The Dark Side of Assessment Literacy: Avoiding the Perils of Accountability

Thomas R. Guskey, PhD
Senior Research Scholar
College of Education and Human Development
University of Louisville
Lexington, KY

Abstract

Educational measurement and evaluation experts generally agree that increasing stakeholders' assessment literacy will yield a variety of positive benefits, especially broadening the range of assessment formats teachers use to measure students' mastery of high level, more cognitively complex learning outcomes. But in the context of education accountability as currently structured in American schools, such efforts also may lead teachers to become more sophisticated in test preparation activities and to narrow both their instruction and classroom assessment practices specifically to enhance students' performance on prescribed, annual high-stakes accountability assessments. This article explains why that is so, describes the process by which it occurred in one state, and offers specific suggestions as to how it might be avoided.

Key Words

assessment, assessment literacy, accountability, classroom assessment, high-stakes assessment, matrix sampling, teacher attitudes, teacher commitment.

For nearly three decades, prominent experts in educational measurement have stressed the importance of assessment literacy (Popham, 2006, 2009, 2011; Stiggins, 1991, 1995; Xu & Brown, 2016). Some argue it may be the single most cost-effective way to improve our schools (Popham, 2018a). Assessment literacy is generally thought of as "the knowledge about how to assess what students know and can do, interpret the results of these assessments, and apply these results to improve student learning and program effectiveness" (Webb, 2002, p. 1). More recently Popham (2018b) described it as simply "an individual's understanding of the fundamental assessment concepts and procedures deemed likely to influence educational decisions." (p. 2).

Improving assessment literacy could yield numerous positive benefits. It could broaden the ways teachers gather information on student learning and use that information to improve instruction. It could enhance students' use of assessments so they become more effective learners. It might even expand parents', families', and community members' interpretations of assessment results and encourage greater involvement in education endeavors.

Clearly the more stakeholders know about assessment techniques, interpretation, and use in decision-making, the better will be the educational decisions they make based on assessment results.

Education accountability systems as they are currently structured in the U.S., however, cast assessment literacy in an entirely different light. In the context of high-stakes accountability, increasing educators' assessment literacy could serve an unintended and far a more disconcerting purpose. This article explains that troubling purpose, why it is likely, and what education leaders must do to avoid it.

Structure of Accountability Systems

Accountability systems in the U.S. emerged from increasing political involvement in education. They began with the *No Child Left Behind Act* (U.S. Congress, 2001) that made educators accountable to the general public for specific student achievement outcomes (Anderson, 2005).

Early accountability systems focused primarily on annual measures of student achievement in language arts and mathematic gathered in grades 3 through 8 and one year beyond. As these systems evolved, they expanded to include achievement in science and social studies, and took into account other measures such as attendance, promotion/retention rates, and graduation/dropout rates.

They further required that results be disaggregated to show progress among different subgroups of students (i.e., economically disadvantaged, English learners, ethnic or racial minorities, and students with disabilities) and to confirm reductions in achievement gaps. The *Every Student Succeeds Act* (U.S. Congress, 2015) has preserved annual grade-level testing but is less prescriptive about how the results are used in accountability systems.

The main challenge in modern accountability systems, of course, is how to accurately and reliably measure these student learning outcomes. Policy-makers and legislators typically pose the additional requirements on accountability systems that assessments of student learning not be too costly and be administered and scored efficiently so they do not require inordinate amounts of students' time.

Development of Accountability Measures

States varied in their approach to measuring these student learning outcomes. Most relied on external vendors to develop their assessments, trusting these vendors to ensure the assessments were aligned with the state's standards for student learning (Polikoff, Porter, & Smithson, 2011). Kentucky led the way in these efforts, establishing a statewide assessment and accountability system designed by experienced practitioners and several top experts in educational assessment (see Guskey, 1994).

A central feature of the Kentucky assessment program, known as the Kentucky Instructional Results Information System (KIRIS), was "on demand" performance events designed to assess students' higher level cognitive skills in several subject areas. These performance events required students to work together in teams to explain phenomenon or to find solutions to complex problems.

For each performance event, a small group of three or four students from a class or grade level was selected to engage in the event. Students worked on the tasks as a group but then prepared individual, written responses to specific questions or prompts regarding the event. Each student completed four events in the areas of math, science, and social studies. Some events were made interdisciplinary, however, combining science and math or math and social studies.

