Enduring Issues In Educational Assessment

Assessments were an important tool recommended by *A Nation at Risk*. Mr. Della-Piana argues that now, a generation after that report, many of the problems with assessment still remain.

**BY GABRIEL M. DELLA-PIANA**

It is common to look backward periodically to understand what happened, to generalize about what happens under such conditions, and to learn how to adapt to new possibilities. A quarter century after *A Nation at Risk* appeared, various issues in assessment, the focus of some of the report’s recommendations, endure.

Despite its title, *A Nation at Risk* was always intended to be a forward-looking document. Its graphically enhanced preamble reads:

> All, regardless of race or class or economic status, are entitled to a fair chance and to the tools for developing their individual powers of mind and spirit to the utmost. This promise means that all children by virtue of their own efforts, competently guided, can hope to attain the mature and informed judgment needed to secure gainful employment, and to manage their own lives, thereby serving not only their own interests but also the progress of society itself.¹

Student assessment was regarded as one tool toward this end. The key recommendations relevant to assessment were in part a reaction to the low standards in then-current implementations of “minimum competency” examinations. The report called for standardized tests of achievement to be administered “at major transition points” between the levels of schooling in order to “certify the student’s credentials; identify the need for remedial intervention, and identify the opportunity for advanced or accelerated work.” Moreover, the report called for a national system of state and local tests that should “include other diagnostic procedures” to help teachers and students evaluate student progress (p. 28).

This was a modest set of recommendations for achievement tests, and much of what was called for is part of current practice. However, we now face a new set of challenges. Readers interested in a historical sketch of

---

¹ *A Nation at Risk*, p. 26.
the developments and challenges prior to 1983 and into the 21st century will want to read a number of works by Lorrie Shepard. My aim here is to offer brief reflections on three somewhat overlapping but enduring issues in educational assessment: validity, construct underrepresentation in outcome measures in intervention studies, and the burdens on the teacher of appropriate classroom assessment.

VALIDITY

Validity is the most important issue in educational assessment, and if we are to have the national system of assessments called for in A Nation at Risk, it is clearly a prime concern. Shortly after the report appeared, the Standards for Educational and Psychological Tests (hereafter Standards) in 1985 and Samuel Messick’s chapter on validity in Robert Linn’s Educational Measurement in 1989 clearly distinguished the concept of validity from earlier conceptions that a test was valid to the extent that it measured what it purported to measure. The shift in thinking was toward the validity of inferences made from test scores and the uses and consequences of testing. As defined by Messick, “Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment.” The 1999 edition of the Standards continued to support a unified concept of validity in which all validity is construct validity (a construct is the characteristic or concept that the test is designed to measure), the test professionals (jointly the developer and the user) are expected to specify what “construct interpretation” will be made based on the test score or score patterns, and the process of validation is the marshaling of evidence from multiple sources related to each of the intended inferences to be made from the test scores and patterns and the uses to which they will be put. The approved sources of evidence in the 1999 version reflect developments in research and test use since 1985. These include, for example, evidence based on the response processes of test-takers, convergent or discriminant evidence (e.g., evidence of different methods of measuring the construct), and evidence based on the consequences of the testing. However, validity remains a contested issue, as made clear by a 2007 critique in Educational Researcher attempting to overturn Messick’s definition.

For a comprehensive look at the enduring challenges of validity in large-scale assessment programs, a useful source is the 2002 volume edited by Gerald Tinker and Thomas Haladyna, Large-Scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation. This volume stresses accountability, not just to the larger public but also to effective professional practices and standards.

CONSTRUCT UNDERREPRESENTATION

Messick argued for giving more attention to construct underrepresentation in the criterion measure — the leaving out of an outcome measure some significant part of what it is intended to measure. In 2004, Thomas Cook and Richard Shavelson and Lisa Towne drew attention to this construct underrepresentation as one of the major difficulties in even the best-designed educational experiments using randomized trials to evaluate interventions. For example, a study of inquiry approaches to the teaching of mathematics may have as the measure of outcomes an easily accessible state assessment that does not tap critical aspects of what the inquiry-oriented intervention was intended to accomplish. And when the intervention itself is not assessed, this creates serious difficulty for any practically useful interpretation of the experimental outcome.

