Journal of Research Practice
Volume 14, Issue 2, Article M2, 2018

ISSN: 1712-851X

Main Article:

Toward a Dialogue: Following Professional Standards on Education Achievement Testing

Gabriel M. Della-Piana
Department of Educational Psychology, University of Utah
Salt Lake City, UT 84112-8914, UNITED STATES

Michael K. Gardner
(Same as above)
gardner@ed.utah.edu

Zachary M. Mayne
(Same as above)

Abstract

The authors describe challenges of following professional standards for educational achievement testing due to the complexity of gathering appropriate evidence to support demanding test interpretation and use. Validity evidence has been found to be low for some individual testing standards, leading to the possibility of faulty or impoverished test interpretation and use. In response to this context, measurement professionals have called for a theory of action including behavior changes of multiple agents involved in the testing process. Also, changing roles have been seen for a broad range of agents including test developers, those who influence testing, and those influenced by testing. Some of these roles are discussed and others illustrated with examples from practice. A sociocultural theory of action noted in the literature is proposed as a thematic guide to practice. The paper concludes with a call for dialogue and two research and development tasks that might advance practice.

Index Terms: educational assessment; educational measurement; professional standard; research process; research purpose; researcher's role; research collaboration; trajectory of development

Suggested Citation: Della-Piana, G. M., Gardner, M. K., & Mayne, Z. M. (2018). Toward a dialogue: Following professional standards on education achievement testing. *Journal of Research Practice*, *14*(2), Article M2. Retrieved from http://jrp.icaap.org/index.php/jrp/article/view/602/491

This article briefly describes the challenges of following the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014) (henceforward, only *Standards*) with respect to gathering validity evidence, putting together evidence and theory for test interpretation and use, and taking steps to respond to the consequent demands for improving practice of multiple participants in, or agents associated with, the testing process. Those demands include, but are not limited to, a need for training of participants in the testing process for changing roles in gathering and using appropriate evidence in the problematic context.

There are signs that low compliance with the *Standards* leaves gaps in evidence and theory to support intended uses of educational achievement measures. The potential of consequences for faulty or impoverished interpretation of test results for student performance and growth thus becomes problematic. Though the article is focused on issues around standards for achievement assessment in social science research on educational interventions, the authors see the discussion as relevant to issues around professional standards in other areas of scientific practice.

One response from the profession has been calling for detailed documentation of a "theory of action" for what a testing program is intended to accomplish in changed behavior. The profession has proposed a procedural theory based on the educational intervention and outcomes. In addition, in this article, a sociocultural theory suggested in recent literature is described and proposed, largely because it includes both *cognitive knowledge* and *self-beliefs* to account for what happens as a result of achievement testing. This theoretical perspective contributes to clarifying and focusing on the changing roles of participants in the testing process.

The potential roles for multiple agents associated with the educational testing process are noted from the literature and briefly illustrated from the authors' experience in testing and evaluation, thus providing a glimpse into how things might look in these changing roles. Some steps are suggested toward research and development practice that might advance educational assessment practice.

1. Challenge of Following the Standards

Following the *Standards* (AERA et al., 2014) in getting validity evidence to support interpretation and use of test scores for decisions is no mean task. Standards are reminders of our professional ideals. Test validation is determining "the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). It is not the test that is validated, but the proposed interpretation of test scores. Furthermore, if a test score is interpreted for a use that has not been validated, it is the responsibility of the user to provide a rationale for that interpretation and use, and to gather new evidence for that context of use (AERA et al., 2014, p. 24, Standard 1.4). Our own study (Della-Piana, Gardner, & Mayne, 2018) and other studies (Cizek, Rosenberg, & Koons, 2008; Sussman & Wilson, 2018) have documented low compliance or noncompliance with some individual standards, notably "consequential" evidence (i.e., effects of the testing itself on changes in curriculum, individuals, and organizations), "cognitive process evidence" (i.e., the thinking process of the test taker in coming to an answer to test questions), and making a validity argument (integrating evidence and theory

into a test score interpretation). This context of the *Standards* along with the following references to the *Standards* provide a background for understanding challenges facing valid test interpretation and use.

