making classroom ASSESSMENTS reliable & valid

robert j. marzano

Hawker Brownlow
EDUCATION

Table of Contents

Appendix

introduction

The Role of Classroom Assessment

Classroom assessment has been largely ignored in the research and practice of assessment theory. This is not to say that it has been inconsequential to classroom practice. To the contrary, the topic of classroom assessment has become more and more popular in the practitioner literature. For example, the book *Classroom Assessment: What Teachers Need to Know* is in its eighth edition (Popham, 2017). Many other publishers continue to release books on the topic. This trend notwithstanding, technical literature in the 20th century has rarely mentioned classroom assessment. As James McMillan (2013b) notes:

> Throughout most of the 20th century, the research on assessment in education focused on the role of standardized testing . . . . It was clear that the professional educational measurement community was concerned with the role of standardized testing, both from a large-scale assessment perspective as well as with how teachers used test data for instruction in their own classrooms. (p. 4)

As evidence, McMillan (2013b) notes that an entire issue of the *Journal of Educational Measurement* that purported to focus on state-of-the-art testing and instruction did not address teacher-made tests. Additionally, the first three editions of *Educational Measurement* (Lindquist, 1951; Linn, 1993; Thorndike, 1971)—which are designed to summarize the state of the art in measurement research, theory, and practice—paid little if any attention to classroom assessment. Finally, both editions of *The Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, 2014)—which, as their titles indicate, are designed to set standards for testing in both psychology and education—made little explicit reference to classroom assessment. It wasn't until the fourth edition in the first decade of the 21st century (Brennan, 2006) that a chapter was included addressing classroom assessment.

Most recently, the *SAGE Handbook of Research on Classroom Assessment* made a stand for the rightful place of classroom assessment: "This book is based on a single assertion: Classroom assessment (CA) is the most powerful type of measurement in education that influences student learning" (McMillan,

2013a, p. xxiii). Throughout this text, I take the same perspective. I also use the convention of referring to classroom assessment as CA. Since the publication of the *SAGE Handbook*, this abbreviation is now the norm in many technical discussions of classroom assessment theory. My intent is for this book to be both technical and practical.

What, then, is the place of CAs in the current K–12 system of assessment, and what is their future? This resource attempts to lay out a future for CA that will render it the primary source of evidence regarding student learning; this would stand in stark contrast to the current situation in which formal measurements of students are left to interim assessments, end-of-course assessments, and state assessments. In this introduction, I will discuss several topics with regard to CAs.

- The curious history of large-scale assessments
- The place of classroom assessment
- Reliability and validity at the heart of the matter
- The need for new paradigms
- The large-scale assessment paradigm for reliability
- The new CA paradigm for reliability
- The large-scale assessment paradigm for validity
- The new CA paradigm for validity

Before delving directly into the future of CA, it is useful to consider the general history of large-scale assessments in U.S. education since it is the foundation of current practices in CA.

The Curious History of Large-Scale Assessments

The present and future of CA are intimately tied to the past and present of large-scale assessments. In 2001, educational measurement expert Robert Linn published "A Century of Standardized Testing: Controversies and Pendulum Swings." Linn notes that the original purpose of large-scale assessment was comparison and began in the 19th century.

Educators commonly refer to J. M. Rice as the inventor of the comparative large-scale assessment. This assignment is based on his 1895 assessment of the spelling ability of some thirty-three thousand students in grades 4 through 12 for which comparative results were reported (Engelhart & Thomas, 1966). However, assessments that educators administered to several hundred students in seventeen schools in Boston and one school in Roxbury in 1845 predated this comparative large-scale assessment. Because of this, Horace Mann (who initiated the effort) deserves credit as the first to administer large-scale tests. Lorrie A. Shepard (2008) elaborates on the contribution of Horace Mann, noting:

> In 1845, Massachusetts State Superintendent of Instruction, Horace Mann, pressured Boston school trustees to adopt written examinations because large increases in enrollments made oral exams unfeasible. Long

> before IQ tests, these examinations were used to classify pupils . . . and to put comparative information about how schools were doing in the hands of state-level authority. (p. 25)

Educators designed these early large-scale assessments to help solve perceived problems within the K–12 system. For example, in 1909, Leonard P. Ayres published the book *Laggards in Our Schools: A Study of Retardation and Elimination in City School Systems*. Despite the book's lack of sensitivity to labeling large groups of students in unflattering ways, it brought attention to the problems associated with repeated retention of students in grade levels. This helped buttress the goal of reformers who wanted to develop programs that would mitigate failure.

The first half of the 20th century was not a flattering era for large-scale assessments. They focused on natural intelligence, and educators used them to classify examinees. To say the least, this era did not represent the initial or current intent of large-scale assessment. I address this period in more detail shortly.

By the second half of the 20th century, educators began to use large-scale assessments more effectively. Such assessments were a central component of James Bryant Conant's (1953) vision of schools designed to provide students with guidance as to appropriate career paths and support in realizing related careers.

The use of large-scale assessment increased dramatically in the 1960s. According to Shepard (2008), the modern era of large-scale assessment started in the mid-1960s: "Title I of the Elementary and Secondary Education Act (ESEA) of 1965 launched the development of the field of educational evaluation and the school accountability movement" (p. 26). Shepard (2008) explains that it was the ESEA mandate for data with which to scrutinize the reform efforts that compelled the research community to develop more finely tuned evaluation tools: "The American Educational Research Association began a monograph series in 1967 to disseminate the latest thinking in evaluation theory, and several educational evaluation organizations and journals date from this period" (p. 26).

