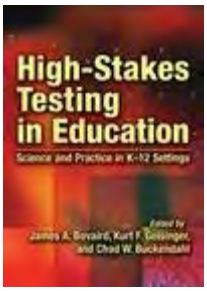
PsycCRITIQUES
October 12, 2011, Vol. 56, No. 41, Article 1



Is Public Education Improved Through High-Stakes Testing? Can It Be?

A Review of

High-Stakes Testing in Education: Science and Practice in K–12 Settings
by James A. Bovaird, Kurt F. Geisinger, and Chad W. Buckendahl (Eds.)
Washington, DC: American Psychological Association, 2011. 263 pp. ISBN 978-1-4338-0973-6. $69.95

doi: 10.1037/a0025467

Reviewed by
Mark D. Shermis

I sit on three technical advisory committees (TACs): one for a state department of education that administers multiple annual statewide assessments (K–12), another for a not-for-profit entity that runs a multipart high-stakes assessment taken by students throughout the world (postsecondary), and one that creates and administers assessments for licensed professionals. For me, the beauty of *High-Stakes Testing in Education: Science and Practice in K–12 Settings,* edited by James Bovaird, Kurt Geisinger, and Chad Buckendahl, is that the 14 chapters in this well-written and edited book correspond closely to my TAC agendas. It is almost like getting a short tutorial on the issues faced daily by those who create and administer tests and attempt to validate test scores. This volume could also be used as a text or text supplement for graduate-level large-scale testing courses.

The first impressive feature of the book is that it is composed of papers delivered at the Contemporary Issues in High-Stakes Testing Conference that served, in part, as a celebration of the many important contributions that Barbara Plake, her students, and collaborators have made to the field of large-scale testing over the years. Aside from trying to ensure that the most salient issues in high-stakes testing are covered, the challenge in editing conference papers is that they all have different levels of specificity, writing styles, and (sometimes) audiences. The style maintained throughout the book is generally nontechnical and to the point, and very even. The editors were very creative in aligning the topics, which begin by addressing very specific concerns and then migrate to more general issues.

I had two general take-away messages from the book. The first is that the technical challenges for producing good high-stakes tests are demanding and relentless, and involve almost as much art as science. So, for example, Chapter 3 (Ferdous, Bechard, and Buckendahl) examines the challenges of setting performance standards on alternate assessments for students with disabilities. The authors describe some of the more commonly accepted standard-setting techniques, many of which rely on expert panelist review to determine cut points and some of which have high-stakes consequences. While the techniques employed in such work may reflect the coalesced wisdom of content and measurement experts, it may fall short of meeting the criteria for “science.”

In Chapter 8 (Cizek, Koons, and Rosenberg), the authors set out to determine how frequently test developers even studied the consequences for using their high-stakes educational tests. After systematically reviewing the educational tests in one edition of the Buros *Mental Measurements Yearbook,* the authors discovered that there was a dearth of information provided by test developers about the possible fallout from failing (or passing) one of their tests. So if a state chooses a particular cutoff for retaining children on the basis of their third-grade reading score, it may easily determine how many children that action would affect, but *consequential validity* would prompt the state with the follow-up, “What ultimately happens to these students as a result?” The authors discuss possible reasons why such questions are not pursued.

The second take-away message is that even those who are intimately familiar with the details of producing high-quality tests have significant concerns about how the enterprise has evolved over time. In Chapter 10 (Bandalos, Ferster, Davis, and Samuelsen) the writers take a hard look at high-stakes testing and accountability systems, beginning with the question, “What is the evidence that the systems are doing what they are supposed to do?”

Indeed, some of the data presented suggest that children in the United States may have lost ground since the advent of high-stakes testing. The authors also raise other questions that seem not to have been adequately addressed or have been overlooked: For example, questions regarding whether high-stakes testing may have narrowed the curriculum, whether there is even a match between instructional time allotted and the domains tested, and the degree to which the assumption of equity of instructional quality across classrooms is reasonable, to name a few.

All of the chapters, from Barbara Plake’s overview of the current state of testing in the United States (Chapter 1) to Kurt Geisinger’s projection of what testing might be like (Chapter 14), engage the reader with topics and coverage that are interesting, useful, and relevant.

A colleague and I were discussing what the fundamental problem is in U.S. K–12 education. (I had earlier performed an Internet search, and the search engine identified more than 700 unique educational problems in the United States.) We agreed that, if we had magical powers, we would banish the 30-kids-in-a-classroom concept, give every student an individual education plan (IEP), and, with a few minor exceptions, put them through a guided learning experience with something like problem-based learning at its core. Moreover, the school year would extend to 210 days of the year instead of 180.

Children would be grouped for music, drama, social events, special presentations, sports, recess, and the like, but age would play less of a role in moving students along through the educational system; rather, it would be students’ performance that would mark their advancement, and teachers would become data-driven managers of student progress. In this fantasy world, high-stakes testing, as it currently functions, would be irrelevant or impossible to implement.

In this alternate reality, what would be optimal is to embed a continuous stream of reliable and valid assessments within the curriculum so that they become part of the overall learning experience. Machine scoring of assessments would provide the backdrop for collecting the vast quantities of data and presenting it to the teacher–manager in a form that not only would provide accountability information but also would be focused primarily on formative assessment. Information would be collected throughout the school year rather than simply at one point in early spring. It is with some degree of irony that I mention that these types of performance evaluations are more likely to be found in licensing exams than they would in a public school setting.

In light of the speculative world above, the problem with U.S. high-stakes testing programs is that they reflect the current model of education and would be yet one more barrier (and have a plethora of lobbyists) to overcome if the individualization of education were to become a reality. (To be fair, states and testing companies would probably jump on the bandwagon with the newer embedded performance assessments if they could figure out a way to implement them on a widespread basis.)

A positive step forward is the work of the two main Race to the Top consortia, PARCC and SMARTER Balance, which are furiously trying to realign or create assessments to match the Common Core State Standards and to move beyond paper-and-pencil tests. But they are still working within the group framework that expects a trajectory of learning within the space of 180 classroom days. The tests produced by the consortia are likely to push some of the boundaries—perhaps even get at higher level thinking skills—but they are not expected to be radically different since the newer assessment techniques have yet to be scrutinized with the same level of exacting expectations as have been multiple-choice tests.

So if the dream of a tailored curriculum for each student were to ever manifest itself, some entity would have to take the first steps. Either testing companies, working in collaboration with those who develop curricula, will need to demonstrate how these new embedded assessments can work, or states will have to develop contracts to move this agenda forward. Otherwise we will be left with business as usual.