The Politics of Testing When Measures "Go Public"

JEFFREY R. HENIG

Teachers College

Background/Context: *Validity issues are often discussed in technical terms, but the context changes when measures enter broad public debate, and a wider range of interests come into play.*

Purpose: *This article, part of a special section of TCR, considers the political dimensions of validity questions as raised by a keynote address and panel discussion originally held at Teachers College in March 2012.*

Research Design: *This is an analytical and reflective piece, based on the author's participation in the panel and drawing on his experience and writing on the broad issue of politics and research.*

Conclusions: *Technical expertise in the construction and interpretation of measurements is important in the new world of high-stakes and "evidence-based" education policy, but the political realities, when measures go public, make the exclusive reliance on such expertise problematic.*

When waters get rough, it is nice to have a harbor to run to and a sturdy anchor that can be embedded in the ground below. As education issues become more prominent on the policy agenda, they are more subject to the swirling currents of partisan and ideological conflict. The so-called "politicization" of education policy debate has been evident across a range of hot button issues, including charter schools, class size, and testing with high stakes for students, teachers, schools, and districts (Henig, 2008). Political and civic leaders are drawn, in such circumstances, to the images of objective science as safe harbor, and to validity—in design and in measurement—as stabilizing anchors. Ironically—but perhaps not so

Teachers College Record Volume 115, 090304, September 2013, 11 pages
Copyright © by Teachers College, Columbia University
0161-4681

surprisingly—experts in measurement and research are typically much more circumspect in their characterization of the power of the tools they bring to bear.

"We have a saying that in God we trust—everybody else has to bring data," New York City Mayor Michael Bloomberg declared while in the midst of that city's controversies over heavy focus on standardized test scores and their incorporation into teacher assessment. "And I know of no ways for a teacher to know whether they are getting through to the child and whether the child understands and is making progress without testing those children," he went on. "And this business of teaching to the test is exactly what we should do, as long as the test reflects what we want them to learn. If the test is 'can you read,' yes, you should find out whether they can read by testing them!" (Cramer, 2002). For Mayor Bloomberg, confident assertion of the validity of test score data signaled that his administration based its decisions on knowledge and expertise, unlike his critics, who, in his characterization, marshaled misinformation in pursuit of political goals.

Measurements and the data they produce seem more solid and consequential when we think of validity as an objective trait of the measures and data, something injected into them by experts following tried-and-true checklists. Some of that sturdy and confidence-building imagery gets shaken when experts themselves challenge the notion that validity is wholly intrinsic to the measures and instead extends to encompass ways in which measures are used and interpreted. As Madhabi Chatterji suggests in her introduction to this special issue, and as elaborated upon in Eva Baker's contribution, many contemporary measurement and testing experts subscribe to Messick's view of validity as "evaluative judgment," not an objective trait (Baker, 2013; Chatterji, 2013). This understanding of validity poses challenges, especially when we consider education policy decisions that intersect with high-stakes conflicts around ideology, political party, and competing interest groups. "A study may be valid by internal criteria and still arouse violent opposition from people against whose interest it goes," Martin Rein observed more than three decades ago. "The crucial issues in a policy debate are not so much matters of fact as questions of interpretation" (Rein, 1976).

What are the implications for collective decision-making if we adopt a notion of validity as the product of theoretically informed judgment, contingent on the specific application, and a matter of degree? How is that affected, in particular, in those instances in which measurement goes public, and how might it be affected in the future world that Baker envisions where testing is done "on the fly and in the wild . . . part of our regular lives, happening in real time, all the time, with lightning speed

analyses, and results stored in multiple repositories" (Baker, 2013)? What happens when the anchor begins to look a little more like a kite, something tossed and turned by political winds? In this article, I offer some reflections on the politics of testing and how that affects and is affected by a less assertively positivistic vision of validity than normally permeates public discourse about research, data, and evidence.