(a) ". . . individual standards should not be considered in isolation" (AERA et al., 2014, p. 7).

(b) What is an applicable individual standard, or even set of standards, in any given context is dependent in part on what is technically feasible (AERA et al., 2014, p. 6).

(c) Professional judgment on appropriate evidence and test interpretation in a given testing context must take into account knowledge of behavioral science, psychometrics, and the standards of the professional field that is the focus of assessment (AERA et al., 2014, p. 7).

(d) Various strands of evidence must be integrated into a sound validity argument in a coherent account of the degree to which evidence and theory support the intended interpretation of test scores for specific uses (AERA et al., 2014, pp. 11-12).

(e) Add to this complexity the difficulties in conducting peer reviews of student achievement test validity, due to the increasing private ownership of education tests and thus proprietary interests that constrain full access to the tests, supporting material for interpretation, the validation processes and data (Burch, 2009).

The challenge of interpretation of education assessments (integrating evidence and theory into a test score interpretation consistent with the *Standards*) is depicted by Pellegrino, Chudowsky, and Glaser (2001) as a process, at a minimum, of reasoning from test scores based on three key elements: "a model of student *cognition* and learning in the domain, a set of beliefs about the kinds of *observations* that will provide evidence of students' competencies, and an *interpretation* process for making sense of the evidence" (p. 44).

In such a challenging context, lacking appropriate validity evidence and facing the complexity of expertise needed for interpretation of test scores, we are left with the possibility in practice of not knowing the validity of proposed interpretations of tests. The consequences for the student are that interpretation of test scores and proposed uses may be faulty or impoverished and action based on interpretations and claims for effects may not be fully evaluated. Thus, the *Standards*and current context of test use provide challenges to both finding and gathering evidence and interpreting test scores (with evidence and theory) to support intended uses. This places demands on university and school district trainers of test developers, testing researchers, test reviewers, test takers, teachers, parents, science writers who address the issues in newspapers and other media, and test users to adapt to the context of testing framed by the *Standards.*Placing responsibility only on test developers, or opting out of testing in response to concerns about frequency of tests or test misuse (for diverse views on the "opt-out movement," see Edelman & Levy, 2016), does not capture the scope and complexity of responsibility. No simple algorithm guides how best to follow the *Standards* in the context of real-life practice.

2. Responses From the Educational Measurement Profession

The educational measurement profession has responded to the problematics of testing outlined above with diverse views on the nature of validity and values underlying appropriate test interpretation (Lissitz, 2009; Markus & Borsboom, 2013; Ryan & Shepard, 2008). However, for purposes of this article, we build on the published professional *Standards* as the best current professional consensus, noting that it recognizes diverse views and, in fact, notes the reasons for difficulty in application, due in part to the professional judgment required for application, the changing environment with respect to knowledge and technology, and the diverse contexts of practice.

2.1. Professional Call for a Theory of Action to Guide Testing Practice

One major response that builds on the *Standards* suggests a procedural documentation in support of intended testing program effects or consequences. The National Council on Measurement in Education, a professional body in the United States, recently published a position paper titled, *Position Statement on Theories of Action for Testing Programs*, with references to research, theory, and examples of this kind of documentation (National Council on Measurement in Education [NCME], 2018). The argument is that, if a testing program is intended to change behavior (e.g., of students, teachers, and school leaders), documentation should be provided for a theory of action including the following elements:

(a) A list of the intended outcomes (both short- and long-term) and the constituent parts of the testing program (i.e., test design and assessment reports)

(b) Causal mechanisms claimed or stated as responsible for the intended change

(c) The ideal implementation that the designers believe is likely to lead to the outcomes

(d) Anticipation of what needs to happen to obtain the desired effects, and what might lead to negative outcomes

(e) A regular evaluation to test the extent to which the program is operating as intended with intended effects and with revision of the theory of action informed by empirical evidence

With this statement, NCME has thus addressed a need for a procedural theory for testing programs to guide work on developing validity evidence. School districts should join with test developers and publishers, and appropriate professionals in measurement and related disciplines, to develop theories of action for their testing context. The NCME proposal is for a procedural theory of action. A social science based theory of action has been proposed in addition to the procedural one.

