This product is part of the RAND Corporation technical report series. Reports may include research findings on a specific topic that is limited in scope; present discussions of the methodology employed in research; provide literature reviews, survey instruments, modeling exercises, guidelines for practitioners and research professionals, and supporting documentation; or deliver preliminary findings. All RAND reports undergo rigorous peer review to ensure that they meet high standards for research quality and objectivity.

REPORT

Expanded Measures of School Performance

Heather L. Schwartz, Laura S. Hamilton,
Brian M. Stecher, Jennifer L. Steele

Prepared for the Sandler Foundation

RAND EDUCATION

This work was prepared for the Sandler Foundation. The research was conducted in RAND Education, a unit of the RAND Corporation.

Published 2011 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
RAND URL: http://www.rand.org
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Preface

Debate continues over the basis by which public schools are deemed to perform adequately under the federal accountability system, No Child Left Behind (NCLB). The question at the center of the debate is which aspects of schooling should inform those determinations—only the percentage of students taking and scoring proficient or higher on standardized math and reading exams, or a broader set of measures about other domains of schooling? If the latter, what are the right categories, and should they inform school accountability decisions? Further, should the federal government mandate, encourage, or leave it to states to decide whether schools should employ a broader set of measures?

The Sandler Foundation asked key federal policymakers involved in reframing the Elementary and Secondary Education Act (ESEA) (Pub. L. 89-10) what information from research they most needed to do their work. Congressional and administration officials and their staff indicated that they are uncertain about what is known regarding expanding measures of school performance. In an effort to address this concern, the Sandler Foundation asked the RAND Corporation to review the evidence regarding expanded measures of school performance beyond those currently required under NCLB and discuss how the federal government might best promote or support the use of such measures. This report documents the findings from our review. The findings and recommendations should be of interest to federal and state legislators and policymakers as they consider upcoming modifications to school accountability systems, such as through the reauthorization of ESEA.

The research sponsor, the Sandler Foundation, is a national foundation that works to improve quality of life. In the area of education, the foundation seeks to further policies that support high-quality learning environments that are equitable for all students.

Contents

Figure and Tables

Figure

Tables

Summary

The upcoming reauthorization of the ESEA, combined with other recent education policy trends, such as improvement to the quality of state data systems and a growing emphasis on data-driven decisionmaking, provides an opportunity to reconsider what factors school performance-reporting systems should include. Critics of NCLB have pointed to the narrowing effects of the law's focus on mathematics and reading achievement, and they have called for efforts to broaden the measures used to rate schools. In this report, we pose and address the following questions regarding expanded measures of school quality:

- What alternative measures of school performance do states currently use in their own accountability systems (in addition to the measures used for NCLB)?
- What are the emerging trends outside the school accountability context in the types of performance measures that districts and states employ to help principals and teachers improve schools?
- What guiding principles can research offer public education agencies about trade-offs to consider when adopting new measures, given limited evidence about whether various school performance measures ultimately lead to improved student outcomes?
- In what ways might the federal government encourage the development and expansion of alternative measures of school performance?

To answer these questions, we convened a panel of five experts on school accountability policies, scanned published research about expanded measures of school performance, conducted ten semistructured phone interviews with staff from local or state education agencies and research institutions, and reviewed the measures employed in each state that publishes its own school ratings in addition to those required under NCLB. After classifying the measures state education agencies (SEAs) use to develop their own school ratings, we then describe categories of measures that research indicates are the most rapidly growing in usage by SEAs and local education agencies (LEAs). We supplement our categories of measures with more detailed examples of localities that have adopted them, examining why they adopted the measures and how the measures are employed.

Rationale for Expanding School Measures

NCLB has focused public attention on student performance on statewide, standardized math and reading exams and, to a lesser extent, the other elements of states' accountability formulae, such as graduation rates. Yet public schools are expected to promote a variety of outcomes, of

which academic achievement as measured by standardized tests is only one. Additional goals of schooling include the preparation of students for life after school, which includes not only readiness for college or the workplace but also social and behavioral outcomes, such as displaying self-regulating behavior, taking personal responsibility, and demonstrating an ability to work in teams. Schools are also expected to promote civic-mindedness (e.g., political knowledge and participation, tolerance, propensity to vote or engage in civic life) and other positive outcomes, such as good physical health and the avoidance of drugs. The adoption of measures that pertain to these other areas of schooling could provide useful information to school-based staff and to the public about how well schools are meeting these collective goals. Further, an expanded set of measures could increase the validity of inferences about schools' effectiveness and offer relevant information to principals and teachers about how to improve their schools' performance.

Additional Measures Currently in Use

In response to NCLB, in 2002, states either established new school accountability systems, revised their existing ones to comply with federal requirements, or operated dual accountability systems that included their own measures as well as those required by federal law. We identified a total of 20 states that publish their own ratings of schools as of the 2008–2009 or 2009–2010 school year that were in addition to the federal annual accountability ratings. Among these 20 states, the most common categories of school performance that were included in state ratings and went beyond NCLB include the following:

- student performance in additional tested subjects (most often, history or social studies)
- measures of growth in student performance over time
- indexes to assign increasing weight to test scores along the entire spectrum of low to high performance instead of the NCLB focus on only proficiency or above
- college-readiness measures, such as American College Testing (ACT) scores or Advanced Placement course taking and test scores.

Although almost all 20 states also included information on their school report cards about school inputs, such as student demographics or school resources, and three states provided information about school processes, such as the quality of student life as reported on student surveys, in almost all cases, state accountability ratings were based exclusively on student outcomes, such as test scores, dropping out, or course taking.

In addition to considering the measures used by states in their own accountability ratings of school performance, we also identified three categories of measures that are rapidly becoming more common in state reporting:

- establishing a safe and supportive school environment
- identifying students who are at risk of failing
- improving student outcomes through more frequent assessments or advanced coursework.

Examples of measures within these categories include students' perceptions of their schools' climate and indicators to predict which students are at greatest risk of failing to com-

plete high school on time. A number of public education agencies are also expanding their measures of student outcomes beyond annual, summative math and reading scores to include additional measures of college readiness, such as advanced course taking, and scores from periodic assessments intended to provide timely information to school-based staff to allow for instructional adjustments during the school year.

Collectively, these measures indicate the additional aspects of school performance to which public education agencies most commonly attend. A number of the measures, such as periodic assessments, at-risk indicators, and student satisfaction, are designed as leading indicators of student achievement or graduation, which are currently the primary measures that determine a school's rating under NCLB. As such, they illustrate the profound influence the federal accountability system has had not only on the development of data systems that have enabled the creation of additional measures but also on the prioritization of certain aspects of schooling that align with NCLB outcomes.

What We Know from Research on Measures of School Performance

Although we identified considerable descriptive information about types of measures and their uses, we found, with a few notable exceptions, almost no published research about the technical quality of the measures,[1] the theories of action that instigated their adoption, the utility of the measures for promoting improved decisionmaking, or the effects of the measures on school practice or student outcomes. Admittedly, assessing their quality, utility, or effects is complicated because these measures are typically used in combination with other new and existing measures and because of other constraints on their use (e.g., the inability to identify or create an appropriate comparison group that is not included in the measurement system). As a result, there is no consensus yet regarding the overall quality of most measures or their utility for improving school performance. However, there is research on the effects of test-based accountability that provides a rationale for developing and adopting additional measures.

Research on test-based accountability systems reinforces the common-sense notion that what gets tested is what gets taught. In particular, high-stakes testing can lead to a narrowed curriculum and other potentially undesirable consequences (such as a focus on students at the threshold of proficiency, in the case of NCLB). But research on the effect of adopting additional measures to broaden ratings of school performance is quite limited, partly because many of the systems adopting such measures are in their early stages. The potential benefits of an expanded set of measures are that they could do the following:

- Allow for a more accurate assessment of the school characteristics widely valued.
- Promote more valid inferences about school performance by offering opportunities to compare performance on multiple overlapping dimensions.
- Provide a more balanced set of incentives to teachers and principals to improve performance in multiple areas.

[1] Exceptions include technical documentation on achievement tests and some surveys.

But there are also risks and trade-offs associated with the adoption of new measures. For example, the proliferation of measures could be a costly reform that could potentially dilute rather than focus attention on core aspects of schooling.

Ultimately, the selection of measures should be informed by the purposes of the measurement system—e.g., whether it will be used solely for monitoring, in a diagnostic or prescriptive way to guide school improvement decisions, or whether it will be included in an accountability system with explicit stakes attached to results. Aside from technical considerations about the construction of measures, the major decisions to make when adopting new measures of school performance include how narrowly the system should be focused, how to balance complexity versus transparency, how to create an affordable system that is still reasonably comprehensive, whether to allow flexibility in choice or use of measures across units, how much to emphasize formative and summative purposes, and whether to adjust for differences in school inputs.

Recommendations for a Federal Role to Promote Improved Measurement of School Performance

The federal government has traditionally played a limited role in shaping state and local education policy, but the NCLB experience provides an example of how the federal government can exert a powerful influence on state and local policy and practice through new accountability requirements. To prompt policymakers' thinking about actions the federal government might take to encourage the development of more comprehensive school measurement systems, we offer three recommendations:

- In the ESEA reauthorization, incorporate a broader range of measures as a basis for accountability decisions than is currently mandated under NCLB. Although there is currently insufficient evidence to make specific choices about which measures should be used, evidence from research on high-stakes testing indicates that educators tend to shift their focus away from what is not measured and toward what is. A federal mandate that states (or state consortia) select their own measures within a broader set of predefined categories might mitigate this risk and might allow stakeholders to draw more valid inferences regarding school performance that better reflect the multiple goals of schooling. We suggest the following five domains of expanded measures as places to start:
  - Expand the measures of achievement and attainment to account for both status and growth and to capture a broader range of academic outcomes in subjects besides math and English and language arts (ELA), as well as in advanced course taking.
  - Promote a positive school culture, including indicators, such as student and teacher satisfaction, academic challenge, engagement, safety, or orderliness.
  - Adopt leading indicators, such as measures of being on track for high school graduation, that provide schools information about students as they progress toward college and career readiness.
  - Promote positive behavioral, emotional, and physical health outcomes for students, including indicators of suspensions, expulsion, and physical health.
  - Augment unadjusted performance indicators with indicators that adjust for discrepancies in resources that children and, by extension, schools have available.

- Avoid creating an immediate new federal mandate to adopt specific measures. As states begin to validate additional measures, these can be gradually integrated into a refined federal system for measuring school performance. States should be required to conduct an evaluation of the technical quality and the effects of the inclusion of new measures within an ESEA accountability framework on student outcomes and school resource allocation. For that, they might require technical assistance or collaboration, which leads to our third recommendation.
- Incorporate the development and evaluation of additional school performance measures as an area of focus within existing competitively awarded federal grants. In light of the variance in state capacity to develop and test new measures and the desirability of developing measures that are consistent across states, offering federal grants for such development could create incentives for states to coordinate their efforts, as through interstate consortia.

The reauthorization of ESEA should be informed by lessons learned from NCLB and other efforts to promote school-level measurement and accountability. Although there are a number of limitations to the NCLB approach, the path toward improving federal reporting and accountability provisions is not always clear. This report describes promising directions for expanding the set of measures that schools have at their disposal while acknowledging the need for more research on the effects of new policies and for a careful consideration of trade-offs involved in designing a new system.

Acknowledgments

We wish to thank Susan Sandler for the support and insight she provided to this project, as well as the Sandler Foundation for its financial support of this endeavor. We also wish to thank Linda Darling-Hammond, Robert Linn, Joan Herman, Brian Gong, and Elaine Allensworth for serving as expert advisers and providing valuable input to guide the search for measures and research. We also thank the researchers and staff we interviewed who provided us with detailed information about state and local accountability systems and measures of school performance. We thank them for their time and patience in sharing documents and information. Cathy Stasz, Catherine Augustine, Vi-Nhuan Le, and Scott Marion provided reviews of earlier drafts, and the final report was substantially improved as a result of their input. We are also indebted to Kate Barker, who provided research support, and Robert Hickam, who formatted the document. Finally, we wish to thank the Broad Foundation for sharing information about its data-tool initiative for school district leaders.

Abbreviations

ACT	American College Testing
AP	Advanced Placement
AYP	Adequate Yearly Progress
CCSR	Consortium on Chicago School Research
CCSS	Common Core State Standards
CSI-USC	Charter School Indicators–University of Southern California
DIBELS	Dynamic Indicators of Basic Early Literacy Skills
ED	U.S. Department of Education
ELA	English and language arts
ERIC	Education Resources Information Center
ESEA	Elementary and Secondary Education Act
GED	General Educational Development Tests
IB	International Baccalaureate
KIPP	Knowledge Is Power Program
LEA	local education agency
NCLB	No Child Left Behind
Ofsted	Office for Standards in Education, Children's Services and Skills
SEA	state education agency

CHAPTER ONE

Introduction

A common criticism of the No Child Left Behind (NCLB) legislation is that it defines school quality using a set of measures that is too narrow. Critics assert that, in so doing, the federal accountability system overlooks important student outcomes and other factors that school leaders and citizens should consider in judging their schools' performance. The system's other reported shortcomings include encouraging teachers to distort instructional content to prioritize tested skills over nontested skills and to emphasize students' proficiency levels rather than improvement (Economic Policy Institute, 2008; Hargreaves and Shirley, 2008).