WHAT IT MEANS WHEN MEASURES "GO PUBLIC"

Discussions of measurement validity are sometimes contained within relatively tight circles of well-informed participants operating on a narrowly and well-defined problem. For example, a local foundation might orient its mission around improving college readiness within a defined set of local high schools and contract with a research organization to develop a suitable index against which to monitor progress for the purpose of summative evaluations of funded interventions that will be shared only among the foundation leadership and board. In that kind of environment, not only is the number of participants limited, making communication manageable, but the range of contrasting values, perspectives, and operating assumptions that need to be given due consideration is also constrained. Those outside the organization might disagree about whether college readiness should be highlighted above goals such as job readiness or long-term life satisfaction and well-being, or about whether the most appropriate interventions should focus on schools or on neighborhood and family conditions, or about whether it is better to partner with existing schools or create new ones. They might have different insights and views about whether available measures of socioeconomic status or student test scores capture idiosyncrasies across racial and ethnic subgroups, about whether failure to complete college within four or five years constitutes evidence of lack of proper preparation, about whether a quantitative metric can substitute for in-depth and qualitative analysis of graduates' knowledge and attitudes. But even if the discussion might benefit from realizing that these other perspectives exist, those within the room have no need to take them all into account, to ponder their implications, to risk losing focus on organizational mission and the task at hand. Disagreements and complexities regarding facts and values can legitimately be stopped at the door

There are at least two important ways in which issues regarding measurement validity can be said to "go public," and in each there are implications for the way the discussion ensues. One way they go public is when the audience of those who might attend to and use the measures is broadly expanded, whether by design or by accident. The broad controversy regarding the public release of individual teacher's value-added

scores provides a good example. In both Los Angeles and New York City, individual teachers' value-added scores were published in local newspapers despite intense opposition from teacher unions and others. While the validity of the measurements was not the sole focus—for example, some arguments against releasing the individual teacher scores could apply even if the measures were accurate—this became the primary axis around which much of the ensuing debate was framed.

This debate about the validity of value-added measures of teacher performance occurred in multiple forums, including departments of education, major media, courts, and the court of public opinion, as evidenced in blog posts and polling. The voices of measurement experts were a part of the public discourse. The LA Times, for example, hired Richard Buddin, a RAND economist, to conduct its analysis. The National Education Policy Center commissioned and actively disseminated technical critiques and re-analyses by experts (Briggs & Dominique, 2011; Durso, 2012). And education journalists scrambling to handle the complex issues reached out to an array of respected analysts such as Sean Corcoran (NYU), Douglas Harris (Wisconsin), Daniel Koretz (Harvard), Douglas Staiger (Dartmouth), William Sanders (UNC; SAS Institute), and others. Inevitably, though, broadly expanding the audience for the discussion brings in voices that lack such technical authority.

Local blog posts about the teacher assessment issue in LA and NYC, for example, reveal a less sophisticated mode of discourse. Some posts appear to be sincere efforts to grapple with complex issues. "I read the article as a biostatistician and was wondering if they were correcting for the obvious bias in that this is not a linear scale," reads one comment to the LA Times. Many others link uber-confident claims about validity to harsh attacks on those who disagree. "Reading some of the 'teacher' comments below, it's clear that a lot of you are either illiterate, statistically illiterate, or are just too lazy to read an article before posting. I'd like to attempt to clarify things for you, but I doubt any of this will get through your thick skulls, so I suppose I'm merely posting this for my own satisfaction," reads a February 2012 post on the NYT Schoolbook site. In reply, another wrote, "If you are smart enough to understand statistics and Value Added data, then you should be smart enough to know that there are many fine teachers in the NY schools."

What is striking is that so much of the public discourse ostensibly focuses on competing claims about validity—despite the fact that many of those posting have very limited familiarity with the measurement details and reveal strong political and ideological predispositions that are likely the real source of the positions they adopt. "This 'data' is as reliable as the 'data' used to persecute people during the Salem witch trials,"

one reader posts, in what is a fairly common example of simple political rhetoric dressed in the garb of methodological critique.

The second way measures can be said to go public is when they get incorporated into policy decisions that have real implications for the creation, distribution, and redistribution of benefits and costs. Political scientist David Easton once famously defined politics as the "authoritative allocation of values" (Easton, 1953). Even when measures are not widely disseminated—even, in the extreme, when they are self-consciously barred from public release—decision-makers may rely upon them to steer decisions about who gets and loses jobs and contracts, about where schools get opened and which ones get closed, about who is eligible for special programs or which programs should be discontinued in tight fiscal times. From the standpoint of democratic theory, this form of going public raises important questions about responsibility and responsiveness of decision-making.

