

Challenges for instructionally supportive accountability tests

Chad W. Buckendahl

Buros Center for Testing
University of Nebraska, Lincoln

Paper presented at the annual meeting of the
Northern Rocky Mountain Educational Research Association
Jackson Hole, WY

October 6, 2005

Challenges for instructionally supportive accountability tests

Current models of state or local educational reform rely on accountability systems that generally include widespread use of standards-based assessments. There have been different interpretations of how these accountability systems are developed and implemented. With the federal No Child Left Behind (2002) legislation driving many discussions and changes at the state level, there have been different responses for addressing the requirements of this policy. Although many have called for a reworking of the policy (e.g., Linn, 2005; Popham, 2004, Olson, 2005), there is often not opportunities for researchers to directly influence policy. Other research has recommended changes to the tests that are used in these accountability systems (Commission for Instructionally Supportive Accountability Tests, 2001; Keller, Moulding, Pellegrino, Popham, & Sandifer, 2004) as a way to influence the utility of the information produced. It is this second line of research that is the focus of this paper.

Discussions of testing programs often begin with questions about the validity evidence that supports the intended interpretations of scores in the program. There are many considerations when attempting to use test scores for multiple purposes. With the perceived diverse purposes of instructional utility and accountability, designing and collecting evidence to support this dual validity framework is obviously challenging. Keller et al.'s (2004) recommendations reiterate those advanced by the Commission (2001) for developing assessments that would inform accountability goals, but also be meaningful for educators. There are nine total recommendations; however, this paper focuses on two that have broad implications for educators and measurement practitioners.

The two recommendations that will be discussed here include: 1) measure mastery of a few priority standards and 2) report results for every important element of these priority standards. Each of these recommendations is discussed below.

Priority standards

For many states it is reasonable and advisable to consider prioritization of standards as a necessary first step in designing any manageable assessment or accountability system. Some states have been very detailed and discrete with how they have defined their standards (e.g., California, Texas). Other states (e.g. Iowa, South Dakota) have characterized their content standards more broadly with illustrative benchmarks or indicators of student performance of how students may demonstrate mastery of the standard. Any system that asks stakeholders to focus on a reduced set of standards will raise negative perceptions of narrowing the curriculum or “teaching to the test”. In these situations, the communicated message to stakeholders is that these priority standards are more important than the other ones. This may or may not be the message we wish to communicate.

When students’ performance on these prioritized standards is then used to make decisions about students, schools, or districts in an accountability system, it is almost assured that greater efforts will be placed on those elements that are influencing factors. If the incentives are created to perform on certain standards and not others, we can predict the behavior. The expectation that teachers will spend appreciable time on additional standards when they are not part of the accountability system is probably unrealistic. Furthermore, adherence to this recommendation appears to encourage the curricular reduction that Popham identified as an “assessment-sired instructional sin”

because educators would likely perceive articulation of prioritized content standards as an effort to minimize their broader view of the curriculum. This is not to suggest that following the recommendation would require a curricular reduction, but the perception that would be powerful and speaks to another challenge. How then do we combat these perceptions?

The measurement community is challenged to better educate the public and policymakers about basic assessment literacy skills. Braun and Mislevey (2005) describe how the public's and policymakers' naïve understanding of particular disciplines can lead to dogmatic beliefs that are difficult to overcome (e.g. 70% is passing) in the context of psychometric science. For example, the public has a fairly negative attitude regarding teachers "teaching to the test." In the national sample, 54% of respondents indicated that this was a "bad thing" (Rose & Gallup, 2005). However, is this really their perception? Had the question been phrased as "Is it a good thing or bad thing to test students' abilities on content or skills they have not been taught," we may have observed a different response. Crocker (2003) made a similar argument about this perception and the gap between how educators view the roles of assessment and instruction in her presidential address to NCME. Correcting these misunderstood perceptions is an ongoing battle.

Another concern about prioritization speaks to the level of detail or specificity of the content (or in some cases, process) standards. As mentioned above, some states have been more general in their adoption of content standards, whereas other states have chosen to be more discreet. In attempting to balance the tensions among content specialists, assessment specialists, and policymakers, there may be different responses to this goal. One positive response may be a more concise description of the most critical

knowledge, skills, or abilities that students should be able to demonstrate. This assumes that each stakeholder group agrees with the description. Another response, though, may be to simply consolidate a number of more specific standards into a larger, more general one. This action may result in an unintended consequence of creating standards that are perceived as too vague by users. Regardless of the shift in focus for a given state (responses that may range from more general to more specific), the challenge of appropriate representation of the domain in the curriculum, instruction, or assessment would still lead to interesting discussions.