2.2. Choice of a Social Science Theory of Action

A recent special issue of *Educational Measurement: Issues and Practice* (Spring 2018, Vol. 37, No. 1) on the connection between large-scale assessment and classroom assessment, centered on student learning, includes an argument for use of a research-based sociocultural theory to guide classroom assessment rather than (or in addition to) procedural designs based on objectives embedded in interim and end-of-year standardized tests alone

(Shepard, Penuel, & Pellegrino, 2018). The choice of sociocultural theory used by these authors to guide classroom assessment is based on its integration of motivational aspects of learning, including self-beliefs entwined with cognitive development. One considers from this theoretical perspective how cognitive knowledge (recall, explanation, and problem-solving) and self-beliefs (self-efficacy, belonging, identity, ability to self-regulate activity) are jointly developed in *communities of practice* and how students may be harmed in an environment where they are incorrectly labeled by the assessment as unable or deficient. In a sociocultural theory of assessment, student-relevant interests, experiences, identities, and long-term trajectories of development inform instructional practice. Shepard et al. (2018) discuss much more including how learning in one context transfers to other contexts, coherence across levels of the system, avoiding grading to motivate, instead of as feedback to improve student work, and responsibility for development placed at the district level where resources are available and professional development is planned.

From the perspective of a sociocultural theory applied to testing, one looks at a student in a *trajectory* toward varied possible future realizations, knowing that test results can not only support cognitive learning, but also shape a student's sense of self and efficacy, and possible educational and professional futures over a long period of time. This too changes the role of trainers and also of teachers and students.

2.3. Selected Literature Reflecting Changing Roles of Students, Teachers, and Professionals

Without attempting an exhaustive review, it is worth noting that the educational measurement profession has not been silent on considering changing roles of all participants in the testing process. The professional body, NCME, also produced a *Position Statement on Student Participation in State Assessment* (NCME, 2017). The statement encourages parents and others to support student participation in state testing programs as a response to the drop in participation (in part due to the opt-out movement, Edelman & Levy, 2016) below the 95% level required by federal education law, and provides a rationale based on usefulness of data.

Susan Brookhart (2018) puts student participation in a larger context. Brookhart notes that the typical, and intuitive, notion many educational professionals follow is: How do we make large-scale assessments useful to classroom teaching and learning? In contrast, the common insight that Brookhart sees is that classroom assessment, tied to learning and decisions by teacher and student, and vertically coherent with district level theories, should be the foundation for large-scale assessment and other achievement-based assessment. Brookhart adds the following insights that frame the role of participants more specifically:

(a) Learning and learners must be the center of classroom/large-scale coherence assessment.

(b) Learners should be able to see assessments as "something we were supposed to learn" and see how to get there and where to go next.

(c) Learners should participate in setting some learning goals, and apply assessment criteria to their own learning, to current interests, and to peer assessment.

(d) Teachers must be seen as more than implementers of assessments and instructional resources provided by experts. Things are always changing, and teachers' work goes on while researchers and resource providers catch up with changes. Also, no matter how sound the resources, they are mediated through teachers with different abilities in their craft knowledge, wisdom of practice, and beliefs. That knowledge should inform work going forward. That work should also contribute to understanding validity in the face of multiple purposes of assessment, in the context of learning progressions, and with concern for the challenges of who is responsible for different parts of the work to be done in test validation.

Another commentary in the *Educational Measurement* special issue (William, 2018) reminds the profession (including all those involved in the testing process) that student assessment should tap what the student can-do (with *scaffolding* and *affordances*) rather than, or in addition to, what the student does-do (under common testing procedures), and measure performances that generalize to other formats and what follows in a learning progression. "Affordance" is a term that is intended to capture all those environmental conditions that help a student perform on a test, including preparation for the test in the form of subject matter training and awareness of testing conditions. "Scaffolding" is a special kind of cognitive affordance that allows students to draw on their deep knowledge that they otherwise would not be able to do. For example, asking a student to edit a piece of writing without any scaffolding might miss what a student is capable of doing. On the other hand if the student is given, as scaffolding, a check list of things to look for (including such things as: "This doesn't sound quite right here" or "People won't be interested in this part") the student may demonstrate an editing ability far beyond what she did without the scaffolding.

