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. Falling somewhat out of favor during the progressive era, testing for school accountability took on a heightened intensity beginning in the 1970s. 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. 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. Nonetheless, much of this latitude is in the finer details—the core of test-based accountability remains universal.

This material is provided free of cost to NEPC’s readers, who may make non-commercial use of the material as long as NEPC and its author(s) are credited as the source. For inquiries about commercial use, please contact NEPC at nepc@colorado.edu.
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,\(^10\) turnaround,\(^11\) 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. 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. 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.

Advocates of multiple measures often speak of a “dashboard” of decision data. 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. 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.

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. 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. 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. 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


   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.


17 See, for example:


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.


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/