Given that prioritization of goals and activities due to available resources must occur within any organization system, the incentives built into the accountability model of those systems will often guide how those priorities are set. Therefore, it could be perceived as unfair for a state to expect schools to provide students a full curricular opportunity to learn, yet only include measurement of the priority standards in the assessment and accountability system. When only a subset of the full curriculum (standards) are tested, guessing about which standards will show up on the state's tests remains a concern. However, if these sampled standards are revealed to schools to allow them to adequately prepare, the incentive system has effectively overridden the broader goal. If the standards that would be measured on a state's test are not revealed in advance to encourage broader instruction across the curriculum, security will then play a larger role because of the incentive to discover what would be tested to spend additional time on these elements.

Results for each curricular aim

My reaction to this recommendation is related to the Commission's initial recommendation that specifies mastery as the target ability level. This is an important discussion point because current federal requirements and most states require reporting students' performance at multiple ability levels (e.g., Advanced, Proficient, Basic, and Below Basic). In measurement, we recognize that when we want to make one decision, we need enough information about the construct to make the decision within a certain tolerable error band. However, when the number of decisions increases, so does the demand on the amount of information necessary to support those decisions. For example, states may be reporting students' abilities at levels like those suggested above, but only have enough measurement information (breadth and depth of content representation) to make a distinction between the Basic and Proficient levels. In this instance we would have little confidence in the additional classification decisions between the Below Basic and Basic levels or the Proficient and Advanced levels.

The challenge of sufficient measurement information and its connection to the ability to report information was illustrated recently in two studies. Buckendahl (2004) presented results from an empirical study that relied on educators' judgments in a Midwestern state that evaluated the capacity for local districts' assessments to provide information on multiple proficiency levels. The conclusions from this study suggested that some of the districts sampled in the study may be able to produce confident results when classifying students into Proficient or Not Proficient, but that few would be able to confidently make the multiple classifications that are required in NCLB. A second study by Ryan (2004) evaluated the distribution of content strands across ability levels in a Southern state. Results of this study suggested that there was a range of abilities

represented across content, but that within a given content strand items may cluster in a particular range of theta and not be evenly distributed across the range. Evaluating this source of validity evidence becomes more meaningful at the reporting level because it is at this point that scores are interpreted by users. Although it may necessitate a change in the regulations, focusing on mastery as the only decision point would likely offer opportunities for states to measure additional standards.

Although the opportunity to measure additional standards expands the sampling of a domain, there is still a concern about the amount of information needed to provide the level of detail that the Commission recommended with respect to reporting. More importantly, using results to meaningfully impact instruction suggests that the information is available to do so. Educators hold fast to the contention that each year's class of students has a range of abilities that differ from the previous year's class. Thus, if data for this year's class are not available until the subsequent year decisions about instructional interventions may be made after the student leaves the classroom or by individuals who may not be as familiar with a given student's abilities. Fall administration has been suggested as a strategy to combat this concern, yet data may still be unavailable until the end of the first semester. Altering the administration strategy also may not align with current instructional, assessment, and reporting requirements for some states. Meeting these challenges requires that we consider other strategies to meet what are often perceived as competing goals.

Alternative directions

Responding to the needs of both policymakers and educators can be daunting as the accountability expectations may not match with those that are instructionally sound.

Our focus though, should be on implementing more efficient strategies to collect meaningful information to serve both purposes. Technology, particularly in the adaptive testing arena, has the capacity to assist us in this goal. Just as the use of intelligent agents has made the internet increasingly customizable, years of adaptive testing research may now be at a point where the benefits outweigh the costs of implementation. Some states (e.g., North Carolina, Virginia) have begun to pilot test the use of computer-delivered adaptive tests. Another state assessment program in Idaho uses a hybrid computer-based testing approach with a fixed form test to report on standards that is then supplemented with an adaptive portion to provide greater information on students' abilities to teachers. These efforts are encouraging as the perceived amount of testing increases. However, recommendations to better use computer assisted technology in measurement are not new.

Approximately twenty years ago Cole (1986) suggested three uses of technology in educational assessment that today have become common practice. First, computers have been used extensively as data entry clerks for scoring tests, keeping track of students' progress, and general data management related testing programs. Second, computers have seen greater use as a tool that supports adaptive testing where each individual is tested with different questions depending on previous responses. Although somewhat limited in educational settings, these tests have been adopted more by large scale admissions (e.g., Graduate Management Admissions Test, Graduate Records Examination) or licensure (e.g., National Council of State Boards of Nursing licensure exams). This is not surprising given the costs associated with developing, administering,

and maintaining these programs. The third use would be to use computers to develop more complex stimuli (e.g., graphics, audio/video clips, interactive decisions).

