CONSEQUENTIAL VALIDITY AND THE TRANSFORMATION OF TESTS FROM MEASUREMENT TOOLS TO POLICY TOOLS

Kevin G. Welner

Originally published in the journal, *Teachers College Record*. Cite as:

Consequential Validity and the Transformation of Tests from Measurement Tools to Policy Tools. *Teachers College Record*, *115*(9).

When used as measurement tools, tests help teachers and others reach judgments about the nature, scope, and extent of students' learning. This information is used for summative purposes, such as grading, placement, or admission. It is also used for formative purposes, such as the tailoring of subsequent instruction. The design and validation of tests for such purposes can be difficult and even problematic, but these problems are well-known and the subject of a great deal of constructive attention. This reaction paper accordingly starts with the premise, albeit a contested premise, that when tests are used for the primary purpose of measuring the learning of individual students, they are—as a policy matter—on reasonably solid ground.

Notably, however, the enormous expansion of test administration and test use over the past couple decades has not been driven by a mere desire to better measure and understand student learning. Rather, the intent of policies like No Child Left Behind (NCLB) has been to use the measurement of student learning to drive broad policy decisions and to change the behavior of teachers, principals and others. The key object of measurement has correspondingly shifted from students to their teachers, principals, schools, and districts.

While tests have always been policy tools in an indirect or secondary sense, these recent policy shifts in using assessments have brought a parallel shift in how tests have been approached and used. What were once primarily measurement tools have now become policy levers—part of what professor Baker calls "purpose creep" (Baker, in press, this issue).

This reaction paper delves into the implications of that purpose creep, focused on one extremely important element: the use of students' assessment scores for high-stakes teacher evaluations. Although Baker's article covers a broader set of issues raised "when measures go public"—affecting students and schools, as well as teachers and principals—the issue of high-stakes employment evaluations has now taken center stage nationally, becoming a primary exemplar of the transformation of testing from measures to levers.

Student Tests as Teacher Evaluators

In large part because of federal pressure exercised through policies such as Race to the Top and the School Improvement Grant program (see Welner and Burris, forthcoming), there has been an enormous shift in state policies regarding the evaluation of teachers. According to a 2012 report from the "National Council on Teacher Quality," an advocacy group that favors using students' test scores for teacher evaluations (National Council for Teacher Quality, 2012, p. 2): "In 2009, only four states were using student achievement as an important criterion in how teacher performance was assessed. In 2012, 20 states require student achievement to be a significant or the most significant factor in judging teacher."

The rationale for these policies is simple: past attempts to measure teacher quality have been largely ineffective. Administrative observations, the most commonly used summative evaluation method, are criticized as subjective and as applied unevenly and irregularly (Weisberg, Sexton, Mulhern, and Keeling, 2009). In contrast, evaluations based on growth in students' standardized test scores offer three advantages: annual scores are readily available; the test scores are widely accepted as legitimate school outcomes; and the resulting numbers are perceived as objective measures.

The strong focus on students' scores on standardized assessments is a relatively recent phenomenon. The main turning point is found in NCLB's Adequate Yearly Progress provisions, which attached to students' scores an escalating series of high-stakes consequences for schools and districts. Before that point, most states used standardized testing mainly for student-level evaluation and for public reporting of schools' progress. Race to the Top and similar policies from the Obama administration have extended those stakes to teachers and principals, even while the administration's "flexibility" waivers have eased some of the stakes faced by schools and districts (see U.S. Department of Education, 2012).

Purpose Creep and Its Implications

Two much-discussed implications of this purpose creep are explained by Campbell's Law. As applied to education, this Law tells us that when quantitative measures such as test scores are used to make key decisions, the measures themselves are subject to corruption pressures and, in addition, the high stakes distort and corrupt teaching and student learning (see Nichols and Berliner, 2007). But beyond the weakening of tests' measurement capacity and the often-harmful effects on classroom practice, the core meaning of tests and testing has been changing in profound ways. This is perhaps most clear from the perspective of teachers in the states that have adopted evaluation laws requiring students' scores to play a key role in high-stakes employment decisions. For these teachers, the standardized testing of their students has now become equivalent to a job evaluation. The corollary, of course, is indeed—as Donald Campbell suggested—that the classroom instruction leading up to that standardized test is now preparation for the teacher's job evaluation. As the test becomes a core part of the job evaluation policy, preparation for the test becomes the job itself.

Moreover, these job evaluations must feel rather arbitrary, given the research finding that teacher differences account for less than 20 percent of the variance in students' test scores (see Hanushek, Kai, & Rivken, 1998; Rowan, Correnti, & Miller, 2002). Teachers teach and students learn; the two are connected, but they are not the same. Although this variance issue raises

concerns best thought of in terms of construct validity, the main issue raised here is one of consequential validity (Messick, 1989; Shepard, 1997), but with a twist. The policy landscape has changed a great deal in the 15 years since Shepard and others (see Green, 1998; Mehrens, 1997; Popham, 1997; Reckase, 1998) were debating whether or not consequential validity should be included within the core idea of validity. For instance, consider Mehrens' contention that it would be "unwise" to confound "inferences about measurement quality with treatment efficacy (or decision-making wisdom)" (1997, p. 17)—a contention made in the context of tests administered for the immediate and overwhelming purpose of measuring student learning. At the time, the broader policy consequences, whether intended or unintended, were secondary to that measurement purpose.