For example, a group of four students might be asked to observe and record data measuring the distance balls made of different materials bounce when dropped from a specific height. Based on their observations, the group would produce specified data tables or other products. From this information, each student was then asked to answer questions

individually that would depend on how well the group worked together to make the observations and record the data (Trimble, 1994).

Matrix Sampling

Research at that time showed that to get an accurate depiction of students' achievement of higher level cognitive skills in science or other subjects requires completion of 10 to 12 well-constructed performance tasks (Shavelson, Baxter, & Pine, 1991, 1992; Dunbar, Koretz, & Hoover, 1991; Messick, 1992). If each task in science took just ten minutes for students to complete, that would require two hours of testing time in science alone. Therefore, to economize the assessment process, the decision was made to use a strategy of "matrix sampling" for the performance events.

In matrix sampling, a substantial number of exemplary performance events, typically 12 or more, are designed for each grade level. Groups of three or four students randomly selected from each class or grade level complete four of the events, with each group completing different events Although no student completed every event, this allowed all events to be completed by some students at each grade level and all students to be involved in the assessment.

Results yielded fairly accurate and reliable estimates of students' achievement of higher level skills in science *at the school level*. If tasks and prompts from each event were well calibrated and reasonable numbers of students in various subgroups (i.e., ten or more) at each level completed events, it also permitted disaggregation of results for meaningful comparisons among student subgroups. Furthermore, because each student completed only four events, testing time in science was drastically reduced. But because each student completed only a limited number of events,

scores were not reliable at the individual student level; *only* at the school level. Since accountability focused on the school level, however, this issue was of little consequence.

Commitment of Teachers

Teachers want their students to succeed in school and to be confident in themselves as learners. They also want to feel they can influence students' learning and contribute to that success. These aspirations extend to students' performance on assessments that are part of accountability systems. Because of the important consequences attached to results from these assessments for students, for their families, for school leaders, and for the teachers themselves, students' performance on these assessments typically becomes a vital concern.

The Kentucky Instructional Results Information System (KIRIS) was clearly high-stakes for schools, school leaders, and teachers. It included financial rewards for schools that showed improved results and sanctions for schools that were not improving. State officials encouraged schools to provide teachers with the training necessary to prepare students for the new challenges of these performance-based assessments in science and other subjects.

Policy with Consequences Drives Practice

The effects on teachers' instructional activities of attaching high-stakes consequences to the results of performance assessments in science were profound. Not only did teachers begin to allocate more time to science lessons, they altered the way they taught science and the way they measured student learning on classroom assessments. Science lessons at all levels included more experiments and lab projects, and assessments involved data summary and interpretation, often integrating mathematics skills (Oldham, 1994).

The pressure for improvement in scores prompted many schools to devise professional development programs focused on the assessment formats and scoring procedures included in the accountability program (Cody & Guskey, 1997). A Rand investigation showed, for example, that all surveyed principals reported encouraging teachers to use materials specifically designed to guide students in inquiry-based events (Koretz, Barron, Mitchel, & Stecher, 1996). As a result, teachers included more performance tasks and authentic experiments as part of their instruction in science. They also taught students strategies for adapting their reporting based on specific scoring rubrics (Guskey & Oldham, 1996).

Funding Drives Policy

Unfortunately, these changes in teachers' instructional practices were short-lived. A newly elected group of state legislators who did not fully understand the matrix sampling procedures and were not particularly assessment literate raised concerns about assessment costs. Developing and piloting the performance events was costly. Scoring students' written responses to the science performance tasks was both time-consuming and expensive. In addition, although accountability remained focused at the school level, these legislators were concerned about the lack of reliability of scores at the individual student level.

Their response to these concerns was to impose drastic changes in the science assessments. Specifically, they wanted the assessments to require less time to administer and score in order to reduce the per-student costs. In addition, they wanted the assessment program to yield reliable data at the individual student level rather than just the school level.

Meeting these demands from legislators left the educational measurement experts who directed KIRIS with few options. The performance events were eliminated from the science assessments, as were the portfolios of student work that had been a foundational component of the language arts assessments. The statewide accountability assessments were returned to a more limited response format consisting of mostly multiple-choice items with a few extended-response items in each subject area.