Another advance is in the area of computer automated scoring of constructed response items. Computer algorithms and scoring programs have been developed to model human scoring patterns and to replicate their characteristics electronically without concerns of rater bias or fatigue (Yang, Buckendahl, Juszkievicz, & Bhola, 2002). This seemingly removes some of the subjective elements in large scale constructed response scoring. With the speed of the automated scoring, the time lag is greatly reduced allowing users to have access to the information more quickly than can be offered with human scoring. However, there continues to be concerns about interpreting the scores produced by these programs. For example, Keith (2003) describes the different sources of validity evidence from content and criterion dimensions (e.g., correlations with students' performance on achievement tests, GRE exam) that may be needed to support decisions. There is also a lingering fear of replacing something that is inherently subjective with an automated approach.

Wainer and Eignor (2000) also point out that adaptive testing is not a panacea and that there are some challenges to avoid. First examinee access (i.e., more test takers than available computers) may present a barrier to particular groups. Access may also mean previous exposure to computers and general technological literacy. Item pool usage and security are also issues because of the on-demand features. With these characteristics there needs to be a sufficiently large pool, item exposure controls, fraud detection, re-take policies, item bank suspension, expulsion, and replenishment policies to consider as part of the development and maintenance. There are also economic realities of developing or

implementing a computer adaptive test that may influence the feasibility for all but the largest programs.

Conclusion

This paper discussed two recommendations related to developing and implementing instructionally supportive accountability tests. Challenges to these recommendations were detailed along with an alternative strategy for attempting to meet these challenges. Criticisms are often levied against policymakers about the current uses of tests in reducing curriculum (primarily from educators) and reporting and using scores that may not be supported by the program's validity evidence (primarily from psychometricians). An alternative strategy that suggests an increased use of technology to address these challenges was also briefly discussed.

To move beyond the current model in many state testing programs of treating assessment as an event rather than part of a basic educational model that integrates curriculum, instruction, and assessment (Shepard, 2000), we need to move assessment into the curriculum sequence. There is also a need to encourage alternative practices in this area. If not, we will perpetuate educators' and policymakers' beliefs about the role of assessment in education.

References

- Braun, H. I. & Mislevy, R. (2005). Intuitive test theory. Phi Delta Kappan, 86(7), 489-497.
- Buckendahl, C. (2004). Evaluating sufficiency of measurement in state assessment: A judgmental approach. Presentation at the Large Scale Assessment Conference, Boston, MA.
- Cole, N. S. (1986). Future directions for educational achievement and ability testing. B. S. Plake, J.C. Witt, and J. V. Mitchell, Jr. (Eds.) The Future of Testing (pp. 73-88). Lincoln, NE: Buros Institute of Mental Measurements.
- Commission on Instructionally Supportive Assessment (2001). Building tests that support instruction and accountability: A guide for policymakers. Washington, DC: Author.
- Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. Educational Measurement: Issues and Practice, 22(3), 5-11.
- Keith, T. Z. (2003). Validity of automated essay scoring systems. In M. D. Shermis and J. C. Burstein (Eds.) Automated Essay Scoring: A Cross-Disciplinary Perspective (pp. 147-167). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Keller, T., Moulding, B., Pellegrino, J., Popham, W. J., & Sandifer, P. (2004). Instructionally supportive accountability tests in science: A viable assessment option? Washington, DC: National Academy of Sciences.

- Linn, R. L. (2005, April). Test-based educational accountability in the era of No Child Left Behind. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Olson, L. (2005). Requests win more leeway under NCLB. Education Week, 24(42), 1.
- Popham, W. J. (2004). Shaping up the “No Child” Act: Is edge-softening enough? Education Week, 23(38), 40.
- Popham, W. J. (2003). Seeking redemption for our psychometric sins. Educational Measurement: Issues and Practice, 22(1), 45-48.
- Rose, L. C. & Gallup, A. M. (2005). The 37th Annual Phi Delta Kappa/Gallup Poll of the public’s attitudes toward the public schools. Phi Delta Kappan, 87(1), 41-57.
- Ryan, J. (2004). Evaluating sufficiency of measurement in state assessment: An empirical approach. Presentation at the Large Scale Assessment Conference. Boston, MA.
- Shepard, L. A. (2000). The role of assessment in a learning culture. Educational Researcher, 29(7), 4-14.
- Wainer, H. & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer (Ed.) Computerized Adaptive Testing: A Primer 2nd ed (pp. 271-300). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Yang, Y., Buckendahl, C., Juskiewicz, P., & Bhola, D. (2002). A review of strategies for validating computer-automated scoring. Applied Measurement in Education, 15(4), 391-412.