In view of these limitations, a careful exploration of measures of school performance is timely for a number of reasons. Primary among them is the upcoming reauthorization of the Elementary and Secondary Education Act (ESEA), of which NCLB is the latest iteration. Second, the majority of states have endorsed the Common Core State Standards (CCSS) in reading and mathematics, which create more uniform expectations for student performance in these subjects and could make many existing assessments of student performance obsolete. Third, the large federal investment in the Race to the Top assessment consortia has generated momentum to revise and expand existing student achievement assessments. Fourth, the increasingly widespread practice of gathering interim or benchmark assessment data throughout the school year provides an expanded set of information on student performance that could be utilized in a variety of ways. Finally, rapid advances in data systems that offer teachers and school leaders real-time information about individual students have facilitated the generation and application of new school performance measures.

The proliferation of student and school performance measures has also led to the development of new analytic methods that could facilitate the development of more useful estimates of school and teacher performance by better adjusting for differences in school and student inputs. For example, to create incentives for continuous improvement, most states with accountability systems that go beyond the NCLB requirements now factor growth in students' test scores over time into their school ratings.[1]

In this report, we pose and address the following questions regarding expanded measures of school quality:

- What alternative measures of school performance do states currently use in their own accountability systems (in addition to the measures used for NCLB)?

[1] This is also a feature that is becoming increasingly common within the NCLB context. As of 2009, 13 states reported the use of student-level test-score growth models as part of a pilot with the U.S. Department of Education (ED), while an additional 21 states reported considering it for accountability purposes (Altman et al., 2010). Only ten states indicated that they were not considering the use of growth models for either informational or accountability purposes.

- What are the emerging trends outside the school accountability context in the types of performance measures that districts and states employ to help principals and teachers improve schools?
- What guiding principles can research offer public education agencies about trade-offs to consider when adopting new measures, given limited evidence about whether various school performance measures ultimately lead to improved student outcomes?
- In what ways might the federal government encourage the development and expansion of alternative measures of school performance?

Our research proceeded in four steps to identify alternative measures in use by local or state education agencies (LEAs and SEAs, respectively) and research about those measures. First, in August 2010, we convened a conference call with five school accountability experts to solicit their guidance on our research questions and suggestions for localities to examine.[2] Second, we identified research about multiple or expanded measures of school performance using the online databases Education Resources Information Center (ERIC) and the ISI Web of Science. We expanded our literature search to include online research from organizations, such as the Baltimore Education Research Consortium, the Consortium on Chicago School Research, or foundations that have sponsored work on this topic. In this scan of published research, we identified localities that use alternative measures of school performance beyond those required by NCLB, and we sought evidence regarding the effects that adopting a specific measure or set of measures could have on school inputs, processes, or outcomes.[3] Third, we reviewed the accountability websites of each of the 20 states that publish their own ratings of their schools' performance separately from NCLB. From these 20 states, we identified and categorized the types of non-NCLB measures they employ. We consider these measures an indication of the broader set of school outcomes SEAs deem relevant for public information and for imposing consequences on schools. Finally, we conducted semistructured phone interviews with ten staff persons from localities identified as using expanded measures. Specifically, we identified these localities or persons using a snowball sample to identify localities reputed to use innovative measures of school performance. Our sample started with the recommendations we received in the August 2010 call with experts and proceeded with the recommendations obtained from each subsequent interviewee. In the end, we interviewed staff at three SEAs, four LEAs, and two research institutions to identify the reasons for their locality's use of expanded measures. In particular, we interviewed staff working for departments of education in the states of Ohio, South Carolina, and Rhode Island and in the districts of Atlanta, Georgia; Cincinnati, Ohio; Charlotte-Mecklenburg, North Carolina; and Prince Georges County, Maryland. We also interviewed staff working for the American Institutes for Research and the Baltimore Education Research Consortium.

In these interviews, we posed questions about types of alternative measures the systems employed, the timing of the measures' adoption, and the motivation for their use. For each

[2] These experts are Linda Darling-Hammond, Brian Gong, Robert Linn, Joan Herman, and Elaine Allensworth.

[3] Note that we did not catalog the much larger literature establishing correlations within secondary data between specific measures and outcomes of interest. Although this research is relevant, the evidence of relationships between one indicator and another (e.g., between student satisfaction and attendance) does not provide information about the effect the collection and reporting of those indicators have, when adopted by LEAs or SEAs, on improving school performance. Thus these studies are not included in this report.

locality, we also conducted web searches to obtain technical documentation about the measures (where available) and to confirm how they are reported to the public. We note the specific sources of particular data points in the chapters that follow.

A comprehensive review of measures developed and used by each school, each organization working with schools, or each of the approximately 13,350 U.S. school districts was beyond the scope of this study. In our scan of alternative measures of school performance, we focused on those currently in use by SEAs or LEAs to provide information for principals and teachers. Although a number of these measures can and do have additional purposes, such as informing parents about school conditions or influencing a superintendent's resource-allocation decisions, we prioritized in our search those measures that provide information on which principals and teachers could theoretically act (as opposed to indicators, such as state per-pupil funding allocations that are beyond the control of school-based staff). Based on our literature review, web searches, and interviews, we describe categories of measures that research indicates are the most rapidly growing in terms of SEA or LEA use.

In describing these categories of measures, we provide more detailed examples of localities that have adopted them, including information about why they adopted the measures and how the measures are used. Our examples draw primarily from U.S. districts and states, but we also include two from the United Kingdom and Australia because they are developed countries with school performance-measurement systems based on a broad set of measures. Although the discussion below does not comprehensively sample from all types of public education agencies,[4] it does include the most current innovative cases we could find.

Although we identified considerable descriptive information about types of measures and their uses, we found, with a few notable exceptions, almost no published research about the technical quality of the measures,[5] the theories of action that instigated their adoption, the utility of the measures for promoting improved decisionmaking, or the effects of the measures on school practice or student outcomes. Admittedly, assessing their quality, utility, or effects is complicated because these measures are typically used in combination with other new and existing measures and because of other constraints on their use (e.g., the inability to identify or create an appropriate comparison group that is not included in the measurement system). As a result, there is no consensus yet regarding the overall quality of most measures or their utility for improving school performance. In lieu of this ideal, we highlight where there is emerging evidence and focus on the trade-offs that need to be considered when thinking about expanding the set of measures used to assess school quality and support school improvement.

As a final note, this report attends to the question of expanding measures of school performance and not to the separate but important topics of evaluating individual teachers or school principals.[6] Nor do we cover measures available at only the district or the state level (e.g., some indicators in the Annie E. Casey Foundation KIDS COUNT database or the Schott Foundation's Opportunity to Learn index, such as the percentage of children in poverty or the number of teens not in high school). Given the vast literatures devoted to these topics, we also do not discuss the numerous technical issues related to developing measures, such as guid-

[4] For example, we do not systematically sample across rural, suburban, and urban districts, nor do we select for regional diversity.

[5] With the exception of technical documentation on achievement tests and some surveys.

[6] For a recent overview of technical considerations in measuring teachers' contributions to improving students' test scores, see McCaffrey, Sass, et al. (2009) and Baker, Barton, et al. (2010).

ance for building data systems,[7] developing and validating measures, approaches to assigning weights to various measures and combining them into a single index, or specific approaches on how to use data to promote school improvement.[8] All of these are important considerations that relate to the development and use of school measures, but a thorough discussion of them is beyond the scope of this study.

Evidence about the rationale for and availability of a variety of types of school measures and the trade-offs involved in their use could help policymakers think about additional options as they consider revising federal legislation. Chapter Two presents some of the rationales for expanding the set of measures of school performance. Chapter Three presents a summary of additional measures currently in use in selected states, districts, and other countries. Empirical evidence about the benefits of additional measures is summarized in Chapter Four, along with a discussion of the trade-offs inherent in their use. Finally, Chapter Five presents recommendations for expanding measures and the role the federal government could play.

---

[7] In addition to legal considerations about the merging and sharing of data about individual students and teachers, there are numerous technical considerations in developing a data-management system. See ED and the Chief Council of State School Officers' efforts to develop a comprehensive P–20 National Education Data Model (National Center for Education Statistics, undated).

[8] Guidance regarding the technical quality of measures used in accountability systems can be found in published standards for testing (Joint Committee on Standards for Educational and Psychological Testing, 1999) and for accountability systems (Baker, Linn, et al., 2002). Several sources provide guidance on effective data use; these include the Data Quality Campaign (undated) and a What Works Clearinghouse Practice Guide (Hamilton, Halverson, et al., 2009).

CHAPTER TWO

The Rationale for Expanding the Set of School Performance Measures

NCLB has focused the public's attention on measures of school performance that are based on student achievement tests. The law's reporting requirements have led to an unprecedented availability of information on student achievement in mathematics and language arts, student graduation rates, and teacher qualifications. At the same time, most users of this information recognize that it is inadequate for understanding what is happening within schools or the extent to which schools are promoting a variety of other outcomes that states and citizens value.[1] In this chapter, we present an expanded view of school performance, and we summarize the broad categories of measures that could be used to support this expanded view. We also discuss the purposes for which these measures might be used by school-based staff.

There are a variety of educational outcomes that society values and expects its public schools to promote. Student achievement is paramount in most policy debates concerning the goals of schooling, but there is disagreement about how much weight should be given to achievement in subjects other than math and reading. There is also disagreement about whether and how to incorporate achievement data from assessments, such as Advanced Placement (AP) exams, which are given to only a subset of students. An additional but related set of goals for schools relates to attainment—including preparing students for life after high school, such as postsecondary education, work, or service in the military. Key milestones toward these postsecondary goals include normal progression to the next grade level, enrollment in college-preparatory coursework, and graduation. Readiness for college or the workplace can also include social and behavioral outcomes, such as displaying self-regulating behavior, taking personal responsibility, and developing an ability to work in teams. Schools are also expected to promote civic-mindedness (e.g., political knowledge and participation, tolerance, propensity to vote or engage in civic life) and the adoption of positive personal and social behaviors, such as the promotion of health and the avoidance of drugs. Many citizens, policymakers, and educators have assigned high priorities to attaining these broader outcomes, including productivity and citizenship, because they are crucial to a well-functioning democracy. Therefore, including these outcomes in a broad school indicator system could provide valuable information about how well schools are meeting society's goals for them, and excluding them could both signal their unimportance and preclude investigation of their status.

Information about the extent to which schools are promoting important outcomes for children is essential, but the effectiveness of a school also depends on the quality of the environment it provides for its students and how well it functions as an organization. Therefore, it

[1] We take as evidence of the claim that many other outcomes are valued the fact that states have adopted academic standards and curriculum frameworks in these areas.

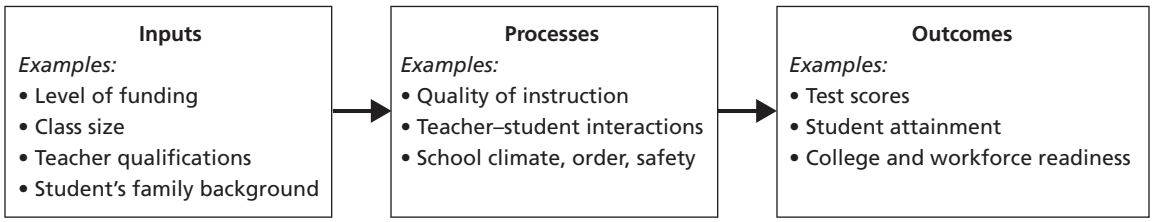
might be beneficial to supplement outcome measures with indicators relating to school inputs and processes. By inputs, we refer to the social and fiscal resources available to the school—resources, such as the level of funding, parental involvement, class size, course offerings, teacher qualifications, conditions of school facilities, and the health of students.[2] By processes, we mean the activities that occur during the school day and the environment in which learning takes place. Examples include the quality of instruction, teacher and student attendance, teacher–student interactions, school safety and order, and students' sense of belonging. Invariably, school inputs and school processes influence one another, and they collectively comprise the conditions for learning. Taken together, they can provide school-based staff with information that theoretically improves desired outcomes. Figure 2.1 provides a schematic to illustrate these concepts.

The school performance data that are widely available to support NCLB are lacking in a variety of ways: They include only a small subset of school outcomes, and they tell us virtually nothing about what is happening within schools (processes) or what resources are at the school's disposal (inputs). The emphasis in NCLB on outcomes is consistent with a view of standards-based accountability as a means to motivate improvement by attaching stakes to outcome measures, but it provides little information that can help educators determine how best to improve their practices. To support reasonably valid inferences about a school's organizational effectiveness and provide useful information for improving school performance, it is therefore necessary to collect additional measures. In particular, measuring these goals requires richer, more comprehensive information about what schools are doing and what outcomes they are producing, along with information to permit possible adjustments to reflect their differing levels of resources. These measures would not necessarily have stakes attached and, as we point out later, probably should not if their main purpose is to provide information to inform improvement efforts. Throughout this report, we employ the terms *measure*, *metric*, and *indicator* synonymously and refer to this broadened set of measures as a school indicator system.

