Mid-Western Educational Researcher

Volume 24 | Issue 1 Article 5

September 2023

Politics, Economics, and Testing: Some Reflections

Michael J. Feuer
George Washington University

Luncheon Keynote Address

Politics, Economics, and Testing: Some Reflections

Michael J. Feuer
George Washington University

Good afternoon, and thank you for such a warm welcome and kind introduction. It is a pleasure to be here, and I've been looking forward to visiting with you—and seeing Columbus—ever since Cindy proposed this many months ago.

My topic today is assessment and accountability, surely not new words or concepts to anyone here, although perhaps my arguments will provoke some new thinking. Let me start with some data that probably will sound familiar, if not in its detail then its gloomy underlying message. This relates to some "recent" test results:

> Out of 57,873 possible answers, students answered only 17,216 correctly and accumulated 35, 947 errors in punctuation in the process. Bloopers abounded: one child said that rivers in North Carolina and Tennessee run in opposite directions because of the will of God. (U.S. Congress, Office of Technology Assessment, 1992, p. 109)

If you're wondering how you missed this important news item, don't worry: it's not from the *most* recent NAEP or SAT or Ohio Achievement Assessment. Rather it's from one of the first instances of large scale written educational testing, *circa 1840*, a time of great reform in American schooling, a period that later became known as the Common School reform movement and was associated with Horace Mann and others who spent their lives trying to broaden the franchise of educational opportunity and raise standards, all at the same time. [Let me digress just for a moment here to cite the great historian of education, Lawrence Cremin, who noted in a brilliant and short book he wrote just before his untimely death, the idea that we could (and should) raise quality standards and increase access simultaneously was a uniquely American ideal and one that we are still, in many ways, pursuing. Cremin's (1990) book should be required reading for anyone contemplating venturing into the turbulent world of education reform…].

But back to my story line…which is about the importance of history in considering contemporary educational challenges. We sometimes forget that some of our most vexing problems are, in their fundamental aspects, not new. We are a relatively young country (to paraphrase from Tom Lehrer's memorable line about Mozart, by the time the US National Academy of Sciences was founded in 1863, the first King of England had already been dead for about 1000 years…). But we do have history here, and our history of education is marinated in flavorful juices of the great American experiment with divided government, with a certain exceptionalism that steered us away from other systems that had seduced so many other societies. Our allergy to centralized authority, coupled with a deeply held aspiration for fairness, are two elements in our unique political culture that have had and continue to have great effect on education policy, reform, and learning.

Part of my message today is a simple one but I hope not simplistic: Our penchant for accountability and our appetite for standardized testing are, in the language of statistics and psychometrics, highly collinear.

But first I want to address a specific aspect of the history, as it relates to testing and accountability. It is sometimes tempting to demonize the testing community for all kinds of perceived evils: bringing us the wondrous frustrations of multiple choice test items that seem to bear little relation to what we really value in teaching and learning, being so ready and willing to market more and more tests that can be scored at greater and greater speed, and for not being terribly concerned with the deeper meanings of test results or their behavioral consequences as long as the results meet certain standards of statistical reliability. We find it easy and convenient to blame the test makers for everything from adverse impact in higher education to the horrors of teaching-to-the-test in K-12. I've actually heard one good friend of mine, in a rather extreme fit of anger, attribute a teacher's suicide to NCLB requirements for student testing!

It's all rather easy, and somewhat enjoyable, this test bashing, and I admit at times I've tasted the Kool-Aid. But let's not forget (and as a recovering economist I cannot forget) that there is usually a *demand* side that at least partially explains why certain strange or undesirable things appear on the market. In this case, i.e., the emergence of uniform written exams, the forces that converged to enable and propel testing as perhaps the most persistent and arguably powerful tool for assessment of educational quality and governance of educational change, had its roots in fundamental aspects of the unique experiment in democracy that was taking shape in our new republic. Why should we be surprised, really, that by 1975 one of the great minds of mental measurement and educational assessment, was lamenting five decades of controversy over mental testing while noting, perhaps immodestly but certainly with scientific validity, that psychometrics had become one of the greatest contributions of psychology to human affairs (Cronbach, 1975)?

