NATIONAL EDUCATION POLICY CENTER

School of Education, University of Colorado Boulder
Boulder, CO 80309-0249
Telephone: 802-383-0058

NEPC@colorado.edu
http://nepc.colorado.edu

Research-Based Options For Education Policymaking

School Accountability, Multiple Measures and Inspectorates in a Post-NCLB World[1]

William J. Mathis, University of Colorado Boulder
October 2015

While the concept of accountability for public schools has been an issue for as long as we have had universal education, policymakers have struggled to find a successful approach. Standardized student testing with published teacher and student test scores, as an accountability mechanism, can be traced to the 1870s.[2] Falling somewhat out of favor during the progressive era, testing for school accountability took on a heightened intensity beginning in the 1970s.[3] The No Child Left Behind Act of 2001 (NCLB) dramatically strengthened the test-based method of school evaluation, prescribing interventions and penalties for schools not meeting fixed test-score targets set by each state.

NCLB is the current version of the federal Elementary and Secondary Education Act (ESEA), and hope waxes and wanes for ESEA reauthorization, even though the law technically expired in 2007. One impact of NCLB has been a centralization of power in the federal government, yet political signs point to a partial return of school accountability mechanisms to state decision-makers.[4] In fact, despite the prescriptiveness of federal law, considerable variation in school approval systems has already taken place, as a result of the federal waiver process.[5] Nonetheless, much of this latitude is in the finer details—the core of test-based accountability remains universal.[6]

With the federal testing mandates and the granting of waivers only when a plan meets the Education Department's ideological criteria, little room has been left for school evaluation approaches that seek meaningful alternative approaches, even though these alternatives may seem far more appropriate in a nation with a strong tradition of local educational governance.

As discussed below, two alternatives are particularly worthy of consideration: (a) combining multiple measures that include inputs as well as outputs; and (b) inspectorate systems incorporating self-evaluations coupled with site visits conducted by disinterested but qualified visitors representing the state or an accreditation group.

Test-Based Models

Test-based school accountability systems consist of three simple components: "testing students, public reporting of school performance, and rewards or sanctions based on some measure of school performance or improvement" (p. 91)[7]. Following the period during the 1970s where tests focused on minimum basic skills, test-based models gained a federal endorsement in the Goals 2000 effort and took on greater prescriptiveness in the NCLB law, which defined the grades to be tested as well a set of interventions or penalties for schools failing to meet test-based proficiency cut-offs.[8] Since NCLB did not seriously address resources or capacity-building, and since almost all children were to meet high standards by 2014, the effort was doomed to fail.[9] The law's prescribed interventions for inadequate progress (reconstitution,[10] turnarounds,[11] restart,[12] and school closure[13]) all shared the problem of little or no evidence of effectiveness at any scalable or practical level.

As it became increasingly clear that a student's test scores in a given year were strongly predicted by that student's scores from the previous year, policies shifted in many states toward reliance on so-called Growth Models. Most commonly, to Value-Added Models (VAM), which attempt to control for prior scores and other measured factors and then attribute the residual—the growth not accounted for by these other factors—to schools or teachers. While this method assumes a causal relationship, the American Statistical Association has cautioned against the high-stakes use of such measures.[14] Several concerns have been raised by researchers: the assumptions underlying these models are problematic; the growth scores assigned to teachers are unstable and are not valid measures of teacher quality; and the test-driven narrowing of teaching and learning remains.[15]

By 2015, test-based standards and accountability policies could show little or no evidence of effectiveness. In fact, they generated unintended and negative consequences such as teaching to the test, curriculum narrowing and drill-and-practice.[16]

Multiple Measures

One of the key criticisms of the test-based model is that standardized testing does not measure all the important aspects of a successful school. Coupled with a growing backlash by parents and policy makers against what they considered to be excessive testing, the logical evolution was toward "multiple measures."[17] The reasoning is straightforward; a more comprehensive set of measures will more validly capture the broader set of cognitive

and affective learning goals of schooling.[18] Unfortunately, "multiple measures" is an elastic term that includes an eclectic variety of elements. Depending upon the speaker and whatever pre-existing data are at hand in a given state, the term can mean many different things and thus result in many different policy approaches.