Purposes of a School Indicator System

Under NCLB, student test scores and graduation rates are primarily used to identify schools for required interventions and, to a lesser extent, to help diagnose weaknesses as a basis for

Figure 2.1
Relationship Between School Inputs, Processes, and Short-Term Outcomes



RAND *TR968-2.1*

[2] Note that, to the degree that schools can affect student-level inputs, such as student health (not shown in Figure 2.1), they might also be considered outcomes of schooling.

school improvement planning. However, it is important to realize that measures can serve broader purposes than they do in the NCLB context. Before discussing specific measures that might be included in a school indicator system, it is important to consider the four main purposes it could serve. An understanding of purposes should guide the choice of measures and the structure of research needed to establish the validity of inferences drawn from the measures. These are as follows:

- Monitoring. A school indicator system can be used as a simple temperature-taking mechanism, a way for policymakers or the public to get periodic snapshots of how schools are doing in terms of outcomes and processes.
- Diagnosis and prescription. An indicator system can provide evidence to help district leaders or others identify areas in which schools might be struggling, as a way of diagnosing the problems that might be contributing to lack of success. Just as diagnostic student assessment systems produce subscores indicating areas of relative strength and weakness, an indicator system that is intended for diagnostic purposes would collect multiple measures of different aspects of schooling to help users pinpoint problems. Moving beyond diagnosis, the system can guide educators in deciding what steps should be taken to improve school performance. These steps might include the adoption of targeted improvement strategies, such as improved professional development or revisions to curricula. Prescriptive information can provide a basis for discussion among school or district staff who are responsible for implementing changes in schools and can also help parents identify areas on which they might want to focus their energies.
- Signaling. Simply by virtue of being tracked, measures included in an indicator system can encourage school or district staff to focus on particular practices or outcomes. For example, a system that incorporates measures related to high school graduation or course taking sends a message to educators that these outcomes are valued and that educators should pay attention to them; absence of such measures can diminish the salience of these outcomes. Whereas the monitoring and diagnostic or prescriptive purposes are served primarily through the scores that are produced on the measures, a signaling purpose can be achieved even in the absence of any provision of information about performance. It is the content of the measure, rather than the scores produced by the measure, that sends the signal.
- Accountability. Finally, an indicator system can create incentives for quality, either through explicit rewards and sanctions or through the public scrutiny that accompanies the publication of school-level performance information. Most of the policy debate around accountability in recent years has focused on outcomes, but stakes could also be attached to the provision of certain kinds of services or other process-related measures. For example, an indicator system might be designed explicitly to incentivize particular kinds of practices, such as the use of data from formative assessments to alter instruction and thereby reduce the risks associated with exclusively test-based systems (see Hamilton, Stecher, and Yuan, 2009, for a discussion of how indicator systems might improve accountability policies).

It is critical that developers and users of school indicator systems understand the purposes for which they are intended. An assessment that has been validated for a particular purpose (such as a student's placement into a mathematics course) is not necessarily valid for a different

purpose (such as evaluating teacher effectiveness) without additional evidence related to that use (Joint Committee on Standards for Educational and Psychological Testing, 1999; Kane, 2006). Similarly, an indicator system that proves to be appropriate for monitoring or diagnosis might not be appropriate for a different purpose, such as accountability. In particular, attaching consequences to information can lead to score corruption and undesirable narrowing of effort to focus on measured outcomes or processes at the expense of those that are not measured. Such risks and limitations are discussed in more detail in Chapter Four.

School indicator systems are often designed to serve multiple purposes, in which case it is essential to obtain validity evidence related to each of those purposes and to recognize trade-offs that are likely to be introduced when a given set of measures is put to diverse uses. Including a process-related measure of instructional practices in an indicator system could prove valuable for informing decisions about professional development, but, if that measure gets incorporated into a formal teacher evaluation system, teachers might respond in ways that raise their scores without fundamentally altering the nature of their practices. As this example illustrates, using a measure for a new purpose might diminish the validity and utility of the measure for its original purpose, so decisions about how to use measures should involve a careful analysis of the possible consequences and the development of strategies to mitigate undesirable consequences.

In addition, the design of an indicator system should be informed by an understanding of who will use the system for each of its purposes. A school indicator system that primarily measures school resources might not provide actionable information to teachers, who have little power to alter the level of resources their school receives, but it would provide information on which superintendents or parents could act. Potential users of school indicator systems could include policymakers, district- and school-level administrators, teachers, parents, students, and the business or higher education communities. Each of these groups has different needs and interests, and the design of the system needs to recognize these differences. It is unlikely that any one school indicator system could effectively serve the purposes of every audience. As stated earlier, in this report, we focus primarily on those measures that, at a minimum, school-based staff can use.

An overview of existing indicator systems reveals a range of purposes and intended users. In the next chapter, we briefly describe the data requirements of NCLB and present some examples of alternative measurement systems that states, districts, and other nations have adopted that go beyond the NCLB requirements. We did not find sufficient evidence to recommend particular measures that should be used for specific purposes, but attention to purposes should influence the final selection of measures in a school indicator system.

CHAPTER THREE

Which Additional Measures Are Currently in Use?

To identify the range of measures currently used in school indicator systems, we first noted which states operate their own school accountability systems that go beyond NCLB and then categorized the measures that these systems include. We consider the inclusion of these measures as an indication of what non-NCLB aspects of schooling states deem important to weigh when assessing school performance. Indicators currently required under the federal law include measures of highly qualified teachers, persistently dangerous schools, graduation rates, and scores from math, science, and reading on statewide standardized tests.[1] We find that the preponderance of measures in states' supplemental systems derive from summative statewide exams, as they do for NCLB. However, the states' non-NCLB measures typically focus on growth in performance over time (as compared with achieving threshold levels of proficiency in a given year) and include tests for an expanded set of academic subjects.

Looking beyond state accountability systems, we also explored through a literature search and a snowball sample of LEAs and SEAs other increasingly popular categories of school-level measures that are not necessarily used for school accountability purposes. We identified three broad categories of measures whose collection is expanding most rapidly: establishing a safe and supportive school environment, identifying students who are at risk of failing, and improving student achievement. Although several of these measures are based on data obtained from individuals (e.g., a parent's satisfaction rating, or the probability that an individual student will not complete high school on time), the ones we discuss are also aggregated to the school level to judge school performance, and are thus included here.

As the following examples illustrate, there are a plethora of measures in use by public education agencies for purposes of school monitoring, diagnosis, prescription, signaling, and accountability. Unfortunately, as noted earlier, with a few exceptions, there is scant research available about the technical quality of most measures other than standardized achievement test scores (and some student and teacher surveys), the utility of the school measures for promoting decisionmaking, or the effects of school measures on student outcomes or school practice (Faubert, 2009). One reason for this lack of evidence is that many of the measures reviewed in

[1] The two primary sets of measures in NCLB are the percentage of students within each school who (1) take the state accountability tests and (2) score proficient or above. The participation-rate measure is intended to prevent the exclusion of students from the accountability system, while the minimum proficiency rates provide clear benchmarks that schools must meet to avoid sanctions. Beginning in 2007–2008, states had to assess the math and reading proficiency of all third through eighth graders and at least one level for grades 10 to 12, and science proficiency for at least one level within three respective grade bands: 3 to 5, 6 to 9, and 10 to 12. States' Adequate Yearly Progress (AYP) definitions also include additional measures, such as graduation rates for high schools, but these tend to be less-heavily scrutinized. Finally, under NCLB, states must publicly report progress toward meeting the goal of all teachers being "highly qualified," along with school safety data.

this report have been developed in recent years, and research about their effects is not yet available. We therefore do not assess the quality of the individual measures discussed in this report. However, research suggests that indicators, such as test scores, that have been tracked for a long time, or class size, that do not require a lot of inference for interpretation tend to be of higher quality than measures of more complex constructs, such as pedagogy and school leadership, that often lack common definitions and are less-widely available (Mayer, Mullens, and Moore, 2000). Where research about individual measures is available, we note it in the discussion.

Common Measures Found in State Accountability Systems That Supplement NCLB

State Accountability Systems in Addition to NCLB

At the time that No Child Left Behind was first adopted in 2002, many states already operated their own accountability systems. In response to NCLB, states established new systems, revised their existing ones to comply with federal requirements, or operated dual accountability systems that included their own measures as well as those required by federal law. In this section, we focus on those states that opted to continue a second school rating system with elements that go beyond those required for NCLB. These systems include measures to which states attach their own rewards or sanctions, as well as measures provided solely for public information or monitoring.

State (and local) accountability ratings supplement rather than supplant federal accountability designations. The federal rules for identifying schools as "in need of improvement" take precedence, and state accountability ratings, where present, typically further specify the type and level of intervention required in those schools. In addition, in some states, the state ratings also determine which schools receive state-based rewards. Most often, the state school indicator systems are used to help prescribe responses to failings identified under NCLB.

The Education Counts database, maintained by the national trade newspaper, *Education Week*, reports that, in 2010, 24 of 50 states operated their own independent accountability systems in which the states assigned ratings to all of their schools based on state-developed criteria. We consulted the websites of each of the 24 states to obtain school report cards and technical documentation regarding the methodology of each rating system. We were able to confirm that 20 of them assigned and published school rankings on the basis of their state system as of 2008–2009 or 2009–2010, although not all rankings were attached to accountability requirements in addition to the federally required designations.[2] Table 3.1 sets out additional criteria that states considered in their accountability systems. Note that, if a measure applies to either a reward or a sanction (or both), we classify it as an accountable measure (A) in Table 3.1. If the indicator is included on the state school report card where the state rating is provided but does not factor into the rating, we classify it as an informational measure (I). Although it is likely that these measures are used for multiple purposes, those that count toward accountability ratings typically determine sanctions, act as signals to school staff of high-priority areas

[2] For example, Tennessee assigns A–F letter grades to its schools, but these are not attached to accountability consequences or rewards separate from NCLB.

It is possible that some states do have their own school accountability systems that diverge from NCLB requirements but are not so indicated in the Education Counts database. In such a case, the state is not included in Table 3.1.

Table 3.1
Twenty States' School Rating Systems

State	Additional Tested Subjects[a]	Outcomes Not Derived from State Tests[b]	Growth Scores Calculated at Individual Level	Growth Scores Calculated at Group Level	Scores Weighted for Entire Distribution[c]	Ratings Are Relative[d]	School Demographics	Learning Conditions[e]
Arizona		A	A	A				
Arkansas		I	A		A		I	I
California	A			A	A	I		
Colorado		A	A			A	I	I
Delaware	A			A	A		I	I
Florida		A	A					I
Indiana	A	I	A	A		A	I	
Louisiana	A	A		A			I	I
Massachusetts			A	A	A	A	I	I
Michigan	A		A[f]	A	A			A
Mississippi	A	A	A		A		I	
North Carolina	A	A	A			A	I	I
Ohio	A		A	A	A		I	
Oklahoma	I	A			A		I	I
Oregon		I	A	A	A		I	
South Carolina	A		A		A	A	I	I
Tennessee	I	I	I		I	I	I	I
Texas	A	A	A	A		A	I	
Utah		I	A				I	I
Virginia	A	I					I	

Table 3.1—Continued

State	Additional Tested Subjects[a]	Outcomes Not Derived from State Tests[b]	Growth Scores Calculated at Individual Level	Growth Scores Calculated at Group Level	Scores Weighted for Entire Distribution[c]	Ratings Are Relative[d]	School Demographics	Learning Conditions[e]

SOURCE: Accountability information from SEA websites.

NOTE: A = accountable; the measure is included in the state accountability framework for interventions, sanctions, or rewards. I = informational; the measure is reported on the school report card and used for monitoring, diagnostic, or prescriptive purposes but not in an accountability rating.

[a] Refers to subjects outside of math, reading, English and language arts (ELA), writing, or science. Examples include history and social studies.

[b] Examples include dropout rates and American College Testing (ACT) scores.

[c] An example would be a performance index.

[d] That is, compared with peer schools or students.

[e] An example would be course offerings.

[f] Growth scores were slated to begin in the 2010–2011 school year.

for attention, or trigger districts' and states' intervention to help "failing" schools. Measures that we term *informational*, on the other hand, are typically used for monitoring (e.g., to the end of redistributing resources) or for the diagnosis of reasons for schools' underperformance.

In this section, we discuss which measures shown in Table 3.1 are used for monitoring, diagnostic, or prescriptive purposes and which are used for accountability purposes. In many state systems, school ratings either trigger rewards (e.g., blue ribbon school status or financial rewards) or determine the level of state or district interventions or sanctions, such as the provision of outside experts for technical assistance, a review committee required to approve a school improvement plan, or decreased management authority at the school level. We turn to these informational measures first.

Measures Provided for Monitoring, Diagnosis, and Prescription. This section describes measures provided for monitoring, diagnosis, and prescription.

- State accountability systems often provide contextual information about school inputs or processes alongside performance ratings, but rarely does this information have consequences for schools. Most states shown in Table 3.1 provided information in school report cards about school conditions, such as student demographics, attendance, and mobility. We identified a wide array of measures that were used as indicators of context. Some examples include school per-pupil spending, course offerings, technology in the classroom, parental attendance at teacher conferences, community demographics, "prime instructional time," the availability of arts programming, the number of first graders who attended full-day kindergarten, grade inflation, on-time progression through school, school principal turnover rate, the presence of a written student code of conduct, a crisis-management plan, and availability of public school choice. In a unique case, North Carolina included prompts for schools to provide information about student health and food offered within the school. However, almost none of these measures were factored into states' ratings. The exception was in Michigan, where one-third of the basis for a school's letter grade was derived from a rating by school teams on a 40-item set of indicators that included both school process and performance measures.
- College-readiness measures were the second-most common indicator in state accountability systems (after test scores). Nine of the 14 states that provided college-readiness measures attached stakes to them; the other five used them strictly for informational purposes. As with school input and process indicators, there are a wide variety of measures related to college readiness, including participation rates and average scores on the SAT, ACT, and AP tests,[3] as well as participation in advanced coursework more generally (including dual-enrollment courses), and rates at which high school graduates enroll in remedial courses upon entering college.[4] Dropout rates were among the most common of these indicators (where dropping out signals lack of readiness for college and is separate from the federally required graduation-rate indicator). Several states weighted dropout

[3] Research examining states' efforts to increase AP enrollment and success found that, although incentives, such as AP exam-fee exemption, increased the likelihood of AP course enrollees taking the exam, the performance-based incentives for schools, such as including AP participation and pass rates in its rating system, were not associated with improved AP participation rates and performance (Jeong, 2009).

