Issues in High-Stakes Testing Programs

Finbarr C. Sloane, Anthony E. Kelly

Theory Into Practice, Volume 42, Number 1, Winter 2003, pp. 12-17
(Article)

Published by Ohio State University College of Education

Finbarr C. Sloane
Anthony E. Kelly

Issues in High-Stakes Testing Programs

The debate over high-stakes testing programs plays out daily in newspapers, on TV, and in the business, education, legal, political, and research communities. This article examines some of the issues at the heart of this debate. Four main areas are covered: the types of tests used, the effects on student motivation and morale, the degree of alignment between the test and the curriculum, and the distinction between assessment of *learning and assessment* for *learning. The article concludes by highlighting the need for teacher input in crafting testing programs that maximize benefits in each of these areas.*

THE CURRENT EMPHASIS ON high-stakes testing has emerged during what might be called the "learning through standards and accountability" era of American education. This era, which is developing during the confluence of a number of important social and political factors, supersedes and combines the prior emphases of minimum competency testing of the 1980s and higher-order thinking of the 1990s (for a review, see Hamilton & Koretz, 2002).

Finbarr C. Sloane is a program director at The National Science Foundation, and Anthony E. Kelly is a professor of education at George Mason University.

What makes this era different is a political climate at the national level that views the apparently poor academic performance[1] of American students on, for example, international assessments such as the Third International Mathematics and Science Study (TIMSS) as a national problem appropriate for federal intervention (Schmidt, McKnight, & Raizen, 1997). The unprecedented calls for school accountability at the federal level coincide with the existence of almost universal state-level testing of student performance. Unlike the performance goals of the minimum competency era (Office of Technology Assessment, 1992), current performance goals typically include a variety of challenging content standards that began most noticeably in mathematics (NCTM, 1989), and now extend across the curriculum. Finally, unlike the efforts of the Clinton administration, there is as yet no attempt to design a single national test; rather, there is a mandate, in the 2001 No Child Left Behind (NCLB) legislation for national *testing*, but with the format of the test left up to individual states.

This article discusses some of the issues that are fueling the debate over current high-stakes testing programs. Four issues in particular are examined: the types of tests used, the effects on student motivation and morale, the degree of alignment between the test and the curriculum, and the distinction between assessment *of* learning and assessment *for* learning.

THEORY INTO PRACTICE, Volume 42, Number 1, Winter 2003

Types of Tests Used

There is disagreement over what constitutes an appropriate form of assessment for meeting tough content standards in a political climate in which demands for accountability in education are prevalent. Part of the tension comes from reconciling in one form of testing two important, but distinct, goals:

1. learning important content to internationally accepted standards; and
2. knowing how schools and students rank locally, statewide, and even nationally

The content *standards* goal demands criterion-referenced testing; the school or student *ranking* goal demands norm-referenced testing. It is difficult, if not impossible, to address both of these demands simultaneously, especially if a state is retrenching during tough economic times and does not wish to fund a variety of complementary testing programs (Heubert & Hauser, 1999; Office of Technology Assessment, 1992).

These forms of testing (criterion- and norm-referenced) are described in other sources (e.g., Airasian, 1996; Nitko, 1996; and articles in this issue). What is important for the reader to note is that the psychometrics (i.e., the technical and conceptual underpinnings) of large-scale high-stakes testing programs are primarily focused on serving the goals of norming and selection. By contrast, teacher-made tests and the "authentic," or portfolio, testing movement have directed their attention toward individual student mastery of specified content and problem-solving skills; that is, the standards goal (Pellegrino, Chudowsky, & Glaser, 2001; Romberg, 1992). There have been very few attempts to meld these goals, but some progress is being made, particularly the cognitive psychometrics of the Tatsuokas (Pellegrino, Chudowsky, & Glaser, 2001; Tatsuoka, 1990).

Furthermore, the construction of tests for ranking students or for norming purposes tends to involve the administration of objective (usually multiple-choice) items to large samples of students. This process produces economically tractable and defensible reliability indices for these ranking/norming purposes. By contrast, authentic test construction, especially for measuring the reform goals of creative problem solving by students of challenging subject matter, tends to be more difficult to do well since the format is often essay or open-ended, requiring well-formulated scoring rubrics. These formats are also more difficult to design, time consuming to administer, and costly to score. Often, such tests fail to meet the reliability criteria of more objective multiple-choice measures (Nitko, 1996). At the same time, defenders of these authentic tests claim that they measure knowledge that is purportedly broader and more applicable to life than the knowledge measured by multiple-choice tests (Black & Wiliam, 1998). This is not to say that one form of testing is better than the other; rather, that different goals bring different tradeoffs. The trick is to be clear about the policy goals; know the strengths and weaknesses of all testing instruments; recognize the political, social, and educational trade-offs involved in using one form of assessment over another; and, most importantly, not demand of any testing instrument performance for which it was not designed.