In looking at the federal "waivers," 24 of 27 applying states proposed a wide variety of multiple measures.[19] In 2009, individual states identified from four to 22 different measures, characterized by a strong collection of outcome measures and a virtual absence of opportunity, input, or process measures.[20]

Advocates of multiple measures often speak of a "dashboard" of decision data.[21] In order to have consistency across schools, the proposed dashboards are composed almost exclusively of empirical measures with data elements such as truancy, graduation rates, and disciplinary referrals. These have the advantage of being highly reliable because they have a standard meaning across schools. But their validity, as a measure of school quality, is open to question.

If a composite (or "report card") score is constructed from these multiple measures, a particular problem is the assignment of weights to the various measures.[22] For example, can 70% passing a math test be added to a 10% decrease in disciplinary referrals, and should this be adjusted for socio-economic factors and school history? While a number of statistical techniques (such as factor analysis) show promise for addressing these concerns, current decisions appear to be based on the judgment of individuals or working groups. There is no optimal answer to this dilemma.[23]

Yet, "multiple measures" has served as a bridging concept between different policy camps. Linda Darling-Hammond and Paul Hill, for instance, released companion reports addressing elements to be included in the next generation of school evaluation systems.[24] While agreeing on vague generalities such as the need for assessment of "college and career ready" standards, the use of evaluation consequences at the school level, outside intervention where needed, and the proper role of government; these agreements are at such a high level of abstraction that "multiple measures" remains more a rhetorical consensus than a verifiable accountability model.

School Self-Evaluations Plus Inspectorates

While eclipsed by test-based models in the United States, self-evaluation combined with inspectorate systems continue to be the norm in most OECD countries. The closest parallel in the United States are regional accreditation organizations that guide self-evaluations and organize visiting teams. The method is particularly used in higher education. Basically, the school conducts a structured self-evaluation and then, in systems combined with an inspectorate, a visiting review team validates the self-evaluation report. That is, the self-evaluation report becomes a foundational document for the inspection team.[25] Through interviews and data review, the team seeks to verify such things as express student expectations, the comprehensiveness of assessment, curricular adequacy, professional development, and available supports and interventions for high needs children.[26] Depending on the particular variation of this approach used, differences may shape the length of advance warning (if any) given to the school, the size of the visiting team, and the degree of disruption to school activities.

The advantages of a self-evaluation and inspection model are that the evaluation can include subjective components that are not easily measured by test scores or the aggregation of quantitative data. Thus, it can be broader and more inclusive, and it is less likely to distort teaching and learning. Also, a self-evaluation can be more revealing of needs than a staged show for visitors. However, subjective goals can be too loosely defined and subjectively presented. Cost is also a concern.[27]

As for evaluating the evaluation system, "Despite its long history and ubiquity, inspection has existed until comparatively recently in an a-theoretical limbo with practices and procedures assessed on little more than the commonsense of those who commend or criticize them" (p. 10).[28] The evaluation problem is that cause and effect are hard to nail down. For example, did the new textbooks recommended by the team result in better teaching and learning? Would the school have purchased the materials anyway? One clear finding, however, is that interviews of participants show a positive perception toward self-evaluations and inspectorates, with 90% of Great Britain principals and teachers reporting being satisfied with the system.[29]

The Threshold Question: Adequate Inputs and the Opportunity Gap

> *[I]f schools are being held accountable for improving teaching and student learning, policymakers at all levels of the educational system, regional and state levels as well as the national level, should also be expected to support the capacity required to produce improved teaching and learning (p. 21).*[30]

The greatest conceptual mistake of test-based accountability systems has been the pretense that poorly supported schools could systemically overcome the effects of poverty by rigorous instruction and testing.[31] The system has inadequately supported teachers and students, has imposed astronomically high goals, and has then inflicted punishment on the most needy.

School evaluation systems will only succeed with all around accountability.[32] This includes holding state and federal governments accountable for ensuring that children have the opportunities to learn necessary for success, inside schools and in their communities. Ultimately, a child denied opportunities will arrive at school with very high needs, and a school denied adequate resources will not effectively address those high needs. No evaluation system, by itself, is capable of overcoming such deficiencies.