[4] Charlotte-Mecklenburg offers an interesting related measure in its accountability system: adjusted participation and pass rates in AP that are reported for only those students with scores sufficiently high on the PSAT to predict AP passage.

and graduation measures by students who are at risk to create an incentive for schools to keep those students in school. Some (rural) states also applied weights to the type of high school diploma received (e.g., technical or vocational, General Educational Development Tests [GED], or standard), as a means of giving schools an incentive to encourage their students to pursue rigorous coursework. Other states tracked students one or two years beyond high school to develop measures of "transition to adult life," such as the percentages of graduates who were enrolled in two- or four-year colleges, remedial classes within college, or vocational training; who were engaged in full-time work, part-time work, or military service; or who were unemployed. As one might expect, college-readiness measures generally applied only to high schools, with the consequence that secondary schools had to meet more criteria than elementary schools to obtain an acceptable state rating. We are not aware of states that included nontest measures beyond attendance in their elementary school ratings, although such metrics, such as students' on-time promotion rates, could be employed for monitoring or diagnostic purposes.

State Test-Related Measures for Accountability. This section describes test-related measures for accountability.

- Half of states' accountability systems included more tested subjects than the federal requirement to test math, reading, and science. In almost all cases, the additional subjects included in state accountability frameworks were among what is often termed the *core courses*—i.e., social studies or history in addition to math, reading, and science. In rare cases, additional accountable subjects include civics, economics, or geography.
- Most states that maintained their own accountability systems incorporated into their rating some consideration of growth in student test scores over time. As Table 3.1 indicates, there was variety in the types of growth scores that states adopted. Fifteen of 20 states calculated the growth in individual students' test scores from one year to the next. These student-level scores were then aggregated to the student subgroup and school level for school accountability ratings. Ten of 20 states calculated the average growth in groups of students' test scores from one year to the next by comparing, for example, fourth-grade proficiency rates in 2010 to fourth-grade proficiency rates in 2009 within the same school. Sometimes, the same states calculated both the individual- and group-level growth rates. The calculation of growth in the average test scores of groups of children is less costly, complex, and data intensive than the calculation of change to individual children's scores over time. However, these kinds of group-level growth scores do not necessarily reflect improvement in children's performance because they compare scores from two different cohorts of students and could thus be an artifact of differences in the characteristics of these students rather than a reflection of true changes in achievement. Despite the prevalence of growth ratings in state accountability systems, in most states, growth alone does not determine a school's score; rather, school ratings are typically jointly determined by proficiency rates taken from a single point in time that are compared to an absolute standard (per NCLB) and growth over time.
- Although most states established uniform expectations for rates of growth in test scores, a number constructed relative standards for growth rates by comparing a student or school only to "peer" students or schools. This feature takes into consideration that students are likely to experience different rates of growth in performance over time and that schools

serve student populations with different demographic characteristics. Comparisons of growth to "peer" schools or students (where peer is typically defined through an algorithm that weights academic performance, free-lunch status, racial/ethnic groups, or other student demographics) make this assumption explicit. Note that, although several states designed their growth-score measures to be relative (by comparing, for example, a focal student to 40 other demographically similar students in the state), in no case was a school's rating entirely based on a relative measure of performance. In other words, a school's rating always included some indicators that held it accountable to an absolute, statewide standard, such as a threshold level for a minimum percentage of students that must score proficient or higher.

- Many states with their own accountability systems weighted student performance along the entire spectrum of low to high performance rather than apply a single proficiency threshold. In 11 of 20 instances, states created a performance index score that assigned an increasing number of points to student scores by performance level. In most of these 11 states, points were awarded within four or five performance levels (e.g., below basic, basic, proficient, advanced) as an incentive for schools to move children up the entire test-score distribution. In other cases, states weighted individual scores based on their continuous scale scores.

Beyond Accountability Systems: Broadening School Indicator Systems

Summative test scores are the primary factor determining a school's rating within the 20 state accountability systems we reviewed. Yet the number of school performance measures in use outside the context of formal accountability systems is growing at a fast pace (see, for example, Broad Foundation, undated [a]; Sparks, 2010; Hartman et al., 2011). In this section, we discuss the most common or most–rapidly expanding categories of such measures. These include input, process, and outcome measures, and they represent a variety of approaches to expanding information systems and using data to improve decisionmaking and promote school improvement.

In developing these categories of measures, we relied on information obtained from a snowball sample of districts recommended by school accountability experts as having innovative measures (see discussion of methods in Chapter Two). From our own literature review and interviews, we identified approximately 130 individual metrics. In addition, we reviewed measures cataloged as part of a Broad Foundation initiative to provide performance-management tools for school districts. The foundation posts on its website a school performance-metric data bank (undated [a]) with 873 metrics that have each been developed in 2005 and beyond. These 14 localities include 12 LEAs, one charter school management organization (Green Dot), and one SEA (South Carolina). In addition, the foundation posts a list of 2,381 survey items (undated [b]) found on parent, teacher, or student surveys administered by 13 organizations. A number of these organizations are also represented in the metric data bank.[5]

The sheer number of metrics prevents a detailed listing of each one. But among the measures we collected and those that the Broad Foundation has gathered, we focus on those mea-

[5] The public school districts in the survey data bank include Oakland, Chicago, New York City, Charlotte-Mecklenburg, and Denver. The remaining seven organizations are charter management organizations, such as Knowledge Is Power Program (KIPP), Green Dot, Aspire, or charter schools.

sures of school performance on which teachers or principals could, in theory, act. We classified these measures into three broad categories: establishing a safe and supportive school environment, identifying students who are at risk of failing, and improving student outcomes.

The Broad Foundation data bank underscores the general trend we also noted in our purposive sample of districts: Most LEAs and SEAs are still in the process of developing, piloting, or refining additional measures of school performance. A number of studies have identified correlations between the leading indicators, such as being at risk of dropping out, and outcomes of interest, such as on-time graduation.[6] Due to the relatively recent adoption of most of these measures, we are not aware of published studies that document the effect these new measures have on teacher and principal practice or student outcomes.[7] However, there are several ongoing studies of such measures, and more information about their effects should be forthcoming (Osher, 2010; Connolly et al., 2010).

Establishing Safe and Supportive School Environments

Beyond the federal requirements in this area, a number of localities in the Broad Foundation data bank and LEAs or SEAs we interviewed have developed measures of school inputs and processes that are alternatively referred to as students' "opportunity to learn," their "conditions for learning," the "school climate," or the "school environment." Although there is no consensus about how to define these terms, alternative definitions draw on similar constructs, such as a safe school climate, high academic expectations, and a supportive environment from teachers and peers.[8] According to surveys of parents, taxpayers, and educators conducted prior to the passage of NCLB, school safety and teacher qualifications were two elements most commonly desired on school report cards, ranking above test scores (Brown, 1999). However, these aspects of schooling are not uniformly measured and reported at the school level. Further, in a 2006 scan of the 50 states, the National School Climate Center noted that few states incorporated their climate-related measures into a general accountability system (McGabe and Cohen, 2006). This comports with our finding that, among the 20 state accountability systems

[6] As stated previously, a review of the literature establishing associations between specific process measures and outcomes is beyond the scope of this report. For examples, see Allensworth and Easton (2005, 2007) on Chicago's "on-track" measure of ninth graders, or the strong association between chronic absenteeism and student performance (Chang and Romero, 2008). Establishing correlations is a necessary first step in selecting measures to help improve school performance, but correlations between measures and outcomes do not ensure that the adoption of a measure will, in fact, alter staff or student behavior or improve outcomes.

[7] A notable exception is the preliminary findings from the ongoing Measures of Effective Teaching project funded by the Bill and Melinda Gates Foundation showing that student perceptions of their teachers (particularly their perceptions of their teachers' ability to control the classroom and to provide challenging material, as measured by questions on Tripod student surveys) are related to the gains in academic achievement of that teacher's students in other classrooms (Bill and Melinda Gates Foundation, 2010). See Rothstein (2011) for a review that questions the strength of the reported relationships.

[8] For example, Osher and colleagues (2008) identify four primary factors that, according to research, can help establish necessary "conditions for learning": (1) a school climate in which students feel physically and emotionally safe; (2) a supportive, engaging community with challenging academic expectations; (3) students who feel that their teachers support them; and (4) social and emotional learning about how to empathize with others, establish positive relationships, recognize and manage emotions, such as anger, and handle challenging situations effectively. Alternatively, the National School Climate Center says that the following dimensions contribute to school climate (McGabe and Cohen, 2006): (a) cleanliness, adequate space, inviting aesthetic quality of school; (b) school course offerings and size of school; (c) socioemotional and physical safety; (d) high expectations for students, individualization of instruction; (e) positive, connected relationships between students and teachers; (f) a sense that there is a school community; (g) high morale among teachers and students; (h) peer norms that learning is important; and (i) home–school–community partnerships with ongoing communication.

reviewed above, none reported social and emotional learning or student–teacher interactions measures for either informational or accountability purposes. This could change in upcoming years, however, because, in October 2010, ED awarded $39 million to 11 states to develop measures of safe and supportive schools.[9]

Where measures of school climate and opportunities to learn are in place, data are typically gathered from surveys, school inspections, or existing administrative records. As we discuss later, our interviews of districts and states suggest that these measures are often collected for the diagnostic purpose of identifying the sources of academic success or failure in schools. We describe each of the three data sources in this section. In describing surveys, we focus on several examples from the United States and note that the use of school-climate surveys has grown among states and districts (Pinkus, 2009; Ho, 2008). Because school inspections are relatively rare in the United States, we also highlight two examples from England and Australia, where attention to school processes plays a much greater role in accountability systems. Finally, we list metrics using administrative data already provided by the 20 states with school rating systems.

School Climate and Input Measures from School Surveys. Numerous cities, states, and charter management organizations, such as those in Washington state, Rhode Island, North Carolina, Anchorage, Cincinnati, Cleveland, New York City, Chicago, Charlotte-Mecklenburg, Denver, Oakland, KIPP, Green Dot, and Aspire, gather survey data from students, teachers, or principals related to school climate and opportunities for learning (Broad Foundation, undated [b]; Osher, 2010; Cobitz, 2010; New York City Department of Education, undated; Votta, 2010). This is by no means a comprehensive list. With the low costs of online survey administration, it is likely that the number of individual schools and districts administering surveys will expand.

With a few exceptions that we discuss in this section, climate and opportunity measures are typically not included in a formal school accountability system but rather are used for diagnostic purposes.[10] This is their primary use in Rhode Island, for example, which is the only state of which we are aware that imposes a near mandate that all its public school teachers and students must complete surveys (Votta, 2010).[11] The state recently replaced its former SALT surveys, which were administered since 1998, with Surveyworks! surveys that are currently administered to all students in grades 4–12. The state expects to administer the surveys in spring 2011 to all K–12 teachers and principals and to all parents of students in grades K–12 in the state. The purpose of the surveys is to provide diagnostic information to superintendents and to school staff to understand the "why" of their academic performance indicators. For example, if a school is having truancy problems, which depresses academic performance, the surveys are intended to help administrators uncover the reasons for and the means by which to address them. They will also serve as the survey component for school accreditation and as a planning tool for the accreditation visits, which occur once every ten years for each public school.

---

[9] Note that, although this discussion focuses on the use of input and process measures to assess school performance, another purpose might also be their use by district, state, or even federal leaders to allocate resources to schools.

[10] See Table 3.1 for a review of the types of measures used in states' school accountability ratings.

[11] Parents and students can opt out, but the state does not tell teachers or principals that the survey is voluntary.

The Rhode Island surveys are a direct product of the state legislature's emphasis on assessing school climate in Article 31, which was first enacted in 1997. Based on its experience over the past decade, the state, in its redesign of the school surveys last year, moved to an all online administration (with a 95-percent participation rate for the student survey this year) and is also abbreviating both the teacher and parent surveys to increase participation. The Rhode Island Department of Education finds that, in general, NCLB accountability crowds out school attention to the survey results, which have low or no stakes attached. However, they find that their larger, urban districts in particular pay close attention to survey results to help inform policy changes to reduce test failure rates and to understand whether and why students do not feel safe or supported at school.

Chicago has among the oldest and most comprehensive practice of administering school surveys (Consortium on Chicago School Research, undated). Since 1991, the Consortium on Chicago School Research (CCSR) has surveyed principals, teachers, and students in the district. As of 2006–2007, Chicago Public Schools has expanded the student survey and created a separate parent survey to produce indicators of school climate and parent involvement for Chicago Public Schools' scorecards. These indicators do not influence the school district's rating of its schools (which is based primarily on test scores), but they are provided to parents and students as information on which to base school choice decisions. The surveys solicit information about more aspects of school climate than are shown in school report cards, but those reported alongside school accountability ratings are parental overall satisfaction and satisfaction with opportunity for involvement. Likewise, student survey data are used to create four rating scales for the report cards: safe and respectful climate, social and emotional learning, academic rigor, and student support.

