TESTWISE

Understanding Educational Assessment, Volume 1

NORA VIVIAN ODENDAHL

Testwise

Understanding Educational Assessment

Volume 1

Nora Vivian Odendahl

ROWMAN & LITTLEFIELD EDUCATION

A division of
ROWMAN & LITTLEFIELD PUBLISHERS, INC.
Lanham • New York • Toronto • Plymouth, UK

Published by Rowman & Littlefield Education
A division of Rowman & Littlefield Publishers, Inc.
A wholly owned subsidiary of The Rowman & Littlefield Publishing Group, Inc.
4501 Forbes Boulevard, Suite 200, Lanham, Maryland 20706
http://www.rowmaneducation.com

Estover Road, Plymouth PL6 7PY, United Kingdom

Copyright © 2011 by Nora Vivian Odendahl

All rights reserved. No part of this book may be reproduced in any form or by any electronic or mechanical means, including information storage and retrieval systems, without written permission from the publisher, except by a reviewer who may quote passages in a review.

British Library Cataloguing in Publication Information Available

Library of Congress Cataloging-in-Publication Data

Odendahl, Nora Vivian.
Testwise : understanding educational assessment, Volume 1 / Nora Vivian Odendahl.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-61048-011-6 (cloth : alk. paper) — ISBN 978-1-61048-012-3 (pbk. : alk. paper) — ISBN 978-1-61048-013-0 (electronic)
1. Educational tests and measurements. I. Title.
LB3051.O34 2010
371.26—dc22 2010041389

™ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI/NISO Z39.48-1992.

Printed in the United States of America

Contents

Prologue

To understand the ways in which our nation assesses learning is to understand something about ourselves. Why?

Tests and other forms of *assessment* influence what we are taught and what we remember.* They can affect our patterns of behavior and our self-perceptions. They even play a role in shaping the social, political, and economic landscape that we inhabit. At the same time, they can exist only with public consent, however tacit.

Yet the mechanisms of this force in our lives often remain a mystery.

To begin with, what does assessment mean? An umbrella term, it includes but is not limited to the formal tests or quantitative measurement with which this book is primarily concerned. It is "any systematic method of obtaining information . . . used to draw inferences about characteristics of people, objects, or programs" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, p. 172). Thus, in education, assessment can span diverse procedures for eliciting evidence of students' skills and knowledge, from nationwide examinations to teachers' informal methods for finding out what their students have learned and are able to do.

The purposes of educational assessment may be to diagnose, classify, select, place, predict, monitor, or change. Assessment may focus on the achievement or instructional needs of individuals, or it may provide data about groups categorized by classroom, school, state, nation, gender, ethnicity, language, disability, family income level, and so on. In turn, decision

*See the Glossary for definitions of specialized terms. The first mention of each such term is in italics.

makers at different levels use this information for crafting policies and influencing individual destinies.

Being assessed is an everyday experience; anyone who has gone to school has taken countless quizzes, course exams, and *standardized tests* with *multiple-choice* and essay questions. Moreover, the news media regularly cover controversies about using test scores either to hold teachers and schools accountable or to make decisions about students' grade-to-grade promotion, high school diplomas, or college and graduate admissions.

Yet such everyday familiarity does not necessarily bring insight. Often the only question debated is whether there is "too much" testing in schools.

Perhaps the realm of testing seems too forbidding to enter—a maze of complexities, situated in a minefield of controversies. No wonder, because although testing is a routine of education, students and the public hear little about what it means and how it works. Even policy makers who make decisions about testing may perceive it as beyond their ken, an arcane discipline requiring a specialized language (Madaus, 1993). Members of the assessment profession themselves can have difficulty communicating across their silos of subspecialties.

Testwise: Understanding Educational Assessment is intended to help improve this situation by lifting the curtain to explain principles and practices, issues and challenges, and inevitable trade-offs. The literature on testing is rich but often abstruse, and I hope to make this wisdom more accessible. The goal is to help readers think, talk, and make decisions about assessment in educational settings.