Recommendations

1. Along with efforts to evaluate schools and impose consequential penalties, each state should assure that students have adequate opportunities, funding and resources to achieve that state's goals.[33]

2. Continued development of multiple-measure and dashboard approaches should strive for comprehensiveness, balance between inputs and outcomes, clarity, and measurability. As contrasted with a convenient collection of available data, the information should accurately and validly reflect the desired learning outcomes and the input resources needed.

3. Standardized test scores should be used cautiously and only in combination with other data, to avoid creating incentives for narrowed and distorted teaching and learning.[34]

4. The aggregation of data into a single score or grade should be avoided. Such procedures hide valuable information while invalidly combining disparate and unrelated objects.[35]

5. States should develop, train and implement school visitation teams. In order to be economical, sites most in need of improvement should be prioritized. Standardized test scores can be validly used to establish initial priorities.[36]

6. External reviews should focus on providing guidance and support for school development and improvement, rather than on imposing sanctions.

7. External reviewers should be qualified experts who meet prescribed standards. Robust training should be compulsory, with retraining required on a periodic basis.

8. Multiple stakeholders (administrators, teachers, students, parents, community leaders, and researchers) should be involved in the design of the state's evaluation/ inspectorate program.

Notes and References

1 Ryan, K. E., Gandha, T., & Ahn, J. (2013). *School Self-evaluation and Inspection for Improving U.S. Schools?* Boulder, CO: National Education Policy Center. Retrieved October 1, 2015 from http://nepc.colorado.edu/publication/school-self-evaluation

> With the rapid evolution of thought on accountability issues, this research summary is expanded from the original policy brief. Two major themes have been added: (1) if schools are to be held accountable for improving teaching and student learning, policymakers at all levels are first obligated to provide the necessary capacity to reach their goals; (2) the concept of "multiple measures" has gained broader acceptance in accountability models and is made a section of this brief.

2 Tyack, D. (1974). The One Best System: A History of American Urban Education. Harvard University Press: Cambridge, MA pp 47-49

3 Baker, S. (June 22, 2013). The Origins, Evolution, and Effects of Test Based Accountability: North Carolina and the Nation, 1976-2009. Scholar Commons. Retrieved October 13, 2015 from http://scholarcommons.usf.edu/compaccountability-2013/Papers/PreConferenceSubmissions/5/

Whitehurst, G. (July 10, 2014). The Future of Test-based Accountability. Brown Center Chalkboard. Retrieved October 13, 2015 from http://www.brookings.edu/research/papers/2014/07/10-accountability-whitehurst

4 Aldeman, C., Robson, K. & Smarick, A. (June 29, 2015). Pacts Americana: Balancing National Interests, State Autonomy, and Education Accountability. Retrieved September 30, 2015 from http://bellwethereducation.org/publication/Pacts-Americana

5 Center on Education Policy (October, 2012). What Impact Will NCLB Waivers Have on the Consistency, Complexity and Transparency of State Accountability Systems? George Washington University. Washington, D.C.

6 Center on Education Policy (October 2012). What Impact Will NCLB Waivers Have on the Consistency, Complexity and Transparency of State Accountability Systems? Georgetown University. Washington, D.C.

7 Kane, T. J. & Staiger, D.O. (Fall 2002). The Promise and Pitfalls of Using Imprecise School Accountability Measures. *Journal of Economic Perspectives*. Volume 16, No.4. pp 91-114.

8 U. S Department of Education. Executive Summary NCLB. Retrieved September 30, 2015 from http://www2.ed.gov/nclb/overview/intro/execsumm.html

9 Mintrop, H. & Sunderman, G. L. (June 2009). The Predictable Failure of Federal Sanctions-Driven Accountability for School Improvement – And Why We May Retain It Anyway.*Educational Researcher*. vol. 38 no. 5, pp353-364

10 Rice, J.K. & Malen, B. (2010). *School reconstitution as an education reform strategy: A synopsis of the evidence*. Washington, DC: National Education Association.

Malen, B. & Rice, J.K. (2004). A framework for assessing the impact of education reforms on school capacity: Insights from studies of high-stakes accountability initiatives. *Educational Policy*, 18 (5), 631-660.

11 Rice, J.K. & Malen, B. (2010). *School reconstitution as an education reform strategy: A synopsis of the evidence*. Washington, DC: National Education Association.

