factor appears to represent items that assess comprehension of fictional text; the second factor appears to represent items related to comprehension of nonfiction

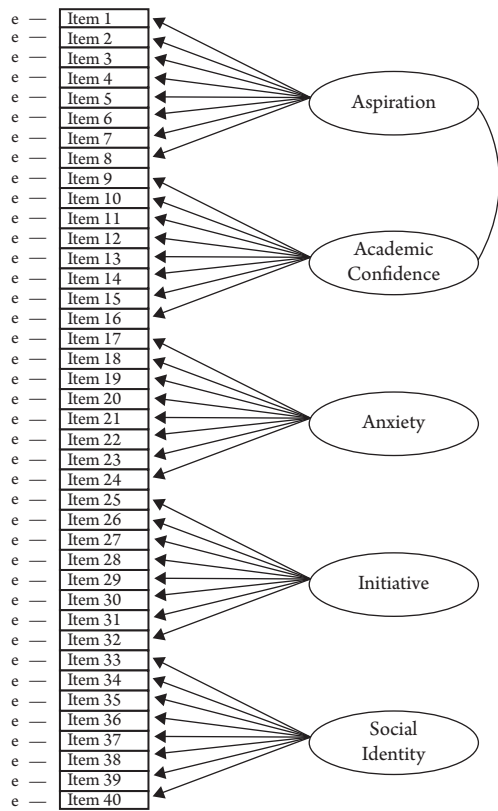


Figure 6–1 Structural Equation Model 1—Five-Factor Self-Concept Scale

text. At Time 2, higher-order reading skills load on the first factor, regardless of the type of passage; literal comprehension loads on the second factor, regardless of the type of passage. Results such as these should lead assessment users to be cautious when interpreting the sub-scores from the assessment. Assessment developers and assessment users should conduct studies to verify that score meaning has not changed over time. If score meaning *has* changed, research supporting the change should be documented, and score reports should reflect the changes.

### Investigating Score Interpretations Across Settings

The question of consistency in score meaning across settings is similar to the question about score meaning over time. In the hypothetical example of the reading assessment, students changed in their patterns of responses between Time 1 and Time 2. Similar

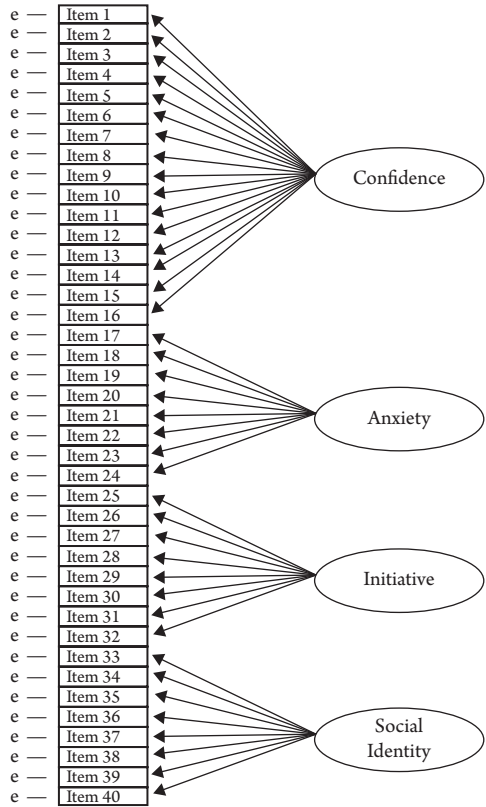


Figure 6–2 Structural Equation Model 2—Four-Factor Self-Concept Scale

results might be found if students are exposed to very different instructional programs. Scores from an assessment of mathematical problem-solving could actually have very different meanings in different school districts. Suppose Anderson School District uses a mathematics instructional program focused on problem-solving, and Baker School District uses a mathematics program focused on memorization and application of algorithms. Scores on a test of problem-solving from Anderson School District might represent students’ learned problem-solving skills. Scores from Baker School District might represent general cognitive abilities. Similar results could occur when using psychological or career assessments. For example, behaviors that are viewed as “resilient” in one community might be considered “defensive” in a different community. Such a difference could affect interpretation of scores from an assessment of coping skills.

Table 6-1 Differing Factor Structures for Two Samples at Two Times				
	Time 1		Time 2	
	Factor 1	Factor 2	Factor 1	Factor 2
Literal Comprehension—Fiction	0.27	0.77	0.25	0.73
Interpretation—Fiction	0.16	0.81	0.69	0.29
Critical Reading—Fiction	0.16	0.76	0.69	0.24
Literal Comprehension—Non-Fiction	0.75	0.14	0.20	0.69
Interpretation—Non-Fiction	0.79	0.20	0.73	0.27
Critical Reading—Non-Fiction	0.63	0.28	0.75	0.40

Assessment developers should provide very clear information about the characteristics of examinees (e.g., gender, ethnicity, education level), contexts (e.g., region, type of community, public or private setting), and other relevant variables (e.g., educational programs, therapeutic models, training programs) for the samples that are used in the development of assessments. Assessment users should evaluate these data to determine whether score interpretations are likely to generalize to their own examinees.

If users have doubts about the generalizability of interpretations, they can conduct their own studies to evaluate score interpretations, especially when interpretations may differ due to cultural differences, educational experiences, social mores, employment contexts, etc. Studies might include item analyses (to determine whether the item parameters are the same as or similar to those for the original samples), factor analyses (to look at the comparability of factor structures across groups), correlations with related measures or other assessments of the same constructs, correlations with non-test behaviors, differential item functioning analyses, and so forth.

## Investigating Score Interpretations Across Languages and Cultures

Another validation issue occurs when assessments must be translated or adapted for use with a different language or cultural group. The International Test Commission (ITC) has developed guidelines for adapting tests for languages and cultures other than those for which they were developed (ITC, 2010). These guidelines apply when the goal of the development or adaptation is to obtain scores that have the same meaning across different populations. Table 6–2 presents the main points of their adaptation guidelines. The table identifies four categories that are important to address in validation studies: context, development and adaptation, administration, and score interpretation. The ideas in the table suggest specific studies that should be done to investigate whether the scores from a test can be interpreted the same way in different populations.

Research related to guidelines D6 through D10 is most salient to questions of score meaning. These guidelines focus on the statistical properties of scores. To investigate equivalence of scores, data are needed from all populations for whom the assessment scores will be used. Studies are needed to determine whether items, subscale scores, and total scores have equivalent properties across different populations. The following examples highlight the types of research that are done related to guidelines D6 through D10.

*Confirmatory Factor Analyses* Two researchers, du Toit and de Bruin (2002), examined the degree to which John Holland's six-factor career model generalized to South African youth. They found that the six-factor model did not transfer to these youth. Contreras, Fernandez, Malcarne, Ingram, and Vaccarino (2004) investigated whether the factor structure of the *Beck Depression and Anxiety Scales* applied to Latino students. They found that Beck's two-factor structure fit the data and that the internal consistency estimates of reliability were acceptable for both scales. Yan, Tang, and Chung (2010) examined the factor structure of the *Perinatal Grief Scale* for Chinese women. The researchers found that, although the items were appropriate for Chinese women, the scales were not. The original scale—normed in the United States—had three factors (active grief, difficulty coping, and despair); however, for Chinese women, the items loaded on three different



Table 6–2 <b>International Guidelines for Translating and Adapting Tests</b>	
Development and Adaptation	
D1	Test developers and publishers should insure that the adaptation process takes full account of linguistic and cultural differences among the populations for whom adapted versions of the test are intended.
D2	Test developers and publishers should provide evidence that the language use in the directions, rubrics, and items themselves, as well as in any handbooks, are appropriate for all cultural and language populations for whom the test or instrument is intended.
D3	Test developers and publishers should provide evidence that the choice of testing techniques, item formats, test conventions, and procedures are familiar to all intended populations.
D4	Test developers and publishers should provide evidence that item content and stimulus materials are familiar to all intended populations.
D5	Test developers and publishers should obtain judgmental evidence, both linguistic and psychological, to improve the accuracy of the adaptation process and acquire evidence of the equivalence of all language versions.
D6	Test developers and publishers should ensure that the data collection design permits the use of appropriate statistical techniques to establish item equivalence between the different language versions of the test or instrument.
D7	Test developers and publishers should apply appropriate statistical techniques to (1) establish the equivalence of the different versions of [scores from] the test or instrument, and (2) identify problematic components or aspects of the test or instrument that may be inadequate to one or more of the intended populations.
D8	Test developers and publishers should provide information on the evaluation of validity in all target populations for whom the adapted versions are intended.

*(continued)*

Table 6–2

**(Continued)****Development and Adaptation**

D9	Test developers and publishers should provide statistical evidence of the equivalence of questions for all intended populations.
D10	Non-equivalent questions between versions of the test or instrument intended for different populations should not be used in preparing a common scale or in comparing these populations. However, they may be useful in enhancing content-related evidence for the validity of scores reported for each population separately.

**Context**

C1	Effects of cultural differences that are not relevant or important to the main purpose of the test should be minimized to the extent possible.
C2	The amount of overlap in the construct measured by the test or instrument in the populations of interest should be assessed.

**Administration**

A1	Test developers and administrators should try to anticipate the types of problems that can be expected, and take appropriate actions to remedy these problems through preparation of appropriate materials and instructions.
A2	Test administrators should be sensitive to a number of factors related to the stimulus materials, administration procedures, and response modes that can moderate the validity of the inferences drawn from the scores.
A3	The aspects of the environment that influence the administration of a test or instrument should be made as similar as possible across populations of interest.
A4	Test administration instructions should be in the source and target languages to minimize the influence of unwanted sources of variation across populations.
A5	The test manual should specify all aspects of the administration that require scrutiny in a new cultural context.

*(continued)*

Table 6-2 <b>(Continued)</b>	
<b>Administration</b>	
A6	The administrator should be unobtrusive, and the administrator-examinee interaction should be minimized. Explicit rules that are described in the manual for administration should be followed.
<b>Score Interpretation</b>	
I1	When a test or instrument is adapted for use in another population, documentation of the changes should be provided, along with evidence of the equivalence.
I2	Score differences among samples of populations administered the test or instrument should not be taken at face value. The researcher has the responsibility to substantiate the differences with other empirical evidence.
I3	Comparisons across populations can only be made at the level of invariance that has been established for the scale on which scores are reported.
I4	The test developer should provide specific information on the ways in which the socio-cultural and ecological contexts of the populations might affect performance, and should suggest procedures to account for these effects in the interpretation of results.

factors (sense of worthlessness, social detachment, and painful recollection). Internal consistency estimates for the three new scales were as strong as for the three original scales. These differences do not mean that the scores from the assessment are invalid but do suggest that interpretations of scores will be different for the different groups.

*Item Analyses* Thi-Xuan-Hanh, Guillemin, Dinh-Cong, and colleagues (2005) investigated the applicability of the *Adolescent Duke Health Profile* for Vietnamese adolescents. They found that they had to adapt two items in the scale in order to validate scores for Vietnamese adolescents. Lau, Cummins, and McPherson (2005) investigated the validity of the *Personal Wellbeing Index* for individuals from Australia and Hong Kong. The scale was intended to serve as an international measure of personal well-being; however,

the researchers found one item that was a strong contributor to the scale for Australians served no function for individuals from Hong Kong.

These guidelines address issues of comparability of scores. If scores from assessments will be used only within one culture and no cross-cultural comparisons are relevant to research, culturally specific scales and derived scores can be developed. However, if assessments will be used to make cross-cultural comparisons, the validity of the score comparisons should be investigated.

Clearly, given the number of guidelines for ensuring cross-cultural/cross-linguistic comparability of scores, the guidelines place a significant responsibility on both assessment developers and assessment users in terms of what score interpretations can be made from scores (across different populations) and what research is needed to validate cross-cultural claims about score interpretations. Assessment users should review technical reports to determine whether translation/adaptation processes are well documented and that appropriate studies have been conducted to warrant the proposed interpretations of assessment scores when comparing cultural/linguistic groups. Any limitations to comparability should be clearly stated in technical reports.