One way to demystify testing might be to turn directly to its technical aspects and explain such terms as *validity* and *reliability*. But before taking that step, we can start by picking up a familiar toolset, the six questions that journalists ask. The table on page ix, "Asking Journalistic Questions About Educational Tests," outlines possible types of queries.

These questions are, in fact, ones that we will pursue in greater depth (and I will suggest further resources to be found in articles, books, and websites). As the list shows, understanding any given test involves many activities.

Volume 1 of *Testwise* is concerned with two fundamental areas:

—*Considering historical and social contexts.* Every test is a creature of its time and place. Tracing the evolution of educational testing reveals the different ways in which people have defined and tried to make "good" tests. Important issues recur and have no easy resolutions, but recognizing historical precedents and parallels to today's challenges may help in finding better answers.

Underlying this history are also deeper philosophical implications, which matter as much as the technical quality of testing. How a nation assesses learning both reflects and perpetuates certain cultural values. What do our tests teach students about such values as scientific inquiry and objectivity, or societal opportunity and justice?

Asking Journalistic Questions About Educational Tests

Who?

- Who created the test? Who chose the test? Who is giving the test? Who is scoring the test? Who is using the scores to make decisions?
- Who are the students taking the test, in terms of their backgrounds and characteristics? What does each student believe about his or her ability to succeed?

Where?

- Under what physical conditions is the test administered?
- In what kind of cultural and societal context does the testing take place?

When?

- At what point in the student's intellectual development does the testing occur? What opportunities has the student had to learn the tested material?
- On what date did the testing occur? What else was happening at that time?

What?

- What knowledge and skills is the test intended to measure? Are these the same ones that the test actually measures?
- What changes are made to accommodate students with special needs?
- What other efforts are made to treat all students fairly and equitably?
- What do the test scores really mean? What *don't* the scores tell us?
- What are both the intended effects and the actual consequences of the test?

Why?

- Why is the test being given? Why use a test rather than some other means?
- If scores are reported as numbers, why? If reported as performance categories, why?
- Why do individual students or certain groups of students perform in particular ways on the test? How does one investigate the different possible explanations?

How?

- How does the test measure knowledge and skills (e.g., using multiple-choice or open-ended questions)?
- How is the test scored—by a computer application or by a person? What rules are used for assigning scores?
- How consistent would a student's scores be if the student were to take a different version of the same test? If the student were to take the test on a different occasion?
- How will the scores be used in making educational decisions about students, curricula, schools, etc.? What other information should also be considered?
- How well do the testing procedures, and the ways in which scores are interpreted and used, meet professional standards?
- How much does the test cost in terms of time, money, and other resources? Do the benefits outweigh the costs?

—*Understanding assessment principles*. Although influenced by public opinion, educators' views, and governmental policies, educational testing has limited external regulation. It does, however, possess its own body of theory and guidelines.

The most important theoretical principle is validity, which includes but goes beyond matters of sound test construction. The *validation* process requires making a comprehensive, logical, and well-supported argument about how a particular test's scores should be understood and used. Validation includes scrutinizing the assessment procedures for relevance, accuracy, consistency, usefulness, accessibility, and *fairness*.

Nevertheless, even the highest-quality testing elicits only samples of student performance, not complete information—and test scores are only ways of interpreting these samples, not absolute truths.

Building on these and other concepts explored in the first volume, Volume 2 of *Testwise* takes a closer look at:

—*Applying principle to practice*. Abstract principles have to be translated into real tests for real students; test designers have to decide which skills and knowledge to assess and how to assess them. "How" decisions include question format: *selected-response* tasks, in which the student selects from among answers that are provided on the test; or *constructed-response* tasks, in which the student generates a product or performance. Each approach entails its own rules and conventions as well as its own advantages and disadvantages for particular purposes.

