JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Spring 2013, Vol. 50, No. 1, pp. 91–98

Validity in Action: Lessons From Studies of Data Use

Pamela A. Moss
University of Michigan

Studies of data use illuminate ways in which education professionals have used test scores and other evidence relevant to students' learning—in action in their own contexts of work—to make decisions about their practice. These studies raise instructive challenges for a validity theory that focuses on intended interpretations and uses of test scores as Kane's (this issue) does. This commentary explores implications of data use studies for elaborating Kane's approach to validation to accommodate the ways test scores are used with other sources of evidence to address users' questions.

Kane's Approach to Validating Test Score Uses

The focus of my commentary is on those elements of Kane's (this issue) approach to validation that relate most directly to the uses of test scores. This is a particularly important area of theory and research in light of the growing emphasis on test use—and data use more generally—in federal legislation, in state and district policies, and in research funding priorities. As Kane notes, not only are testing programs expected to provide information that education professionals can use in their practice to support students' learning, they are now commonly seen as strategies of educational reform. In analyzing Kane's contribution to validity theory, I draw on the growing empirical literature on data use in educational systems (Coburn & Turner, 2011, 2012; Moss, 2007). These studies illuminate the ways in which education professionals have used test scores and other evidence relevant to students' learning—in action in their own contexts of work—to make decisions about their practice. I sketch possible implications of this research agenda for elaborating Kane's approach to validation.

The framework for validation provided by Kane has made a substantial contribution to validity theory in educational measurement. Like any theory, Kane's theory of validation provides an intellectual framework or set of conceptual tools that shape our understanding and our actions. It illuminates some aspects of social phenomena for consideration and leaves others in the background. The focus of validity in Kane's work (Kane, 2006, this issue)—and in much of the validity literature in measurement—has been on validity of *intended* interpretations and uses of scores based on tests or other standardized forms of assessment. "Interpretations involve claims about test takers or other units of analysis (e.g., teachers, schools), and score uses involve decisions about these units of analysis" (p. 2). His key tools—the interpretation/use argument (IUA) which lays out the inferences to be evaluated, the validity argument which evaluates the IUA, and the listing of common inferences likely to need evaluation (including scoring, generalization, extrapolation, theoretical or causal, and decision inferences)—foreground a distinctive set of concepts and strategies. Although Kane's approach is coherent with those of other widely cited validity theorists, it draws our attention more explicitly to the chain of

inferences needed to move from the specific performances on a particular test to the "target domain" or theory about which we want to draw conclusions and the decisions those conclusions inform. Importantly, he notes that specifying an IUA helps protect against inappropriate interpretations and uses by making gaps in the evidence harder to ignore.

Kane orients his discussion of test use on a "decision rule, which stipulates that certain actions be taken given certain test scores" (p. 46). He argues that decision inferences must be evaluated in terms of their consequences: they "require evidence that the procedure achieves its goals without unacceptable negative consequences" (p. 15). He notes further (citing Cronbach) that "test users tend to be in the best position to evaluate the consequences of their own decision rules . . . though the developer of a test should help the user in any practicable way" (p. 57). Kane is careful to acknowledge that these intended inferences and uses are presumptions that make no guarantee about their validity in individual cases: validity arguments should be qualified in terms of the likelihood that they hold in individual cases and should include mechanisms (primarily local) for identifying those cases where the argument does not hold. Kane speaks more briefly to the situation where testing programs are seen as and need to be evaluated as interventions that drive reform. Elsewhere, Bennett, Kane, and Bridgeman (2011) point to the value of specifying a "theory of action" to evaluate testing programs as interventions—a point I will return to later.

As the many citations to his work suggest, Kane's approach to validation has well served members of the measurement community who develop and evaluate tests. However, the focus on decision rules and, indeed, on the test or testing program as the primary source of evidence, underanticipates the complexity of how test scores are being used locally, in action, by teachers and other education professionals in different contexts for their own purposes.

Lessons From Data Use Studies

Evidence from studies of data use or from studies of professional practice where data are used suggest a number of productive challenges to validity theories that focus exclusively on test-based information in evaluating test uses. In sketching these challenges, I draw heavily on special issues of *Teachers College Record* (2012) and the *American Journal of Education* (2011), edited by Cynthia Coburn and Andrea Bueschel and sponsored by the Spencer Foundation, along with a yearbook I edited (2007) on *Evidence and Decision Making* for the National Society for the Study of Education. This collection of articles—mostly systematic reviews of literature or theorized examples of practice—provide a relatively comprehensive introduction to data use studies and the empirical literature on which they are based. Although many of the authors acknowledge that the research base is nascent—surprisingly so in light of the policy press for data use (Turner & Coburn, 2012)—the articles point to a research agenda focused on professional and organizational capacity to use data to enhance students' learning.