We had better recognize that this tension would not have been possible if there hadn't been, for a long time and

for many legitimate reasons, a powerful demand side in the production and distribution of tests, an appetite for standardization that had its roots in the coinciding principles of democratic accountability and efficiency in the expansion of educational opportunity.

In sharing the news of how poorly students performed on that 1840 assessment and of how charmingly wrong some of their answers really were I could be making a perhaps simpler point, namely that rumors of the golden age—that elusive and transitory period in history when things were fine as compared to how awful things have become—are more than a tad exaggerated. Now, it has been empirically documented frequently (most recently in an extraordinary book by two Harvard economists, Goldin & Katz, 2010) that at least until the last quarter of the 20th century our remarkable educational system was, indeed, in something of a golden age, largely responsible for advancing the general economic and social welfare of our nation and for uplifting the quality of life and standard of living to levels well above any other nation in the world. (It is important that we keep this historical record in mind as we contemplate the future, and though much of the doom and gloom rhetoric based on cross-sectional evidence from international comparative assessments is exaggerated, there is reason to fear the ill effects of complacency borne of prior success.)

But my main reason for recalling the 19th century experience with testing is to make a different historical point: it is to emphasize that standardized educational tests have been a staple of public accountability in education for almost two centuries, and that from their inception they have been popular devices used for both good and mischief. Horace Mann and his partners in the great reform movement were not only brilliant social reformers intent on expanding the educational franchise, but they were shrewd politicians too, who understood long before the ascendance of professional communications experts and policy wonks that by including certain questions on the tests they could expose the failures of school masters they were battling with, and, as one of our preeminent educational historians noted, use testing as a "bludgeon of reform…"(Tyack, 1974). In a word, if you think some teachers and principals are feeling pressured by NCLB testing, you are right: but based on the historical evidence one cannot help think that today's test-based accountability pales in its ferocity when compared with the earliest episodes of the "bludgeoning…"

We've been testing for a long time. My point is that it's not so surprising when viewed in the context of the American experiment. There is a deeply American quality to this reliance on tests: they were a remarkable invention of social engineering in large part because they did not appear to require a tradeoff between efficiency and fairness—they rather spectacularly seemed to achieve both goals at once. I would argue that standardized testing became a symbol of the aspiration for fairness and universal access that distinguished American schools from European and Asian schools.

Moreover, as the tests grew more sophisticated, both in their format and in models for scoring and interpreting of results, they increasingly were viewed as tools of rational—scientific—management. Let me elaborate just a bit on this comment and tie testing to more generic properties of technology and society. In a country and culture already beginning to exhibit a certain fascination with the possibilities of technology—which would of course characterize the extraordinary transformation of the American society and economy over the remainder of the century—here was one, standardized testing, with genuinely *dual* uses. The duality, at the time, was already comprised of measurement (i.e., describing what the kids are learning) and reform (i.e., motivating change to improve their learning). And since then that duality has blossomed from two branches into a rather more complex system with multiple purposes, multiple designs, and a highly complex interweaving of goals and constraints that makes most rational policy analysts run for something simpler (mapping the origins of the universe, for example.)[1]

Let me try to underscore some of the key ideas embedded in this brief historical prelude:

- Neither NCLB nor its recent antecedents (Goals 2000: remember that?) are new attempts to rely (and perhaps over-rely) on testing as a technology of reform, nor is the evidence of arguably irreconcilable multiple uses of test results;
- Tests—like most if not all technologies—are imperfect, which means that some results will overstate and other results will understate the "true" state of a child's learning or potential;
- The fact that we continue to rely on tests is to a large extent attributable to our unflagging pursuit of at least some "objectivity" in the way we evaluate teaching and learning, which is rooted in the framing principles and philosophy of the American democratic experiment and, in particular, our aversion to centralized authority;[2] and
- What has been missing from the often heated debate over assessment and its multiple uses has been a kind of rational and dispassionate analytical framework for assessing its benefits and costs, perhaps similar to the analytical frames we apply to other complex phenomena in which there needs to be attention to both the good and the bad, a framework that could perhaps inform policy makers and the public about the strengths and limitations of testing and stimulate the kind of research needed to increase the benefits and reduce the costs.