Classroom assessment is related to externally imposed testing but also has differences. Here, teachers' goals include adapting instruction appropriately and helping students develop the ability to evaluate their own learning processes and their own work products. Skills in self-monitoring are useful beyond school, and a society that understands basic principles of assessment—as a form of reasoning from evidence—may make better decisions about testing (and other matters).

—*Thinking about tomorrow.* While heeding the lessons of the past, our approaches to assessment of learning should also probe new ideas, experiment, innovate, anticipate the future, and look beyond the borders of our own country. As we go forward, new forms of technology and research findings in different disciplines will shape the investigation of skills, knowledge, and other attributes relevant to academic achievement.

The two volumes of *Testwise* are intended to make the reader just that, by offering a guide to critical thinking about educational assessment. I hope that this approach will be helpful to anyone concerned about the ways in which we define and demonstrate learning.

Chapter One

The Origins of Educational Measurement: A Quest to Quantify

> *Units in which to measure the changes wrought by education are essential to an adequate science of education.*
>
> —Edward L. Thorndike, 1910 (p. 8)

With the emergence of educational measurement as a discipline came daunting challenges, thorny issues, and heated debates, some of which were specific to their era but many of which still resonate today. Yet if there is one unifying theme for the founding of the field in the late nineteenth and early twentieth centuries, it is the attempt to use scientific methods to capture and quantify what seems ineffable: knowledge and intellectual skills.

Bringing the concept of scientific measurement into education and psychology was a paradigm shift from the centuries-long tradition of qualitative assessment. Different lines of inquiry led to this same turning point. One focused on uncovering what Edward L. Thorndike (1874–1949) called "the changes wrought by education"; the other on investigating mental capacities as biological phenomena. How these two approaches originally arose and eventually influenced each other is the subject of this chapter.

BEFORE 1900: QUALITATIVE ASSESSMENT IN EDUCATION; MEASUREMENT IN THE SCIENCES

The First Two Millennia of Evaluating Learning

Although the earliest known program for systematically assessing learning was actually an employment test, it foreshadowed many later developments in educational testing. In 210 B.C.E., under the Han dynasty, the Chinese

government introduced competitive civil service examinations that assessed both scholarship and military skills. This program, which continued with many changes over time until 1905, is especially noteworthy for having pioneered the practice of using tests for sociopolitical goals. The tests helped broaden eligibility for government posts, limit the power of the hereditary aristocracy, and inculcate values that reinforced the existing political structures (Elman, 1991).

At the same time, the program illustrated many recurrent challenges in testing. For example, cheating and gaming the system were common and went beyond such obvious strategies as smuggling answers into the test administration or hiring substitutes to take the exam. By 681 C.E., it became clear that some examinees were using memorized essays. To elicit original responses displaying the examinees' actual knowledge of history and philosophy, the examiners added another section in which examinees had to compose poetry on these subjects, but again, some examinees adapted by relying on memorized poetic conventions (Suen & Yu, 2006).

Another perennial issue arose during the Sung Dynasty (960–1279), when examiners tried to assess more than simple recall of classic literary works. Instead, they asked examinees to demonstrate analytic skills by discussing particular issues in these texts. But examiners ultimately abandoned this experiment in measuring higher-order reasoning and thinking because government officials worried that the scoring of examinees' responses would be too subjective. Even today, similar concerns about subjectivity affect decisions about question formats and scoring procedures (Madaus & O'Dwyer, 1999).