Koretz (2017) joined other professionals who had issued warnings over several decades that the pressure to raise test scores would lead to cheating, finding other ways to cut corners, or failing, all of which are counter to the purpose of testing for understanding student performance and influencing cognitive development and a student's sense of self and efficacy. Koretz also notes that teachers would be evaluated on faulty measures. And that participants in the testing process would not be able to see the difference between test preparation and good instruction. Finally, Koretz contends that education, and testing in education, is a complicated and complex system that should lead those involved in the process to approach reforms with humility and some trepidation in trying out reforms informed by a wide range of participants in the testing process. He also notes that change will be difficult, expensive in time and labor costs, will include mistakes, and will require room for argumentation and the deliberation of those who influence testing and are influenced by it.

Perhaps most important from a teacher's point of view (as well as the point of view of those who influence the process) is that faulty or paltry test use must be countered by looking at a student in a trajectory of development, rather than point-in-time status. It changes how a teacher might assess and support achievement beyond the rubric of the formal test. For example, in a unit on scientific argumentation, assessment beyond the content of the argumentation might consider possible educational and work futures of students, such as the following:

(a) Scientist doing science and science report writing
(b) Scientist doing peer review
(c) Science story-teller
(d) Science illustrator
(e) Science poetry/lyrics/fiction writer
(f) Science policy shaper
(g) Science patent attorney
(h) Science-related public legislator
(i) Science-informed citizen

Then there is the problem of accessible expertise. The profession has been concerned about the training for needed expertise in testing and assessment for at least three decades. It has been argued that even if appropriate validity evidence were accessible for tests in use, the training and availability of needed expertise in the complex and changing environment is challenging (Brennan & Plake, 1991; Brookhart, 2011; Herszenhorn, 2006; Packman, Camara, & Huff, 2010; Sireci, 2000). That changing environment has also been addressed by a look into the future by Bennett (2018) on what to watch for in the changing world of assessment (to be revisited in the final section of the article).

2.4. What Might Changing Roles Look Like in Practice: Examples of Actions by Multiple Agents in the Testing Process

Given the complexities of establishing appropriate, valid, and useful testing within the constraints suggested above, and the varied roles suggested by the profession, it becomes clear that there are contributions to be made to the testing process (including validity evidence and test interpretation) by multiple agents or participants in the process. Given the implied variation in ways of contributing, each professional and participant in the testing process should pursue their part of the work of assessment depending on demands of their tasks and their skills and interest, asking themselves the question: What is my best role in the current context of assessment as it relates to student learning, especially from a sociocultural theory perspective. Drawing on our own experience we consider the following real scenarios:

Scenario 1. Peer Feedback on Writing

A teacher, randomly pairing students to give feedback to each other on their writing, notices that a low scoring boy on the state writing assessment was paired with a high scoring girl and moves over to observe them. Students had a scaffolding guide for critique in the form of seven statements, such as: (i) People won't be interested in this part; and (ii) This doesn't sound quite right here. The low scoring boy points to the girl's paper and says, "This doesn't sound quite right here". The girl says, "No it is fine. It is a new paragraph." Then on rereading it she says, "Wow. Transition. He found my error in transition between paragraphs." The teacher gets a better sense of a student's aptitude (see Corno et al., 2002, for research backing), one student gets a sense of self-efficacy, and another sees others in less of a one-dimensional way.

Scenario 2. Teachers' Response to State Assessment

A school principal looked at a survey of her fifth grade teachers and noted that most teachers reported following the new experimental program in mathematics including giving end-of-unit tests. But on her periodic "walk-through," she observed some teachers were so focused on the state assessment that they had changed end-of-unit test items to make them more like the state assessment. In continuing the walk-through and talking with teachers, she found that though many changed the end-of-unit tests, a few decided to use Monday and Wednesday to focus on the state assessment and the remaining days for the experimental program. The principal now has information to keep the school focused on both state assessment and the experimental program by finding out what made teachers change end-of-unit tests. She decided it was difficulty teachers had in working with students in small groups, as the experimental program called for.