In sum, all tests, whether used for purposes of ranking, norming, or cognitive diagnosis, provide a particular form of evidence to support certain claims by certain groups. No test is valid or reliable in an absolute sense; and no test data are above criticism (Airasian, 1996). The question is, Given certain social goals and the limitations of any testing format, does the test provide useful data to support decision makers in pursuit of these goals? In other words, the use of a test (of any kind) does not relieve the user from the responsibility of decision making or from placing an interpretation on a score or set of scores.

Effects on Student Motivation and Morale

Much of the debate over high-stakes testing programs is fueled by arguments regarding their effects on students. The impact on student motivation and morale is at the center of this discussion. In a recent RAND publication (Hamilton, Stecher, & Klein, 2002), Stecher discusses the potential effects of high-stakes testing for students, teachers, school administrators, and policy makers. Table 1 outlines the positive and negative effects he describes for students.

Table 1
Potential Effects of High-Stakes Testing on Students

Positive Effects	Negative Effects
Provide students with clearer information about their own knowledge and skills	Frustrate students and discourage them from trying
Motivate students to work harder in school	Make students more competitive
Send clearer signals to students about what to study	Cause students to devalue grades and school assessments
Help students associate and align personal effort with rewards	

Source: Adapted from Stecher (2002)

The potential effects listed include changes to students' motivation and morale, both positive and negative. However, it should be noted that while high-stakes testing is a potential explanatory factor for these effects, it is not the only one. Thus, it is unlikely that these kinds of student outcomes can be explained solely (or even primarily) by the introduction and use of high-stakes testing (i.e., the tests will interact with the schooling contexts in which they occur).

In our opinion, tests (especially high-stakes tests) are sometimes unfairly criticized for their effects on student motivation for learning without sufficient recognition of the complexities of the research findings in this area (Kohn, 1999). For example, it is not always clear if the anxiety that students may show in a high-stakes testing situation is due to the tests themselves, or to generally inadequate preparation for learning (attributable to a variety of causes, including, perhaps inadequate instruction). Thus, it is important that teachers and policy makers not blame the thermometer for the fever.

In the debate over the impact of high-stakes testing on student motivation, it is critical that teachers stay abreast of research findings and not be swayed by rhetoric. For example, a study in Chicago found that for 102 low-achieving sixth and eighth graders who were placed in a high-stakes testing context, the majority of the students showed increased work efforts which, in turn, translated into higher gains in learning. At the same time, it should be noted that one third of the students showed no change (Roderick & Engel, 2001). In addition, a study of higher education students showed that frequent testing was more effective than frequent homework for improving their retention of information—particularly among low-achieving students (Tuckman, 2000). The important point to take away from these findings is that the impact of high-stakes testing on student motivation is not monolithic. Thus, understanding the effects of high-stakes testing on student motivation will involve understanding the complex interactions with many student, teacher, and school context variables.

Alignment Between the Test and the Curriculum

One of the criticisms of high-stakes testing during the 1980s was that the emphasis on minimal competency levels for students resulted in schools teaching directly to these minimal competencies rather than the broader curriculum (Madaus, 1983). By contrast, those involved in the standards-based curricular reforms of the 1990s advocated the use of authentic, or portfolio, assessments so that teaching to the test was more or less equivalent to good instruction. In other words, this reform movement set high standards and then demanded assessment practices that were aligned with the standards.

But the solution is not simple. Frustratingly, when students are solving complex cognitive challenges in difficult subject matter that are project-based and perhaps span many weeks (or even months), it becomes difficult to assign academic credit to individual students in a nonsubjective manner. This is even truer when the same projects require this effort to be executed in cooperative learning groups. This factor, along with the NCLB legislation, was part of the reason that states such as Maryland opted to change from a performance-based assessment to more objective tests similar to those used by Virginia and other states. On the other hand, even when objective measures are used (in which the alignment with the curriculum is, if not perfect, less fuzzy), states may not wish to face the inference that low scores on reliable tests imply that students have not mastered the curriculum

and so should be denied graduation, or additionally, that their schools are failing.