New York City is the only locality we examined that factors survey responses into its own school accountability rating (New York City Department of Education, undated). As of 2007, New York City began to release annual school progress reports that assign a letter grade of A to F to each public school. Parent, teacher, and student responses to surveys about the school environment account for 10 percent of the 100-point scale. According to the 2010 survey results, about half of eligible parents completed the survey, while 82 percent of students and 76 percent of teachers did so. Survey items pertain to academic expectations, communication, engagement, safety, and respect.

These topic areas generally align with those found in the 13 organizations (which include five public school districts and eight charter school management organizations) compiled by the Broad Foundation (undated [b]). Among these 13 organizations, additional categories on student surveys include satisfaction with course offerings, extracurricular activities, school facilities, technology, food, and school discipline rules. Teacher surveys address some of the same categories but also include questions about the principal, the quality of the professional community, professional self-development, and resources at their disposal. Finally, parent surveys focus on communication with teachers and their children about school, academic expectations, school safety, and satisfaction with teachers.

Surveys solicit important information about perceptions, which can shed light on sources of satisfaction and dissatisfaction and areas in which school processes could be improved. Direct observations of classroom or school activity by outsiders can, on the other hand, offer independent information about other crucial school processes, such as the quality of instruction, curriculum, or interactions.

School Inspections. England and Australia offer examples of school indicator systems that place a much greater focus on direct observation of school processes by experts external to the school. In England, for example, the curriculum, standardized tests, assessment procedures for teachers and school leaders, annual school report cards, and school inspection reports are all nationally administered and regulated (Huber, Moorman, and Pont, 2008). The Office for Standards in Education, Children's Services and Skills (Ofsted) conducts school inspections at least once every three years and develops a composite score by which it rates schools on a four-point scale.[12] As with student achievement in schools, all school inspection reports are posted annually for public access. The school head (principal) is responsible for addressing the remedies or recommendations proposed in the inspection report within a set timetable. The report's score carries sanctions such that schools receiving an unsatisfactory "special measures" rating cannot hire new teachers without Ofsted's prior permission. Further, because (similar to in the United States) schools receive funding based on the number of pupils they enroll, negative ratings can produce enrollment declines and consequent reductions in school resources (Huber, Moorman, and Pont, 2008).

School inspections are elaborate and relatively lengthy (lasting up to five days) and consequently require considerable preparation on the part of schools, including preparation of a written school-wide self-evaluation (Huber, Moorman, and Pont, 2008). During the inspection, observers examine the school's self-evaluation, as well as its student performance data (including standardized test scores) and examples of student work. Observers also conduct classroom observations, analyze parent survey data, and interview pupils and staff. Among other things, the inspection reports evaluate schools along 11 dimensions, including student behavior; student safety; student enjoyment of learning; development of workforce and other skills; spiritual, moral, and cultural development; effectiveness of care, guidance, and support; and effectiveness of leadership in driving improvement.

A second interesting example of school inspectorate systems comes from the state of Victoria in southeastern Australia, which is home to the city of Melbourne and is the most populated state in the country. Approximately two-thirds of the state's 850,000 students attend public schools (Matthews, Moorman, and Nusche, 2008). Australian public schools, like those in the UK, have national standards and assessments, but states have considerable autonomy over other aspects of schooling. As set out in a blueprint first published by the department in 2003, Victoria's Ministry of Education and Training has focused its school-improvement efforts on capacity building through development of teacher knowledge, professional leadership, and establishing a shared organizational vision that focuses on high expectations, purposeful teaching, and creating safe and stimulating learning environments for students (Victoria Department of Education and Training, 2006).

Each year, Victoria's government issues school performance summaries for parents and school staff that group data about schools into three categories:

- student learning, including national examination scores
- student engagement and well-being

[12] Quantitative and qualitative research about the effects of inspectorates in the United Kingdom and the Netherlands on school performance yields inconclusive results. In their overview of the research, Dedering and Müller (2010) report that studies generally find a relationship between inspections and changes to school processes, such as management and instruction, but little to no evidence of effects of inspections on student achievement.

- student pathways and transitions.

Each measure is compared with those from all other schools in Victoria, although the performance summaries are not tied to specific sanctions. As of 2009, in its annual reports, student performance is both reported in absolute terms and adjusted to schools with similar students.

Within the three categories of student learning, engagement, and student transitions, the Victoria Ministry of Education and Early Childhood Development inspects all of its schools at least once every four years (Victoria Department of Education and Early Childhood Development, 2009). Depending on a school's performance level, the ministry applies one of four successively detailed reviews. In the lowest-performing schools, the ministry conducts "extended diagnostic review" that can occur more than once every four years and entails four days of visitation that involves a previsit, a panel meeting with the principal and school council, and a report to the staff and school council. Successively higher-performing schools receive a two-day diagnostic review, a one-day continuous improvement review, and (for highest-performing schools) a "flexible and focused" review on an area the school self-identifies as one in which improvement is needed. Following its review, the ministry prepares a report with recommendations that the school is expected to share with parents and students. Its recommendations, however, are nonbinding, and no explicit sanctions are tied to the reviews.

Relatively few districts or states in the United States engage in this level of extended inspections for all of their schools. Charlotte-Mecklenburg is one example among dozens of public education agencies in the United States that have recently engaged the services of Cambridge Education to oversee inspections somewhat similar to those just described in the UK and Victoria (Charlotte-Mecklenburg Schools, undated). As reported in interviews with the district, schools in Charlotte can volunteer to participate in the school quality reviews that are led by an external team (one Cambridge Education staff person, several Charlotte district personnel) for a two-day visit. Eventually, all schools will be expected to receive a school quality review. Prior to the visit, schools are to fill out a self-evaluation form, and the visit itself yields a report that focuses on six criteria: student achievement, teaching and learning, curriculum, leadership and management, learning environment, and involvement of parents and the community. The report is not a part of the district's accountability system but is rather intended as a diagnostic tool for both the school staff and district staff in distribution of resources to schools.

School inspections can gather information about the instructional core of schools, but they are not incorporated into the SEA or LEA performance-management systems we examined. Instead, the most common form of measures we found in states' supplemental accountability systems or district performance-management systems that related to school inputs and processes came from existing administrative data. In most cases that we identified, this information was gathered and publicly reported because either state legislatures or school boards required it (Votta, 2010; Cobitz, 2010; Busbee, 2010). We turn to this category of measures next.

School Climate and Input Measures from Administrative Data. Currently, the most common factors that states report on school report cards that pertain to conditions for learning relate to school inputs.[13] These include indicators, such as course offerings, safety, school facilities, and fiscal resources. Though such indicators are rarely included in the 20 states' school

[13] See Table 3.1 for detail and data sources.

accountability ratings, we found a wide variety of input measures on state or district school report cards. They included the categories and data points illustrated in Table 3.2.

These measures, all of which are derived from administrative data, focus on aggregated school-level information and almost exclusively pertain to school inputs rather than processes or outcomes. We are not aware of major innovations to this class of measures. They also tend to be the least actionable category of measures by teachers and principals because these inputs are usually (but not always) determined by factors outside the school, such as residential sorting patterns and school district decisions regarding resource allocation.

The next subsection discusses a category of measures that also utilizes administrative data but employs the data prescriptively so that schools can intervene to assist struggling students.

Table 3.2
School Climate and Input Measures from Administrative Data

Category	Example of Measure
Student readiness	Percentage of entering students who are ready for school according to DIBELS Percentage of first graders who attended full-day kindergarten Rate at which students are retained in grade
Technology	Number of computers per classroom or number of students per computer Average age of media center Percentage of classrooms connected to the Internet
Adequate resources	District-level (or, in one case, school-level) expenditures Average teacher salary within the school Property valuation per student Percentage of total school funds spent on instruction
Staff resources	Number of staff, by type (e.g., counselors, social workers, teachers) Average age range of teachers as an indicator of experience
Courses	Average class size Course offerings or curriculum highlights Opportunities in the arts Access to college-preparatory curricula Amount of prime instructional time Presence of a character-development program
Student qualifications	Students who are older than usual for their grade level Percentage of students still present in the second semester of school Eligibility for gifted and talented programming Percentage of students receiving reading remediation
Parental involvement and investments	Parental participation rates and number of parent volunteer hours Percentage of students enrolled in extracurricular activities Percentage of parent calls returned within three days Number of grievance calls made to the district Presence of written statements regarding crisis management, parent involvement, or student code of conduct
Safety	Ratio of juvenile offenders Out-of-school suspensions and expulsions Acts of crime per 100 students
Inputs relative to outputs	ROI, meaning test-score growth adjusted to account for the level of spending at a school or for instructional spending

SOURCE: Authors' categorizations from accountability information on SEA websites. See Table 3.1 for list of SEAs.

NOTE: DIBELS = Dynamic Indicators of Basic Early Literacy Skills. ROI = return on investment.

Identification of Students at Risk

There is considerable interest and activity devoted to developing leading indicators of which children are at greatest risk of either dropping out from school or failing to graduate on time (see, for example, Hartman et al., 2011; Massachusetts Department of Elementary and Secondary Education, 2010; Allensworth and Easton, 2005). This class of measures has gained urgency because of recent changes to NCLB requiring schools to measure graduation rates, which is a high-stakes indicator for AYP, using a uniform method. It is both in this area and in assessment that we discern the most rapid developments in school measures taking place. Many of the at-risk measures are intended for the diagnostic purpose of identifying at an early stage which students most need extra services. However, some of these measures have been aggregated to the school level and even included in school accountability ratings, as is discussed below.

We identified four innovations occurring in this area: (1) developing individual-level predictions and thus creating ways to provide school leaders with lists of individual students for interventions; (2) more frequently refreshing these predictions, which requires an integrated database that synchronizes daily attendance data with student transcript data; (3) testing combinations of factors to collectively predict a child's likelihood to stay in school and progress through grades on time; and (4) tailoring these predictions to apply even to children in the early grades.

Fueled by the increased availability of student-level longitudinal data, a spate of research about leading indicators, such as credit accumulation, grade-promotion rates, attendance, mobility, enrollment in college-preparatory coursework, and course failures, have established correlations between the measure(s) of interest and outcomes, such as on-time graduation (Pinkus, 2009). For example, based on a combination of student discipline events, student mobility, and unexcused absences, Ohio provides each school annually a school-level "risk factor index" (Cohen, 2010). Some of the promising measures we identified in our literature reviews and interviews include the following:

- On track to complete high school. CCSR has developed what has since become a relatively widely used indicator of whether a ninth grader is "on track" to complete high school.[14] Based on its analysis of student patterns in course taking and performance, CCSR has determined that first-time freshman students are considered on track at the end of their freshman year if they have accumulated at least five course credits and failed no more than one semester of a course in a core subject (English, math, social science, or science). This definition has since been incorporated into the Chicago Public Schools' performance policy for high schools. In addition, the district provides weekly student-by-student reports to schools that identify students at risk of being off-track by the end of the year to use in monitoring and targeting intervention.
- Chronic absenteeism. Given that chronic absenteeism (usually defined as missing 20 or more days of school in a single school year) is highly predictive of subsequent academic failure, Baltimore has been aggressive in documenting and attempting to reduce absences. A recent report (Mac Iver, 2010) found that 42 percent of Baltimore high school students missed at least one month of school in 2008–2009. Increasing rates of absenteeism were also predictive of leaving school altogether. In response, the district has encouraged

[14] See Allensworth and Easton (2005, 2007) for evidence of its predictive power of on-time graduation.

schools to develop policies to increase attendance, such as assigning an attendance monitor or introducing incentives for students to attend school.

- The 10 percent of students most at risk. As of 2010–2011, Charlotte-Mecklenburg has developed an at-risk measure that identifies the 10 percent of students within each grade level (from 1 through 12) with the highest at-risk index score (Cobitz, 2010). When principals log into their data portal, any students on this list who are enrolled in their particular schools are identified on the front page. Likewise, teachers can see which students in their classrooms (if any) are among this group. Depending on the grade level of the student, the score is based on a weighted combination of the following factors: the percentage of days absent since the beginning of the school year; the number of siblings in the district and who have dropped out; and whether the child is over-age for his or her grade, is limited English proficient, has special needs within certain categories, has low test scores, or has a lack of extracurricular activities. Although a number of these metrics are not mutable, the purpose of the index is to direct the attention of principals, social workers, counselors, and teachers to students at an early stage who might not otherwise be identified as high risk based on any one single characteristic.
- Early warning indicator index. Massachusetts assigns each incoming ninth grader to one of five levels of risk based on their eighth-grade attendance, math, and ELA score (Massachusetts Department of Elementary and Secondary Education, 2010). These were selected from among 11 variables initially considered or used in prior years for the index (factors, such as student mobility, limited English proficiency, and special-education status). The state provides student rosters along with their respective index values to districts for them to use or refine as desired. The most recent report identified a little more than one-third of incoming freshmen in Massachusetts as being at risk of dropping out (Vaznis, 2010).