This second way in which validity issues go public often leaves the key discourse in the hands of policy and political elites, and therefore may not degrade the median level of information and expertise on the part of those using the measures to the same extent as the first. But it brings some other important political dynamics into play. Policy decisions are made by elected representatives who often bring to bear important substantive knowledge, as well as years of experience with the slippery translation of policy ideas into on-the-ground practice. But, with exceptions, these political leaders rarely have deep knowledge of measurement issues, either in general or in the specific arena of education. While elected officials can and do tap into the more technical expertise of executive agencies and departments, the principal-agent literature in political science suggests that bureaucracies may use their both their real expertise and their reputation for expertise to pursue their own organizational interests and policy agenda, taking advantage of and sometimes deliberately contributing to general-purpose politicians' limited grasp of the finer details. And, of course, some elected officials rival the least informed citizens in their obliviousness. Consider, for instance, the Republican congressman from Florida who, in arguing for de-funding the federal American Community Survey—an annual survey of about three million Americans collecting data on basic demographics and other important issues like income and benefits, health insurance, education, residence, work, commuting, and what people spend for basic essentials—relied on his own very unscientific understanding of the word "random" to argue that "in the end this is not a scientific survey. It's a random survey" (Rampell, 2012).

Quite apart from the levels of resident knowledge and expertise, other political dynamics come into play when measurement issues go public.

Test scores and other measures—as well as data and research more generally—are not simply utilized to inform better policy decisions, they are also incorporated into the narrower calculations of political actors: whether elected officials seeking public support or interest groups maneuvering for advantage (Henig, 2012). Politicians find it useful, for example, to hew to the image of measurement as objective and straightforward precisely because it absolves them of responsibility for outcomes that some elements of their potential constituency might oppose. "Assumptions about objectivity and impartiality become even more important when some schools and students are being rewarded and others are being sanctioned," Lorraine McDonnell observes. "When 'winners and losers' are created through policy actions, decisions need to be based on what appear to be objective grounds. For that reason, the myth of objective test data persists in the minds of both policy elites and the public, despite numerous expert critiques seeking to dispel it" (McDonnell, 2004). Interest groups wield measurement and data less to absolve themselves from responsibility than as a political weapon: a way to convince policymakers that their claims are (or can be sold to the public as) objective and to dismiss arguments made by competing groups. "When you can summarize a whole bunch of complicated things in a single number, that has a lot of power and it's hard to ignore, especially when it tells a story that you want to promote," one expert put it in reflecting on his two decades of experience with student and teaching assessment. "And that's where it gets really twisted" (Hawkins, 2012).

If going public introduces the likelihood of misunderstanding and political manipulation, why not turn to an alternative model in which discussion is more constrained? Wouldn't the level of discourse be raised if we made participants display their measurement and testing credentials at the door?

WHY MEASUREMENT EXPERTS THEMSELVES SAY IT'S NOT AN OPTION TO RESTRICT VALIDITY DISCUSSION TO THOSE IN THE KNOW

Political scientists normally expect the collective behavior of experts to be predictable in the same way as that of other interest groups. Based on that, we'd anticipate that measurement experts would try to maximize their power and influence and therefore would be in the lead of the movement to return discourse about measurement to those in the know: themselves. In this instance, though, it is experts in measurement and validity who are taking the lead in redefining validity in ways that make it more problematic to hand them a decisive role. Their reconceptualization of validity makes it more complex and multidimensional, but

the complexities and dimensions it adds demand that the circle of those involved and taking responsibility be expanded, not contracted.

Let's be clear: Measurement and testing experts do not doubt that they have technical expertise and special knowledge – tools that are scarce and important. "Testing is by its nature a highly technical enterprise that rests on a foundation of complex mathematics, much of which is not generally understood even by quantitative social scientists in other fields," Koretz writes. The "core principles and concepts are truly essential. Without an understanding of validity, reliability, bias, scaling, and standard setting, for example, one cannot fully make sense of the information yielded by tests or find sensible resolutions to the currently bitter controversies about testing in American education" (Koretz, 2008).

