The Epidemiology of Modern Test Score Use: Anticipating Aggregation, Adjustment, and Equating

The Epidemiology of Modern Test Score Use: Anticipating Aggregation, Adjustment, and Equating

Andrew Ho
Harvard Graduate School of Education

In his thoughtful feature article, Haertel (2013) pushes testing experts to broaden the scope of their validation efforts and to invite scholars from other disciplines to join them. He credits existing validation frameworks for helping the measurement community to identify incomplete or nonexistent validity arguments. However, he notes his sense that something is missing in these frameworks, particularly as they seem to identify questionable, poorly articulated, or inappropriate uses only after the milk, as it were, has been spilled. I found his description of these uses helpful, particularly his identification of "indirect actions" of tests that have been a blind spot, by accident or by design, for validation efforts. His piece represents a call to action for testing experts to embrace more responsibility for validation and forge new alliances to get the work done.

In this brief response, I try to maintain the momentum that Haertel (2013) initiates by identifying and then loosening what I see as a blockage point in current efforts to do what he proposes. Like Haertel (2013), I believe that we as testing experts have been limiting our vision by waiting for others to articulate and validate interpretive arguments for rapidly proliferating test uses. I agree that the comforting but often toothless guidance of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), that, "the ultimate responsibility for appropriate test use and interpretation lies predominantly with the test user" (p. 112), is insufficient and relegates us to being reactive instead of proactive. So what is holding us back?

I argue that validation frameworks are missing a theory about how test uses change. The perspective on modern test score usage that I find most helpful is epidemiological. Numbers travel. They aggregate and spread. This spread is not random. Numbers travel and aggregate along vectors, and their spread is facilitated by known risk factors. The more we understand about the tendency of test usage to propagate along known vectors, the better we are able to anticipate or require validation. I acknowledge that the comparison of the proliferation of test score uses to an influenza epidemic is not perfect, but I hope it is instructive nonetheless.

This perspective on the spread of test use is implicit in many validation frameworks. Haertel and Herman (2005) provide a historical account of interpretive arguments for accountability testing, but they pay less attention to how scores from a single testing programs can support different uses over time. Kane (2006) and Messick (1989) provide illustrative categories of test uses but do not map the ebb and flow of a single testing program's purpose over time. The key question that I think Haertel's (2013) article forces us to ask is not, "When are you getting back to me with an interpretive argument about how testing improves schooling?" It is, instead, "Why didn't we see this coming?" And then, "Anticipating that similar uses are likely to proliferate in the future, how can we make sure they are defensible or discouraged?"

In the remainder of this commentary, I present three tendencies toward which I (and certainly many others) have observed modern test score use drifting over time. These loosely relate to the framework of test uses that Haertel (2013) presents in his Table 1. If his table is a map, then these three tendencies illustrate patterns of test use as they spread across this map.

My thesis is that validation efforts will continue to be reactive, and always a step behind, until we develop useful models for predicting changes in test score use over time.

The first tendency is toward aggregation. Examples of interpretations of aggregated test scores include simple "percentage proficient"-type descriptions as well as trends, gaps, gap trends, and far more complex school "Adequate Yearly Progress" calculations. The key point is that this aggregation can occur at a substantial distance in space and time from the design and development of the original test. An epidemiological perspective on test score use allows us to anticipate that a test designed for individual-level inferences and uses is likely to be extended by downstream users. These downstream uses will tend towards higher levels of aggregation, to support trends, gaps, and gap trends, as well as school, district, and state aggregates.

In Haertel's (2013) table, aggregation generally proceeds downwards in the table, from learners to methods and actors. Certainly many tests that initially supported individual placement and selection decisions have been aggregated to support inferences about methods and actors. The increasingly widespread use of SAT and ACT scores as state-level achievement indicators is but one example. Aggregation also reflects a shift from left to right in the table, from direct action to indirect action, as the description of a testing initiative at an aggregate level "shapes perceptions" and "focuses the system" (Haertel, 2013, Table 1). In many ways, the shift from direct action to indirect action is one that is synonymous with aggregation or the impact of an aggregate-level test use.

The second tendency is toward adjustment. By adjustment, I refer loosely to the combination of a score with information from a different scale or construct. As a common example, individual test scores can be predicted by other variables, like a socioeconomic status

index. The difference between the observed and predicted scores are interpretable as test scores above and beyond what is predicted by socioeconomic status. A popular score adjustment procedure is the widely used Student Growth Percentile metric (Betebenner, 2008) that predicts current scores from past scores. The combination of current scores with past scores, any or all of which may be on different test score scales, allows for different, conditional interpretations. Again, the key point is that this adjustment can occur at a substantial distance in space and time from the design and development of the original test. Value-added metrics for teachers can be seen as an example of both aggregation (from student-level scores to classroom-level scores) and adjustment (using predictions from past scores).