### **Cut Scores in Validation of Score Interpretation and Use**

Cut scores were introduced in Chapter 5 as they relate to minimizing errors in decision-making. One might reasonably ask, “Decisions about what?” Cut scores go beyond inference and give the assessment scores another layer of meaning (e.g., the patient is at risk, the student is gifted, the potential employee is qualified). Cut scores are used with assessments in education, psychology, and employment. One or more individuals (generally experts in the relevant field or profession) use a deliberative process to set a cut score, usually involving data regarding normative performance on the assessment and the likely results of decisions that arise from the cut scores.<sup>1</sup>

For example, throughout the United States, achievement tests are administered to assess whether students are learning the

---

1. Cizek and Bunch (2007) have written a very useful handbook that describes a range of deliberative methods for setting cut scores.

mathematics knowledge and skills described in state and national curriculum standards. Low scores are interpreted to mean low mathematics achievement; high scores are interpreted to mean high mathematics achievement. When cut scores designate proficiency levels (e.g., proficient/not proficient), the cut scores have added to the meaning of the total scores. What does “proficient” mean? States must develop detailed performance-level descriptors (PLDs) to describe the academic performance characteristics of a proficient student. Unfortunately, states are not required to formally investigate the appropriateness of the PLDs as interpretations of test scores. To investigate PLDs, validation researchers would have to directly observe the work of students who perform at the proficient and non-proficient levels. When proficiency-level interpretation is generalized beyond the test scores, investigations would include non-test behaviors as well as test behaviors.

Cut scores are used in many fields to identify people who are at risk in some way. For example, “Dr. Jackson” develops a new scale to assess depression and suicidal ideation. It is to be used in hospital settings with patients who have been hospitalized for depression and/or attempted suicide. Prior research shows a correlation of  $r = 0.71$  between scores on the new scale and scores from the Beck Depression Inventory.

Dr. Jackson wants to set a cut score on his depression scale to flag patients who are at risk of suicide. He contracts with hospitals to assess their patients twice within the first four days of hospitalization, with two days between assessment events. He uses data from the first day of testing to set a cut score. He sets the cut score at the score on the depression scale that is most likely to identify patients who have attempted suicide and to discriminate between patients who have attempted suicide and those who are depressed but have not attempted suicide (see Table 6–3). To assess the

Table 6–3 Relationship Between Scores on Depression Scale and Suicide Attempts		
	Attempted Suicide	Did Not Attempt Suicide
At or above cut score	98	13
Below cut score	2	87

Table 6-4

**Decision-Consistency for Cut Score for Suicide Risk on Depression Scale**

	<b>At or Above Cut Score</b>	<b>Below Cut Score</b>
At or above cut score	130	8
Below cut score	5	57

reliability of the cut score, he uses the data from the second day of testing to compute the percent of agreement between the testing Time 1 and testing Time 2 (see Table 6-4). With a percentage of exact agreement at 94 percent, Dr. Jackson claims that the cut score will result in reliable identification of patients who are at risk of suicide. Dr. Jackson's data show that his cut score identified 98 out of 100 patients who attempted suicide and only 8 out of 100 patients who were hospitalized for depression but did not attempt suicide.

Cross-validation research is often used to provide support for the validity of score interpretations. The goal of cross-validation research is to determine whether the scores are likely to have the same meaning with different samples from the same population. "Dr. Nolen" collects data from 200 patients in a different region of the country who have been hospitalized for depression and/or suicide. Table 6-5 presents Dr. Nolen's cross-validation data using Dr. Jackson's new depression scale.

Her data show that the cut score identifies 96 out of 100 patients who attempted suicide. Note, however, that the cut score also flags 38 patients who were hospitalized for depression but did not attempt suicide. These data do not provide support for

Table 6-5

**Cross-Validation of the Relationship Between Scores on Depression Scale and Suicide Attempts**

	<b>Attempted Suicide</b>	<b>Did Not Attempt Suicide</b>
At or above cut score	96	38
Below cut score	4	62

Dr. Jackson’s claim that the scores from the depression scale can be interpreted in terms of suicide risk.<sup>2</sup>

“Dr. Jones” does a two-year follow-up study with the patients from Dr. Jackson’s initial study to determine whether these patients’ depression scores during hospitalization predicted future suicide attempts. The data are shown in Table 6–6. These data provide little support for the claim that scores above the cut score can be interpreted as predictive of future suicidal behaviors. These examples show that, to trust score interpretations based on cut scores, a single study may be inadequate to provide support for the cut scores provided in the user’s manual.

In summary, many different types of studies can be used to investigate score interpretations. The studies described here are examples but do not exhaust the range of possible studies. What is most important is that assessment developers are very clear about the interpretations that are intended to be made from test scores and that the studies are designed to investigate whether those interpretations can be made. Some studies may be replications within a

Table 6–6 Follow-up Predictive Study of the Relationship Between Scores on Depression Scale and Suicide Attempts		
	Attempted Suicide	Did Not Attempt Suicide
At or above cut score	37	74
Below cut score	1	89

2. Note here that a distinction must be made between score *interpretation* and score *use*. Each cell of Tables 6–3, 6–5, and 6–6 can be interpreted in terms of error. The upper-left-hand cell and the lower-right-hand cell give the number of examinees who are correctly diagnosed with the depression scale (i.e., no error). The upper-right-hand corner represents Type 1 error—false positives. The lower-left-hand corner represents Type 2 error—false negatives. When a cut score is intended to predict life-threatening behaviors, health care providers are generally more comfortable with false positives than false negatives. In other words, they would rather be concerned about people who are not at risk than be complaisant about people who are at risk. For this scenario, a score *interpretation* of “at risk of suicide” is not supported by the data. The data provide no evidence regarding the degree to which the scale provides a valid measure of depression or suicidal ideation.

population at different times and in different contexts; others may require investigations of the fidelity of translations and the cultural relevance of items and score dimensions for different populations.

### Scores Are Appropriate for Intended Uses

Assuming that the structure of scores supports score interpretations, the second argument to support Claim 4 (in Table 5–1) is that the *uses* of assessment scores are appropriate. When assessments are developed, their intended uses are generally known. However, assessment purposes may change due to time or context. Assessment developers may create assessment tools for one purpose, and assessment users may use the scores for a different purpose. A criterion-referenced test might be designed for assessing general achievement and policy makers may decide to set a passing-score for the test and use the passing score as a high school graduation requirement. An assessment of on-the-job problem-solving skills, designed for purposes of training, may become job-placement test. An assessment designed to diagnose psychopathology in a hospital setting may be used in a community counseling center to screen entering clients. Whether assessment scores are being used for their intended purpose or for a new purpose, research is needed to support (or refute) the claim that scores are appropriate for the given use. Research should support two facets of the argument: (1) score inferences and interpretations are *useful* for the identified purpose, and (2) using scores for the identified purpose is *appropriate*.

### Support for Intended Uses

When assessment developers construct assessment tools, they have an intended use in mind and should provide evidence to support any intended uses for their assessments. If the use is selection, placement, or identification, research should focus on efficacy or accuracy of selection, placement, or identification. If the use is diagnosis, research should focus on the accuracy of diagnosis. If the use is to monitor progress over time, research should focus on whether scores change in predictable ways over time. The following scenarios highlight a variety uses of test scores and evidence that supports or does not support the intended use of the scores. Note that these are illustrative and do not reflect the plethora



of possible uses of test scores and the types of studies needed to support those uses.

### **Using Scores to Monitor Change over Time**

Suppose “Dr. Kim” has developed a test to assess the progress of dementia in patients with Alzheimer’s disease. First, Dr. Kim develops a set of items and tasks that measure short- to long-term memory—twenty items require recall of information (e.g., address, telephone number, names of family members), and twenty items require patients to complete simple tasks (tying shoes, making a sandwich). The items and tasks range from easy to difficult. Dr. Kim uses item response theory (IRT) to create an interval scale that ranges from easy (e.g., recalling a telephone number) to difficult (e.g., recalling the name of the president of the United States in 1969).

Since the assessment is designed to assess progress of dementia, she conducts a study to determine whether the scores are sensitive to progress of the disease. She administers the assessment to patients with different degrees of dementia to verify that the scores change systematically in the expected direction. Dr. Kim finds that patients with more serious dementia have lower scores on her assessment than patients with early-onset Alzheimer’s disease and that there is a fairly steady decrease in scores as dementia increases. This study supports the intended use of her assessment—to assess progress of the disease.

### **Using Scores in Selection**

ABC Manufacturing Company contracts with a testing company to develop a 30-item screening test for potential employees. To evaluate the usefulness of the test in selecting new employees, the company administers the test to all employees who have been hired within the past three years. The company compares test performance with the employees’ most recent job performance ratings of Exceeds Expectations, Meets Expectations, and Does Not Meet Expectations.

Table 6–7 shows the results of the study. Based on the results, the scores from the employment screener may not be useful in selecting new employees. The pattern of scores shows that there is a relationship between scores on the screener and performance ratings. However, mean scores of the three groups are not very

Table 6–7

**Means and Standard Deviations for Employees at Different Performance Levels**

<b>Group</b>	<b>Mean</b>	<b>Standard Deviation</b>
Exceeds expectations	21.35	3.21
Meets expectations	20.37	4.72
Does not meet expectations	19.78	5.24

different, and the standard deviations suggest there is a significant overlap in score distributions for the groups.

### Using Scores to Focus Instructional Interventions

XYZ Testing Company claims that its mathematics assessments are aligned with a state's mathematics content standards and can be used to diagnose students' proficiencies related to those standards. Their tests are to be administered three times per year to chart students' progress on tested knowledge and skills. The company claims that, if students' scores improve on the interim assessments, their chances of doing well on the state standardized test are improved. State mathematics test scores for Wickham School District did not improve between 2010 and 2011. The district decides to adopt the interim assessment system to help teachers focus their instruction and improve their students' achievement. The district provides instructional materials related to the skills on the interim assessments and urges teachers to use the materials in their classrooms to address the needs of students who do poorly on the interim assessments.

The school district conducts a study to determine whether targeting instruction based on the results of the interim diagnostic test scores actually improves achievement on the state's standardized achievement test. The district charts scores on the interim assessment and on the state's standardized mathematics tests. Figure 6–3 shows the progress of the students' scores on the diagnostic assessment during one school year. Figure 6–4 shows the score changes on the state mathematics test. Although scores have increased on the interim diagnostic assessment over the course of the year, scores on the state mathematics test have decreased. The district concludes that use of the interim diagnostic assessment

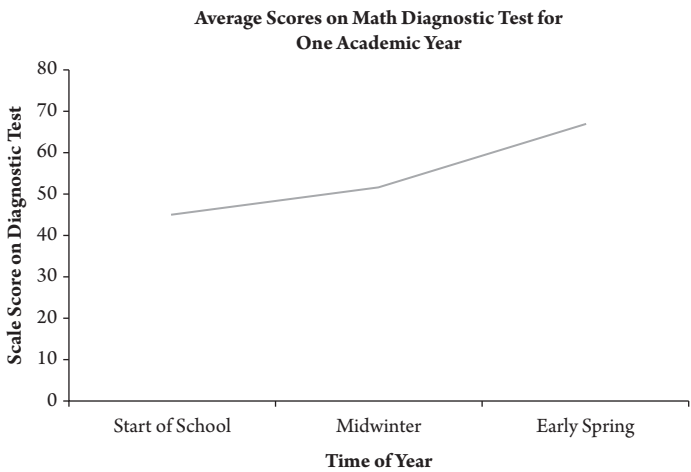


Figure 6-3 Scores on Diagnostic Test over Time

has not helped their students achieve state standards and may have done more harm than good.

Assessment developers have an obligation to conduct studies to investigate any claims they make regarding the use of the scores from their assessments. Evidence should be brought to bear to support those claims, and assessment users should ask for that evidence before using an assessment for the purpose intended by the test developer.