The response of teachers to these changes in assessment format was predictable and immediate. Wanting to ensure their students did well on the new, restricted-response format science assessments, teachers revised their classroom assessments to more closely parallel the state assessments in science. Instructional strategies that resembled the performance events were abandoned in favor of activities and practices that prepared students for the more limited response format of multiple-choice items and brief, extended-response items.

As numerous studies have shown, teachers focus on the content tested and the way it is tested (Herman, 2004; Herman & Linn 2014). Arguments posed by state leaders in science education that students would do well on these restricted-response assessments when taught through a more inquiry-based approach to science fell on deaf ears. The teachers felt compelled to prepare their students for precisely what they would be asked to do on the new restricted-response, accountability assessments.

New Focus on Assessment Literacy

So what will result today from increasing stakeholders' assessment literacy? Ideally it will broaden teachers' understanding of how to construct authentic assessments that tap student's performance in real-world contexts. It will help teachers design assessments that yield reliable results and are well-aligned with high level, cognitively complex student learning goals. Teachers will also know better how to gain valuable evidence from demonstrations, performances, projects, exhibits, and digital portfolios that can be used to guide improvements in instruction and student learning.

Increasing students' assessment literacy will improve their use of assessment results to guide the correction of learning errors and help them become better managers and self-regulators of their own learning. Enhancing the assessment literacy of parents, families, and community members will inform their interpretations of assessment results. They will better understand what assessment results mean and the limitations of those results when drawing conclusions about the quality of instructional programs and schools.

But in the context of high-stakes accountability, where assessment-based decisions have serious and sometimes irreversible impact on the lives of students and their teachers both during school and afterward, increased assessment literacy also may lead teachers on a very different path. It may help them target their instruction and classroom assessments even more specifically on test preparation tasks.

Instead of broadening the array of assessment formats they employ, it actually may narrow what they teach, how they teach, and how they assess student learning to align more directly with the content and processes of those high-stakes assessments. It may make them even more highly skilled at focusing their instruction and classroom assessments on ways to improve students' performance on the limited but less expensive assessment formats

that provide the foundation for many of today's education accountability systems. And teachers will do this for noble reasons: because they care about the consequences attached to performance on those high-stakes assessments for their students, for them as teachers, and for their schools.

The Solution

This is not to suggest that efforts to improve the assessment literacy of all stakeholders should be abandoned. Teachers especially need help to broaden the ways they gather information on student learning and use that information to design effective instructional activities. They also need guidance in how to involve students in the assessment process so that students become insightful judges of their own performance and better self-regulators of their learning progress.

To avoid the unintended and potentially negative consequences that might accompany these efforts to improve assessment literacy, however, we must do two things. First, we must focus increased attention on perhaps the most influential but often most neglected group of stakeholders: policy-makers and legislators (see White, 2018). School leaders at all levels must make efforts to help these important decision-makers become more literate in every aspect of the assessment process.

In particular, policy makers and legislators need to understand that accountability assessments should model the types of assessment formats we hope teachers will use in their classrooms both to measure student achievement and to guide improvements in teaching and learning. In this way, teachers can teach to tests that are truly worth teaching to, and test preparation becomes a valuable instructional practice.

Credible high-stakes accountability assessments should focus on important 21st century learning goals, such as solving complex problems, reasoning and applying what is learned in new and different situations, communicating effectively, working collaboratively with classmates, and using higher cognitive processes. The best accountability assessments will also reflect authentic tasks and real-world contexts.

Assessments composed of multiple-choice and short, extended-response items certainly have their place and purpose. They offer an efficient and relatively inexpensive way to gather information about an important but fairly narrow range of student learning outcomes. Nevertheless, their limitations in measuring complex reasoning, communication, creativity, problem-solving, and other important learning goals must also be recognized.

Second, we must ensure the development of high-stakes accountability assessments is guided by valued learning goals rather than simply efficiency and cost. Cheap tests that don't measure the right things will not help us improve education. They are a waste of time and money, and a disservice both to educators and the students they teach. Increasing stakeholders' knowledge of the most valid means of capturing evidence on students' achievement of important 21st century learning goals will lead to more purposeful accountability assessments.