Malen, B. & Rice, J.K. (2004). A framework for assessing the impact of education reforms on school capacity: Insights from studies of high-stakes accountability initiatives. *Educational Policy*, 18 (5), 631-660.

Trujillo, T. & Renée, M. (2012). *Democratic School Turnarounds: Pursuing Equity and Learning from Evidence*. Boulder, CO: National Education Policy Center. Retrieved October 13, 2015 from http://nepc.colorado.edu/publication/democratic-school-turnarounds.

12 Miron, G., Evergreen, S. & Urschel, J.L. (2008). *The impact of school choice reforms on student*

achievement. Boulder, CO: National Education Policy Center. Retrieved October 13, 2015 from http://epsl.asu.edu/epru/documents/EPSL-0803-262-EPRU.pdf

13 Kirschner, B. & Van Steenis, E. (2016, in press)."The Costs and Benefits of School Closure for Students". in Mathis, W. & Trujillo, T. (eds.) *Test-Based Education Reforms: Lessons from a Failed Agenda, Promises for Success*. Information Age Publishing. Charlotte, N.C.

14 American Statistical Association (April 8, 2014). ASA Statement on Using Value Added Models for Educational Assessment.

15 Rubin, D.B., Stuart, E. A., Zanutto, E. L. (Spring 2004). A Potential Outcomes View of Value-Added Assessment in Education. Journal of Educational and Behavioral Statistics Vol. 29, No. 1, pp. 103-116.

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. Retrieved October 17, 2015 from http://nepc.colorado.edu/publication/analysis-la-times-2011.

16 Welner, K. G. & Mathis, W. J. (February 2015). Reauthorization of the Elementary and Secondary Education Act: Time to Move Beyond Test-Focused Policies. NEPC Policy Memo. Retrieved October 13, 2015 from http://nepc.colorado.edu/publication/esea

17 See, for example:

Koretz, D. (2005). Using Multiple Measures to Address Perverse Incentives and Score Inflation. Educational Measurement: Issues and Practice. Volume 22, Issue 2, pages 18–26, June 2003

ASCD (June, 2013). Multiple Measures of Accountability. Policy Points.

18 Morton, B. A. & Dalton, B.(May 2007). Changes in Instructional Hours in Four Subjects by Public School Teachers of Grades 1 Through 4 (Issue Brief). Retrieved October 1, 2015 from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007305

Center on Education Policy (2008). Instructional Time in Elementary Schools: A Closer look at changes for specific subjects. A follow-up report to the 2007 CEP report, *Choices, changes, and challenges: Curriculum and instruction in the NCLB era*. *Georgetown University*.

19 Riddle, W. (May 8, 2012). Major Accountability Themes of Second-Round State Applications for NCLB Waivers. Center on Education Policy. Retrieved October 9, 2015 from http://www.cep-dc.org/displayDocument.cfm?DocumentID=404

20 Susan M. Brookhart (November 2009). The Many Meanings of "Multiple Measures" *Educational Leadership*. Retrieved September 30, 2025 from http://www.ascd.org/publications/educational-leadership/nov09/vol67/num03/The-Many-Meanings-of-%C2%A3Multiple-Measures%C2%A3.aspx

21 The U. S Education Department's "dashboard" can be found at http://dashboard.ed.gov/ There is a wide variety of commercial dashboard programs on the market.

22 Education Commission of the States (December 2013). School Accountability "Report Cards" What gets Measured? Retrieved September 30, 2015 from http://ecs.force.com/mbdata/mbquestRTL?Rep=AR03

23 Institute of Education Sciences (May 2008). Weighting Options for constructing composite domain outcomes. Retrieved October 2, 2015 from https://ies.ed.gov/ncee/pubs/20084018/app_c.asp

24 Darling-Hammond, L. & Hill, P. T. (June 7, 2015). Is there a third way for ESEA? *HuffPost Education*. Retrieved October 1, 2015 from http://www.huffingtonpost.com/linda-darlinghammond/is-there-a-third-way-for-_b_7013634.htm

25 Ryan, K.E., Gandha, T., & Ahn, J. (2013). *School Self-evaluation and Inspection for Improving U.S. Schools?* Boulder, CO: National Education Policy Center. Retrieved October 1, 2015 from http://nepc.colorado.edu/publication/school-self-evaluation.