Across these texts, we see that test scores and other data relevant to student learning are used for variety of purposes in different contexts by professionals in different roles at different levels of the educational system. Many of the purposes to which

test-based information is put cannot be served by test-based information alone. Different decisions require different kinds and configurations of data. Knapp, Copland, and Swinnerton (2007) characterize the variety of data potentially relevant to guiding, directing, assessing, and supporting teaching and learning as information that

- represents the content or conduct of instruction or its effects on student learning and the student experience, as well as the factors and conditions that most immediately affect these matters; and
- is, or could be, used in leadership actions aimed directly at the improvement of instruction, learning, and the student experience, or the organizational conditions that support instructional improvement. (Knapp et al., 2007, p. 80)

If the goal is to make decisions about how to improve teaching and learning or to make choices among alternative courses of action or policies, evidence of student outcomes alone is insufficient; one must consider information about the conceptual and material resources, the teaching processes and practices, and the organizational routines and cultures that shape or influence those outcomes.

Information relevant to student learning is needed by teachers and other professionals to identify problems of various sorts and develop accounts of their causes; to develop and tryout specific action implications and trace their consequences; and to account for their actions to others (Jennings, 2012). Studies of the most resourceful uses of data at the school and system level suggest that teachers and leaders should engage in cycles of question-driven inquiry drawing on multiple sources of evidence (Ikemoto & Marsh, 2007; Knapp et al., 2007; Marsh, 2012; Supovitz, 2012).

An important dimension that arises is the issue of grain size or timescale. Decisions relate to different timescales, and the evidence needed to support the decision should be relevant to its timescale. Some decisions require data that inform midcourse shifts in teaching practice while it is happening; other decisions, such as resource allocations or planning professional development, happen on slower timescales and can benefit from information at a larger grain size. Supovitz (2012), for instance, characterizes iterative inquiry cycles involving assessment of student learning and responsive actions as occurring on different timescales—within lesson, within unit across lesson, across units, and annual—and as serving different purposes and different stakeholders' needs.

Although sources of evidence illustrated in this literature include standardized assessments of student learning, they also include a wide range of other sources of evidence—formal and informal, quantitative and qualitative—to which educators have access or that they can develop themselves. These range from the information available in ongoing classroom interactions and samples of students' work; teacher accounts of classroom practice, instructional artifacts, and discussions of standardized test results; data from videotapes, interviews, and surveys; various indicators of resources and social structures; and published research reports. A given cycle of inquiry might well require data of different grain sizes from the identification of a problem area that state-level testing informs to finer-grained information as relevant teaching strategies are tried out and evaluated.

The quality of the data use depends, as well, on the capacity of professionals to make sense of the data in addressing their own problems. Information does not become evidence until people "notice, frame, and interpret" it as relevant to a problem or decision (Spillane & Miele, 2007). How test information is put to work depends on the sense people make of it (including whether they attend to it at all), the resources used to interpret it, the locally relevant questions it provokes or provides evidence for, the other evidence brought to bear, and local supports/constraints for developing meaningful interpretations and decisions (Little, 2012; Phillips, 2007; Spillane & Miele, 2007). Roderick (2012), for instance, in her work with data users through the Consortium on Chicago School Research, noted:

> The work was not giving the principals, counselor, and teacher the data. It was making the data understandable, useable, and relevant to the central problems they face. It was grappling with their questions, posing questions back, and always presenting the research evidence about effective ways to respond to the data. It was about bringing educators in like roles across schools together to identify and solve common problems and, using data, develop strategies for improvement. Even then, it took multiple experiences looking at data before the principals in the group could easily begin to analyze the data, talk about how the problems played out in their building, and identify strategies. It took time for educators not to look at every piece of data about their school as a judgment on the quality of their work. And it took even more time for educators to begin making the data live in their school by supporting teams in ongoing monitoring and data use.

Organization theorists Brown and Duguid (2000) draw a productive distinction between information and knowledge. "Attending too closely to information overlooks the social context that helps people understand what the information might mean and why it matters" (p. 6). Reports from large-scale assessments provide *information*, and information that serves some purposes better than others. But, as Brown and Duguid note, "The same stream of information directed at different people doesn't produce the same knowledge in each. If the people are engaged in different practices, if they are learning to be different kinds of people, then they will respond to the information in different ways" (p. 129). For example, Coburn and Talbert (2006), following their 2-year study of evidence use in a large urban district, concluded:

> Our research suggests that moving toward a coherent systemic strategy for evidence based practice may require a system of evidence use that allows for and supports access to different kinds of evidence for different purposes at different levels of the system. Individuals with different work roles have substantively different data needs. A strategy for evidence-based district reform must acknowledge these differences and create mechanisms to bring productive dialogue and coordination across them. Here, congruence does not mean the same conceptions of evidence; rather, it means creating complementary approaches at different levels and functions. (p. 491)