In medieval Europe, where opportunities for education were scarce, systematic assessment procedures were also rare. However, theological examinations at the University of Paris and the University of Bologna began in the late twelfth century. They took the form of oral disputations conducted in Latin, often in front of an audience, and covering knowledge of specified texts. Even though written tests were introduced in the 1500s when paper became more widely available and Chinese influence reached westward, the oral *viva voce* ("by or with the living voice") expanded to encompass a larger range of disciplines and predominated in European universities until the Enlightenment (Madaus & Kellaghan, 1993).[1]

For centuries, an examinee's performance in the *viva voce* was described qualitatively or simply classified as pass versus fail. In the mid-1700s, the University of Cambridge broke with tradition, by instituting written sections for the "Mathematical Tripos" exam and ranking the examinees. Ranking was further facilitated in the 1790s when a mathematics professor, William Farish (1759–1837), began assigning numerical scores to students for their performance (Madaus & O'Dwyer, 1999).

Farish's innovation would have far-reaching implications for policy and society. Numerical scores could be accumulated, aggregated, and then "classified, averaged, and normed. They could be used to describe and compare groups or institutions, and to fix individuals, groups, and institutions in statistical distributions" (Madaus & Kellaghan, 1993, under "Brief History of Testing," ¶ 4).

In the Nineteenth Century, New Purposes for Assessing Learning

Until the 1840s, educational assessment in both Europe and the United States typically focused on individual achievement and was still conducted orally in elementary and secondary schools. American practices began to depart from European ones in 1845 when Horace Mann (1796–1859), a lawyer and advocate for universal public education, introduced changes in the format and purpose of school testing.

As secretary of the board of education for the state of Massachusetts, Mann criticized the Boston public schools' use of individual oral examinations and urged replacing them with written examinations that had prescribed timing for each question. He argued that written tests would allow teachers to pose more questions; give students a better opportunity to display their knowledge; be more impartial; and offer fairer comparisons of student achievement across classrooms and schools (although, in fact, teachers chose students to be tested) (Witte, Trachsel, & Walters, 1986).

To emphasize the modernity, convenience, and objectivity of the written exams, Mann compared them to another technology that had just emerged: photography. His tests captured "a sort of Daguerreotype likeness, as it were, of the state and condition of the pupils' minds" that could be "taken and carried away, for general inspection" (1845, p. 334, quoted in Witte et al., 1986, p. 19).

Mann's program also pioneered the use of tests for judging the performance of someone other than the actual examinee. Because government officials viewed the scores as measures of schools' effectiveness, the tests served to make teachers and principals accountable to these officials. Such *accountability testing* became common enough by the 1870s that the head of the National Education Association felt the need to warn against the practice (Resnick, 1982).

In the mid-1890s, a physician interested in children's development, Dr. Joseph Rice (1857–1934), helped introduce statistical methods into the scrutiny of curricula and teaching methods. Rice's best-known work involved giving a standardized spelling test to elementary school students and comparing the scores with the amount of classroom time that students spent on spelling.

The results were striking. After testing 30,000 students, Rice determined that those who spent about fifteen minutes per day on spelling performed as well as those who worked on spelling for an hour a day (Pulliam, 1991). These findings would subsequently influence how spelling was taught. Thus Rice set a precedent for using evidence gained from assessment to inform instruction.

As is now apparent, Mann's and Rice's innovations were in keeping with the spirit of their age. They reflected the trend of industrial capitalism toward achieving greater standardization, precision, and efficiency (Madaus & Kellaghan, 1993).

Measuring the Mind with Tools From the Natural Sciences

Despite differences in purpose and method, the historical examples mentioned so far were all concerned with investigating knowledge and academic skills that examinees could presumably acquire through study. However, during the nineteenth century some researchers sought to extend the natural and physical sciences by measuring intellectual attributes as biological phenomena—thereby equating the mind with the brain.

One seed of this approach was planted in the early 1800s, when Johann Friedrich Herbart (1776–1841) proposed that mathematics could be used in psychology. Comparing investigation of the mind to that of the physical universe, Herbart (n.d./1877) argued that because mental phenomena such as perception or emotion could exist in greater or lesser degrees, they could be observed and quantified just as physical phenomena could. In fact, such measurement might reveal the underlying principles of the mind. This perspective helped promote psychology's evolution from a philosophical and conceptual discipline to an experimental and quantitative one, with the first psychological laboratory established by Wilhelm Wundt (1832–1920) in 1879 (Hatfield, 2007).