1 A Nobel-prize winning physicist once confessed that after working on education reform for a few years he decided to go back to estimating the ages and chemical composition of the planets… which he said was much easier.

2 For a description of how American policy makers at times envy their counterparts in more centralized systems, see my discussion of French Education Minister Claude Allegre's visit to the National Academy of Sciences, in Feuer, M. (2006), *Moderating the Debate: Rationality and the Promise of American Education.* Cambridge: Harvard Education Press.

This leads me to the main lesson I'd like to impart today, namely that we need to start thinking about the future of testing and accountability through a lens of potential benefits, potential costs (or risks), and perhaps most important, the pursuit of reasonable rather than optimal solutions to the problems of testing and accountability. I'm going to revert to some core principles of economics in advancing us toward such a framing of the issues.

It is a staple of economic theory that individually rational and self-interest seeking behavior can lead to disastrous or at least seriously suboptimal social outcomes. Anyone who has driven on a highway and has confronted the frustration of "rubbernecking," for example, has first hand experience with the failure of individual rationality in terms of its collective results (see for example, Schelling, 1974).

What does traffic flow have to do with testing and accountability? In a nutshell, the fact that certain behaviors, or technologies, lead to unintended or undesirable consequences, is not, in itself, a sufficient basis for banning the technology; rather, understanding the sources of what are sometimes called "externalities" in the literature of political economy, is an important foundation upon which to build appropriate policy remedies.[3] The lesson is that

- there have always been and continue to be justifiable arguments for accountability generally and for the use of tests as one tool of accountability;
- there are unintended negative consequences of testing for accountability that need to be anticipated as best as possible; but
- the undesired risks or costs associated with testing as a tool of accountability need to be weighed against the potential and measureable benefits as well as against the counterfactual case of eliminating testing from the toolbox of acceptable accountability practices.

So, just as we would not prohibit either traffic flows or the rights of drivers to look out their windows, we should not lurch toward a prohibition of testing just because it obviously (and not so obviously) entails downside risks and some unarguably bad behavior. A good example of public policy analysis that hinges on this approach to dealing with benefits and costs is in the environmental movement. Strategies for remedying the ill effects of individually-self interested behavior that results in water and air pollution have evolved from the naïve view that damaging the environment was in some ways analogous to crimes warranting rigid and coercive policing, to the development of more sophisticated political, legal, regulatory and incentives-based approaches. A prominent economist working in this area offered this contrast:

> …the police power approach … is appropriate when a certain kind of behavior is perceived as a terrible social threat and it is felt the behavior must be stopped even at great cost…but there is a world of difference between hijacking [and other such crimes] and pollution. Hijacking is a threat to life and property without redeeming features, whereas pollution is a by-product of thousands of individual decisions in the course of very desirable activities—production and consumption of commodities and services. Hijacking should be prevented if possible, whereas with pollution, the goal is to induce people to continue the desirable activities in ways that reduce and alter environmental discharges…. (Mills, 1978, p. 204)

Embedded here is the notion that simply prohibiting polluting behavior is likely to be inefficient, counterproductive, and insensitive to negative consequences that could be even more damaging than the pollution itself. A range of strategies have been devised over the years, with varying success, all aimed at inducing changes in behavior and collectively reducing the pace and magnitude of a perceived and real set of externalities. Regulatory approaches, for example, can be costly to design and implement, and though still in wide use have exhibited mixed levels of success; variations on taxation schemes, which are intended to curb polluting behavior by imposing monetary charges, have become more popular, to economists at least, although such programs also can be costly to design and enforce.