26 Poon, J. D. & Carr, K. T. (January 2015).Evolving Coherent Systems of Accountability for Next Genration Learning: A Decision Framework. Council of Chief State School Officers. Retrieved October 1, 2015 from http://www.ccsso.org/Documents/Accountability%20Decision%20Tree-EXEC%20SUMM-Portrait-DigitalVersion(0).pdf

Rothstein, R., Jacobson, R., & Wilder, T. (2008). Grading Education: Getting Accountability Right. Washington D.C. and New York, NY. Economic Policy Institute. Teachers College Press.

27 Poon, J. D. & Carr, K. T. (January 2015).Evolving Coherent Systems of Accountability for next Genration Learning: A Decision Framework. Council of Chief State School Officers. Retrieved October 1, 2015 from http://www.ccsso.org/Documents/Accountability%20Decision%20Tree-EXEC%20SUMM-Portrait-DigitalVersion(0).pdf

28 Wilcox, B. (2000). Making School Inspection Visits more Effective: the English experience. International Institute of Education Planning. UNESCO. Paris

29 Wilcox, B. (2000). Making School Inspection Visits more Effective: the English experience. International Institute of Education Planning. UNESCO. Paris

30 Ryan, K.E., Gandha, T., & Ahn, J. (2013). *School Self-evaluation and Inspection for Improving U.S. Schools?* Boulder, CO: National Education Policy Center. Retrieved October 1, 2015 from http://nepc.colorado.edu/publication/school-self-evaluation.

31 Berliner, D. Our Impoverished View of Educational Reform. *Teachers College Record*. Volume 108,Numbe r6, June2006, pp.949–995.

32 Gebhardt, K. (2013). Model legislative language for comprehensive assessment and accountability. Boulder, CO: National Education Policy Center. Retrieved October 2, 2015 from http://nepc.colorado.edu/publication/data-driven-improvement-accountability/.

33 Ryan, K.E., Gandha, T., & Ahn, J. (2013). *School Self-evaluation and Inspection for Improving U.S. Schools?* Boulder, CO: National Education Policy Center. Retrieved October 1, 2015 from http://nepc.colorado.edu/publication/school-self-evaluation.

Rothstein, R., Jacobson, R., & Wilder, T. (2008). Grading Education: Getting Accountability Right. Washington D.C. and New York, NY Economic Policy Institute. Teachers College Press.

34 Howe, K.R. & Murray, K. (2015). Why School Report Cards Merit a Failing Grade. Boulder, CO: National Education Policy Center. Retrieved October 13, 2015 from http://nepc.colorado.edu/publication/why-school-report-cards-fail.

35 Howe, K.R. & Murray, K. (2015). Why School Report Cards Merit a Failing Grade. Boulder, CO: National Education Policy Center. Retrieved October 13, 2015 from http://nepc.colorado.edu/publication/why-school-report-cards-fail.

36 Ratner, G. & Neill, M (December 15, 2009) “Integrating ‘Helping Schools Improve’ With ‘Accountability’ Under ESEA: The Key Role For Qualitative, As Well As Quantitative, Evaluations And The Use Of “Inspectorates” - Working Paper II. Retrieved October 13, 2015 from http://www.fairtest.org/sites/default/files/SQR-Inspectorate_working_paper_2.pdf

Hussain, L. (Summer 2013). The School inspector Calls. *Education Next*. Retrieved October 13, 2015 from http://educationnext.org/the-school-inspector-calls/

---

*This is a section of **Research-Based Options for Education Policymaking**, a multipart brief that takes up a number of important policy issues and identifies policies supported by research. Each section focuses on a different issue, and its recommendations to policymakers are based on the latest scholarship. **Research-Based Options for Education Policymaking** is published by The National Education Policy Center, housed at the University Of Colorado Boulder, and is made possible in part by funding from the Great Lakes Center for Education Research and Practice.*

*The mission of the **National Education Policy Center** is to produce and disseminate high-quality, peer-reviewed research to inform education policy discussions. We are guided by the belief that the democratic governance of public education is strengthened when policies are based on sound evidence. For more information on NEPC, please visit http://nepc.colorado.edu/.*