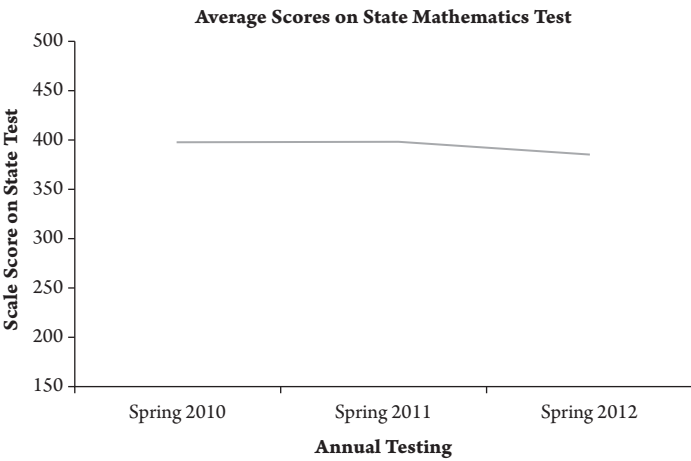


Figure 6-4 Scores on State Mathematics Test over Time

### Using Assessment Scores for New Purposes

Scores from standardized assessments are sometimes trusted more than human judgments. The efficiency of standardized assessment and the perceived trustworthiness of scores interact and lead to a wide range of assessment uses for which there is no support (Green, 1998). Unfortunately, assessment policies often preempt research. Assessments are used for placement, promotion, graduation, or job selection when no research has substantiated these uses of the assessment scores.

For example, during the 1980s, under pressure to raise achievement, many U.S. states implemented kindergarten and first-grade readiness testing policies. Although the purpose of the readiness tests was to support instructional planning, local educational agencies used test scores to deny children entrance to kindergarten or the first grade based on their readiness test scores. Yet, of 16 controlled studies investigating the impact of retention policies, one found gains, another found long-term, significant negative impacts, and the rest found no academic benefits for children (Shepard, 1989).

With the passage of the No Child Left Behind Act (2003) in the United States, all states were required to develop standards-based assessments and set cut scores to identify students who were proficient on the targeted content. Many states began to use their standards-based achievement tests to award or deny high school diplomas (Dietz, 2010) or to make grade promotion decisions. Assessing proficiency in relation to a set of curriculum standards is not the same as assessing whether students have learned what they need to know to be successful beyond high school or in the next higher grade. If test scores are to be used for purposes such as high school graduation or promotion to a higher grade level, researchers should collect scores from students at all score levels and document whether examinees with higher scores perform better in successive grades, in post-high school careers, and in college than do examinees with lower scores. Such studies should be conducted *before* the assessment scores are used for such high-stakes decisions.

Whenever assessment scores are used for a new purpose, studies similar to the ones described above should be conducted before the assessment scores are used for a new purpose. Without adequate empirical support for the efficacy of using the scores for a new purpose, new uses are not validated.

### Challenges in Evaluating the Validity of Scores for Selection Decisions

One of the challenges of investigating the appropriateness of using assessment scores for selection purposes is that it may not be possible to obtain a good measure of the degree to which scores on an assessment predict criterion performance. For example, if an assessment is used for job placement, it is not possible to obtain information about job performance for individuals who were not hired. Correlations between assessment scores and job performance are likely to be lower than would be expected if all job candidates were hired. In addition, if on-the-job training makes it possible for all employees to meet or exceed employer expectations, assessment scores will not be predictive of actual job performance.<sup>3</sup>

To support the validity of the claim that scores from an assessment can be used in job selection decisions, assessment developers would gather data on all job candidates, allow all potential candidates to do the job, and assess whether the scores were useful in differentiating between successful and unsuccessful candidates.<sup>4</sup>

Similar issues arise for any situation in which assessment scores are used for selection. For example, scores from an achievement test might be used to select students for a gifted education program. Achievement is highly related to opportunity to learn. Therefore, students with effective teachers are more likely to be selected for gifted programs than are students with ineffective teachers. Educators would need to conduct studies to evaluate whether students with high achievement scores benefit from gifted education programs more than students with moderate or low achievement scores.

Despite the challenges in conducting research to validate the use of assessment scores for selection purposes, assessment users are responsible for gathering evidence to support claims that the scores from a given assessment are appropriate for any selection

- 
3. This scenario raises an important validation question: If on-the-job training results in high job performance for all employees, is the assessment needed to select employees? Probably not.
  4. Few employers would accept this as a model for research. Employers would be likely to complain about the costs of wages, benefits, and training for low-performing employees. Therefore, the studies needed to ensure fair use of assessment scores in employment selection decisions would probably involve assessment of existing employees.

decision (e.g., AERA, NCME, APA, 1999; Equal Employment Opportunity Commission, 1978).

#### Summary of Claim 4

Validation of score interpretations and uses begins with clear ideas about the intended interpretations and uses of assessment scores. Researchers should consider these purposes when designing studies. Any score interpretations should be supported with studies showing that the interpretations are stable over time, across groups within the target population, and in varied settings. When an assessment is developed for one population and is used with examinees who have different linguistic or cultural backgrounds than the original population, studies should show that translations and adaptations are appropriate and that scales and scores retain their meaning in the new population. If the purpose of an assessment is to identify students who are at risk of failing in school, research should focus on whether assessment scores are adequate for this purpose. If the purpose of an assessment is to select individuals for employment, research should focus on the adequacy of assessment scores for making appropriate selections. If the purpose of an assessment is to diagnose strengths and weaknesses of learners, validation research should focus on the adequacy of assessment scores in serving this purpose. The *Standards* (AERA, NCME, APA, 1999) are clear that assessment users should conduct research to validate proposed score interpretations and uses for which no validity evidence is available. Some would say that validation of the interpretation and uses of assessment scores is the most central task of all assessment research (Shepard, 1993).

#### **Claim 5: Consequences of Score Interpretation and Use Are Appropriate**

*Finally, we can trace the social consequences of interpreting and using test scores in particular ways, scrutinizing not only the intended outcomes but also unintended side effects.*  
(Messick, 1989, p. 16)

Validity theorists, assessment developers, and assessment users generally agree that multiple lines of evidence are needed to support (or refute) any intended score interpretations and uses.

Within the past twenty years, many validity theorists have accepted the idea that validation research must also take into account the *consequences* of score interpretation and use. Assessment developers and users are likely to consider intended consequences as they approach validation research—looking for evidence to support the intended consequences of score interpretations and uses (Messick, 1989). However, current validity theorists (Shepard, 1993; Messick, 1989; Moss, 1998; Kane, 2006) agree that validation work must go beyond intended consequences and consider unintended consequences—particularly unintended consequences of construct-irrelevant variance, under-representation of the construct, and unintended uses of assessment scores.

Consideration of the social consequences of score interpretation and use has long been an important issue in the field of assessment; however, validity theorists are divided in terms of whether social consequences should be a validity issue. Some argue that consequences are out of the control of assessment developers and, therefore, should not be considered a validity issue (Lissitz & Samuels, 2007; Mehrens, 1997; Popham, 1997; Wiley, 1991). Wiley (1991) argued that, although inappropriate use of scores is an important social issue, it “involves social and moral analyses beyond the scope of test validation...and would needlessly complicate the conception and definition of test validity” (p. 88). Mehrens (1997) argues that consequences of test score use have more to do with the appropriateness of decisions and actions based on test scores than with the validity of score-based inferences. “The accuracy of the inference that the person has a fever is separable from the consequences of any treatment decision that may follow it” (p. 16). Popham (1997) states that assessment developers and users should investigate both actual and potential consequences of score uses. However, he argues that if consequences are considered an aspect of validity, this will draw attention away from the central validity issue—the accuracy of score-based inferences.

Other validity theorists claim that it is the *use* of assessment scores that frames the central validity issue—whether inferences from test scores are useful for their intended purpose (Cronbach, 1971, 1989; Linn, 1997; Moss, 1998; Shepard, 1997). Linn (1997) counters Popham’s (1997) notion that the accuracy of score-based inferences is the central validity issue. He states, “...it should be clear that the accuracy of test interpretation does not guarantee

usefulness or that uses made of test results will be appropriate” (p. 14). Cronbach (1971) argues:

The justification for any [assessment-based] decision is a prediction that the outcome will be more satisfactory under one course of action than another. Testing is intended to reduce the number of incorrect predictions and hence the number of decisions that will be regretted later. . . . What is the payoff when decisions are made in the proposed way, and how does this compare with the payoff resulting when decisions are made without these data? (p. 448)

Cronbach (1988) suggests that any score interpretation is open to validity challenges if adverse consequences arise. Shepard’s (1997) argument is consistent with Cronbach’s (1971) ideas about payoff. Shepard claims that investigations of intended and unintended consequences are simply an aspect of the validation of score interpretation and use and not a separate category for validation research. For example, if scores from assessments are used for selection or placement, studies must demonstrate that the all parties benefit from the selection or placement process. She notes that the relationship between scores on a selection or placement test and individual experiences based on subsequent placement or selection is central to the meaning of test scores. Shepard (1993) claims that, as soon as test scores are used for any test-based decision, validation of the consequences of the decision is required.

Moss (1998) argues that testing occurs in a social context, supports social purposes, and the meanings of assessment scores are co-constructed by assessment developers and assessment users.

To the extent that the practices in which we engage change the social reality we study, the study of consequences becomes an essential aspect of validity even for those who choose to limit the scope of validity to a test-based interpretation. (p. 7)

Messick (1989) considers consequences only in terms of whether they are due to construct-under-representation or construct-irrelevant variance. Other validity theorists (Kane, 2006; Moss, 1998; Shepard, 1993) consider any unintended social consequences as a focus for validity research.



Part of the controversy about the role of consequences in validation research has to do with responsibility for validity research. Who is responsible for investigating the consequences of test score interpretation and use? Shepard (1997) distinguishes between consequences of the *planned use* of assessment scores and *misuse* of scores. She makes it clear that when scores are used for any unintended purpose, assessment users rather than assessment developers are responsible for validation research. Green (1998) notes there are many more uses of assessment scores than were ever intended by assessment developers, making research on consequences of the interpretation and use of scores an impossible task for developers. For example, Lane, Park, and Stone (1998) summarize the intended purposes of state level, standards-based achievement tests:

In general, statewide assessment programs are intended to have an impact on the following: the implemented curriculum; the instructional content and strategies; the content and format of classroom assessments; student, teacher, and administrator motivation and effort; the improvement of learning for all students; the nature of professional development support; teacher participation in the administration, development, and scoring of the assessment; student, teacher, administrator, and public awareness and beliefs about the assessment, criteria for judging performance, and the use of the assessment results; and the use and nature of test preparation materials. (p. 25)

Moss (1998) suggests that assessment developers and assessment researchers work in partnership with other social science researchers to develop a long-term research agenda that helps us understand “the potential slippage between what we well-meaningly intend and what we in fact effect” (p. 11) in assessment.

If we assume that the consequences of score interpretation and use are central validity issues, what strategies can be used to investigate social consequences? Are these strategies the same for score interpretation as they are for the use of scores?

### Value Implications of Score Interpretations

The first argument for Claim 5 is that the value implications of score interpretations are relevant to construct or criterion

performance. In Chapter 1 of this book, three types of claims were introduced: inferences, interpretations, and conclusions. Inferences are narrowly defined and tied closely to data; interpretations add a value component to the inference; conclusions draw upon multiple data points to make summary statements. It is the second type of claim that is the focus of this section—the implications of score interpretations.

In a way, this is the simpler of the consequential validity issues. Assessments are administered. Examinees complete the assessments. Scores or other summary statements are made. Inferences are drawn from scores to the construct. Interpretations of scores apply value judgments to the inferences. For example, a therapist administers a battery of tests to all patients who enter a psychiatric hospital. The battery includes the *Beck Depression Inventory* (BDI), the *Taylor Manifest Anxiety Scale* (TMAS), and the *Minnesota Multiphasic Personality Inventory* (MMPI). Once the patients complete the tests, the scores from the tests are used to determine each patient's profile on the MMPI, as well as their levels of depression and anxiety. An inference statement might be, "This patient has a high depression score; therefore, the patient is very depressed." An interpretive statement might be, "This patient is critically depressed and at risk of suicide." A conclusion statement might be,

Based on this patient's profile of scores for the MMPI, the TMAS, and the BDI, this patient is severely depressed with suicidal tendencies. She is socially introverted and expresses many symptoms of pain and discomfort. Most of her physical symptoms are probably psychosomatic due to her depression. Psycho-pharmaceutical intervention is recommended with close monitoring of her behavior over the next 48 hours.