The Partnership for Assessment of Readiness for College and Careers (PARCC) assessments are a positive step in that direction. Although developing, administering and scoring these types of assessments will be somewhat more costly, the payoffs in terms of

students better prepared for success in school and beyond are vitally important.

With greater assessment literacy, policy-makers and legislators can demand better quality products from the vendors they hire to develop their state's accountability assessments.

They will understand the diverse assessment formats this requires, particularly performance events, projects, demonstrations, and portfolios of students' work. They also will understand the difference between

reliability at the school level versus the individual student level, and know how school level reliability opens up a broader range of authentic assessment formats that can be employed with reasonable cost.

Increasing assessment literacy among stakeholders in the assessment process will help improve our schools, but only if efforts also target the policy-makers and legislators who make the important decisions about the format and structure of high-stakes accountability assessments.

Author Biography

Thomas Guskey is a senior research scholar in the University of Louisville's College of Education and Human Development. He is Professor Emeritus at the University of Kentucky. E-mail: guskey@uky.edu

References

Anderson, J. A. (2005). *Accountability in education*. Paris, France: International Academy of Education and International Institute for Educational Planning (United Nations Educational, Scientific and Cultural Organization).

Cody, C. B., & Guskey, T. R. (1997). Professional development. In J. C. Lindle, J. M. Petrosko, & R. S. Pankratz (Eds.), *1996 Review of Research on the Kentucky Education Reform Act* (pp. 191-209). Frankfort, KY: The Kentucky Institute for Education Research.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*(4), 289-303.

Guskey, T. R. (Ed.) (1994). *High Stakes Performance Assessment: Perspectives on Kentucky's Educational Reform.* Thousand Oaks, CA: Corwin Press.

Guskey, T. R. (2008, March 16). Selecting school tests a matter of competence. *The Lexington Herald-Leader*, p. D-1 & 6. Available at: http://www.kentucky.com/589/story/348307.html

Guskey, T. R., & Oldham, B. R. (1996). *Despite the best intentions: Inconsistencies among components of Kentucky's systemic reform*. Paper presented at the annual meeting of the American Educational Research Association, New York City, NY.

Herman, J. (2004). The effects of testing on instruction. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 141-166). New York, NY: Teachers College Press.

Herman, J., & Linn, R. (2014). New assessments, new rigor. *Educational Leadership*, *71*(6), 34-37.

Koretz, D. M., Barron, S., Mitchell, K. J., & Stecher, B. M. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: Rand Corporation.

Messick, S. (1992, June). The interplay of evidence and consequences in the validation of performance assessments. *ETS Research Report Series*, *1992*(1), i-42.

Oldham, B. R. (1994). A school district's perspective. In T. R. Guskey (Ed.), *High stakes performance assessment: Perspectives on Kentucky's educational reform* (pp. 55-63). Thousand Oaks, CA: Corwin Press.

Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal*, *48*(4), 965-995.

Popham, W. J. (2006). Needed: A Dose of assessment literacy. *Educational Leadership*, *63*(6), 84-85.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice*, *48*(1), 4–11.

Popham, W. J. (2011). Transformational assessment in action. Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W. J. (2018a, June). *Expanding assessment literacy: A pitch to American publishers*. Presentation to the Council of Chief State School Officers 2018 National Conference on Student Assessment, San Diego, CA.

Popham, W. J. (2018b). *Assessment literacy for educators in a hurry.* Alexandria, VA: Association for Supervision and Curriculum Development.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, *4*(4), 347-362.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Research news and comment: Performance assessments: Political rhetoric and measurement reality. *Educational Researcher*, *21*(2), 22-27.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, *72*(7), 534-539.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, *77*(3), 238.

Trimble, C. S. (1994). Ensuring educational accountability. In T. R. Guskey (ed.), *High stakes performance assessment: Perspectives on Kentucky's educational reform* (pp. 37-54). Thousand Oaks, CA: Corwin Press.

U.S. Congress (2001). *No Child Left Behind Act of 2001*. Washington, DC: Author.

U.S. Congress (2015). *Every Student Succeeds Act of 2015*. Washington, DC: Author.

Webb, N. L. (2002). Assessment literacy in a standards-based, urban education setting. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

White, R. S. (2018). Who do state policy makers listen to? The case of teacher evaluation. *Phi Delta Kappan*, *99*(8), 13-18.

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, *58*(1), 149-162.