Today, conversely and perversely, we must consider what happens when tests are transformed from being primarily measurement tools to being primarily policy levers. In particular, what happens when tests become policy levers that allocate responsibility for students' scores to others—to teachers, educational leaders or public education systems—in rigid and largely indefensible ways? Accordingly, we are no longer really considering consequential validity as the consequences of the use of educational tests for measurement of student learning. The use has become the minion of its own consequences.

When educational assessments are used as policy levers more so than as measures of student learning, their usefulness and their merits have relatively little to do with classic notions of content validity, criterion-related validity, or even construct validity. Questions about whether, for instance, the assessments measure what they were intended to measure become subordinated to the key question of whether the policy use of the test is driving recognized goals.

One can imagine, from the perspective of a policy maker, a high-stakes assessment that is poor in validity based on traditional psychometric criteria but which is nevertheless focused on complex, applied learning. As felt at the school level, such a test would be relatively likely to drive more complex, applied instruction. Contrast this with a high-stakes assessment that is sound by the same psychometric standards but measures learning that is more superficial. As compared to the first option, it would be more likely to drive superficial instruction.

While neither of these options is ideal, the second option would be preferable if the assessment were primarily a measurement tool. If, however, the purpose of the assessment shifts to becoming primarily a policy lever, the key validity issue correspondingly shifts, and the first option becomes much more attractive. The issue is no longer whether the assessment measures what it is purporting to measure in teaching and learning contexts. It's whether the measure as a policy tool is accomplishing what it is intended to accomplish.

Conclusion

Meaning is created by use. The all-encompassing push for test-based accountability policies in the U.S. has fundamentally changed the nature of test use. The core meaning of tests and testing has accordingly been qualitatively changed. This semiotic shift has already been felt on several fronts, but much of the fallout is likely still forthcoming. The ripples will be felt throughout the nation's schools, making it difficult to fully comprehend what the future holds for changes in a variety of important areas: in prospective teachers' career decisions, in teacher preparation

practices, in the development of legal liability frameworks, in collegial relationships within schools, and in continued changes to classroom practice.

Over a half-century ago, Cronbach and Meehl (1955) connected the idea of validity to the idea of "inference-making." Those inferences, they explained, depend on how a test is used. And current policy uses are troubling. When teacher identity can predict less than 20 percent of the variance in test results, but those results are being used in policies that depend on a much greater predictive capacity, there is an inference-making problem. Validity, like meaning, arises from use.

Accordingly, the gravity of the recent transformations in test use is about much more than words or about definitions of validity. That is, good reasons may, notwithstanding those transformations, nonetheless counsel against incorporating consequential validity within the core idea of validity. But the terminology is much less important than the implications. As a policy matter, validity discussions are important because of the basic dictate that tests should be validated; important decisions shouldn't be based on invalid instruments. Bringing consequential validity into the discussion is another way of saying that tests should not be used when the consequences of their use go against recognized goals. Reframing this to account for the new policy context, the key contention is that tests should not be used to drive policy unless the consequences of that process itself have been validated.

The measurement validity of the instrument tells us only a fraction of what we need to know. Appropriating the wording of Mehrens (1997), it's the "treatment efficacy" and the "decision-making wisdom" that should be validated (along with the assessment itself) before the tests are used as policy levers. The research and evaluation underlying such an evaluation will not always be clear-cut, but they are indeed clearly necessary. In the brave new world of test-based accountability policies, anything less places convention above need.

References

Baker, E. L. (in press). The chimera of validity. *Teachers College Record, 115(9).*

Briggs, D. C., & Domingue, B. D. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness ranking of Los Angeles Unified School District Teachers by the Los Angeles Times. Boulder, CO: National Education Policy Center. From http://nepc.colorado.edu/publication/due-dilligence.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

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. From http://nepc.colorado.edu/publication/analysis-la-times-2011.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. Educational Measurement: Issues and Practice, 17(2), 16-19, 34.

Hanushek, E. A., Kai, J. F., & Rivken, S. J. (1998). Teachers, Schools and Academic Achievement. NBER working paper 6691. Cambridge MA. From http://www.cgp.upenn.edu/pdf/Hanushek_NBER.PDF

Mehrens, W. A. (1997). The consequences of consequential validity. Educational Measurement: *Issues and Practice, 16*(2), 16-18.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

National Council on Teacher Quality (2011). *NCTQ State Teacher Policy Yearbook, Brief Area 3: Identifying Effective Teacher Effectiveness Policies*. Washington, DC: Author.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools.* Cambridge, MA: Harvard Education Press.

Popham, W. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Reckase, M. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13-16.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What Large-scale, Survey Research Tells Us about Teacher effects on Student Achievement: Insights from the Prospects Study of Elementary Schools. CRPE Research Report RR-051. From http://cw.marianuniversity.edu/mreardon/755/document%20repository/Teacher%20Effects%20on%20Student%20Achievement.pdf

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–8, 13, 24.

U.S. Department of Education. (2012, June 7 [updated]). ESEA Flexibility. From http://www.ed.gov/esea/flexibility/documents/esea-flexibility.doc.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Differences. Brooklyn*, NY: The New Teacher Project.

Welner, K. G., & Burris, C. C. (forthcoming, 2013). "NCLB's Intensifying Makeover: Race to the Top's Troubling Changes to Rules, Incentives, and Practice." In Sonya Horsford & Camille Wilson (Eds.), A Nation of Students At Risk: Advancing Equity and Achievement in America's Diversifying Schools. New York: Routledge.

Winerip, M. (2011, November 7). In Tennessee, following the rules for evaluations off a cliff. *The New York Times*. From http://www.nytimes.com/2011/11/07/education/tennessees-rules-on-teacher-evaluations-bring-frustration.html.