It is evident that the first statement is closely tied to a single test score—an inference is made about the level of depression based on responses to test items. The second statement is an interpretation of that test score. The third is a conclusion based on multiple sources of data. When investigating whether the value implications of assessment scores are relevant to the construct or criterion performance, the validity questions focus on intended and unintended consequences of score interpretations. Intended consequences

are generally spelled out in assessment manuals or handbooks. Unintended consequences may be positive or negative.

For example, in 1972, Senator Thomas Eagleton was a candidate for the vice-presidency of the United States. It was revealed that he had been hospitalized for manic depression with suicidal tendencies. The presidential candidate who had selected Eagleton to be his running mate, Senator George McGovern, asked Senator Eagleton to step down, and invited another person, Sargent Shriver, to be his vice-presidential running mate. The intended consequence of Senator Eagleton’s diagnosis was to ensure that he received appropriate interventions for his depression. The unintended consequence was that a diagnosis of manic depression with suicidal tendencies harmed his political career.

Evaluation of the consequences of score interpretations begins with investigations of the interpretations themselves. Suppose “Dr. McKinley” wants to create an assessment of extroversion. She might write items like the ones in Figure 6–5. Suppose “Dr. Hsu” creates an assessment of hyperactive disorder. He, too, might write items similar to the ones in Figure 6–5. Investigations of the validity of inferences from assessment scores might support the use of these items for an assessment of extroversion or an assessment of hyperactive disorder. Yet, the interpretation of scores from an assessment of extroversion would probably be quite different from the interpretation of scores from an assessment of hyperactive disorder.

Kane (2006) and Shepard (1997) would suggest that negative social consequences such as lower academic self-esteem and lower achievement motivation are sufficient reasons to question the validity of score interpretations. In contrast, according to Messick (1989), this outcome is not a threat to the validity of scores unless the lower scores are caused by construct-irrelevant variance or construct under-representation.

I’m the life of the party	SA	A	D	SD
I like to be the center of attention	SA	A	D	SD
I like to watch exciting movies	SA	A	D	SD
I enjoy playing individual sports	SA	A	D	SD
I enjoy playing team sports	SA	A	D	SD
I have a lot of energy	SA	A	D	SD

Figure 6–5 Items from a Questionnaire

An example of this distinction can be found in studies comparing the mathematical performance of males and females. Many studies suggest that females do not perform as well as males on mathematics assessments involving non-routine problems (Backman, 1972; Benbow & Stanley, 1980, 1983; deWolf, 1981; Fennema & Carpenter, 1981; Fennema & Sherman, 1977; Hanna & Sonnenschein, 1985; Pallas & Alexander, 1983; Pattison & Grieve, 1984). Studies also suggest that females do better than males with computation (Fennema & Carpenter, 1981; Fennema & Sherman, 1978; Gallagher & DeLisi, 1994; Marshall, 1984; Threadgill-Sowder & Sowder, 1985). Results from studies such as these have led to characterizations of females as having lower mathematical abilities than males. The consequences of these interpretations have had implications for girls to view mathematics as a field for males (Ambady, Shih, Kim, & Pittinsky, 2001; Fear-Fenn & Kapostasy, 1992; Watt, 2000) and to influence females' course-taking patterns (Hill, Corbett, & St. Rose, 2010).

In a differential item functioning study by Taylor and Lee (2012), the researchers compared mathematics item performances of males and females over five years and three grade levels. They found that females did well on non-routine mathematics problems if the items were in constructed-response formats. In a DIF study by Lawrence (1995), she found that females performed as well as males on constructed-response Scholastic Aptitude Test I (SAT I) mathematics items. In addition to differential item functioning studies, research on stereotype threat (Spencer, Steele, & Quinn, 1999) suggests that, when females feel that they must perform well to prove themselves, they perform less well than females who complete the same tests without the stereotype threat.

The threat can be easily induced by asking students to indicate their gender before a test or simply having a larger ratio of men to women in a testing situation (Inzlicht & Ben-Zeev, 2000). Research consistently finds that stereotype threat adversely affects women's math performance to a modest degree (Nguyen & Ryan, 2008) and may account for as much as 20 points on the math portion of the SAT. (Walton & Spencer, 2009) (AAUW, 2010, p. 40)

These studies suggest that earlier interpretations of female students' mathematical problem-solving skills may have been due, in part, to construct-irrelevant variance. The purposes of the mathematics assessments were not to undermine females' academic performance; however, the value implications of low mathematics test performance are clearly visible in the studies that focused on why girls do not perform well.

Validity researchers must determine how they will conceptualize research on the value implications of score interpretations. Research should focus on threats to validity due to construct-irrelevant variance and construct under-representation. If low performance on a test is due to limited English proficiency, unfamiliarity with item formats, or anxiety, score inferences are likely to be faulty and could lead to inaccurate interpretations about examinees. In addition, researchers should consider alternate explanations of results. For the example in Figure 6-3, a label of "hyperactive disorder" would have very different value implications than a label of "extroverted" or "energetic." A close look at the demands of items and tasks, through both qualitative and quantitative means, is necessary. Studies that examine the relationship between items and criterion behaviors would help identify items that better predict hyperactivity versus extroversion.

The critical issue in terms of value implications is that assessment interpretations take on meaning beyond the assessment. As the studies on girls and mathematics suggest, the cultural interpretations of scores must be anticipated and examined. Whenever possible, assessment developers should anticipate the value implications of score interpretations, consider the possible consequences, and use caution when naming constructs in technical reports and score reports. Assessment users should exercise extreme caution when drawing conclusions based on test scores.<sup>5</sup> Finally, assessment researchers should consider negative consequences of score interpretations as potential focuses for validation research.

---

5. Studies of the consequences of score interpretations are likely to be found in journal articles and presentations at professional conferences rather than in technical reports provided by assessment developers.

## Consequences of the Use of Scores Are Appropriate

The second argument for Claim 5 is that the consequences of the uses of assessment scores are appropriate. Based on the debates about consequential bases for validation during the late twentieth century, it is evident that studies investigating the consequences of score use require collaboration among assessment developers, assessment researchers, and assessment users. Assessment developers cannot anticipate all possible uses of assessments; assessment users cannot anticipate all possible consequences. Critical here is evaluation of the intended consequences of intended uses of score interpretations.

If a school readiness test claims to measure which children are ready for regular kindergarten and which would benefit by waiting a year, then validity requires more than a correlation between test scores and school performance. It must be shown specifically that low-scoring children who spend an extra year in preschool or developmental kindergarten are more successful when they enter regular kindergarten (are better adjusted, learn more, etc.) than they would have been without the extra year. (Shepard, 1993, p. 406)

The point is that assessment developers and users need to show, not only a positive relationship between performance on a test and performance on the criterion (e.g., success in school), but also that the intended use benefits all parties.

Also critical are investigations into the *unintended* consequences of score uses. For example, suppose scores from a mathematics test are used to place middle-school students in mathematics courses. Moderate to low scores are interpreted to mean that students are not ready to take a pre-algebra course and should continue to study basic arithmetic. A potential negative social consequence of this score use is that students who are denied entrance to pre-algebra courses feel a social stigma, are frustrated by reiteration of basic mathematics content, lose motivation, and fall further behind their peers in mathematics.

Suppose a diagnostic reading comprehension test suggests that a group of students should be placed in a below-grade-level reading group focused on the development of reading fluency. A decision

to place students in a reading group focused on the pre-reading skills could restrict students' opportunities to learn reading comprehension and interpretation skills—which could have long-term negative effects on students' achievement. A validation study focused on social consequences would examine the appropriateness of the intervention and the consequences for students. For example, Riddle-Buly and Valencia (2002) administered a battery of reading assessments (e.g., word-attack skills, oral reading fluency, and comprehension) to students who did not meet state standards on a reading comprehension test. They conducted a cluster analysis of students' scores on the battery of tests and found that low-achieving students clustered into five or six different groups. Two of the groups had the expected pattern of reading skills; the patterns of reading skills for the other groups were quite different from what was expected by theory. For example, one group had excellent fluency and accuracy skills but did not comprehend what they read. Placement in reading instruction focused on reading fluency would provide a significant disservice to these students.

Unintended social consequences could also benefit examinees. Scores from the *Scholastic Assessment Test* (SAT; formerly *Scholastic Aptitude Test*) are used as indicators for admission decisions in many colleges and universities. An unintended consequence of this use is a plethora of programs that are available to help students prepare for the SAT. Most of these programs focus on test-taking skills and strategies, although some also include content reviews. Some preparation programs also guarantee that they will improve test scores. For example, Princeton Review (2013) guarantees examinees that, if students take the Princeton Review "Ultimate Classroom Course," their SAT scores will increase by at least 150 points.<sup>6</sup> Although these courses may assist students in raising their SAT scores, do their scores truly represent their scholastic achievement? What about students who do not have the funds for SAT preparation programs? Should their scores

---

6. Note that a report from the National Association for College Admissions Counseling (Griggs, 2009) reviews the results of over 30 studies on the effects of SAT coaching. The report suggests that score increases from coaching are between 10 and 20 points. The report also notes that score increases of 10 to 20 points can have a significant impact on the probability of admissions in a more- and less-selective universities.

be compared with the scores of those who have taken preparation courses—do they have the same meaning? Since this practice is widespread and well known, colleges and universities, as SAT score users, should conduct studies to determine whether the scores of students who have taken test preparation courses are as predictive of college success (e.g., freshman GPA) as the scores of students who did not take preparation courses.

The relationship between socio-economic status, ethnicity and gender, and SAT performance also has resulted in negative social consequences. For example, Gandara and Lopez (1998) found that high-achieving Latino students who were denied entry to college based on SAT scores were emotionally traumatized. A report from the California State Legislature (1999) suggested that SAT scores were closely tied to socio-economic status and under-predicted college performance for females and minorities. Other researchers have found similar results (e.g., Fincher, 2000; Kobrin, Milewski, Everson, & Zhou, 2003; Mattson, 2007; Rampell, 2009).

The story of the SAT is a prototypical example of the consequential issues related to assessment—both in terms of value implications and social consequences. The original intention of the SAT was to provide an unbiased measure of academic aptitude so that socio-economic status would not be such a strong factor in college admissions decisions. The value implications of SAT score interpretations are evident in the plethora of studies focused on the predictive value of SAT scores (e.g., Cohn, Cohn, Balch, & Bradley, 2004; Gayles, 2006; Hunter & Samter, 2000; Kobrin & Patterson, 2011; Mattern, Shaw, & Kobrin, 2011; Zwick & Schlemer, 2004; Zwick & Sklar, 2005), the influence of test preparation on SAT scores (see Griggs, 2009, for a summary), and anecdotal reports on the effect of SAT test preparation on students' thinking skills and perspectives about learning (Atkins, 2009; Ruenzel, 2004).

Despite the fact that empirical studies of SAT scores suggest the scores add little to the predictive power of college grade point average and that SAT scores are not equally predictive for all students, the United States' cultural obsession with SAT scores suggests value implications that have mythical proportions. Colleges have an obligation to consider the studies already conducted, as well as results from their own data, regarding whether scores from admission tests should play such a significant role in college admissions



decisions. To the extent that socio-economic status, ethnicity, and gender<sup>7</sup> play a role in SAT scores, data suggest that construct-irrelevant variance may contribute to students' scores. The impact of the SAT scores on high-school students' behaviors and on college admissions decisions are clear examples of intended and unintended social consequences of score interpretation and use.