Scenario 3. Reanalyzing Test Scores for Policy Insights

A school district assessment division used existing data to search for a new understanding, following a kind of "principled discovery" or looking back at data in a completed study for reanalysis from a theoretically driven question (Mark, Henry, & Julnes, 2000, pp. 258-265). The project reported high gains for point-in-time analyses of achievement disaggregated by demographics. Reanalysis for cohorts (rather than point-in-time) overall, and for Hispanic free lunch versus Hispanic non-free lunch, told a different story—smaller gains overall and non-free lunch flat, with free-lunch gaining. Naturally, if the data are not there, the reanalysis cannot be done. However, it illustrates what a district level research group can do to make sense of test scores by simply snooping around the results to come up with new questions and policy insights from the same data.

Scenario 4. Pitfalls of Generalization Across Test Formats

A district measurement professional and a university researcher combined efforts to ask and answer new questions about student achievement. They decide to use a sample of released items from TIMSS (Trends in International Mathematics and Science Study) with a group of 279 seventh grade students to compare with the state assessment. The state assessment had a number series item to "select a number sentence" to determine the eighth number in the series 7, 14, 21, 28 . . . A large majority (85%) of the students correctly selected $7 \times 8 = 56$. However for the TIMSS item, "Correct" responses were 3%, "No response" was 49%, and the "International norm" was 18% correct. The format of the corresponding TIMSS item was different; it was a sequence of triangles: One small triangle, followed by two small triangles embedded in a larger one, followed by four small triangles embedded in a larger one (i.e., the series 1, 3, 5 . . .). The task was to determine how many triangles would be in triangle 8 if the series were extended. Many students tried drawing the triangles to solve the problem. From this, researchers learned about the pitfalls of generalization of a concept across test formats. Following up this finding with *think-aloud* or *cognitive process tracing* could have produced additional insights into students' problem solving skills.

These scenarios are illustrative of how multiple participants can influence the testing process in appropriate ways. Each participant in the process contributes something unique and necessary to the valid use of assessments in education.

3. Where to From Here?

We can only sketch some actions that might move the profession along paths suggested by the problematics and perspectives outlined above. Our major intent has been to start a dialogue on the role of multiple agents in the testing process given the problematics of testing in the current context. To move things along from the perspectives of the professionals we drew upon in the paper, we propose two research and development possibilities.

3.1. Casebooks and Addenda to the Standards

Linn (2006) made a suggestion for research and development at the time when the 1999 *Standards* were being considered for revision and professional inputs were solicited. Linn contended that practice would be advanced more by providing an addendum for clarification of, or extensions of, the *Standards* to meet current demands originating from developments in technology, curriculum, and professional ethics. In addition, Linn proposed casebooks that might be developed providing realistic examples of application of the *Standards* to specific contexts (e.g., design and development of educational assessments; use and interpretation of educational assessments; administration, scoring, and reporting of educational assessments). The *Standards,* representing the best available professional consensus, makes a strong case that standards cannot be seen as algorithms since they are so much dependent upon context and professional judgment. Since professional judgment depends upon available skills, knowledge, values, and the context of application, Linn's suggestion makes sense. What is needed is multiple examples of practice, clustered in different areas of application, from which some practical rules may evolve. Perhaps the agencies that should lead such an effort are the three professional organizations that produced the *Standards*.