The test-curricular alignment problem necessitates our revisiting the distinction between objective and authentic. Content mastery at some level is a *cognitive* event: the understanding of powerful, complex, and sometimes fuzzy ideas. For that reason, at least for challenging content, it may be difficult to write clear and simple standards, thereby making their operationalization for curriculum development, test construction (of any genre, objective or authentic), and alignment between the two, problematic (Gronlund, 1998). Moreover, even when clear standards can be written, teaching to the test may inflate perceptions of student learning of the broader curriculum (Koretz, Linn, Dunbar, & Shepard, 1991). This can occur because content not covered in some state standards (but covered in others) may be neglected, or because students trained in one test format are less able to answer the same question in another format (Shepard, 1993).

Assessment *of* Learning Versus Assessment *for* Learning

Any test that demonstrates to an individual student that he or she is failing, is a high-stakes test (Heubert & Hauser, 1999). For low-performing children, a label of *low ability* can adhere for a lifetime and negatively impact their confidence in their learning ability. The fact that a state or the federal government can aggregate a child's score with others and denounce an entire school as failing provides little service to either the child or the school, unless it comes with substantial remedial resources.

In addition, high-stakes tests are given late in the school year. Rarely do they provide useful diagnostic information for the student or the teacher or diagnostic information that is available in a timely fashion. In any event, until the field of psychometric theory matures enough to generate reliable, valid, and technological solutions in real time, the teacher will likely have to draw more heavily on formative assessment techniques in order to foster ongoing student learning.

The distinction between high-stakes testing and these formative assessment techniques can be described as assessment *of* learning compared to assessment *for* learning. In the case of the former, the goal of the test is to measure what students know or can do. There is less, or no, emphasis on providing information to improve student learning. In the case of the latter, the goal of the test (always only partially achieved) is to provide information that will improve student learning. Thus, of interest here is not so much whether the student scores at a certain level, but why they do so and what can be done to help them move to the next level.

Students can be effective instruments in their own learning if the teacher is clear on the learning goals, and the students are informed of their current performance and given clear steps for remediation. This observation applies equally to homework (Black & Wiliam, 1998). The goal of formative assessment is to help the student learn, not to compare the student to others; most particularly, its goal is not to place labels on students. Formative assessment can take place continuously—and does—during formal instruction when teachers ask questions. It can also occur between students within small groups and among small groups if the tasks are well designed (e.g., Kelly & Lesh, 2000). The task for teachers is to know and understand their state's standards, and then translate this knowledge to continuously help students learn and self-assess to meet those standards. This concerted and focused effort should involve cooperation among school administrators, teachers, and parents.

At the classroom level, Cronbach (1977) provided four principles useful for guiding these assessments: (a) there should be learning targets that students seek to attain, (b) students should believe they can achieve these learning targets, (c) students should understand the degree to which they are attaining these learning targets, and (d) attaining classroom learning targets should lead students to apply their learning in authentic settings. Cronbach implicitly implies that grades can serve as the goals that some students seek to attain, although he also notes that (a) grades do not motivate all students, particularly those who feel that high grades are out of their reach, (b) students can use grades only to judge their progress when they are given appropriate feedback about what has been attained and how the grades were assigned (additionally, students will

need to know what the final performance state looks like), and (c) grades tend to be holistic judgments rather than descriptions of strengths and weaknesses. As such they fail to inform students of what they need to do better so that their current performance can be changed to meet and more closely resemble the final performance state. We close by offering Cronbach (1977), who notes: "Teachers must evaluate, but in my view they should do as little comparison and as little summary rating of individuals as the institutional setting allows." (p. 687)

Conclusion[2]

In line with the authors of a recent National Research Council report (Chudowsky, Pellegrino, & Glaser, 2001) we believe that the future of testing research lies in the interaction of individual cognition, instruction, and a new cognitively based psychometrics. When these three research fields fruitfully combine, we will more than likely see better tests, ones that are based on a cognitive model of student learning and development. These tests will be able to better capture diagnostic and instructional consequences. However, the fields as yet have only minimal overlap, and this has suboptimal consequences for students, teachers, administrators, and policy makers. In fact, we are still a long way from this National Research Council goal, given the real time needs of teachers and students. Until then, teachers and their students are left with a large information gap. However, this is a gap they must take more active responsibility for filling themselves.