### Summary of Validity Claim 5

To investigate the social consequences of assessment score interpretation and use, validation researchers must frame the research in terms of Messick's (1989) facets of validity or the validity claims from Table 5-1 (adapted from Kane, 2006). First, the researcher (or teacher, or policymaker, or assessment user) should look to the evidence (both logical arguments and empirical evidence) that supports the validity claims that scores can be trusted, that scores are reliable, and that one can extrapolate from the scores to the construct or criterion performance. In addition, test users should scrutinize data regarding alternative explanations for scores. Although close examination of item-level data or scoring models may seem burdensome, if assessment developers have not done their due diligence in the foundational work of assessment development, all interpretations and uses are suspect. Next, test users should consider whether the evidence provides substantive support for score meanings, interpretations, and uses. This requires a clear understanding of what interpretations are intended and whether the evidence provided supports those interpretations. It also requires clear and public statements about intended score uses. If it appears that assessment developers or users have "raked together a collection of correlations" (Cronbach, 1989, p. 155) with no substantive rationale for those correlations, if the evidence does not take into account both the arguments for and the arguments against the intended use, adverse consequences become more possible.

If the test user is satisfied that assessment developers have built a strong case for the intended interpretations and uses of assessment

---

7. For mathematics, females' SAT mathematics scores are significantly lower than males' SAT scores, despite the fact that females tend to have higher high-school mathematics grades and higher high-school and college grade point averages.

scores, research on consequences can focus on intended and unintended consequences. Studies might focus on the consequences of unintended uses (e.g., the observational assessment designed to assess kindergarten students' learning needs that is used to evaluate the quality of pre-kindergarten programs; the assessment designed to diagnose workplace efficiencies that is used to evaluate workers' skills) or the consequences of unintended interpretations (e.g., the assessment that is intended to diagnose depression and anxiety that is used to identify post-traumatic stress disorder; the assessment that is developed in one country and used, without adaptation and verification of internal structure, in another country). Given that assessments are so plentiful and those who are trained to understand their development, interpretations, and uses are so few, the validation researcher will never lack work to do in any field.

## Summary

This chapter has provided an overview and examples of the types of research that are needed to validate the most critical aspects of assessment—interpretation and uses of assessment scores. The chapter also addresses some examples designed to show how tracing the social consequences of interpretation and use is the responsibility of assessment developers and users. Assessment development work does not cease once assessments are published and in widespread use. Most assessment developers are fully aware that the structure and functions of assessments evolve as cultural mores and values change and as the demands of daily life change. Keeping a finger on the pulse of those changes is essential for scores to retain meaning or change meaning in intentional (rather than accidental or wildly inappropriate) ways.

Validation of scores, from documentation of the initial conception of the constructs through the test development processes to the studies that provide evidence to support score meaning, score interpretations, and score uses, requires care and intentionality. The work begins with clear ideas about the ultimate purpose of the test and the intended uses of test scores. Validation work continues as long as the test is in use. Chapter 7 provides resources that may be helpful to individuals who are interested in reading more or in learning more about validity theory and the validation methods described in Chapters 1 through 6.

## References

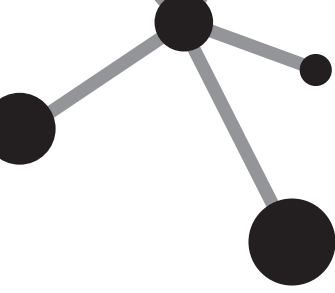
- Ambady, N., Shih, M., Kim, A., & Pittinsky, T. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science*, 12(5), 385–390.
- American Association of University Women (2010). *Why so few? Women in science, technology, engineering, and mathematics* Washington, DC: AAUW.
- American Educational Research Association, National Council on Measurement in Education, & American Psychological Association (1999). *Standards for Educational and Psychological Testing*. Washington, DC: AERA, NCME, & APA.
- Arndt, J., Greenberg, J., Pyszczynski, T., Solomon, S. (1997). Subliminal exposure to death-related stimuli increases defense of the cultural worldview. *Psychological Science*, 8, 379–385.
- Atkins, R. C. (2009, April). The new SAT: A test at war with itself. Paper presented at the annual meeting of the American Educational Research Association, San Diego.
- Backman, M. E. (1972). Patterns of mental abilities: Ethnic, socioeconomic, and sex differences. *American Educational Research Journal*, 9, 1–12.
- Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science*, 210, 1262–1264.
- Benbow, C. P., & Stanley, J. C. (1983). Sex differences in mathematical reasoning ability: More facts. *Science*, 222, 1029–1030.
- Ben-Shakkar, G., & Sanai, Y. (1991). Gender differences in multiple-choice tests: The role of guessing tendencies. *Journal of Educational Measurement*, 28, 23–35.
- Byrnes, J. P., & Takahira, S. (1993). Explaining gender differences on SAT-math items. *Developmental Psychology*, 29, 805–810.
- California State Legislature (1999, Oct.). The danger in overemphasizing the use of Scholastic Assessment Tests (SATs) as a tool for college admissions. Joint Hearing of the Senate Select Committee on Higher Education Admissions and Outreach and the Senate Committee on Education, Sacramento, CA.
- Carlton, S. T., & Harris, A. M. (1992). Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons. *ETS Research Report 92-64*. Princeton, NJ: Educational Testing Service.
- Cizek, G. J., & Bunch, M. (2007). *Standard Setting: A Practitioner's Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.
- Cohn, E., Cohn, S., Balch, D. C., & Bradley, J., Jr. (2004). Determinants of undergraduate GPAs: SAT scores, high school GPA and high school rank. *Economics of Education Review*, 23, 577–586.
- Contreras, S., Fernandez, S., Malcarne, V. L., Ingram, R. E., & Vaccarino, V. R. (2004). Reliability and validity of the Beck Depression and Anxiety Inventories in Caucasian Americans and Latinos. *Hispanic Journal of Behavioral Sciences*, 26, 446–462.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.; pp. 443–507). Washington, DC: American Council on Education.

- Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement Theory and Public Policy* (Proceedings of a symposium in honor of Lloyd G. Humphreys, pp. 147–171). Urbana: University of Illinois Press.
- deWolf, V. A. (1981). High school mathematics preparation and sex differences in quantitative abilities. *Psychology of Women Quarterly*, 5, 555–567.
- Dietz, S. (2010, Dec.). *State High School Tests: Exit Exams and Other Assessments*. Washington, DC: Center for Educational Policy.
- Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential problem performance in mathematics achievement problems. *Journal of Educational Measurement*, 24, 157–166.
- du Toit, R., & de Bruin, G. P. (2002). The structural validity of Holland's R-I-A-S-E-C model of vocational personality types for young black South African men and women. *Journal of Career Assessment*, 10, 62–77.
- Equal Employment Opportunity Commission (1978). *Uniform Guidelines on Employment Selection Procedures*. Washington, DC: EEOC.
- Fear-Fenn, M., & Kapostasy, K. K. (1992). *Math + Science + Technology = Vocational Preparation: A Difficult Equation to Balance. Report from the Ohio State University, Columbus, Center for Sex Equity* Columbus, OH: Ohio State Department of Education.
- Fennema, E., & Carpenter, T. E. (1981). Sex-related differences in mathematics: Results from the national assessment. *Mathematics Teacher*, 74, 554–559.
- Fennema, E., & Sherman, J. (1977). Sex related differences in mathematics achievement, spatial visualization and affective factors. *American Educational Research Journal*, 14, 51–71.
- Fincher, C. (2000, July). Assessment uses of the SAT in the university system of Georgia: IHE perspectives. Paper presented at the annual meeting of the International Association for Educational Assessment, Oxford, England.
- Gallagher, A. M. & De Lisi, R. (1994). Gender differences in Scholastic Aptitude Test – mathematics problem solving among high-ability students. *Journal of Educational Measurement*, 86, 204–211.
- Gandara, P., & Lopez, E. (1998). Latino students and college entrance exams: How much do they really matter? *Hispanic Journal of Behavioral Sciences*, 20, 17–38.
- Garner, M., & Engelhard, G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, 12, 29–51.
- Gayler, K. (2003). *State High School Exams 2003 Annual Report: Put to the Test*. Washington, DC: Center on Education Policy.
- Gayles, J. (2006). Race, graduating performance, and admissions: Georgia State University's freshman index. *College and University*, 82, 27–34.
- Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practices*, 17(2), 16–19.
- Griggs, D. C. (2009). *Preparation for College Admissions Exams: 2009 NACAC Discussion Paper*. Arlington, VA: National Association of College Admissions Counselors.

- Hanna, G. S., & Sonnenschein, J. L. (1985). Relative validity of the Orleans-Hanna Algebra Prognosis Test in the prediction of girls' and boys' grades in first-year algebra. *Educational and Psychological Measurement*, 45, 361–367.
- Hill, C., Corbett, C., & St. Rose, A. (2010). *Women in Science, Technology, Engineering and Mathematics*. Washington, DC: American Association of University Women.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item Response Theory: Application to Psychological Measurement*. Homewood, IL: Dow Jones-Irwin.
- Hunter, J. G., & Samter, W. (2000). A college admission test protocol to mitigate the effects of false negative SAT scores. *Journal of College Admission*, 168, 22–29.
- International Test Commission (2010). Guidelines for translating and adapting tests. Retrieved April 26, 2013 from <http://www.intestcom.org/Guidelines/Adapting+Tests.php>
- Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science*, 11(5), 365–371.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.; pp. 17–64). Washington, DC: American Council on Education.
- Kobrin, J. L., & Patterson, B. F. (2011). Contextual factors associated with the validity of SAT scores and high school GPA for predicting first-year college grades. *Educational Assessment*, 16, 207–226.
- Kobrin, J. L., Milewski, G. B., Everson, H., & Zhou, Y. (2003, April). An investigation of school-level factors for students with discrepant high school GPA and SAT scores. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.
- Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice*, 17(2) Summer, 24–28.
- Lane, S., Wang, N., & Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practices*, 15(4), 21–27, 31.
- Lau, A. L. D., Cummins, R. A., & McPherson, W. (2005). An investigation into the cross-cultural equivalence of the Personal Wellbeing Index. *Social Indicators Research*, 72, 403–430.
- Lawrence, I. M. (1995). *DIF Data on Free-Response SAT I Mathematical Items*. Princeton, NJ: Educational Testing Service Report Number ETS-RR-95–22.
- Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, 16(2), 14–16.
- Lissitz, R. W. & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36, 437–448.
- Mattern, K. D., Shaw, E. J., & Kobrin, J. L. (2011). An alternative presentation of incremental validity: Discrepant SAT and HSGPA performance. *Educational & Psychological Measurement*, 71, 638–662.
- Marshall, S. E (1984). Sex differences in children's mathematics achievement: Solving computations and story problems. *Journal of Educational Psychology*, 76, 194–204.