All these initiatives share a basic proposition, namely that *the goal of reducing a negative externality requires attention to benefits and costs—in the estimation of the effects of the polluting technology and the effects on the economy and society of curbing the pollution, and in the estimation of the costs associated with designing and implementing the policy strategy itself.*

Perhaps this schema help us untangle the problems of testing and accountability. The main point is that test based accountability systems have benefits and costs, and I'll start with the latter. Here is an abridged list of the things that can and do go wrong when tests are used inappropriately:

1. Tests are imprecise tools of estimation that provide only a partial view of selected aspects of what students know ("domain sampling"). Using tests as a basis for more comprehensive judgments is usually inappropriate.
2. Tests alone offer preliminary clues, at best, as to how students learned whatever it is they demonstrate on the test. Inferences about teachers, schools, principals, class size, and other possible causes require substantially more data than score reports.
3. Most of the tests available "off the shelf" are not well-suited to providing teachers with useful information on the cognitive or intellectual barriers their students face, the special work they need to improve, or the ways teachers can shape their lessons to help kids overcome specific learning gaps.
4. When test results are used as a basis for making significant decisions ("high stakes" decisions) the validity of

3 This argument is expanded in Feuer, M. "Externalities of Testing: Lessons from the Blizzard of 2010," *Measurement*, 8: 59–69, 2010.

the scores can be compromised.

5. As a corollary to #4, teaching to the test is usually a bad idea, no matter what the test looks like, unless of course we don't care much about the validity or reliability of the information the test was originally designed to produce.
6. Decisions based on any cut score methodology will result in misclassification, assuming tests are imperfect estimators of the underlying domain of interest. (Most people worry about false negatives, i.e., kids erroneously being identified as "below basic" when in fact they're not. Adverse impact issues arise from this type of error. But as important is the problem of false positives, i.e., kids (or schools) that are branded as passing when in fact they're not.)
7. Using tests as the sole basis for measuring adequate yearly progress can lead to huge numbers of schools being misclassified, even when the source of their failure can be quite random. The effects of such misclassification on resource allocation, student mobility, parental support for schools, and morale can be costly in ways we don't really know how to measure.
8. Too much emphasis on test results naturally leads to distortions in the way both good and bad teachers allocate their time. We simply don't know how many good teachers will (a) develop reactive strategies that undermine their otherwise good instincts, (b) find ways to game the system just so they can go on with their good teaching, or (c) give up and leave the system to only those teachers for whom the testing doesn't make much difference!
9. Tests used to compare schools across states ought to be designed with enough similarity of content and format to permit valid comparisons. Reconciling this simple dictum, distilled from the literature on test equivalence and linking, with the historical and contemporary insistence on state and grass roots control of curricula and pedagogy is a full-time job.
10. There is a risk that as tests become both more important and more similar across states and jurisdictions they will become a de facto national curriculum. Few Americans seem ready for that.

Against this impressive array of good reasons to curb our enthusiasm for testing, what can possibly be said in its defense? I will offer a few general answers and some more directly tied to our current situation.

1. The alternative to standardized assessment of student learning is a return to subjectivity and intuition, neither of which should be viewed as a curse except if they are attributes of decisions that a) would benefit from more rigor and precision and b) are the basis for actions that can seriously affect children, teachers, or the public trust. Although we've all been in classrooms where inspired teachers are doing wonderful things, relying on a "know it when you see it" criterion for evaluating teaching and learning would be insulting to the profession not to mention hazardous to the learning opportunities of generations of children.
2. Given the complexities embedded in the words "education" and "learning," it is important to agree on at least some basic approximations and on some metrics to inform parents and others of whether anything of value is actually taking place. A culture that is capable of digesting and interpreting, with exquisite subtlety, the massive quantities of statistics that are collected about, for example, major league baseball, has clearly expressed its appetite for quantitative information about progress of education.
3. When designed and implemented properly, tests can provide useful information to teachers, principals, and school officials striving for improved policies and practices. The fact that test results are often expropriated for uses that go beyond their technological capacity and beyond the aims for which they have been validated is not, in itself, a sufficient argument against their use for the purposes for which they were designed and validated.
4. Without some agreed-upon quantitative benchmarks, the good embedded in so many of our schools and the high quality of professionalism exercised by so many of our teachers will a) be suspect and b) not become the basis for learning and improvement elsewhere, where it is needed most. It's not just that test scores provide a certain kind of braking function on public and political jockeying and the impulse to make extravagant claims of success prematurely; it goes the other way too, in terms of providing evidence of genuine progress that can be the foundation for scaling up progress beyond specific cases.
5. The inverse of point 4 should be obvious: without agreed-upon metrics the concept of public accountability is fundamentally undermined. Recall the initial use of uniform examinations at the birth of the common school reform, i.e., the idea of giving parents on "both sides of the tracks" information about how their children were faring. Identifying schools or school systems that are in trouble, using well designed tests of student achievement, should be an acceptable basis for further investigation, and most important, design of programmatic or policy remedies.