The National Assessment of Educational Progress (NAEP) began in 1969 and "was part of the same general trend toward large-scale data gathering" (Shepard, 2008, p. 27). However, researchers and policymakers designed NAEP for program evaluation as opposed to individual student performance evaluation.

The need to gather and utilize data about individual students started minimum competency testing in the United States. This spread quickly, and by 1980 "all states had a minimum competency testing program or a state testing program of some kind" (Shepard, 2008, p. 31). But this, too, ran aground because of the amount of time and resources necessary for large-scale competency tests.

The next wave of school reform was the "excellence movement" spawned by the high visibility report *A Nation at Risk* (National Commission on Excellence in Education, 1983). It cited low standards and a watered-down curriculum as reasons for the lackluster performance of U.S. schools. It also faulted the minimum competency movement, noting that focusing on minimum requirements distracted educators from the more noble and appropriate goal of maximizing students' competencies.

Fueled by these criticisms, researchers and policymakers focused on the identification of rigorous and challenging standards for all students in the core subject areas. Standards work in mathematics set the tone for the reform:

> Leading the way, the National Council of Teachers of Mathematics report on *Curriculum and Evaluation Standards for School Mathematics* (1989) expanded the purview of elementary school mathematics to include geometry and spatial sense, measurement, statistics and probability, and patterns and relationships, and at the same time emphasized problem solving, communication, mathematical reasoning, and mathematical connections rather than computation and rote activities. (Shepard, 2008, p. 35)

By the early 1990s, virtually every major academic subject area had sample standards for K–12 education.

Shepard (2008) notes that standards-based reform, begun in the 1990s, "is the most enduring of test-based accountability reforms" (p. 37). However, she also cautioned that the version of this reform enacted in No Child Left Behind (NCLB) "contradicts core principles of the standards movement" mostly because the assessments associated with NCLB did not place ample focus on the application and use of knowledge reflected in the standards researchers developed (Shepard, 2008, p. 37). Also, the accountability system that accompanied NCLB focused on rewards and punishments.

The beginning of the new century saw an emphasis on testing that was highly focused on standards. In 2009, the National Governors Association Center for Best Practices (NGA) and the Council of Chief State School Officers (CCSSO) partnered in "a state-led process that [drew] evidence and [led] to development and adoption of a common core of state standards . . . in English language arts and mathematics for grades K–12" (as cited in Rothman, 2011, p. 62). This effort, referred to as the *Common Core State Standards (CCSS)*, resulted in the establishment of two state consortia that were tasked with designing new assessments aligned to the standards. One consortium was the Partnership for Assessment of Readiness for College and Careers (PARCC); the other was the Smarter Balanced Assessment Consortium (SBAC):

> Each consortium planned to offer several different kinds of assessments aligned to the CCSS, including year-end summative assessments, interim or benchmark assessments (used throughout the school year), and resources that teachers could use for formative assessment in the classroom. In addition to being computer-administered, these new assessments would include performance tasks, which require students to demonstrate a skill or procedure or create a product. (Marzano, Yanoski, Hoegh, & Simms, 2013, p. 7)

These efforts are still under way although with less widespread use than in their initiation.

Next, I discuss previous abuses of large-scale assessments that occurred in the first half of the 20th century (Houts, 1977). To illustrate the nature and extent of these abuses, consider the first intelligence test usable for groups that Alfred Binet developed in 1905. It was grounded in the theory that intelligence was not a fixed entity. Rather, educators could remediate low intelligence if they identified it. As Leon J. Kamin (1977) notes in his book on the nature and use of his IQ test, Binet includes a chapter, "The Training of Intelligence," in which he outlines educational interventions for those who scored low on his test. There was clearly an implied focus on helping low-performing students. It wasn't until the Americanized version of the Stanford-Binet test (by Lewis M. Terman, 1916) that the concept of IQ solidified as a fixed entity with little or no chance of improvement. Consequently, educators would use the IQ test to identify students with low intelligence so they could monitor and deal with them accordingly. Terman (1916) notes:

> In the near future intelligence tests will bring tens of thousands of these high-grade defectives under the surveillance and protection of society. This will ultimately result in curtailing the reproduction of feeble-mindedness and in the elimination of an enormous amount of crime, pauperism, and industrial inefficiency. It is hardly necessary to emphasize that the high-grade cases, of the type now so frequently overlooked, are precisely the ones whose guardianship it is most important for the State to assume. (pp. 6–7)

The perspective that Lewis Terman articulated became widespread in the United States and led to the development of Arthur Otis's (one of Terman's students) Army Alpha test. According to Kamin (1977), performance scores for 125,000 draftees were analyzed and published in 1921 by the National Academy of Sciences, titled *Memoirs of the National Academy of Sciences: Psychological Examining in the United States Army* (Yerkes, 1921). The report contains the chapter "Relation of Intelligence Ratings to Nativity," which focuses on an analysis of about twelve thousand draftees who reported that they were born outside of the United States. Educators assigned a letter grade from A to E for each of the draftees, and the distribution of these letter grades was analyzed for each country. The report notes:

> The range of differences between the countries is a very wide one . . . . In general, the Scandinavian and English speaking countries stand high in the list, while the Slavic and Latin countries stand low . . . the countries tend to fall into two groups: Canada, Great Britain, the Scandinavian and Teutonic countries . . . [as opposed to] the Latin and Slavic countries. (Yerkes, 1921, p. 699)

Clearly, the perspective regarding intelligence has changed dramatically and large-scale assessments have come a long way in their use of scores on tests since the early part of the 20th century. Yet even now, the mere mention of the terms *large-scale assessment* or *standardized assessment* prompts criticisms to which assessment experts must respond (see Phelps, 2009).