- Mattson, C. E., (2007). Beyond admission: Understanding pre-college variables and the success of at-risk students. *Journal of College Admission*, 196, 8–13.
- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16(2), 16–18.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed.; pp. 13–103). Washington DC: American Council on Education.
- Moss, P. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practices*, 17(2), 6–12.
- Nguyen, H.-H. H., & Ryan, A. M. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, 93(6), 1314–1334.
- Pallas, A. M., & Alexander, K. L. (1983). Sex differences in quantitative SAT performance: New evidence on the differential coursework hypothesis. *American Educational Research Journal*, 20, 165–182.
- Pattison, P., & Grieve, N. (1984). Do spatial skills contribute to sex differences in different types of mathematical problems? *Journal of Educational Psychology*, 76, 678–689.
- Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice*, 16, 9–13.
- Princeton Review (2013). *Princeton Review money back guarantee* Retrieved April 27, 2013 from <http://www.princetonreview.com/guarantee-sat-act.aspx>.
- Rampell, C. (2009, Aug.). SAT scores and family income. *New York Times*. Retrieved March 18, 2012, from <http://economix.blogs.nytimes.com/2009/08/27/sat-scores-and-family-income/>.
- Riddle-Buly, M., & Valencia, S. (2002). Below the bar: Profiles of students to fail state reading assessments. *Educational Evaluation and Policy Analysis*, 24, 219–239.
- Ruenzel, D. (2004). Point of view—The SAT and the assault on literature. *Phi Delta Kappan*, 86, 247–248.
- Shepard, L. A. (1989). A review of research on kindergarten retention. In L. A. Shepard & M. L. Smith (Eds.), *Flunking Grades: Research and Policies on Retention*. London: Falmer Press.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405–450.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16, 5–8, 13, 24.
- Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 35(1), 4–28.
- Taylor, C. S., & Lee, Y. (2004, Dec.). Is reading a dimension in the WASL mathematics test? Using differential item functioning to examine the dimensions of WASL mathematics. Presented at the Washington State Assessment Conference, Seattle, WA.
- Taylor, C. S., & Lee, Y. (2012). Gender DIF in tests with mixed item formats. *Applied Measurement in Education*, 25, 246–280.
- Thi-Xuan-Hanh, V., Guillemin, F., Dinh-Cong, D., Parkerson, G. R., Bach-Thu, P., Tu-Quynh, P., & Briancon, S. (2005). Health related quality of life of adolescents in Vietnam: Cross-cultural adaptation and validation of the Adolescent Duke Health Profile. *Journal of Adolescence*, 28, 127–146.

- Threadgill-Sowder, J., Sowder, L., Moyer, J. C., & Moyer, M. B. (1985). Cognitive variables and performance on mathematical story problems. *Journal of Experimental Education*, 54, 56–62.
- United States Congress (2003). No Child Left Behind Act—Reauthorization of the Elementary and Secondary Education Act. Retrieved September 14, 2008, from <http://www2.ed.gov/nclb/landing.jhtml>.
- Walker, C., & Beretvas, N. S. (2000, April). Using multidimensional versus unidimensional ability estimates to determine student proficiency in mathematics. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*, 20(9), 1132–1139.
- Watt, H. M. G. (2000, Dec.). Exploring perceived personal and social gender stereotypes of maths with secondary students: An explanation for continued gender differences in participation? Paper presented at the annual meeting of the Australian Association for Research in Education, Sydney.
- Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), *Improving Inquiry in the Social Sciences: A Volume in the Honor of Lee J. Cronbach* (pp. 75–107). Hillsdale, NJ: Erlbaum.
- Yan, E., Tang, C. S., & Chung, T. (2010). Validation of the Perinatal Grief Scale for use in Chinese women who have experienced recent reproductive loss. *Death Studies*, 34, 151–171.
- Zwick, R., & Schlemer, L. (2004). SAT validity for linguistic minorities at the University of California, Santa Barbara. *Educational Measurement: Issues and Practice*, 23(1), 6–16.
- Zwick, R., & Sklar, J. C. (2005). Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language. *American Educational Research Journal*, 42, 439–464.



---

## VALIDITY THEORY AND VALIDATION RESOURCES

IN THIS CHAPTER, I provide a list of resources that can help validity researchers along their way to developing strong validation programs. The sections are organized along the lines of the chapters in this book: philosophy of science, validity theory, validation in research, construct-related evidence for validity, and validation of score interpretation, use, and consequences. Most of the resources described here were referenced in relevant sections of the book.

### **Philosophy of Science**

The following are key philosophers who have influenced our thinking about the nature of validity. Their perspectives bear upon the strategies we use to gather evidence for validity.

#### Logical Positivism and Instrumentalism

Auguste Comte is considered the founder of sociology and positivism. His focus was on theoretical relationships among constructs. The key idea of positivism is that evidence should be



brought to bear to support theoretical claims. The main work that outlines his views is *A General View of Positivism* (originally published in 1844; reissued by Cambridge University Press in 2009).

Ernst Mach was a major contributor to ideas about positivism. He focused on the use of experimental events to define scientific laws as well as the mathematical relationships among variables. His writings can be found in a volume by Blackmore (1972).

Carl Hempel was known for the deductive nomological model of scientific explanation, wherein theoretical propositions are arranged for purposes of explanation and investigation. His best-known work was *Philosophy of Natural Science* (1966).

### Empirical Falsification

Karl Popper was a critic of the “conformationist” (Messick, 1989) view of the scientific endeavor. His main stance was that theories could never be proven but could be subjected to scientific tests. If scientific tests do not support theoretical claims, the theory is disproven. Popper is sometimes considered a rationalist and a realist. Popper’s main works related to philosophy of science were *Realism and the Aim of Science* (originally published in 1983), *Conjectures and Refutations: The Growth of Scientific Knowledge* (1963), and *Objective Knowledge: An Evolutionary Approach* (1972, 1979).

### Rationalism

René Descartes focused on the role of reasoning in scientific knowledge. In *Discourse on the Method* (1637), Descartes proposed fundamental principles that can be known without a doubt. He focused on logical reasoning and rejected ideas that could be doubted or logically disproven. Another important work related to philosophy of science by Descartes was *Principles of Philosophy* (1644).

Steven Toulmin focused on the logical-argument aspect of scientific thinking. He proposed a five-part model for a scientific argument:

1. *Claim*: the position or *claim* being argued for
2. *Grounds*: *reasons* or supporting *evidence* that bolster the claim

3. *Warrant*: the principle, provision, or *chain of reasoning* that connects the grounds/reason to the claim
4. *Backing*: *support, justification, reasons* to back up the warrant
5. *Rebuttal/reservation*: exceptions to the claim; description and *rebuttal of counter-examples* and *counter-arguments*

Many of these terms should be familiar from reading the earlier chapters of this book. Throughout this book, the term *claim* has been used to describe any inference, interpretation, or conclusion that is being posited. The idea of using evidence to support a claim is essential. The notion of a *chain of reasoning* is suggested in previous chapters, both in terms of the theoretical rationales that are the foundation of validity claims and in terms of the interpretations that support validation data. *Support* and *justification* are referenced throughout the preceding chapters. And, finally, the idea of investigating counter-arguments (alternate explanations) has been discussed throughout the text. Toulmin's key works are *An Examination of the Place of Reason in Ethics* (1950) and *The Uses of Argument*, 2nd ed. (1958, 2003).

## Relativism

Paul Karl Feyerabend was a radical relativist. He rejected the idea of any universal methodological rules and focused his writings on the ways in which human perspectives and methodologies influenced the results of investigations. His key works were *Against Method* (1975), *Farewell to Reason* (1987), *Problems of Empiricism* (1981), and *Knowledge, Science, and Relativism* (1999).

Thomas Kuhn was a relativist who introduced the idea of scientific revolution (cataclysmic change caused by the shifting of scientific paradigms). He claimed that, as anomalies to theoretical predictions mount, scientists face fierce resistance to new theories. Alternate explanations (Kuhn's competing paradigms) are frequently incompatible. When a new theory prevails, it is not an evolutionary process but, rather, a transformational process resulting in a whole new view of phenomena. He also raised the relativist perspective that methodologies are bound by theory and that results of investigations are influenced by theory-bound methodologies. His most famous work is *The Structure of Scientific Revolution* (1996).

## Rationalism

David Hume focused on understanding human behavior and was known for his empiricism and skepticism. His most famous work was *A Treatise on Human Nature* (1739).

Immanuel Kant attempted to resolve conflicts between purely empirical and purely rational (theoretical) approaches to the scientific endeavor. He claimed that empiricists believed that all knowledge comes through experience whereas, rationalists believed that reason and innate ideas come first. Kant argued that experience that is not processed by reason is purely subjective and reason without evidence would only lead to theoretical illusions. His most important related works were *Critique of Pure Reason* (1781, 1999) and *Critique of Practical Reason* (1788, 1997).

Dudley Shapere is well known for his critique of Kuhn's *Structure of Scientific Revolution* (Shapere, 1964). He was particularly critical of the notion of *paradigm*—an all-encompassing frame that guides current research and that must build sufficient force to shatter older paradigms. Shapere claimed that scientists consider multiple competing theories simultaneously. He also declared that theory replacement constitutes a gain in knowledge (1989). He stated:

[An] adequate philosophy of science must show how it is possible that we might know, without guaranteeing that we must know. (1982, p. 83)

Shapere's requirements for a philosophy of science are that the philosophy must:

1. Acknowledge that the evaluative standards and procedures are subject to change.
2. Accommodate the possibility that knowledge is achieved within science.
3. "Preserve the objectivity and rationality of science" (Shapere, 1984) and exhibit rationality and objectivity.

## General Philosophy of Science

One final resource related to the philosophy of science is an edited volume by Frederick Suppe (1977) entitled *Structure of Scientific Theories*. Essays by many of the philosophers discussed here are found in the book.

## Validation in Research

The best sources for both method and philosophy regarding internal and external threats to validity can be found in three volumes *Experimental Design* (Campbell & Stanley, 1966) and *Experimental and Quasi-Experimental Designs for Research* (Cook & Campbell, 1979), and *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Shadish, Cook, & Campbell, 2002). These books provide a background in validity theory from the perspective of research. They include details about a full range of experimental and quasi-experimental designs and a careful look at internal and external threats to validity. In addition, many books on educational and psychological research-methods provide reviews of research designs and how they address threats to validity.

In 2011, a set of articles was published in the journal *New Directions in Evaluation*, edited by Chen, Donaldson, and Mark (2011). The purpose of the articles was to critique Shadish, Cook, and Campbell's work from a variety of perspectives (e.g., program evaluation, measurement). The authors (Chen & Garbe, 2011; Chen, Donaldson, & Mark, 2011; Gargani & Donaldson, 2011; Greene, 2011; House, 2011; Julnes, 2011; Mark, 2011; Reichardt, 2011) consider a range of issues regarding validation work. Shadish (2011) responds to their concerns in the final article of the journal edition.

## Statistical Methods

Throughout this book, I have discussed a variety of research methods. Most of the statistical models (multiple regression, analysis of variance, analysis of covariance) are explained in basic statistics textbooks (e.g., Garner, 2010; Howell, 2011; Urdan 2010). Excellent sources for learning about hierarchical linear modeling (HLM) and structural equation modeling (SEM) are the manuals for HLM and SEM programs (e.g., *Hierarchical Linear Models: Applications and Data Analysis* by Raudenbush & Bryk, 2002; *Structural Equation Modeling with EQS* by Byrne, 2006) as well as textbooks (see, for example, Hox, 2002; Kline, 2010).

## Validity Theory

Resources for reading about validity theory have also been noted in these chapters. Key among them are Cronbach and Meehl's

(1955) article on construct validity; Messick's (1989) landmark chapter entitled "Validity" in the third edition of *Educational Measurement* (Linn, 1989); and Kane's (2006) chapter entitled "Validation" in the fourth edition of *Educational Measurement* (Brennan, 2006). Other key resources are Cronbach's (1971) chapter entitled "Test Validation" in *Educational Measurement*, 2nd ed.; Shepard's (1993) chapter "Evaluating Test Validity" in the *Review of Educational Research*; and Moss's (1994) article "Can There Be Validity Without Reliability?" in *Educational Researcher*.

The key reading in terms of the debate over consequential evidence for validity can be found in several issues of *Educational Measurement: Issues and Practices* (volumes 14, 16, and 17). Perspectives in this journal include those of test publishers (Green, 1998), test developers (Reckase, 1998; Yen, 1998), practitioners (Taleporos, 1998), researchers (Lane, Parke, & Stone, 1998; Lane & Stone, 2002), and philosophers (Popham, 1997; Linn 1997, 1998; Mehrens, 1997; Moss, 1998; Shepard, 1997).

Beyond concerns about consequential issues related to validity, critiques of the "unitarian" view of validity offered by Messick (1989) have been presented (Boorsboom, Mellenbergh, & van Heerden, 2004; Lissitz & Samuelsen, 2007). An entire issue of *Educational Researcher* was dedicated to the critique by Lissitz and Samuelsen and responses from different perspectives in the assessment community (Embretson, 2007; Gorin, 2007; Mislevy, 2007; Moss, 2007; Sireci, 2007).