Improving Academic Performance

College and Career Readiness. As described in the state accountability section, outside of state standardized test scores, college-readiness measures are the most common indicators included in state school accountability systems. The Broad Foundation data bank (undated [a]) corroborates this finding: Twelve localities have a total of 204 metrics pertaining to college and career readiness—one-quarter of the total number of metrics in the data bank. Not all of these 204 metrics are unique; some are quite similar to one another across the 12 localities. The range of measures pertain to participation rates, average scores, and the proportion of students obtaining passing scores on college-readiness tests, such as the SAT, International Baccalaureate (IB), AP, or ACT exams. Other indicators examine dropout rates and predictors of dropping out, such as credit accumulation or withdrawal from required courses. A number of the 12 localities track college entry through the percentage of seniors who have applied for financial aid or had transcripts sent to college, the percentage of graduates enrolled in college or in remedial courses, or the percentage of students who meet a state-developed college-readiness index. Our review suggests that there is greater consistency among districts' and states' measures related to college or workforce readiness than among measures in other categories, such as student or parent satisfaction, school inputs, or student growth on standardized tests. Namely, the majority of the measures in this class focus on participation in challenging coursework (e.g., AP courses, dual-enrollment courses, academically oriented diplomas) and attainment on college-oriented tests, such as the ACT or the AP. Only one of the 20 states

with their own accountability rating systems (Oklahoma) has measures that explicitly relate to workforce readiness in the sense of tracking employment in the years following high school.[15]

Periodic Assessments to Provide Information for Instructional Improvement. Aside from identification of students at risk, the other area of measures in which interviewees or relevant literature indicated the greatest growth is in district or state adoption of interim or formative assessments. Periodic assessments are standardized tests that are designed to be administered regularly during the year to provide more frequent information on student learning. Research suggests that the use of periodic assessments is increasing, and recent federal efforts could make them even more popular (Hamilton, Stecher, Marsh, et al., 2007). ED has awarded two state consortia $160 million to $170 million each to develop assessments that are aligned with the CCSS, and both consortia include periodic assessments in their plans. States are scheduled to first implement new assessments aligned to the CCSS in 2014–2015, and these assessments will likely substantially alter the way in which students are tested and how inferences are drawn from them.

Namely, the two consortia propose to increase the number of assessments administered throughout the school year, a goal that aligns with an ongoing shift in practice among many districts and some states. Cincinnati is one such example (Holtzapple, 2010). For the past seven years, the district has required its teachers to administer quarterly benchmark exams. However, as of 2010–2011, the district has revised this requirement to instead mandate that 14 short-cycle assessments are to be administered throughout the year by all teachers in grades 3–8 of reading, math, science, and social studies. Teachers can decide at what times to administer them, but the assessments are timed to occur roughly once every two weeks. The 14 exams range from five to eight test items each and have both multiple-choice and constructed-response items. The assessments are intended to influence teachers' instructional practice because data from the assessments are available within 24 hours via on online portal for teachers and principals. In addition to the short-cycle assessments, the district also required diagnostic tests in August and in January in the four core subjects. Further, the dashboard indicates the percentage of students within each school who have taken each short assessment so that the district can ensure their widespread use.

[15] Note that, although West Virginia and Kentucky do not issue school ratings separately from federally required NCLB ratings, they also have measures that relate to students in the years beyond high school.

CHAPTER FOUR

What Guidance Does Research Offer for Expanding Measures of School Performance?

As the preceding chapter indicates, numerous states and districts are working to develop and implement school indicator systems for informational and accountability purposes. Because many of these systems are in their early stages, limited evidence exists about their effectiveness as levers for improving school quality or student outcomes. However, research on the use of tests in accountability systems provides some guidance to help us predict the likely consequences of the use of supplemental measures in these systems and to identify possible risks associated with their use. We briefly summarize key findings from this research, and then we describe some of the possible benefits, challenges, and trade-offs associated with school indicator systems.

Lessons Learned from Research on Test-Based Accountability

A growing body of evidence from research on test-based accountability reinforces the commonsense notion that what gets tested is what gets taught. Specifically, high-stakes testing tends to lead teachers to focus more on tested subjects and on tested content within a subject, and less on material that is not tested (for reviews, see Koretz, 2008; Hamilton, 2003; Stecher, 2002). School and district administrators often reinforce this tendency by aligning curricula, pacing guides, professional development, and other resources with the high-stakes tests (Stecher et al., 2008; Hamilton, Stecher, Russell, et al., 2008; Center on Education Policy, 2006).

Although some alignment of curriculum with a state's content priorities might be appropriate, focusing curriculum and instruction narrowly on mastering a specific test rather than on teaching the underlying content (of which the test is merely a small sample) can contribute to score inflation. Score inflation occurs when students' performance gains on a particular test outpace their actual knowledge gains in the underlying construct(s) the test was designed to measure (Koretz, 2008). The NCLB experience is illustrative: Researchers find that, although scores on state accountability tests have increased substantially (Chudowsky and Chudowsky, 2010), scores on low-stakes tests, such as the National Assessment of Educational Progress, have at best shown positive but modest improvement (Chudowsky and Chudowsky, 2010; Reback, Rockoff, and Schwartz, 2011; Wong, Cook, and Steiner, 2009).

Moreover, the metric matters: Systems that reward teachers or schools for moving students from below a cut score to above it, as NCLB does, have been associated with efforts to target instruction to students who are performing just below the proficiency cut score (Booher-Jennings, 2005; Hamilton, Stecher, Marsh, et al., 2007). Although such responses might seem efficient from the point of view of a school wishing to boost its proficiency rate, they are

undesirable because they allow an arbitrary proficiency cutoff to determine students' access to instructional resources. Such strategies could draw resources away from students at lower and higher points in the performance distribution, whose instructional needs might equal or exceed those of students scoring near the proficiency cut score. In addition, reliance on specific metrics can distort conclusions about performance and thereby undermine efforts to use the data for decisionmaking. Presenting scores exclusively in terms of percentages of students above proficient, for instance, not only masks important performance differences at other points in the score distribution but can lead to inaccurate inferences regarding changes in performance over time or differences in performance among subgroups of students (for a discussion of this phenomenon, see Center on Education Policy, 2007, and Holland, 2002). Although there is no conclusive evidence that adopting a different set of measures would necessarily change how educators respond to testing, it is likely that, by measuring a broader range of outcomes, along with inputs and processes, some of the most problematic responses would be mitigated, as we discuss later.

A few studies have examined the consequences associated with adopting a multiple-measure indicator system. Chester (2005), for instance, documented some of the lessons learned from Ohio's use of multiple measures of student achievement at the district and school levels. Although this system focused primarily on test scores, Ohio's experience illustrates some ways in which the use of multiple indicators of performance can affect the utility of the system. The state's accountability system assigned each school to one of five performance categories based on four measures: test performance status, test performance growth, attendance rates, and graduation rates. The study reported that combining schools' data on these measures enhanced the validity of inferences about their performance (e.g., by avoiding the distorting effects of exclusive reliance on percentage proficient, discussed earlier) and improved the consistency of school classifications over what the consistency would have been if a single measure had been used. As Chester points out, the extent to which validity of inferences is improved depends not only on the specific measures that are included in the system but also on the rules for combining information from those measures, a topic to which we return later in this chapter.

Brown, Wohlstetter, and Liu (2008) examined a broader indicator system called the Charter School Indicators–University of Southern California (CSI-USC), which incorporated data on inputs, processes, and outcomes to provide publicly available information on California charter schools. This system was based on the balanced scorecard framework (Kaplan and Norton, 1992), which views organizations' performance through multiple lenses. In the case of schools, such lenses include finances and resources, school quality, student performance, and measures of efficiency. This research provides guidance for others who are involved in developing indicator systems; in particular, it points to the need for stakeholder involvement and for clear presentation of data and findings. However, the system's long-term effects on parents and other members of the public who are interested in this information remain to be seen.

Despite the well-documented problems stemming from high-stakes uses of tests and the lack of research on multiple-measure accountability or indicator systems, reliance on measures and incentives as a means to improve schools continues to enjoy widespread support among policymakers. This support is likely to continue for a number of reasons. Perhaps most importantly, although large-scale assessments are expensive, they tend to be less costly than other approaches to changing what schools do. Reforms, such as the adoption of new curricula or the redesign of professional development, typically require more resources than does the implementation of a new testing program, which is one reason policymakers have favored

test-based accountability as a means to promote school improvement (Linn, 2000). In addition, federal, state, and district investments in new data systems provide an incentive to continue and expand the use of measures for decisionmaking. The lessons learned from research on high-stakes student achievement testing can provide some guidance regarding the promises and pitfalls of measurement-based reform.

Possible Benefits of Expanded Measures

As noted above, much of the impetus for moving toward a more expansive measurement system comes from the concern that, by focusing on a small number of subject areas in a subset of K–12 grades, current accountability metrics provide a limited picture of how well schools are attaining the many goals toward which they are expected to work. If our notion of an effective school includes not only promoting students' achievement and attainment but also such features as providing a safe learning environment and encouraging the development of civic responsibility, an expanded system could allow a more accurate assessment of the school characteristics that society values.

An expanded set of measures could also promote more valid inferences about school performance by offering opportunities to compare performance on multiple overlapping dimensions. For example, an increase in mathematics test scores might be interpreted as evidence that the quality of mathematics instruction is improving, but such an increase could also reflect curriculum narrowing and score inflation. If this increase were accompanied by evidence of improvements in the quality of teaching as measured by direct observations of instruction, the user of the information might be more confident that an inference of improved instructional quality is warranted. Similarly, gains in test scores might be interpreted as evidence that students are learning more, but this inference would be more supportable if gains on NCLB-type tests were accompanied by gains on other measures, such as end-of-course tests that capture improvements in more advanced content. A system that was expanded to include measures of learning in a broader range of subjects, grade levels, and courses could help address the kinds of questions that most parents and members of the public are likely to ask about their schools, including questions about the relative success of students at different achievement levels and with different profiles of interests and goals. In other words, such measures could better serve accountability goals by helping a broad range of stakeholders understand how effective a school is on average and how well it meets the needs of particular kinds of students.

As discussed in Chester's (2005) description of Ohio's multiple-measure system for classifying schools into performance categories, if designed appropriately, these systems might not only improve the validity of inferences about performance but can also improve aspects of reliability, including the consistency of classifications of school into performance categories. Scores from any single measure include some degree of measurement error, and combining scores across measures has the potential to reduce this error, though, as discussed later, the specific rules for combining measures can affect validity and reliability and are not always straightforward.

An additional, related benefit of an expanded system relates to the incentive effects that high-stakes measures impose. As discussed above, high-stakes accountability sometimes leads to an increased focus on tested material at the expense of untested material. Such changes might result not only in important parts of the curriculum being neglected but also in the

neglect of desirable competencies that are difficult to measure reliably on standardized assessments, such as communication and teamwork skills. An emphasis on achievement tests might also lead educators to neglect other aspects of students' educational experiences, such as the school's social and emotional climate. The research on effects of testing on these other aspects of schooling is limited, but there is some evidence that schools have reduced opportunities for students to participate in activities, such as arts and field trips, and that teachers worry about negative effects on students' engagement (Hamilton, Stecher, Marsh, et al., 2007; Stecher et al., 2008). A more balanced set of incentives might mitigate these risks, though it is clearly impractical to measure every input, process, and outcome of interest, and developers of these systems will need to weigh the costs and benefits of an expanded set of measures.

Risks and Trade-Offs Associated with Expanded Measures

Although the arguments in favor of expanded measures are strong, those responsible for setting policy or designing systems must proceed with caution and should be armed with an understanding of the various risks involved. For example, the Center for Public Education (2008) has published a list of "good measures for good schools" that sets out a number of commonsense questions as a guide for school quality. The questions pertain to five domains: student achievement during school, workforce and college readiness as of the end of school, school climate, school resources and staff, and school demographics. Notwithstanding the substantial progress made in developing measures in most of these domains, many localities currently lack the data to assess some of these domains, especially workforce and college readiness and school climate. And even where data are available, the research about the usefulness of the measure for improving aspects of school performance is limited. We know, for example, that class-size reduction can boost achievement in the early grades in some contexts (Mosteller, 1995) but not how it affects behavior, engagement, and student–teacher interactions. So even an understanding of the limited research available will take policymakers only so far.

Ultimately, the selection of measures in an indicator system should be informed by the purposes of the system—e.g., whether it will be used solely for monitoring, in a diagnostic or prescriptive way to guide school-improvement decisions, or whether it will be part of accountability with explicit stakes attached to results. Different uses could suggest different design decisions. For example, measures that are used for high-stakes accountability purposes for school practitioners should focus on factors that the educators are able to influence through their practice (e.g., growth in student learning, attendance, student–teacher interactions) rather than on conditions or outcomes that are not under educators' control (e.g., student demographics and spending). By contrast, a system that is intended to inform school choice might lead to different decisions about what to include. For instance, measures that are outside the control of school practitioners (e.g., per-pupil spending or quality of facilities) might nonetheless be helpful to parents who are trying to decide which school offers the most appropriate environment for their children. Choice of measures should also be informed, to the extent possible, by research on what inputs, processes, and outcomes matter for long-term student success, while keeping in mind that the research in most areas has limitations.

Even a careful consideration of one's purposes and of the existing research does not necessarily lead to a straightforward approach to designing an indicator system. In this section, we briefly discuss several trade-offs that must be addressed, and we describe several additional

considerations that developers and users should keep in mind as they develop, implement, and evaluate new measures of school performance.