3.2. Survey Current Expertise and Practice, and Match to Given Criteria

A line of research and development, perhaps best addressed by those who train users from a university base or school district base, is to update knowledge of what is being taught to test developers and test users since the job task descriptions of Packman, Camara, and Huff (2010) and to match that status picture with the demands of practice such as represented in the following sources: Brookhart's (2011) depiction of assessment knowledge and skills for teachers; the special issue of *Educational Measurement: Issues and Practice* on strengthening the connections between classroom assessment and large-scale assessment (Spring 2018, Vol. 37, No. 1); and Bennett (2018) on what to watch for in the changing world of assessment. Bennett lists a set of eight changes including the following: (i) Make greater use of more complex tasks; (ii) Attempt to improve learning; (iii) Be better at accounting for context; (iv) Use automated scoring; and (v) Provide more effective reporting. If one were to collect syllabi for university courses in educational measurement today, we suspect that the majority would be heavily weighted on psychometrics

(appropriately) but not so much on Bennett's list of future developments nor Brookhart's list of knowledge and skills for teachers.

We hope this article plays a small part in contributing to a dialogue to move practice along the lines the profession has been posing for the roles of multiple agents in the testing process, informed in part by expectations for what the future of educational assessment should be.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bennett, R. E. (2018). Educational assessment: What to watch in a rapidly changing world. *Educational Assessment: Issues and Practice*, *37*(4), 7-15.

Brennan, R. L. & Plake, B.S. (1991). Surveys of programs and employment in educational measurement, *Educational Measurement: Issues and Practice*, *10*(2), 32.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, *30*(1), 3-12.

Brookhart, S. M. (2018). Learning is the primary source of coherence in assessment. *Educational Measurement: Issues and Practice*, *37*(1), 35-38

Burch, P. (2009). *Hidden markets: The new education privatization*. New York, NY: Routledge.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*(3), 397-412.

Corno, L., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W., & Talbert, J. E. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow* (Educational Psychology Series). Mahwah, NJ: Lawrence Erlbaum.

Della-Piana, G. M., Gardner, M. K., & Mayne, Z. R. (2018). *Exploration of validity evidence gaps in science educational achievement testing* [Final report]. National Science Foundation Grant No. 58502176.

Edelman, J. & Levy, S. (2016). Making sense of the opt-out movement. *Education Next*, *16*(4), 55-64.

Herszenhorn, D. M. (2006, May 5). As test-taking grows, test-makers grow rarer. *The New York Times*. Retrieved from https://www.nytimes.com/2006/05/05/education/05testers.html

Koretz, D. (2017). *The testing charade: Pretending to make schools better.* Chicago, IL: University of Chicago Press.

Linn, R. L. (2006). Following the standards: Is it time for another revision? *Educational Assessment: Issues and Practice*, *25*(3), 54-56.

Lissitz, R. W. (2009). *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: An integrated framework for understanding, guiding, and improving policies and programs*. San Francisco, CA: Jossey-Bass.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning.* New York, NY: Routledge.

National Council on Measurement in Education (2017, February 2). *Position statement on student participation in state assessment*. Philadelphia, PA: Author. Retrieved January 9, 2019, from https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/Documents/Board_Approved_Assessment_Participation_Pos.pdf

National Council on Measurement in Education (2018, July 26). *Position statement on theories of action for testing programs*. Philadelphia, PA: Author. Retrieved January 9, 2019, from https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/Documents/NCME_Position_Paper_on_Theories_of_Action_-_Final_July__2018.pdf

Packman, S., Camara, W. J., & Huff, K. (2010). A snapshot of industry and academic professional activities, compensation, and engagement in educational measurement. *Educational Measurement: Issues and Practice*, *29*(3), 15-24.

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

Ryan, K. E., & Shepard, L. A. (2008). *The future of test-based educational accountability*. New York, NY: Routledge.

Shepard, L. A., Penuel, W. R., & Pellegrion, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, *37*(1), 21-34.

Sireci, S. (2000). Recruiting the next generation of measurement professionals. *Educational Measurement: Issues and Practice*, *19*(4), 5-9.

Sussman, J., & Wilson, M. R. (2018). The use and validity of standardized achievement tests for evaluating new curricular interventions in mathematics and science. *American Journal of Evaluation*, Online First.

William, D. (2018). How can assessment support learning? A response to Wilson and Shepard, Penuel, and Pellegrino. *Educational Measurement: Issues and Practice*, *37*(1), 42-44.

Received 31 October 2018 | Accepted 10 January 2019 | Published 15 January 2019