The most important resource for practical validation guidelines is the *Standards for Educational and Psychological Testing* (AERA, NCME, APA, 1999). These standards provide chapters on a range of issues related to assessment—validity, reliability, responsibilities of test takers, and responsibilities of test users. In addition, separate chapters focus on validity issues related to educational and psychological testing, testing of students with disabilities, assessment of English language learners, and assessment for certification and licensure. Interpretations of the *Standards* have been provided by various organizations (e.g., *Principles for the Validation and Use of Personnel Selection Procedures*); however, many are not complete, and some lack the authority and credibility of the *Standards*. A draft of the next edition of the *Standards* is currently under review, and the new edition is due to be released in 2013 or 2014.

## Resources for Assessment Development Work

One of the best sources of information about assessment development and research is the fourth edition of *Educational Measurement* (Brennan, 2006). Each chapter is written by one of the leading thinkers about the assessment issue and contains the most current knowledge about each topic. Resources for the statistical analyses described in Chapter 5 include *Best Test Design* (Wright & Stone, 1979), which explains basic item response theory (IRT), and *Rating Scale Analysis* (Wright & Masters, 1982), which explains partial credit modeling for use with ratings scales and item that are scored with rubrics. More complete treatments of item response theory can be found in *Fundamentals of Item Response Theory* (Hambleton, Swaminathan, & Rogers, 1991), *Handbook of Polytomous Item Response Theory Models* (Nehring & Ostini, 2006), and *Item Response Theory* (Embretson & Reise, 2000). Information about classical test theory (CTT)—the method used in the development of most psychological tests—can be found in *Test Theory: A Unified Treatment* (McDonald, 1999), *Introduction to Classical and Modern Test Theory* (Crocker, 2006), and *Introduction to Measurement Theory* (Allen & Yen 1979).

## Summary

The resources listed here can provide an opportunity for deep study of the concept of validity from a variety of perspectives. The chapters in this book are intended to provide both theory and practical applications of the ideas described in these sources. Considerations of issues of validity are weighty and merit attention. Considerations of validity are critical in making certain that claims from research and assessment can be trusted. Those who think and write about validity probably consider the weight of the issue more than is common in the business of research and assessment development today. Hopefully, reading this book will give you a sense of standards for research and assessment, will make you more thoughtful about your research plans, and will help you be a more conscious consumer of others' research and assessment claims.

## References

- Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Pacific Grove, CA: Brooks-Cole Publishers.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: AERA, APA & NCME.
- Blackmore, J. T. (1972). *Ernst Mach—His Life, Work, and Influence*. Bethesda, MD: Sentinel Open Press.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.
- Brennan, R. L. (2006). *Educational Assessment*, 4th ed Washington, DC: American Council on Education and Oryx Press on Higher Education.
- Byrne, B. M. (2006). *Structural Equation Modeling with EQS*. Newbury Park, CA: Sage Publications.
- Campbell, D., & Stanley, J. (1966). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.
- Chen, H. T., & Garbe, P. (2011). Assessing program outcomes from the bottom-up approach: An innovative perspective to outcome evaluation. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing Validity in Outcome Evaluation: Theory and Practice. New Directions for Evaluation*, 130, 93–106. San Francisco: Wiley Publications.
- Chen, H. T., Donaldson, S. I., & Mark, M. M. (2011). Validity frameworks for outcome evaluation. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing Validity in Outcome Evaluation: Theory and Practice. New Directions for Evaluation*, 130, 5–16. San Francisco: Wiley Publications.
- Comte, A. (1844, 2009). *A General View of Positivism*. Cambridge, England: Cambridge University Press.
- Cook, T., & Campbell, D. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.
- Crocker, L., & Algina, J. (2006). *Introduction to Classical and Modern Test Theory*. Independence, KY: Wadsworth-Cengage Publishing.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.
- Descartes, R. (1637, 2008). *Discourse on the Method*. Translation by I. Maclean. London: Oxford University Press.
- Descartes, R. (1644). *Principles of Philosophy* (J. Bennett, Trans.). Retrieved April 19, 2013 from [http://www.earlymoderntexts.com/f\\_descarte.html](http://www.earlymoderntexts.com/f_descarte.html).
- Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, 36, 449–455.
- Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Feyerabend, P. K. (1975). *Against Method*. London: Verso—New Left Books.



- Feyerabend, P. K. (1981). *Problems with Empiricism: Philosophical Papers*, Vol. 2. Cambridge, England: Cambridge University Press.
- Feyerabend, P. K. (1981). *Realism, Rationalism, and the Scientific Method: Philosophical Papers*, Vol. 1. Cambridge, England: Cambridge University Press.
- Feyerabend, P. K. (1987). *Farewell to Reason*. London: Verso—New Left Books.
- Feyerabend, P. K. (1999). *Knowledge, Science, and Relativism: Philosophical Papers*, Vol. 2. Cambridge, England: Cambridge University Press.
- Garner, R. (2010). *The joy of stats: A short guide to introductory statistics*. Toronto: University of Toronto Press.
- Gargani, J., & Donaldson, S. I. (2011). What works for whom, where, why, for what, and when? Using evaluation evidence to take action in local contexts. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing Validity in Outcome Evaluation: Theory and Practice. New Directions for Evaluation*, 130, 17–30.
- Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher*, 36, 456–462.
- Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice*, 17 (2), 16–19, 34.
- Greene, J. C. (2011). The construct(ion) of validity as argument. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing Validity in Outcome Evaluation: Theory and Practice. New Directions for Evaluation*, 130, 81–91.
- Hambleton, R. J., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.
- Hempel, C. G. (1966). *Philosophy of Natural Science*. Englewood Cliffs, NJ: Prentice Hall.
- House, E. R. (2011). Conflict of interest and Campbellian validity. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing Validity in Outcome Evaluation: Theory and Practice. New Directions for Evaluation*, 130, 69–80.
- Howell, D. C. (2011). *Fundamental statistics for the behavioral sciences* (6th ed.). Belmont, CA: Thompson-Wadsworth.
- Hox, J. J. (2002). *Multilevel Analysis: Techniques and Applications*. New York: Routledge.
- Hume, D. (1739, 2000). *A Treatise of Human Nature*, D. F. Norton and M. J. Norton, Eds. Oxford, England: Oxford University Press.
- Julnes, G. (2011). Reframing validity in research and evaluation: A multidimensional, systematic model of valid inference. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing Validity in Outcome Evaluation: Theory and Practice. New Directions for Evaluation*, 130, 55–67.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education.
- Kant, I. (1781, 1999). *Critique of Pure Reason*. Cambridge, England: Cambridge University Press.
- Kant, I. (1788, 1997). *Critique of Practical Reason*. Cambridge, England: Cambridge University Press.
- Kline, R. (2010). *Principles and Practice of Structural Equation Modeling*, 3rd ed. New York: The Guilford Press.



- Kuhn, T. (1996). *Structure of Scientific Revolution*, 3rd ed. Chicago: University of Chicago Press.
- Lane, S., Parke, C., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice*, 17(2), 24–28.
- Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice*, 21(1), 23–30.
- Linn, R. L. (1989). *Educational assessment* (3rd ed). Washington, DC: American Council on Education.
- Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, 16(2), 14–16.
- Linn, R. L. (1998). Partitioning responsibility for the evaluation of consequences of assessment programs. *Educational Measurement: Issues and Practice*, 16(2), 28–30.
- Lissitz, R. W., & Samuels, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36, 437–448.
- Mark, M. M. (2011). New (and old) directions for validity concerning generalizability. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing Validity in Outcome Evaluation: Theory and Practice. New Directions for Evaluation*, 130, 31–42.
- McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Upper Saddle River, NJ: Lawrence Erlbaum Associates.
- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16(2), 16–18.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). Washington DC: American Council on Education.
- Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, 36, 463–469.
- Moss, P. A. (1994). Can there be validity without reliability. *Educational Researcher*, 23(2), 4–12.
- Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17(2), 6–12.
- Moss, P. A. (2007). Reconstructing validity. *Educational Researcher*, 36, 470–476.
- Nehring, M. L., & Ostini, R. (2006). *Polytomous Item Response Theory Models*. Thousand Oaks, CA: Sage Publications.
- Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice*, 16(2), 9–13.
- Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Google Books. Retrieved March 11, 2012, from [http://books.google.com/books/about/Conjectures\\_and\\_refutations.html?id=IENmxiVBaSoC](http://books.google.com/books/about/Conjectures_and_refutations.html?id=IENmxiVBaSoC).
- Popper, K. (1972, 1979). *Objective Knowledge: An Evolutionary Approach*. Google Books. Retrieved, March 11, 2012, from [http://books.google.com/books/about/Objective\\_knowledge.html?id=o8oPAQAIAAJ](http://books.google.com/books/about/Objective_knowledge.html?id=o8oPAQAIAAJ).
- Popper, K. (1983, 1996). *Realism and the Aim of Science*. Oxford, England: Routledge Press.

- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications.
- Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice*, 17(2), 13–16.
- Reichardt, C. S. (2011). Criticisms of and an alternative to the Shadish, Cook, and Campbell validity typology. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing Validity in Outcome Evaluation: Theory and Practice. New Directions for Evaluation*, 130, 43–53.
- Shadish, W. R. (2011). The truth about validity. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing Validity in Outcome Evaluation: Theory and Practice. New Directions for Evaluation*, 130, 107–117.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.
- Shapere, D. (1964). The structure of scientific revolutions. *The Philosophical Review*, 72, 383–394.
- Shapere, D. (1982). Reason, reference, and the quest for knowledge. *Philosophy of Science*, 49, 1–23.
- Shapere, D. (1984). *Reason and the Search for Knowledge*. Boston: D. Reidel Publishing Company.
- Shapere, D. (1989) Evolution and continuity in scientific change, *Philosophy of Science*, 56, 419–437.
- Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammon (Ed.), *Review of Research in Education*, Vol. 19 (pp. 405–450). Washington, DC: AERA.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5–8.
- Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, 36, 477–481.
- Suppe, F. (1977). *Structure of Scientific Theories*. Chicago: Illini Press.
- Taleporos, E. (1998). Consequential validity: A practitioner's perspective. *Educational Measurement: Issues and Practice*, 17(2), 20–23, 34.
- Toulmin, S. (1950). *An Examination of the Place of Reason in Ethics*. Cambridge, England: Cambridge University Press.
- Toulmin, S. (1958, 2003). *The Uses of Argument*. Cambridge, England: Cambridge University Press.
- Urdan, T. C. (2010). *Statistics in plain English* (3rd ed.). New York: Routledge, Taylor and Francis.
- Wright, B., & Masters, G. (1982). *Rating Scale Analysis*. Chicago: MESA Press.
- Wright, B., & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press.
- Yen, W. M. (1998). Investigating the consequential aspects of validity: Who is responsible and what should they do? *Educational Measurement: Issues and Practice*, 17(2), 5.