- Breadth versus focus. The key value of a more expansive system is that it could more closely align indicators with the goals society holds for schools. Greater breadth might also lessen the likelihood of inappropriate narrowing of curricula or instruction. For example, 11 of the 20 states factored more subjects than math and ELA into their school ratings. At the same time, having additional measures could reduce the utility of the system as a mechanism for helping educators focus their work. Many teachers who participated in the RAND NCLB studies (Hamilton, Stecher, Marsh, et al., 2007; Stecher et al., 2008) expressed concerns about state standards that were too broad or that included more content than could realistically be covered, and an expanded set of measures could exacerbate these concerns. One of the drawbacks associated with NCLB—namely, its emphasis on a small number of tested outcomes—could also be considered a strength in that it enables the system to send a clear message about what outcomes educators are expected to emphasize. Comprehensive measures can potentially scatter rather than focus educators' attention, though this effect might be mitigated through efforts to structure and communicate the information effectively, and through training and ongoing support to help educators figure out how to prioritize and synthesize signals from multiple measures. Developers need to consider the balance between a reasonably comprehensive set of measures and the need to help teachers and other educators understand what goals they are expected to promote.
- Complexity versus transparency. A related concern involves the ways in which information from specific measures is transformed and aggregated to produce a number of other indicators. To incorporate differences in the characteristics of students or school inputs, more complex indicators might be needed, but complexity makes the indicators more difficult to understand. This concern is illustrated by recent work on value-added modeling of teacher effects. Researchers have explored a variety of approaches to developing measures that attempt to isolate the effect of a teacher on students' test scores, and these methods typically involve complex statistical models that are difficult for anyone but highly trained methodologists to understand. Although these approaches might do a better job than simpler measures of supporting accurate inferences, their complexity could limit their utility for helping teachers or others determine how to improve teachers' performance, and could also reduce the likelihood that educators will support the measurement system (Chudowsky, Koenig, and Braun, 2010). Similar concerns apply to highly complex metrics of school-level performance, such as performance indexes (used in ten of the 20 states' rating systems) that assign different weights to student scores along the spectrum of very low to very high performance. For some purposes, more transparent measures might be preferable, particularly when the measures are intended to support improvement efforts, but, for certain kinds of inferences, it might be necessary to sacrifice some transparency in an effort to promote valid inferences. Regardless of purpose, decisions about whether to implement complex modeling approaches require consideration of the resources—both human and technological—available to apply those modeling approaches.
- Comprehensiveness versus affordability. Clearly, the number of measures cannot be expanded without incurring additional costs. Within the domain of student achievement,

efforts to avoid the narrowing effects associated with multiple-choice test items must contend with the added costs of testing more subjects and including a broader range of item formats. Perhaps the most common form of expanded measures is the increasingly frequent use of standardized measures of student performance throughout the school year. Costs associated with administration of additional assessments, as well as the expense of creating and maintaining the data systems, need to be carefully considered in light of alternative uses for those funds. In addition, those who develop or mandate new measures should consider not only the monetary costs but the burdens that additional measurement could impose on educators' and students' time. For example, school quality reviews, such as are done in Charlotte-Mecklenburg, are resource-intensive because they require substantial preparation by school staff and a trained team that observes the school for at least two days and prepares a report with its findings. Yet the team's review touches on aspects of school climate, leadership, and instruction that administrative data alone cannot capture. In all of these cases, it is important to evaluate whether the benefits of these measures outweigh the possible loss of instructional time, especially for assessment systems that are not fully aligned with the curriculum. Finally, one must consider the cost of "not knowing" about performance in areas that are not measured. Although it is cheaper in the short run not to measure higher-order thinking skills, it could be more expensive in the long run if schools ignore those skills because they are not measured.

- Uniformity versus flexibility. Decisions that involve comparing performance across schools will benefit from a uniform set of indicators in each school. At the same time, schools differ in many ways that make it difficult to measure their performance using identical indicators. A uniform set of indicators might not only reduce the utility of information but could also stifle efforts to innovate and adapt to local contextual factors. For example, a set of measures that is appropriate for assessing the performance of a high school science magnet program might not work well for a comprehensive middle school. A hybrid system, with some common measures and some customized information, might be more desirable. A related concern is that measures that are overly prescriptive could lead to unintended consequences, such as reduced staff morale or a resistance to trying new approaches to solving instructional problems. Thus, developers need to examine their measures in light of the ways in which they might be expected to influence educators' actions. To the extent that the system attempts to serve diagnostic or prescriptive purpose, it could be more important to incorporate flexibility, so it can respond to locally identified areas of need.
- Formative versus summative purposes. The annual tests that form the backbone of most existing school performance-measurement systems tend not to be perceived by teachers as useful for day-to-day instructional decisionmaking because the score reports are not sufficiently fine-grained and are available infrequently (Wayman and Stringfield, 2006). Instructional decisionmaking, in contrast, tends to benefit from measures that are given frequently, embedded in local curricula, and linked to guidance for follow-up (Perie et al., 2007), but these measures often lack the technical qualities that are required when using tests to make high-stakes decisions. Moreover, using diagnostic tests for high-stakes purposes could dilute their diagnostic usefulness if teachers begin to treat students' performance on these assessments as end goals rather than as indicators for refining their instruction. A single set of measures might not adequately serve both formative and summative purposes, and those responsible for determining how performance is measured

should avoid the temptation to use the same measures for multiple purposes unless there is validity evidence to support each of the proposed uses.

- Signaling versus preventing corruption of measures. Another trade-off related to the purposes of measures stems from the signaling function that we discussed earlier in this report. The mere decision to include a measure in an accountability system with stakes attached sends a signal to educators and others about what processes or outcomes are valued and can increase the likelihood that the areas that are measured will be a focus of educators' and students' efforts. From a signaling perspective, one could argue for attaching stakes to a broad range of measures, such as school climate surveys and student course taking. As noted above, however, attaching stakes leads to a risk of score corruption, and some kinds of measures might be particularly susceptible to manipulation if educators are under pressure to improve performance on these dimensions. For instance, high stakes attached to AP course enrollment could lead to higher enrollment by students who are not fully prepared for these courses, which could hinder the quality of the experience for these students and for the students who are prepared. System designers must carefully examine the likelihood that high stakes will lead to corruption and decide whether to reduce stakes or whether other steps might be taken to maintain the integrity of the measure.
- Adjusting versus not adjusting for inputs. The era of standards-based reform has been characterized by a widely held view that it is important to hold all schools, and all students, to similar, high standards. Thus, NCLB's primary accountability provision was designed to focus on the percentage of a school's students who score at the proficient level or higher, regardless of how those students performed before entering the school or what kinds of challenges they face outside of school. This approach to accountability is popular among some education reformers because it embodies the ideal that the same high expectations are held for all students and schools. At the same time, by ignoring inputs, the system could inadvertently reduce motivation to improve, both on the part of high-scoring schools that do well because they serve advantaged students and among low-scoring schools that do not have a realistic chance of meeting the target even if they achieve significant improvement (Reback, Rockoff, and Schwartz, 2011). This type of system can produce inaccurate inferences about which schools are effective and can reduce morale and support for the system to the extent that it is seen as unfair. Approaches to addressing this problem include the use of growth models or risk-adjustment procedures that take into account prior performance and other characteristics of students, schools, or neighborhoods that might influence performance. These approaches introduce technical and practical challenges, and some might view their use as a way of lowering standards for some students or schools, but, for some purposes, it is likely that adjusted measures will provide better information and incentives than unadjusted measures.

Additional Technical Considerations

Those who design school indicator systems face a number of significant decisions about the features of the measures and the methods for combining them and attaching them to consequences. These decisions, some of which were discussed earlier, pertain to the methods for creating indices (e.g., which measures will be reported in terms of status, growth, or both,

and how information will be combined across different measures, if at all); whether to apply risk-adjustment procedures to account for differences in inputs; whether to set performance targets and, if so, how ambitious these should be; and how much to customize assessments to address individual students' varying school experiences (e.g., whether to rely primarily on course-specific assessments at the high school level). There are no simple answers to these questions; the decisions need to be informed by an understanding of local context and constraints and by a careful consideration of the goals the system is intended to promote.

Regardless of how the various decisions are resolved, those who develop, mandate, or use school indicator systems need to examine the technical quality of these systems to maximize the likelihood that they will provide accurate information and produce desirable outcomes. An overriding concern is validity, a term that refers to "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (Joint Committee on Standards for Educational and Psychological Testing, 1999, p. 184). The requirement to investigate validity pertains not only to the tests or other measures that are used to create the indicators but also to any derived or summary scores, such as an AYP index. The process of validation involves an accumulation of evidence to support the use of a particular set of test scores for a particular purpose (Kane, 2006). As we noted earlier, users should recognize that a test that might be viewed as having validity evidence for one purpose, such as informing instructional decisions, should not automatically be considered to have validity evidence for a different purpose, such as identifying low-performing schools that are in need of intervention. System developers need to contend with a limited amount of information about the technical properties of many of their measures; many new assessments and surveys have little available information about their technical quality, and even established measures, such as AP exams, have not been validated for use in indicator systems. Moreover, as discussed earlier, attaching stakes to measures can lead to the corruption of scores, and nontest measures are not necessarily immune to this problem. Developers and users should view validation as an ongoing activity that can be informed by the accumulation of evidence as the system is rolled out and implemented over time.

A related concern is reliability, a term that refers to the degree to which test scores are consistent over repeated measurements and are free of errors of measurement. It is important to gather as much information as possible to understand how much error is in the system, in order to help users determine what level of confidence they should place in the information they receive. Errors can result from a variety of sources, including the sampling of specific tasks, the differences in how raters apply scoring criteria, and the specific set of students included in the measures. The appropriate method for estimating reliability depends in part on the possible sources of error (e.g., whether raters are used to score test items) and on the kinds of scores produced (e.g., classification consistency should be examined if schools are assigned to categories rather than awarded a score on a continuous distribution).

Moreover, when individual scores are combined into a composite, reliability should be estimated for the composite and not just for the original scores that it comprises. Some composites, such as school-level averages of test scores, might have higher levels of reliability than the individual scores that are used to calculate them (Hill and DePascale, 2002). But methods for combining scores can also lead to threats to reliability over and above the measurement error associated with the original, individual-level scores. For example, estimates of teaching effectiveness (i.e., value added) that assess average changes of student scores by classroom tend to be unstable from one year to the next, and, although some of this instability reflects actual

changes in teachers' contributions to student achievement, much of it is noise stemming from random classroom-level error in addition to the student-level error (Lankford et al., 2010; McCaffrey, Sass, et al., 2009). As the value-added example illustrates, it is often challenging to determine the extent to which instability in scores over time is due to signal or noise—that is, do changes in performance represent real changes, or do they primarily result from measurement errors?

A third aspect of technical quality that users should consider is fairness, which essentially refers to the extent to which a measure has the same meaning for each individual or organization who receives a score on that measure (Joint Committee on Standards for Educational and Psychological Testing, 1999). Considerations of a measure's fairness are related to both validity and reliability: A measure can lack fairness if it measures something (e.g., student socioeconomic status) that is unrelated to the construct of interest (e.g., student mastery of mathematics) or if scores for some examinees (e.g., English language learners) are subject to greater errors of measurement than those of other examinees. Documenting the fairness of school-level indicators is important for promoting stakeholder buy-in and for maximizing the validity of inferences about school performance.

CHAPTER FIVE

Recommendations for a Federal Role to Promote Improved Measurement of School Performance

As this report has sought to demonstrate, some states and districts have maintained or expanded indicator systems that go beyond the requirements imposed by NCLB. These state and local initiatives include broader measures of school inputs (e.g., school-level instructional expenditures, adjustments to performance to account for variation in resources), processes (e.g., student safety and teacher and student satisfaction), and outcomes (e.g., gains in achievement, graduating students who are college- or career-ready).

However, there is limited research documenting the quality of these measures, their utility for decisionmaking, or their effects on educators' practices and student outcomes. As a result, states' and districts' efforts offer only anecdotal guidance for deciding the appropriate number or the right balance of input, process, and outcome measures to include in a school indicator system. Following feedback from school organizations contributing to its school performance-metric data bank, the Broad Foundation (2009) advises school districts to develop data dashboards for teachers and principals that have ten to 20 metrics presented on the equivalent of two pages. Although this guidance is not derived from empirical testing of threshold levels at which the number of metrics can scatter rather than focus school administrators' attention, it nevertheless represents the collective experience of localities that currently implement a number of measures described above. At the same time, the quality of the measures should outweigh any formulaic determination about their quantity; a smaller number of high-quality measures is preferable to a greater number of low-quality measures.

Although the federal government has historically played a fairly small role in shaping state and local education decisions, NCLB demonstrates that federal legislation can have a substantial impact on what schools and districts do. Among other things, it has prompted the expansion of states' longitudinally linked data systems that allow them to track student progress and link students' achievement-test scores to other information in ways that were not possible only a few years ago. Further, NCLB accountability ratings have had a tremendous influence on state and local rating systems. Anecdotal evidence confirms that NCLB is the primary rating that schools consider, with local rating systems playing the secondary role of providing diagnostic information to respond to the federal rating. NCLB accountability has also prompted the local development of additional measures to predict performance on high-stakes outcomes, such as indicators of at-risk students and the periodic assessments described earlier. The influence of NCLB legislation suggests that the federal government could help effectively create incentives for states and districts to expand on their data systems to develop and test other measures of school performance.