The core modesty of their claims of expertise comes into play, though, when these experts consider the limitations of their craft in handling multi-dimensional, loosely-defined, value-laden concepts and authoritative decisions with potentially severe implications for people, organizations, and communities. Consider again the case I raised earlier of measurement used as an internal organizational tool tightly lashed to organization mission, as one part of a data-feedback system that informs privately held discussions of performance and strategy, and with few spillover costs imposed on those outside the organization. Even in those non-public decisions, technical aspects of measurement validity need to be informed and augmented by other kinds of knowledge and expertise, usually provided by organization leaders who must explicate the mission and core values, draw on lessons learned from past efforts, and make sure measurement experts know how the measures are likely to be used. As measures go public, the kinds of knowledge that are relevant grows, and with that, in expanding concentric circles, comes a need and legitimacy for a wider variety of actors and voices to be heard.

Messick's view of validity calls for consideration of the way that "empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other forms of evidence" (Messick, 1989, as quoted in Chatterji, 2013). Determining the validity of inferences demands judgments about complex causal relationships within the spheres of action to which the measures will be applied. If the goal is to inform classroom practice, the domains of relevant expertise may be restricted largely to those relating to factors like pedagogy, student learning, and motivation, and the particular school and community context. If the goal is the somewhat broader one of meeting a local community's particular vision of what its children should know in order to thrive and contribute, the varieties of relevant knowledge expand to include aspects of values and local context for

which school board members, other elected officials, parents, the business community and other civic leaders are the relevant experts. If the goal is to drive systemic educational change, orchestrated in Washington, DC and implemented in districts and schools nationwide, the domains of relevant knowledge expand much further still.

Not all measurement experts would agree with Messick that validity extends beyond interpretation to considering the range of consequences that might occur. Consequences, after all, may depend not only on unforeseeable events but also on actors who may blunder or deliberately misuse data in order to mislead and subvert. In the wake of the development of the atomic bomb, physicists wrestled mightily with whether their role legitimately extended to considering how national leaders might use the power their knowledge had created. But society certainly does need to think about consequences, and it is reasonable to expect measurement experts to be helpful resources in anticipating probable risks and benefits of particular measurement or testing regimes. For these kinds of deliberations, the relevant experience extends to include broad understanding of multiple organizational systems and how they interact: how economies change, for example, in what they demand of workers, of how social capital within schools and within communities reinforce or tug at one another, and insights into human nature and capacity for understanding, to boot.

Besides requiring theoretically and empirically informed interpretation, and consideration of possible consequences, measurement experts generally recognize that validity is a question of degree. "Validity is a continuum, one end of which is anchored by inferences that simply are not justified," writes Koretz.

> At the other end of the spectrum, however, we are rarely fortunate enough to be able to walk away from the table having decided that an inference is valid, pure and simple. Rather, some inferences are better supported than others, but because the evidence bearing on this point is usually limited, we have to hedge our bets. (Koretz, 2008)

As when we allow for the need for theoretical and evidence based interpretation, acknowledging the need to place bets based on indicators whose validity may be fuzzy at the margins underscores the importance of drawing on a broader array of perspectives. Assessing the relative risks of a Type 1 versus a Type 2 error depends on judgment about probabilities (e.g., how likely it is that this measure understates genuine performance, or is differentially valid for different population subgroups), but also about how to place value on the risks and opportunities described.

In making decisions like these, measurement and social sciences should not be considered to be "a beam of light in a dark room. It is more like a candle in a lighted room," supplementing what policymakers already know, accustomed as they are to integrating substantive knowledge with their understanding of what the public thinks and knows (Weiss & Bucuvalas, 1980, p. 170).

CONCLUDING THOUGHTS

When measures go public, whether by virtue of widespread dissemination or their attachment to significant consequences, the inclination is to imbue those measures with certainty in order to justify action and sidestep responsibility if things go badly. Rightly or wrongly, political actors believe that the broader public has little tolerance for complexity or uncertainty, and so they craft their positions with that in mind. But psychometricians, statisticians, and indeed most social scientists are comfortable with uncertainty—it is the sea in which they regularly swim. Measurement error is their enemy, but an omnipresent one whose impact they seek to minimize but don't expect to defeat. Empirical predictions are framed probabilistically. The mix is not a natural one. Often, policy actors and measurement experts appear to be talking past one another and at cross-purposes, each wishing the other would take a bigger share of responsibility.