*This page intentionally left blank*

---

# INDEX

- achievement tests, standards-based, 174
- alpha level, 67
- alternative hypotheses, 66
  - failing to reject, 66
  - option of rejecting, 66
- ambiguity of results, 12
- assessment. *See also specific topics*
  - defined, 3n2
  - vs. measurement, 3n2
  - validation and, 16–21
- assessment development, validation
  - in, 4–5
- assessment development work,
  - resources for, 195
- assessment scores. *See also scores; specific topics*
  - interpretation, use, and consequences of, 20
  - planned use vs.
    - misuse of, 174
  - used for new purposes, 169
- assessment tools, 84
  - support for intended uses, 165–66
- base model, 50, 50f
- bias and sensitivity reviews, 131–32
- block design
  - controlling for internal threats to validity through, 34f, 34–35, 37f, 37–39
  - covariate design combined with, 37f, 37–39
- claim(s), 190
  - and construct validation, 84–86
  - defined, 2
- comparative fit index (CFI), 49
- Comte, Auguste, 189–90
- conclusions, defined, 2
- confirmatory factor analysis (CFA), 130, 135, 157, 160
- construct-irrelevant variance, 18, 130–41
- construct-related evidence for validity, 19–20
- construct-related validity claims,
  - gathering evidence to support, 88–90, 96. *See also* validation research: claims that should guide

- construct validation, 143–44. *See also*  
    construct-related evidence  
    for validity; construct-related  
    validity claims  
claims and, 84–86  
design of experimental study for,  
    129, 130f  
types of evidence needed in, 130,  
    142
- construct(s), 48, 83. *See also* criterion  
    performance(s)  
definition of, 4  
    evaluation of the, 113–15  
over- and under-representation of,  
    19, 20, 172, 173, 178  
synonyms and related terms, 4n3,  
    83
- convergent evidence, 125
- correlational studies, 10–11, 126–28.  
    *See also under* internal threats  
    to validity
- covariate design, combining block  
    design with, 37f, 37–39
- criterion performance(s), 83–84  
    alignment with, 125–30  
    evaluation of the definition of,  
        113–15  
    extrapolating from score to, 124–43
- criterion-referenced decisions, 122
- criterion-referenced scoring model,  
    101
- criterion-referenced test (CRT), 135,  
    136, 136t, 137, 138t, 139
- Cronbach, Lee, 16, 142, 173
- cultures, score interpretations  
    investigating across, 157–61
- cut scores, 122–24, 161–64
- data collection and treatment,  
    interaction between, 60
- data dredging. *See* “fishing for results”
- decision consistency, 122–24, 123t
- decisions. *See also* score  
    interpretations; selection  
    decisions; *specific topics*  
    criterion-referenced, 122
- demand characteristics, 28–29
- dependent variables, 26
- Descartes, René, 190
- differential item functioning (DIF),  
    132–35, 152–53, 177  
    defined, 133
- diffusion of treatment, 12
- discriminant evidence, 131
- double-blind studies, 29
- effect size, 68–69  
    defined, 68
- empirical falsification. *See* falsification
- endogenous variables, 48
- equalization of treatment, 13
- error, 49, 66–67. *See also specific topics*  
    estimating, 121  
    experiment-wise, 15, 73–74  
    minimizing, 119–20  
    Type I, 66, 70, 71  
    Type II, 67, 70, 73
- estimating error, 121
- examinees, observation of, 125–26
- exogenous variables, 48
- experiment-wise error, 15, 73–74
- experimental design(s)  
    to control internal threats to  
        validity, 27–39  
    limits of, 29–31  
    simple, 27, 27f
- experimental research, 10–11, 129,  
    140–41, 151–52. *See also*  
    quasi-experimental research
- experimenter effect, 29
- explanatory designs, 40
- explanatory methods, 41
- exploratory factor analysis (EFA),  
    135, 137
- external validity, 14–15, 61, 63–64.  
    *See also* generalizability;  
    interaction(s)  
    controlling for, through research  
        designs, 61–63
- factor analysis, 129–30, 135, 136f, 137,  
    157, 160
- factors, 48
- false negative rate ( $\beta$ ), 70

- falsification, 9, 13, 66, 190  
 feasibility (experimental design), 31  
 Feyerabend, Paul Karl, 191  
 “fishing for results,” 74
- generalizability, 121–22. *See also*  
     external validity  
     interactions and, 56–61  
     populations, samples, and, 55–56  
 generalizability studies (G studies),  
     121–22
- Hawthorne effect, 28, 29  
 Hempel, Carl, 190  
 hierarchical linear modeling (HLM),  
     46, 47, 50, 63, 193. *See also*  
     multilevel modeling
- history, 12  
     interaction with treatment, 14  
 Hume, David, 192
- imitation of treatment, 12  
 independent variables, 26  
 inferences, 1–2  
     defined, 1  
 instrumentalism, 9, 189–90  
 instrumentation, 12  
 interaction(s)  
     between data collection and  
         treatment, 60  
     between historical events and  
         treatment, 60–61  
     between history and treatment, 14  
     between selection and treatment,  
         14, 56–57  
     between situation and treatment,  
         57–59  
     between testing and treatment, 14  
     between treatments, 14, 59  
     generalizability and, 56–61  
 internal threats to validity  
     categories of, 26  
     under conditions of random  
         selection and random  
         assignment, 28–29  
     correlational designs for  
         addressing, 40–53  
     experimental designs to control, 27–39  
     quasi-experimental design strategies  
         for controlling, 39–40  
 internal validity, 11–14. *See also*  
     internal threats to validity  
     defined, 11  
 interpretation (of statistical  
     results), 1–2. *See also* score  
     interpretations  
     defined, 1  
     over- and under-interpretation, 65,  
         76–79
- item analyses, 160–61. *See also*  
     representativeness of items  
     and tasks
- item characteristic curves, 134, 134f  
 item difficulty, 104  
 item popularity, 104  
 item response theory (IRT), 100, 104,  
     105, 107, 123, 166, 195
- items and tasks  
     evaluation of item/task  
         specifications, 115–18f,  
         117–18f  
     item/task content reviews, 116,  
         118–19  
     representativeness, 112–19
- Kane, M., 89–90, 96, 112, 142, 143,  
     148, 149  
 Kant, Immanuel, 192  
 Kuhn, Thomas S., 191
- Lane, S., 174  
 language structure, 78  
 languages, score interpretations  
     investigating across, 157–61
- latent variables, 48  
 Linn, R. L., 172  
 logical arguments, 4–5
- Mach, Ernst, 190  
 matching and covariates, controlling  
     internal threats to validity  
     through, 35–37, 36f
- maturation, 11  
 measurement vs. assessment, 3n2

- Meehl, Paul E., 16
- Mehrens, W. A., 172
- Messick, Samuel A., 96, 124–25, 149, 173
- “conformationist” view of the scientific endeavor, 190
  - on consequences of score interpretation and use, 18, 88–90, 171
  - on multiple lines and sources of evidence, 88, 89, 96–97, 130, 142
  - on test content, 112
  - two-dimensional framework on facets of validity, 18, 18f, 88–89, 89f
  - “unitarian” view of validity, 194
  - validation research and, 89–90
  - on validity and validation, 17–18, 147
- meta-analysis, 69
- Moss, P., 173, 174
- multilevel modeling. *See also*
- hierarchical linear modeling
  - to account for potential internal threats to validity, 46–48
- multiple regression to control for internal threats to validity, 41–42, 43t
- mutli-trait/multi-method studies, 137, 138t, 139–40
- nesting, 46
- No Child Left Behind Act (NCLB), 169
- nomological net/nomological network, 4, 5f, 126–27, 127f
- defined, 4
- norm-referenced scoring model, 101
- norm-referenced test (NRT), 135, 136t, 137, 138t, 139
- null hypothesis, 16
- defined, 16, 66
- omitted variable bias, 15, 75
- one-tailed statistical tests, 70–71
- Park, C. S., 174
- path analysis to account for potential internal threats to validity, 42, 43t, 44–46, 45f
- Pearson correlation ( $r$ ), 68
- performance-level descriptors (PLDs), 162
- Popham, W. J., 172
- Popper, Karl, 190
- positivism, 9
- logical, 189–90
- pre-test/post-test quasi-experimental design, 32f, 32–34
- pre-testing, 12
- probability estimates of error, 67
- profile analysis, 128–29
- psychometrics, 21
- quasi-experimental research, 10–11, 129, 151–52
- quasi-experimental strategies for addressing internal threats to validity, 31–32
- random irrelevancies, 12
- random selection and random assignment, 30–31, 56
- internal threats to validity under conditions of, 30–31, 56
- rater drift, 107
- rationalism, 9, 190–92
- reading fluency, 78
- realism, 9, 13, 15
- relativism, 9, 13, 15, 16, 65, 191
- reliability estimates, 120–21
- repeated measures (pre- and post-testing), controlling for internal threats to validity through, 32f, 32–34
- replication to support external validity of causal claims, 62–63
- representativeness of items and tasks, 112–19
- robust statistical tests, 67
- root mean square error of approximation (RMSEA), 49
- sample sizes and external validity, 63
- Scholastic Assessment Test (SAT), 180–82

- science, philosophy of, 189–92
- scientific argument, Toulmin's six-part model of, 190–91
- score interpretations, 149–50. *See also* cut scores; interpretation; score use; scores
- investigating, 150–51
- across languages and cultures, 157–61
- across settings, 154–56
- over time, 153–54
- investigating alternative, 151–53
- properties of observed scores
- supporting, 150–64
- score use. *See also* under assessment scores
- appropriate consequences of, 171–74, 179–83
- value implications, 174–78
- to focus instructional interventions, 167–68
- to monitor change over time, 168
- in selection, 166–67
- unintended consequences of, 171–74, 179–83
- scores, 2n1. *See also* assessment scores; *specific topics*
- appropriate for intended uses, 164–71
- decisions made from. *See* score interpretations
- defined, 2n1, 2n2
- as signs, 87–88
- selection. *See also* random selection and random assignment
- interactions with, 12, 14, 56–57
- selection bias, 11
- selection decisions
- challenges in evaluating the validity of scores for, 170–71
- score use in, 166–67
- sensitivity, 70. *See also* bias and sensitivity reviews
- setting(s)
- interaction of treatment with, 14
- investigating score interpretations across, 154–56
- Shapere, Dudley, 192
- Shepard, L. A., 173, 174, 179
- situation, interaction of treatment and, 57–59
- standardized difference, 68
- standardized root mean square residual (SRMR), 49
- standards-based achievement tests, purposes of, 174
- Standards for Educational and Psychological Testing*, 194
- statistical conclusions
- defined, 15
- validity of, 15–16
- threats to, 65, 79–80. *See also specific threats*
- statistical methods, writings on, 193
- statistical models, over- or under-interpreted differences between alternative, 16
- statistical power, 15, 70–73
- statistical regression, 12
- statistical significance, 67–68
- statistical tests
- one-tailed, 70–71
- robust, 67
- two-tailed, 70, 71, 72t
- use of multiple, 74
- violations of assumptions of, 15, 75–76
- statistics conclusions, factors to consider regarding the validity of, 67–79
- statistics fundamentals, 66–67
- Stone, C. A., 174
- structural equation modeling (SEM), 79t
- components needed to propose causal models in, 48
- to investigate threats to internal validity and consider alternate explanations, 48–52, 50f
- SEM research, 153, 154f, 155f
- tasks. *See* items and tasks
- test scores. *See* scores
- test specifications, evaluation of, 115–16



- testing and treatment, interactions
  - between, 14
- time
  - investigating score interpretations over, 153–54
  - score use to monitor change over, 168
- Toulmin, Steven, 190–91
- translating and adapting tests,
  - international guidelines for, 157, 158–60t
- treatment implementation,
  - unreliability of, 12
- treatments, interactions between, 14
- true scores, 119–21
- two-sample, two-tailed test, 71, 72f, 73
- two-tailed statistical tests, 70, 71, 72t
- Type I error, 66, 70, 71
- Type II error, 67, 70, 73
- “valid,” synonyms and words used to define, 2
- validation. *See also specific topics*
  - defined, 2
  - overview, 2–3
  - in theory building and assessment development, 4–5
  - writings on, 193
- validation research
  - claims that should guide, 90, 91–95t, 96
  - extrapolating from score to domain or construct/criterion performance, 124–43
  - generalizing from scores to a universe of behaviors/responses related to the construct, 112–24
  - scores can be used to make inferences, 96–111
  - writings on, 193
- validation research and evaluation,
  - combined framework for. *See* validation research: claims that should guide
- validity. *See also specific topics*
  - defined, 2, 17
  - Messick’s facets of, 18, 18f, 88–89, 89f
  - overview, 1, 147
  - types and conceptions of, 17. *See also specific types of validity*
  - uses of the term, 1
- validity theory
  - history, 19
  - philosophical foundations, 5, 6–8t, 9–10
  - writings on, 193–94
- variables
  - dependent vs. independent, 26
  - endogenous vs. exogenous, 48
  - omitted variable bias, 15, 75
- variance, 69
- Wiley, D. E., 172