This points to the importance of a research agenda that examines (a) the way professionals in different roles and contexts interact with and use data; (b) the organizational resources at different levels of the system that support or constrain their practice of data use; and (c) the ways in which different approaches impact

the practices of education professionals and organizations as well as the learning of their students. The collection of articles I have cited sketch a multimethodological research agenda that focuses on different aspects of the problem of data use: data-informed interactions and norms in professional learning communities (Little, 2012); teachers' information needs and the sorts of tests that might serve them (Supovitz, 2012); professional community norms and school-level routines intended to support data use (Spillane, 2012); the ways social networks support and constrain the interpretation and use of data (Daly, 2012); the sorts of support districts provide and the ways schools and districts interact in serving one another's information needs (Honig & Venkateswaran, 2012); how "performance metrics" (like the adequate yearly progress [AYP] metric), the incentives and sanctions associated with accountability programs, or alternatively the creation of safe spaces for learning from data shape practice (Colyvas, 2012; Jennings, 2012; Marsh, 2012); and how large- and small-scale interventions and other points of leverage intended to increase data use capacity impact professional, organizational, and student learning (Coburn & Turner, 2011; Marsh, 2012).

Implications for Validity Theory

Studies of data use and the research agenda the use illustrates raise challenges for a validity theory that focuses on intended interpretations and uses of scores from tests or other standardized forms of assessment. They suggest that interpretations are shaped by different users' questions, that most such questions require multiple sources of evidence, and that the validity of their conclusions (interpretations, decisions, and actions) relies in large part on local capacity to use data well. As Kane noted, validity of test use is ultimately the user's responsibility, but the test developer should help the user "in any practicable way" (Kane [quoting Cronbach], this issue, p. 57). How might members of the measurement community practicably respond to these challenges in conceptualizing validity for test use? This clearly is a long-range agenda, but here are some possible directions.

First, I argue that understanding the data needs and uses of educational professionals is a research agenda that the measurement community should embrace as research collaborators or as active consumers of others' research. As Roderick (2012) suggested in her commentary on research on data use interventions: "We need a better, more robust theory of action and a strong evidence base for understanding what data schools need to address different problems and how data can be used effectively at various organizational levels and under what conditions" (p. 4). Even in the short run, addressing these questions would inform the assessments we develop, the sorts of materials we provide, and the advice we give about how to use them.

Second, with respect to our representation of validity theory, I argue that the focus—scope and direction—of validity questions/arguments needs to be able to shift, as needed. Sometimes the focus will remain on an intended interpretation and use of a test score (that is the traditional focus of validity research), and Kane's approach as articulated (2006, this issue) appears to be serving us very well. However,

studies of test use, and data use more generally, suggest that sometimes—more often than not—a broader focus is needed to support local users. This shift in focus entails the inclusion of interpretations/decisions that respond to locally relevant questions and that necessarily draw on a broader range of evidence about students' learning and the factors that shape it. A validity theory supporting such data uses might be most productively conceptualized around the particular questions or problems that evidence is needed to address, rather than around an intended interpretation/use from a particular testing program and the circumscribed evidence it provides (Moss, Girard, & Haniford, 2006). However, because such interpretations and uses are routinely made in professional practice, often in light of evolving evidence relevant to students' learning, it will not be possible to document the validity of each interpretation and decision. Hence, the focus of validity questions will need to shift again to the broader learning or organizational environment and the extent to which it is sufficiently well resourced to support an evidence-based professional practice that enhances student learning (Moss et al., 2006).

Kane's notion of an IUA and validity argument—elaborated to include Bennett, Kane, and Bridgeman's (2011) notion of a theory of action and (program) evaluation—are still useful here, but the IUA/theory of action would be considerably more complex. In the Bennett et al. (2011) example, the theory of action focused on evaluating the impact of the through-course assessments envisioned by the state consortia developing assessments of the Common Core State Standards. In addition to the standard elements of the IUA, it included attention to intermediate and long-range effects, the mechanisms through which the assessments lead to the intended effects, anticipation of unintended negative effects, and strategies through which negative effects might be mitigated. With questions about data use more generally, the theory of action would need to include attention to the mechanisms through which inquiry is routinely carried out and the resources that support it, the intermediate effects on the quality of data use, and the longer-range effects of data use on professional practice and organizational culture as well as student learning. The research agenda sketched in the closing paragraphs of the previous section begins to specify the elements that such a theory of action would need to incorporate [and papers by Coburn and Turner (2011) and Marsh (2012) depict alternative models of such a theory of action focused on data use interventions].