Also emerging in the early-to-mid-nineteenth century was a biological-medical field called *craniometry* that focused on measurement of human skulls. Many craniometrists took the very literal view that the larger the skull, the keener the mind that had once occupied it.*

As Stephen Jay Gould (1941–2002) describes in *The Mismeasure of Man* (1981), these craniometrists included physicians Samuel Morton (1791–1851) and Pierre Paul Broca (1824–1880), who at times selectively chose and

*Today, neuroscientists do not use skull measurements to measure brains; they use magnetic resonance imaging (MRI) to obtain much more accurate information about the size and shape of brains in living humans, including specialized brain areas and structures associated with different aspects of cognitive functioning.

manipulated data of skull measurements in order to claim that white males were more intelligent than women and than men of other races. But even where their skull measurements were precise and accurate, the inferences the craniometrists drew were without warrant. Offering no evidence about their subjects' actual ability to perform intellectual tasks, and ignoring the fact that modern humans have smaller brains than those of whales, elephants, dolphins, or Neanderthals, the craniometrists nonetheless asserted a strict correspondence between intelligence and size of the physical brain.

In Gould's view, craniometry served as a precursor for much of early intelligence testing, relying on flawed assumptions about biological determinism and about intelligence as a single quantity.* In fact, the founder of modern intelligence testing, Alfred Binet (1857–1911), began his investigations in 1898 by trying to continue Broca's work on correlating brain size and intellect. Unsuccessful, Binet would turn to a different method of mental measurement after 1900 (as discussed in the next section).

Yet the biological approach to measuring the mind yielded some concepts and methods still used today. The chief contributors were Francis Galton (1822–1911) and one of his students, Karl Pearson (1857–1936), both working in England. They shared an interest in using quantitative techniques to classify observations of traits and investigate human development systematically. For example, in depicting differences among children, they employed *distribution* curves showing the patterns of age, height, and (presumed) intelligence across a given group or population.

Such techniques revolutionized the natural and social sciences. A few decades later, describing the brand-new field of educational measurement, Ayres (1918) praised Galton for creating "the statistical methods necessary for the quantitative study of material which seemed at the outset entirely qualitative and not at all numerical in nature" (p. 11).[2]

However, Galton also introduced the term "eugenics" in 1883, while advocating that the human species be modified through selective breeding. Both he and Pearson promoted beliefs in hierarchical racial differences. Thus, some of the seminal contributions to quantitative measurement in the social sciences, such as the use of percentiles, the computation of *correlations*, and the concept of statistical significance, were originally associated with sociopolitical agendas that would later be discredited and repudiated (Resnick, 1982).

Elsewhere, in the United States, James Cattell (1860–1944), who had studied with Wundt in Germany and briefly worked with Galton in England,

*Gould (1981) defines biological determinism as the view that "shared behavioral norms, and the social and economic differences between human groups—primarily races, classes, and sexes—arise from inherited, inborn distinctions and that society, in this sense, is an accurate reflection of biology" (p. 20).

decided to focus on measuring individual differences in psychological and physical attributes within an educational setting. Cattell (1890) explained his larger purpose:

> Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement. A step in this direction could be made by applying a series of mental tests and measurements to a large number of individuals. (p. 373)

Accordingly, Cattell tried to apply Galton's theory that intelligence was a matter of "neurological efficiency" by using "psychophysical" tests of reaction times, sensory perceptions, and memory to predict future performance of incoming students at Columbia University. But one of Cattell's own graduate students, Clark Wissler (1870–1947), found that individuals' scores across tests did not show a consistent pattern, nor did they appear to be predictive of a student's subsequent course grades (Plucker, 2003).