The third tendency is toward equating. Equating allows for the mapping of a score on one test onto the score scale of another test (for a formal definition, see Holland and Dorans, 2006). Tracking the origins and the evolution of the use of state NAEP scores provides an illustrative lesson in this tendency toward equating. Comparisons of state proficiency standards began with casual approaches (e.g., Musick, 1996). These received enough attention to warrant academic discouragement of the use of NAEP as a common metric (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999). Ultimately, however, regular reporting of state standards on the NAEP scale proceeded using the best statistical approaches available (Bandeira de Mello, 2011).

Yet again, we see a test use that is separated in time and space from the original design of the assessment. We can discern a pattern, where a compelling question is asked by policymakers, researchers, and practitioners that can only be supported by cross-test comparisons, and it is answered using ad hoc equating approaches. There may be pushback

from the academic community, but ultimately the link is incorporated into standard practice. We can observe the same pattern recurring with current linking efforts between NAEP and TIMSS (National Center for Education Statistics, 2011), where researcher-led reports that used more ad hoc linking procedures (e.g., Hanushek and Woessman, 2011) framed a question that ultimately demanded an answer using the best statistical approaches available. This linking does not meet the standards demanded of "equating" as defined by the measurement community (Holland & Dorans, 2006). However, I would argue that we have ceded an opportunity to advocate for our standards by lacking the foresight to predict what is, in retrospect, a completely unsurprising desired use of a test.

These three tendencies, aggregation, adjustment, and equating, are less vectors in the epidemiological sense than they are itineraries for the journeys of test scores. To get at the root causes, to identify the vectors themselves, it is helpful to identify the intuitions that support test score use outside the measurement community. Braun and Mislevy (2005) identify many of these intuitions in their piece, *Intuitive Test Theory*. They describe naïve theories of test use that explain testing as it is experienced by students in school. Then they note that these naïve theories fail, just as naïve theories of the physical world fail, in supporting test use and test-based analysis at a large scale. Intuitions like "a test is a test is a test," "a score is a score is a score," and "any two tests that measure the same thing can be made interchangeable with a little 'equating' magic" (Braun & Mislevy, 2005,p. 493) describe why aggregation, adjustment, and equating all seem permissible without need of further validation.

Similarly, and in keeping with Haertel's (2013) call for further collaboration with scholars of other disciplines, scholars from many areas have dedicated pages to understanding the contribution of numbers and quantification to the veneer of objectivity (e.g., National Research Council, 2012; Porter, 1996). If these are the causes of the uncritical spread of test use, prevention, as opposed to treatment, requires inoculation in the form of statistical literacy and assessment literacy: a swapping of naïve intuition for expert intuition, where casual aggregation, adjustment, and equating would feel just as suspicious to the public as it does to psychometricians.

Needless to say, test scores can travel faster, farther, and freer than ever before. Anyone with a blog and an internet connection can conduct and disseminate an ad hoc aggregation, an ad hoc adjustment, or an ad hoc equating from publicly available data. Certainly, the measurement community cannot be expected to anticipate every imaginable use of test scores. However, some uses of test scores will drift along known vectors so predictably, that I believe they can and should be addressed proactively and as a matter of course. Haertel's (2013) thoughtful article has called the measurement community to action and helped to deepen our understanding of test use, particularly of the poorly articulated indirect actions that tests are assumed to take. Now, I hope that we would not delay validation while waiting for others to articulate these interpretive arguments. I think that, by now, we know which way the winds are blowing.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association

Bandeira de Mello, V. (2011), *Mapping state proficiency standards onto the NAEP scales: Variation and change in state standards for Reading and Mathematics, 2005–2009* (NCES 2011-458). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC: Government Printing Office.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4): 42–51.

Braun, H, & Mislevy, R. (2005). Intuitive test theory. *Phi Delta Kappan, 86*, 489-497.

Feuer, M.J., Holland, P.W., Green, B.F., Bertenthal, M.W., & Hemphill, F.C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement*. *The 104th Yearbook of the National Society for the Study of Education, Part II* (pp. 1-34). Malden, MA: Blackwell.

Hanushek, E. A., & Woessmann, L. (2011). How much do educational outcomes matter in OECD countries?. *Economic Policy, 26*, 427-491.

Holland, P. W., & Dorans, N. J. (2006). Linking and Equating. In R. Brennan (Ed.), *Educational measurement, 4th ed*. (pp. 187-220). Westport, CT: Praeger.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement, 4th ed*. (pp. 17-64). Westport, CT: Praeger.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement, 3rd ed*. (pp. 13-103). New York: Macmillan.

Musick, M. (1996). *Setting education standards high enough*. Atlanta: Southern Regional Education Board.

National Center for Education Statistics. (2011). *NAEP-TIMSS linking study: Comparing state academic performance against international benchmarks.* Brochure. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/about/schools/2011472.pdf

National Research Council. (2012). *Using Science as Evidence in Public Policy*. Committee on the Use of Social Science Knowledge in Public Policy. K. Prewitt, T. A. Schwandt, and M. L. Straf, Editors. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life.* Princeton, NJ: Princeton University Press.