In their chapter on data use in educational systems, Thorn, Meyer, and Gamoran (2007) pointed to the problem of data "silos" which made it difficult for educational systems to build connections across data sources in addressing educational problems. My commentary is in part driven by the same concern. Of course, the research agenda I have outlined should and could not be the responsibility of any particular test developer or evaluator. And nothing I have said is intended to challenge Kane's approach for validating *intended* interpretations and uses of test scores. However, if a goal of the measurement community is to support professional educators in using test-based information well, we need to elaborate our representation of validity theory to accommodate the ways test scores are used in action with other sources of evidence to address local problems. Kane's contribution to validity theory in measurement provides a generative foundation on which we can build.

References

Bennett, R., Kane, M., & Bridgeman, B. (2011, February). *Theory of action and validity argument in the context of through-course summative assessment*. Paper presented at the Invitational Research Symposium on Through-Course Summative Assessments, Atlanta, GA.

Brown, J. S., & Duguid, P. (2000). *The social life of information*. Boston, MA: Harvard Business School Press.

Coburn, C. E., & Talbert, J. E. (2006). Conceptions of evidence use in school districts: Mapping the terrain. *American Journal of Education*, *112*, 469–495.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research & Perspective*, *9*(4), 173–206. doi:10.1080/15366367.2011.626729

Coburn, C. E., & Turner, E. O. (2012). The practice of data use: An introduction. *American Journal of Education*, *118*, 99–111.

Colyvas, J. A. (2012). Performance metrics as formal structures and through the lens of social mechanisms: When do they work and how do they influence? *American Journal of Education*, *118*, 167–197.

Daly, A. J. (2012). Data, dyads, and dynamics: Exploring data use and social networks in educational improvement. *Teachers College Record*, *114*(11), 1–38.

Honig, M. I., & Venkateswaran, N. (2012). School-central office relationships in evidence use: Understanding evidence use as a systems problem. *American Journal of Education*, *118*, 199–222.

Ikemoto, G. S., & Marsh, J. A. (2007). Cutting through the "data-driven" mantra: Different conceptions of data-driven decision making. In P. A. Moss (Ed.), *Evidence and decision making: 106th Yearbook of the National Society for the Study of Education, Part 1* (pp. 105–131). Malden, MA: Blackwell.

Jennings, J. L. (2012). The effects of accountability system design on teachers' use of test score data. *Teachers College Record*, *114*(11), 1–23.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Knapp, M. S., Copland, M. A., & Swinnerton, J. A. (2007). Understanding the promise and dynamics of data-informed leadership. In P. A. Moss (Ed.), *Evidence and decision making: 106th Yearbook of the National Society for the Study of Education, Part 1* (pp. 74–104). Malden, MA: Blackwell.

Little, J. W. (2012). Understanding data use practice among teachers: The contribution of micro-process studies. *American Journal of Education*, *118*, 143–166.

Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record*, *114*(11), 1–48.

Moss, P. A. (Ed.). (2007). *Evidence and decision making: The 106th Yearbook of the National Society for the Study of Education, Part I*. Malden, MA.: Blackwell.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education*, *30*(1), 109–162. doi:10.3102/0091732×030001109

Phillips, D. C. (2007). Adding complexity: Philosophical perspectives on the relationship between evidence and policy. In P. A. Moss (Ed.), *Evidence and decision making: 106th Yearbook of the National Society for the Study of Education, Part 1* (pp. 376–402). Malden, MA: Blackwell.

Roderick, M. (2012). Drowning in data but thirsty for analysis. *Teachers College Record*, *114*(11), 1–9.

Spillane, J. P. (2012). Data in practice: Conceptualizing the data-based decision-making phenomena. *American Journal of Education*, *118*, 113–141.

Spillane, J. P., & Miele, D. B. (2007). Evidence in practice: A framing of the terrain. *Evidence and decision making: 106th Yearbook of the National Society for the Study of Education, Part 1* (pp. 46–73) Malden, MA: Blackwell.

Supovitz, J. (2012). Getting at student understanding—The key to teachers' use of test data. *Teachers College Record*, *114*(11), 1–29.

Thorn, C., Meyer, R. H., & Gamoran, A. (2007). Evidence and decision in education systems. In P. A. Moss (Ed.), *Evidence and decision making: The 106th Yearbook of the National Society for the Study of Education, Part 1* (pp. 340–361). Malden, MA: Blackwell.

Turner, E. O., & Coburn, C. E. (2012). Interventions to promote data use: An introduction. *Teachers College Record*, *114*(11), 1–13.

Author

PAMELA A. MOSS is Professor at the School of Education, University of Michigan, 610 East University, Ann Arbor, MI 48109; pamoss@umich.edu. Her areas of specialization are at the intersections of educational assessment, validity theory, and interpretive social science.