Even though these particular attempts were not successful, Cattell's and Wissler's methods for designing mental tests and analyzing test results laid the foundations for the field of *psychometrics* (quantitative measurement of psychological attributes). Moreover, in 1892 Cattell helped establish the American Psychological Association, the same professional organization that sets today's standards for psychological testing.

INTELLIGENCE TESTING IN THE EARLY TWENTIETH CENTURY

Intelligence Testing in Schools Prior to World War I

Moving away from the venerable tradition of evaluating what students had learned, many psychologists and educators in the United States began to focus on measuring general intellectual capabilities. Two events of 1904 contributed to this trend, even though they both occurred on the other side of the Atlantic Ocean.

First, French educational officials asked Binet, a lawyer turned psychologist, to find a way to identify students who were struggling and might need special attention. Using tasks intended to require the kinds of everyday knowledge and reasoning that he had observed in children over the years, Binet created an assessment to diagnose learning impairments.

Binet assigned age levels to the tasks—the more difficult the task, the higher the age level. He calculated a child's "mental age" by identifying the age level of the most difficult tasks that the child could perform cor-

rectly, then subtracting the child's actual age. The concept of an *intelligence quotient* or IQ arose later, when, in accordance with the recommendation of German psychologist Wilhelm Stern (1871–1938), the scoring method was adjusted to divide mental age by chronological age (Gould, 1981).

The second significant event in 1904 was when a British student of Wundt's, Charles Spearman (1863–1945), published an experimental study that used only 123 subjects but was grandly titled "'General Intelligence' Objectively Determined and Measured." Rejecting Binet's "practical" method of using problem-solving or reasoning tasks, Spearman followed Galton and Cattell's psychophysical approach by administering tests of sensory discrimination to some schoolchildren. He also gauged "school cleverness" and "common sense" via the students' grades and via interviews with teachers and the students themselves (Spearman, 1904, pp. 241–42, 250–51).[3]

To identify relationships among these sets of data or "factors," Spearman invented the mathematical technique of *factor analysis*. However, Spearman (or at least, his followers) then made the "invalid inference" that positive correlations in performance across tests and other types of measures had a single, unambiguous cause: a person's fixed amount of intelligence (Gould, 1981, pp. 250–57). Spearman's naming of this correlation as "the general factor" or "*g*," implying a "Universal Unity of the Intellective Function," furthered the idea of intelligence as a unitary entity (1904, p. 273).[4]

Although the distinctions have been blurred over time, Spearman's assumptions and conclusions differed fundamentally from those of Binet, who believed that intellectual development occurred at varying rates and could be influenced by environment. Binet did not view IQ as innate intelligence or as a means of ranking all students; instead, Binet viewed his mental-age *scale* as merely a means to identify children who needed more help. Not only did Binet worry that low scores would become self-fulfilling, but he also favored a personalized, individual approach over mass testing (Plucker, 2003; Gould, 1981; Brown, 1992).

Nevertheless, the idea of measuring students' intelligence was rapidly embraced in the United States. Not only were birthrates and immigration rates surging at the beginning of the twentieth century, but also many students were failing schools' existing tests. During the Progressive era of 1900–1917, reformers thought that the best way to improve schools' efficiency would be to separate students of apparently different levels of ability into different educational tracks (Resnick, 1982). However, even those schools that did try to track pupils had no systematic means of classifying them. And administrators did not trust teachers to improve the situation, because most teachers of this era had little formal training and were seldom valued as professionals (Brown, 1992).

Intelligence testing seemed to offer a solution. Influential followers of Galton in the United States argued that grouping students according to "scientifically" measured mental ability would help teachers conduct classes more effectively and be kinder to pupils who might be frustrated by overly challenging curricula (Bracey, 1995).

One such follower, psychologist Lewis Terman (1877–1956) of Stanford University, would become a key figure in intelligence testing. Having conducted experimental testing on California schoolchildren from 1911 to 1915, he published an English-language revision of Binet's test, the Stanford-Binet test, in 1916. He claimed that his examination—lasting less than an hour—could "contribute more to a real understanding of the case than anything else that could be done" to diagnose and classify students according to "native ability" (1916, ¶ 6).