We believe that the issue of assessment *of* learning and *for* learning belongs with those who take direct responsibility for learning: teachers and their students. We further believe (and emerging research supports) that the use of testing for accountability can cause a system to attend to assessments, but that accountability by itself is unlikely to lead to deep, or long-term, changes in teaching practices or student learning (Firestone, Fitz, & Broadfoot, 1999). Consequently, the teaching profession needs to actively engage in the testing debate, demanding more powerful psychometric theories and better instrumentation. Further, it must demand the resources to educate students at levels far beyond the brittle knowledge standards that are expressed in current high-stakes tests. As noted above, the coalition of individual cognition, instruction, and a new cognitively based psychometrics offers the possibility for tests (and assessment tools) to reach their potential, and better serve teaching and learning. Without an active and professionally based teacher voice in this conversation, it is highly unlikely that tests of the future will fully serve the dual goals of cognition and instruction, while, at the same time, responding to legitimate calls for accountability in the educational system.

Notes

The views expressed in this article do not necessarily reflect the views of the National Science Foundation.

1. We use the qualifier *apparently* because individual states' performance on TIMSS often exceeds the poor performance of U.S. children on the average.
2. It is not possible to fully explore the various sources of good ideas for formative assessment presented in this article. The reader is directed to the following as a starting point: the work of Minstrell and his theory of facets of learning (www.talariainc.com); the work of Lesh on model-eliciting problems (Kelly & Lesh, 2000); the work of Mazur on problems that students co-solve as a basis for further instruction (see ConcepTests at galileo.harvard.edu/galileo/lgm/pi); and the work on assessment by the National Institute for Science Education (www.flaguide.org).

References

Airasian, P.W. (1996). *Assessment in the classroom*. New York: McGraw-Hill.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Cronbach, L.J. (1977). *Educational psychology* (3rd ed.). New York: Harcourt Brace Jovanovich.

Firestone, W.W., Fitz, J., & Broadfoot, P. (2000). Power, learning, and legitimation: Assessment implementation across levels in the US and the UK. *American Educational Research Journal, 36*(2), 759-793.

Gronlund, N.E. (1998). *Assessment of student achievement* (6th ed.). Boston, MA: Allyn and Bacon.

Hamilton, L.S., & Koretz, D.M. (2002). Chapter 2: Tests and their uses in test-based accountability systems. In L.S. Hamilton, B.M. Stecher, & S.P. Klein (Eds.), *Making sense of test-based accountability in education*. Santa Monica, CA: RAND.

Hamilton, L.S., Stecher, B.M., & Klein, S.P. (2002). *Making sense of test-based accountability in education*. Santa Monica CA: RAND.

Heubert, J., & Hauser, R. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Kelly, A.E., & Lesh, R.A. (2000). *Handbook of research design in mathematics and science education*. Mahwah, NJ: Erlbaum.

Kohn, A. (1999). The costs of overemphasizing achievement. *School Administrator*, *56*(10), 40-42, 44-46.

Koretz, D., Linn, R.L., Dunbar, S.B., & Shepard, L. (1991, April). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Madaus, G.F. (Ed.). (1983). *The courts, validity, and minimum competency testing*. Boston, MA: Kluwer-Nijhoff.

National Council of Teachers of Mathematics (NCTM). (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Nitko, A.J. (1996). *Educational assessment of students*. Englewood, NJ: Prentice Hall.

Office of Technology Assessment. (1992, February). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Roderick, M., & Engel, M. (2001). The grasshopper and the ant: Motivational responses of low-achieving students to high-stakes testing. *Educational Evaluation and Policy Analysis*, *23*(3), 197-227.

Romberg, T.A. (1992). Assessing mathematics competence and achievement. In H. Berlack, F.M. Newmann, E. Adams, D.A. Archbald, T. Burgess, J. Raven, & T.A. Romberg (Eds.), *Toward a new science of educational testing and assessment*. Albany: State University of New York Press.

Schmidt, W.H., McNight, C., & Raizen, S. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Dordrecht/Boston/London: Kluwer.

Shepard, L.A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405-450.

Stecher, B.M. (2002). Chapter 4: Consequences of large-scale, high stakes testing on school and classroom practice. In L.S. Hamilton, B.M. Stecher, & S.P. Klein (Eds.), *Making sense of test-based accountability in education*. Santa Monica CA: RAND.

Tatsuoka, K.K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.

Tuckman, B.W. (2000, May). *The use of frequent testing to increase students' motivation to achieve*. Paper presented at the 7th Workshop on Achievement and Task Motivation: An International Conference on Motivation, Leuven, Belgium.