And now for a few more specific arguments relevant to our current situation:

- The bad news about NCLB notwithstanding, there are some positive results that are perhaps under-reported. For example:

  > NCLB's focus on students with low achievement seems to have had some short-term positive effects. The percentage of schools meeting Adequate Yearly Progress (AYP) targets increased in 2003-04 from the year before in most states, and the recently released National Assessment of Educational Progress (NAEP) long-term

trend scores have shown some narrowing of achievement gaps. (Linn, 2005, p. 1)

- On the issue of whether teachers understand and are able to align their teaching to state standards, the news is mixed, but on the plus side some research has been illuminating: Many district and state superintendents and teachers have applauded the move toward greater alignment of curriculum to state standards. On the very important matter of the achievement gap, which has been a persistent problem and one that NCLB explicitly seeks to affect, there is also mixed news but on the positive side there is evidence that percentages of students scoring proficient have risen and that gaps between subgroups have narrowed in most states at the elementary, middle, and high school levels, although in a notable minority of cases gaps have widened (see for example, Dietz & Roy, 2010).
- A more comprehensive set of studies point also to significant progress in the narrowing of the achievement gap as a result not specifically of NCLB but of the more general standards movement of which NCLB is the latest example (Gamoran, 2008).

Is there a grand lesson here? Let me suggest that at least three basic conclusions are worthy of consideration by education researchers eager to contribute to improved policy making. First, there is a hardy appetite in the policy world for credible and reliable information derived from empirical study, and we should be proud of our community for its diligence in the pursuit of answers to complex questions. Second, the most interesting questions are, indeed, too complex to expect definitive or optimal solutions, and our goal should be to provide reasonable, rather than perfect recommendations, based on appropriate rather than exhaustive deliberation.[4] And finally, there is merit in analyzing reforms from the standpoint of their potential (and measurable) benefits, intended and unintended, along with their actual risks and downside effects. Our overarching goal as a community should be to engage with policy makers and politicians who are entrusted with making the tough decisions, humbly and carefully outline the pluses and minuses of any particular action, and offer our scientific expertise toward the design and implementation of programs that can help all our children learn.

4 This is the essence of Herbert Simon's definition of "procedural rationality." See Feuer (2006) for application to education policy and research.

References

Cremin, L. (1990). *Popular education and its discontents*. New York: Harper and Row.

Cronbach, L. (1975). Five decades of controversy over mental testing. *American Psychologist*, *30*(1), 1-14.

Dietz, S. & Roy, M. (2010). *How many schools have not made adequate yearly progress under the No Child Left Behind Act?* Washington, DC: Center on Education Policy.

Feuer, M. (2010). Externalities of testing: Lessons from the blizzard of 2010, *Measurement*, *8*, 59-69.

Feuer, M. (2006). *Moderating the debate: Rationality and the promise of American education.* Cambridge: Harvard Education Press.

Gamoran, A. (2008). *Standards-based reform and the poverty gap.* Washington, DC: Brookings Institution.

Goldin, C. & Katz, L. (2010). *The race between education and technology.* Cambridge: Harvard University Press.

Linn, R. (2005). *Fixing the NCLB accountability system*. UCLA: CRESST.

Mills, E. (1978). *The economics of environmental quality*. New York: Norton.

Schelling, T. (1974). *Micromotives and macrobehavior.* New York: Norton.

Tyack, D. (1974). *The one best system.* Cambridge: Harvard University Press.

U.S. Congress, Office of Technology Assessment (1992, February). *Testing in America's schools: Asking the right questions*. Retrieved from http://govinfo.library.unt.edu