Part of the solution lies in attention to research design and implementation as laid out in the Standards. One might also consider choosing from a wider range of research methods, depending on the stage of development of an innovation, such as in the case of development of a new curriculum or assessment tool. Separate validation is required, of course, for each of the wide range of uses of test scores summarized by Thomas Haladyna. Haladyna discusses such uses of test scores as college admissions, scholarship criteria, high school graduation or certification, grade promotion, curriculum evaluation, and diagnosis of student learning. It is the inference and use that is to be validated, not the test. Similarly, separate validation is required for the long-term time scale for some outcomes, as noted by Connie Della-Piana and Gabriel Della-Piana in a recent article.

DEMANDS OF CLASSROOM ASSESSMENTS

Measuring performance on open-ended cognitive processes and problem solving puts heavy cognitive and management demands on the teacher. In Knowing What Students Know, a publication of the National Research Council, the authors draw on advances in cognitive science and measurement to broaden conceptions of what aspects of learning it is important to assess. The authors see current assessments as impoverished with respect

APRIL 2008 591
to what is observed and how it is interpreted. In their place, they would insert an assessment triangle consisting of 1) a model of student cognition and learning in the specific domain of knowledge, 2) a concern for the kinds of observations that best provide evidence of students’ competencies, and 3) a process of interpretation that yields more valid and fair inferences about student achievement. Such assessment would allow researchers to study the impact of these practices on student learning, on teaching, and on decision making about curriculum and schools. I present two simple examples to illustrate the complexity of such assessments and the interpretative and management demands on the teacher.

First, what does it take to “make sense of” writing test scores? In a 1993 article, “Teacher, What Does My Writing Test Score Mean?” I showed how analytically scored writing assessments yield identical scores for two students with different patterns of errors that are not identified, which leaves the diagnostic burden to the teacher. For example, two students may earn identical scores of 1 or 2 on a 4-point scale for problems with “conventions” (spelling, punctuation, capitalization). One student, however, when asked to read her unpunctuated, misspelled writing reads with proper intonation, fluency, and correct pronunciation of misspelled words (such as aprtment, hgeren poroxide, dropped, scraht, chaut for caught, nkt for next); another student does not. Each student clearly needs different remediation, but the test doesn’t offer any help to the teacher.

The second illustration is from validation of problem solving in mathematics, as described in Knowing What Students Know. In one task, students were asked to write an explanation for their responses to the question: “For what would you want your blood checked if you were having a transfusion?” The scoring scheme for analysis of the written explanation was a “concept map” of expert teacher solutions. The concept map was a diagram with “transfused blood” in a rectangle and arrows to such items as HIV/AIDS, disease factors, types of antigen, and blood types. The test “appeared” to get at student understanding, but the analysis and interviews with students revealed that the scoring overstated student understanding and reasoning. One can easily see the demands on a teacher’s skill, time, and management that such analyses require to make sense of student explanations.

However, there are hopeful signs of new assessments and assessment practices in development. And legislation has been proposed by U.S. Sen. Russ Feingold (D-Wis.) for an “Improving Student Testing Act.” In short, while much of the machinery we would employ to reform our schools is driven by the engine of assessment, enduring assessment issues are still with us a generation after A Nation at Risk.


8. E. A. Suchman apparently first proposed adapting the evaluation method to the stage of development of an innovation or intervention, as cited in Lee J. Cronbach et al., Toward Reform of Program Evaluation: Aims, Methods, and Institutional Arrangement (San Francisco: Jossey-Bass, 1980), pp. 107-235. Cronbach and his associates modified Suchman’s stages to: breadboard (in engineering, the first empirical trial open to modification), superrealization (ideal tryout conditions), prototype (realistic conditions of tryout), and established program.


Copyright Notice
Phi Delta Kappa International, Inc., holds copyright to this article, which may be reproduced or otherwise used only in accordance with U.S. law governing fair use. MULTIPLE copies, in print and electronic formats, may not be made or distributed without express permission from Phi Delta Kappa International, Inc. All rights reserved.

Note that photographs, artwork, advertising, and other elements to which Phi Delta Kappa does not hold copyright may have been removed from these pages.

All images included with this document are used with permission and may not be separated from this editorial content or used for any other purpose without the express written permission of the copyright holder.

Please fax permission requests to the attention of KAPPAN Permissions Editor at 812/339-0018 or e-mail permission requests to kappan@pdkintl.org.

For further information, contact:

Phi Delta Kappa International, Inc.
408 N. Union St.
P.O. Box 789
Bloomington, Indiana 47402-0789
812/339-1156 Phone
800/766-1156 Tollfree
812/339-0018 Fax

http://www.pdkintl.org