Technical expertise in the construction and interpretation of measurements clearly and increasingly is important in the new world of high-stakes and "evidence-based" education policy. But the political realities, when measures go public, make the exclusive reliance on such expertise problematic: conceivably possible but technically unsupportable and socially dangerous. One set of potential dangers lies in the over-concentration of power and control in the hands of public agencies. That's the more familiar meme in American political thought and the motivation for a range of regulatory measures designed to limit government monopoly over personal information. The imaginings of Eva Baker, in this volume, summon another possible danger. The proliferation of social media, online transactions, and corporate troves of linked data might open access and democratize use and debate, but it could just as likely spawn a world of testing and assessment where proprietary rights are jealously guarded and data and validity assessment are managed by a privatized technocracy largely buffered from regulation and citizen protection.

Reducing the public pressure on measurement—for example, by lowering the direct and immediate consequences of a small array of tests upon students, teachers, and schools—might take some of the heat off. But even those dismayed by the contemporary manifestation of data-based

reform and high-stakes measurements don't wish for a Luddite-like reaction against measurement and data per se. The basic tension lies in competing visions and values—about what are the important goals of education, about how they should be operationalized and balanced, and about who should be crafting the responses as we learn more about what school practices do and do not affect. We cannot reasonably expect to permanently and consensually resolve those tensions, and in a pluralistic democracy our inclination should be to broaden the conversation. That means lowering the median level of technical expertise among the discussants. But the range of types of knowledge and the advantages of having those compete in the open make that a price we should be willing to pay.

References

Baker, E. L. (2013). The chimera of validity. *Teachers College Record, 115*(9), 1-26.

Briggs, D. C., & Dominique, B. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District Teachers by the Los Angeles Times. Boulder, CO: National Education Policy Center.

Chatterji, M. (2013). Bad tests or bad test use? Why we need stakeholder conversations on validity. *Teachers College Record, (115)*9, 1-10.

Cramer, P. (2012, March 2). Following Bloomberg, Walcott shifts on teacher ratings release. Gotham Schools. Retrieved from http://gothamschools.org/2012/03/02/following-bloomberg-walcott-shifts-on-teacher-ratings-release/

Durso, C. S. (2012). An analysis of the use and validity of test-based teacher evaluations. Reported by the Los Angeles Times, 2011. Boulder, CO: National Education Policy Center.

Easton, D. (1953). *The political system, an inquiry into the state of political science*. New York, NY: Alfred A. Knopf.

Hawkins, B. (2012, June 15). Student-testing pioneer Angermeyr is skeptical about high-stakes trends. *MINNPOST*. Retrieved from http://www.minnpost.com/learning-curve/2012/06/student-testing-pioneer-angermeyr-skeptical-about-high-stakes-trends

Henig, J. R. (2008). *Spin cycle: How research is used in policy debates: The case of charter schools*. New York, NY: Russell Sage Foundation/Century Foundation.

Henig, J. R. (2012). The politics of data use. *Teachers College Record, 114*(11), 1-32.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

McDonnell, L. M. (2004). *Politics, persuasion, and educational testing*. Cambridge, MA: Harvard University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. New York, NY: American Council on Education.

Rampell, C. (2012, May 19). The beginning of the end of the census? New York Times, Retrieved from http://www.nytimes.com/2012/05/20/sunday-review/the-debate-over-the-american-community-survey.html

Rein, M. (1976). *Social science & public policy*. New York, NY: Penguin Books.

Weiss, C. H., & Bucuvalas, M. J. (1980). *Social science research and decision making*. New York, NY: Columbia University Press.

JEFFREY R. HENIG is a professor of political science and education and Chair of the Department of Educational Policy and Social Analysis at Teachers College, and a professor of political science at Columbia University. He is the author or coauthor of eight books, including *The Color of School Reform: Race, Politics and the Challenge of Urban Education* (Princeton, 1999) and *Building Civic Capacity: The Politics of Reforming Urban Schools* (Kansas, 2001), both of which were named—in 1999 and 2001, respectively—the best book written on urban politics by the Urban Politics Section of the American Political Science Association. *Spin Cycle: How Research Gets Used in Policy Debates: The Case of Charter Schools* (Russell Sage, 2008) won the American Educational Research Association's (AERA) Outstanding Book Award in 2010. Most recently, he is co-editor and contributor to *Between Public and Private: Politics, Governance and the New Portfolio Models for Urban School Reform* (Harvard Education Press, 2010).