Terman's revision used Stern's IQ-ratio calculation and provided standardized materials and administration procedures. It also included so-called "norms," statistical tables for comparing a student to the rest of the population. However, the norms supposed to represent typical performance had been obtained by testing only 1,000 students from nearby middle-class schools (Chapman, 1988).

At the time of the First World War, Terman and one of his graduate students, Virgil E. Dickson (1885–1963), were testing public school students in Oakland, California, as part of the effort to classify and track students. In addition to using the Stanford-Binet, Dickson was trying out measures that Terman was helping to develop for the U.S. Army.

The Army Mental Tests in World War I

The relatively late entry of the United States into the war meant that military forces had to be mobilized as quickly as possible, and placing the approximately 1.7 million men recruited by the Army in 1917–1918 into appropriate positions was a formidable challenge. Robert Yerkes (1876–1956), then president of the American Psychological Association, gathered a group of psychologists, including Terman, to develop a test for this purpose. Terman advocated using a paper-and-pencil version of the Stanford-Binet test, which Arthur Otis (1886–1954) had revised to permit group administration rather than the established practice of individual, personalized administration (Chapman, 1988).

From the two-week session devoted to adapting Otis's version emerged the "Alpha" and "Beta" tests. The text-based Alpha, which focused on comprehension, reasoning, and memory, was intended for recruits who could read and write English; the Beta, which relied on "concrete materials" such as

pictures and diagrams, was intended for "foreigners and illiterates" (Terman, 1918, pp. 179–80). To guard against "coaching," the psychologists created several different test "'forms,' each differing entirely in substance from every other 'form,' yet all exactly equal in difficulty and alike psychologically" (pp. 178–79).

Scores for "intelligence" were reported as letter grades and also translated into mental ages based on the Stanford-Binet student norms. For comparison, the testing also included college students, prisoners, manual laborers, and even several hundred female prostitutes (Yoakum & Yerkes, 1920).

Yet looking at Alpha *test forms* reproduced in Yoakum and Yerkes's 1920 account (now available online), one finds many questions that seem irrelevant to any military duties and that, while purporting to measure intelligence, would instead measure cultural experience and knowledge of trivia. Such questions included: the products endorsed by certain celebrities, the locations of Ivy League schools and auto manufacturers, the names of card games and fabrics, and the total number of Henry VIII's wives.

As Gould (1981) describes, no less important to the outcome of the project was the effect of very nonstandardized, even chaotic testing conditions (despite Terman's admonition that "the greatest care must be taken to keep conditions uniform" in accordance with the "Examiner's Guide"). These problems included rooms so noisy that recruits—some of whom had never before taken a test—could not hear the instructions; drastically inadequate time limits; reassignment of men who could not read or speak English to the Alpha tests simply because the waiting lines for the Beta tests were too long; and many other serious violations of protocol.

Then, in interpreting the test results, Yerkes observed that low education levels and ill health were associated with poor performance on the tests—but he failed to consider that adverse environments, rather than hereditary intelligence, might have affected the scores. He also failed to draw the obvious inference from the fact that scores for soldiers who were immigrants rose with each year of residence in the United States (Gould, 1981). Instead, Yerkes concluded that the average mental age of adult recruits was as low as thirteen (although conscientious objectors did earn higher scores, as Yoakum and Yerkes grudgingly noted). One might wonder how these same Americans could ever have helped win the war.

Such a chain of dubious test construction, irregular testing conditions, and flawed score interpretation vividly illustrates the kinds of weak links that can undermine the intended meaning of test results—and shows why standardizing procedures is such a priority in large-scale testing. The fact that a score of zero was the score most frequently attained on six of the eight Alpha test forms speaks for itself (Gould, 1981, p. 214). Even so, the public perception