To prompt federal policymakers' thinking about what form the federal role might take in motivating the development of more comprehensive school measurement systems, we present the following three recommendations:

- In the ESEA reauthorization, incorporate a broader range of measures as a basis for accountability decisions than is currently mandated under NCLB. Although there is currently insufficient evidence to make specific choices about which measures should be used, evidence from research on high-stakes testing indicates that educators tend to shift their focus away from what is not measured and toward what is. Given the broad set of goals for schooling that we discussed at the beginning of this report, it is clear that systems that rely exclusively on standardized tests in a small set of subjects create a risk that some critical goals will be shortchanged. A federal mandate that states (or state consortia) select their own measures within a broader set of predefined categories might mitigate this risk and might allow stakeholders to draw more valid inferences regarding school performance that better reflect the multiple goals of schooling. This form of controlled flexibility builds on the recent precedent of Race to the Top, in which applicant states were required to revise teacher evaluation methods to include multiple measures of their choosing and to base at least a significant part of the evaluation on student growth or test scores. Although broadening accountability measures will require additional resources, the value of the information and the improvement in schools' incentive structures is likely to justify the additional cost. We suggest the following five domains of expanded measures as places to start because they reflect the broader goals of schooling discussed in Chapter One or they address areas in which states have increased their measures to meet perceived needs:
  - Expand the measures of achievement and attainment to account for both status and growth and to capture a broader range of academic outcomes in subjects besides math and ELA, as well as in advanced course taking.
  - Promote a positive school culture, including indicators, such as student and teacher satisfaction, academic challenge, engagement, safety, or orderliness.
  - Adopt leading indicators, such as measures of being on track for high school graduation, that provide schools information about students as they progress toward college and career readiness.
  - Promote positive behavioral, emotional, and physical health outcomes for students, including indicators of suspensions, expulsion, physical health.
  - Augment unadjusted performance indicators with indicators that adjust for discrepancies in resources that children and, by extension, schools have available.
- Avoid creating a new federal mandate to adopt specific measures. Mandating new measures that have not been well evaluated could have unintended effects, such as siphoning resources away from more productive measurement and reform efforts that are already under way. As states begin to validate additional measures, these can be gradually integrated into a refined federal system for measuring school performance. States should be required to conduct an evaluation of the technical quality and the effects of the inclusion of new measures within an ESEA accountability framework on student outcomes and school resource allocation. For that, they might require technical assistance or collaboration, which leads to our third recommendation.

- Incorporate the development and evaluation of additional school performance measures as an area of focus within existing competitively awarded federal grants. Recognizing that states vary in their capacity to develop and test new measures and the desirability of developing measures that are consistent across states, offering federal grants for such development could create incentives for states to coordinate their efforts, as through interstate consortia. Examples of existing grants that have driven such coordination efforts include ED's Enhanced Assessment Grants and the National Science Foundation's Promoting Research and Innovation in Methodologies for Evaluation program. Further, in establishing priority areas for the development and validation of additional measures in certain school goal areas, these funds should also be contingent on the provision of clear, explicit support to teachers and principals so they can interpret the new measures and adapt their practices in response.

Many federal policymakers agree that the reauthorization of ESEA should build on knowledge acquired in the past decade about districts' and schools' responses to NCLB (Dillon, 2010). The question remains how to undertake this reauthorization in a way that fine-tunes the law's performance-measurement requirements rather than replacing one imperfect system with one that is even more unwieldy. At the same time, it is important that the impetus for using measurement to inform educational decisionmaking remains firmly in place. This report has described promising directions for expanding the set of measures that schools have at their disposal while acknowledging the need for more research on how the availability of such measures affects educational practice and student achievement. Even with more research, however, the public will have to weigh carefully the trade-offs in choosing what facets of their public schools should be measured and how those measures should be used to inform high-stakes policy decisions.

Bibliography

Allensworth, Elaine, and John Q. Easton, *The On-Track Indicator as a Predictor of High School Graduation*, Chicago, Ill.: Consortium on Chicago School Research, June 2005. As of February 16, 2011: http://ccsr.uchicago.edu/content/publications.php?pub_id=10

———, *What Matters for Staying On-Track and Graduating in Chicago Public Schools*, Chicago, Ill.: Consortium on Chicago School Research, July 2007. As of February 16, 2011: http://ccsr.uchicago.edu/content/publications.php?pub_id=116

Altman, J. R., S. S. Lazarus, R. F. Quenemoen, J. Kearns, M. Quenemoen, and M. L. Thurlow, *2009 Survey of States: Accomplishments and New Issues at the End of a Decade of Change*, Minneapolis, Minn.: University of Minnesota, National Center on Educational Outcomes, 2010. As of February 16, 2011: http://www.cehd.umn.edu/NCEO/onlinepubs/StateReports/2009_survey_of_states.htm

Baker, Eva L., Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Helen F. Ladd, Robert L. Linn, Diane Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A. Shepard, *Problems with the Use of Student Test Scores to Evaluate Teachers*, Washington, D.C.: Economic Policy Institute, Briefing Paper 278, August 27, 2010. As of February 16, 2011: http://www.epi.org/publications/entry/bp278

Baker, Eva L., Robert L. Linn, Joan L. Herman, and Daniel Koretz, *Standards for Educational Accountability Systems*, Los Angeles, Calif.: National Center for Research on Evaluation, Standards, and Student Testing, Policy Brief 5, Winter 2002. As of February 16, 2011: http://www.cse.ucla.edu/products/policy/cresst_policy5.pdf

Bill and Melinda Gates Foundation, "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project," MET project policy brief, December 2010. As of February 21, 2011: http://www.gatesfoundation.org/college-ready-education/Documents/preliminary-finding-policy-brief.pdf

Booher-Jennings, Jennifer, "Below the Bubble: 'Educational Triage' and the Texas Accountability System," *American Educational Research Journal*, Vol. 42, No. 2, Summer 2005, pp. 231–268. As of February 16, 2011: http://www.jstor.org/stable/3699376

Broad Foundation, school-level metric bank, undated spreadsheet (a). As of December 10, 2010: http://www.broadeducation.org/asset/1344-schoollevelmetricbank.xls

———, stakeholder survey question bank, undated spreadsheet (b). As of December 10, 2010: http://www.broadeducation.org/asset/1344-stakeholdersurveyquestionbank.xls

———, *Metrics Bank: Example Metrics for School-Level Dashboards and Scorecards*, updated December 15, 2009. As of February 16, 2011: http://www.broadeducation.org/asset/1344-introtometricbank.pdf

Brown, Richard S., "Creating School Accountability Reports," *School Administrator*, Vol. 56, No. 10, November 1999, pp. 12–14, 16–17. As of February 16, 2011: http://www.aasa.org/SchoolAdministratorArticle.aspx?id=14928

Brown, Richard S., Priscilla Wohlstetter, and Sunny Liu, "Developing an Indicator System for Schools of Choice: A Balanced Scorecard Approach," *Journal of School Choice*, Vol. 2, No. 4, 2008, pp. 392–414.

Busbee, Nancy, director, Office of Federal and State Accountability, Division of Accountability, South Carolina State Department of Education, telephone interview, October 12, 2010.

Center for Public Education, "Good Measures for Good Schools," 2008.

Center on Education Policy, *From the Capital to the Classroom: Year 4 of the No Child Left Behind Act*, Washington, D.C., 2006.

———, *Answering the Question That Matters Most: Has Student Achievement Increased Since No Child Left Behind?* Washington, D.C., 2007. As of December 22, 2010:
http://www.cep-dc.org/cfcontent_file.cfm?Attachment=CEP_Report_StudentAchievement_053107.pdf

Chang, Hedy N., and Mariajosé Romero, *Present, Engaged, and Accounted For: The Critical Importance of Addressing Chronic Absence in the Early Grades*, New York: National Center for Children in Poverty, September 2008. As of February 16, 2011:
http://www.nccp.org/publications/pub_837.html

Charlotte-Mecklenburg Schools, "School Quality Reviews," undated web page. As of February 8, 2011:
http://www.cms.k12.nc.us/cmsdepartments/accountability/cfsi/Pages/SQRs.aspx

Chester, Mitchell D., "Making Valid and Consistent Inferences About School Effectiveness from Multiple Measures," *Educational Measurement: Issues and Practice*, Vol. 24, No. 4, December 2005, pp. 40–52.

Chudowsky, Naomi, and Victor Chudowksy, *State Test Score Trends Through 2008–09*, Part 1: *Rising Scores on State Tests and NAEP*, Washington, D.C.: Center on Education Policy, revised November 2010. As of February 16, 2011:
http://www.cep-dc.org/
cfcontent_file.cfm?Attachment=Chudowsky_FullReport_2008-09Part1_StateNAEP_Revised113010.pdf

Chudowsky, Naomi, Judith A. Koenig, and Henry I. Braun, *Getting Value Out of Value-Added: Report of a Workshop*, Washington, D.C.: National Academies Press, 2010.

Cobitz, Christopher, executive director, state and federal programs, Charlotte-Mecklenburg Schools, telephone interview, October 5, 2010.

Cohen, Matthew, executive director, Office of Policy and Accountability, Ohio Department of Education, telephone interview, October 5, 2010.

Connolly, Faith, executive director; Stephen Plank, research codirector; Martha Abele Mac Iver, research scientist; and Rachel Durham, assistant research scientist, Baltimore Education Research Consortium, telephone interview, November 9, 2010.

Consortium on Chicago School Research, "Surveys of CPS Schools," undated web page. As of February 16, 2011:
http://ccsr.uchicago.edu/content/page.php?cat=4

Cronin, John, Chester E. Finn, and Michael J. Petrilli, *The Proficiency Illusion*, Washington, D.C.: Thomas B. Fordham Institute, October 4, 2007.

Data Quality Campaign, undated home page. As of February 17, 2011:
http://dataqualitycampaign.org/

Dedering, Kathrin, and Sabine Müller, "School Improvement Through Inspections? First Empirical Insights from Germany," *Journal of Educational Change*, December 3, 2010. As of February 16, 2011:
http://www.springerlink.com/content/64225m06u86355g4/

Dillon, Sam, "New Challenges for Obama's Education Agenda in the Face of a G.O.P.-Led House," *New York Times*, December 11, 2010, p. A36. As of February 16, 2011:
http://www.nytimes.com/2010/12/12/us/politics/12education.html

Economic Policy Institute, "A Broader, Bolder Approach to Education," June 10, 2008. As of February 16, 2011:
http://www.boldapproach.org/statement.html

Elmore, Richard, "Leadership as the Practice of Improvement," in Beatriz Pont, Deborah Nusche, and David Hopkins, eds., *Improving School Leadership*, Volume 2: *Case Studies on System Leadership*, Paris: OECD, 2008, pp. 37–68. As of February 16, 2011:
http://www.oecd-ilibrary.org/education/improving-school-leadership_9789264039551-en

Faubert, Violaine, *School Evaluation: Current Practices in OECD Countries and a Literature Review*, Paris: OECD, Education Working Paper 42, 2009.

Hamilton, Laura S., "Assessment as a Policy Tool," *Review of Research in Education*, Vol. 27, 2003, pp. 25–68. As of February 21, 2011:
http://www.rand.org/pubs/reprints/RP1163.html

Hamilton, Laura S., Richard Halverson, Sharnell S. Jackson, Ellen Mandinach, Jonathan A. Supovitz, Jeffrey C. Wayman, Cassandra Pickens, Emily Sama Martin, and Jennifer L. Steele, *Using Student Achievement Data to Support Instructional Decision Making*, Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, September 2009. As of February 16, 2011:
http://ies.ed.gov/ncee/wwc/pdf/practiceguides/dddm_pg_092909.pdf

Hamilton, Laura S., Brian M. Stecher, Julie A. Marsh, Jennifer Sloan McCombs, Abby Robyn, Jennifer Russell, Scott Naftel, and Heather Barney, *Standards-Based Accountability Under No Child Left Behind: Experiences of Teachers and Administrators in Three States*, Santa Monica, Calif.: RAND Corporation, MG-589-NSF, 2007. As of February 16, 2011:
http://www.rand.org/pubs/monographs/MG589.html

Hamilton, Laura S., Brian M. Stecher, Jennifer Russell, Julie A. Marsh, and J. Miles, "Accountability and Teaching Practices: School-Level Actions and Teacher Responses," in Bruce Fuller, Melissa K. Henne, and Emily Hannum, eds., *Strong States, Weak Schools: The Benefits and Dilemmas of Centralized Accountability*, Bingley, UK: Emerald JAI, 2008, pp. 31–66.

Hamilton, Laura S., Brian M. Stecher, and Kun Yuan, *Standards-Based Reform in the United States: History, Research, and Future Directions*, Washington, D.C.: Center on Education Policy, 2009. As of February 16, 2011:
http://www.rand.org/pubs/reprints/RP1384.html

Hannaway, Jane, and Laura S. Hamilton, *Accountability Policies: Implications for School and Classroom Practices*, Washington, D.C.: Urban Institute, October 16, 2008. As of February 16, 2011:
http://www.urban.org/url.cfm?ID=411779

Hargreaves, Andy, and Dennis Shirley, "Beyond Standardization: Powerful New Principles for Improvement," *Phi Delta Kappan*, Vol. 90, No. 2, October 2008, pp. 135–143.

Hartman, Jenifer, Chuck Wilkins, Lois Gregory, Laura Feagans Gould, and Stephanie D'Souza, *Applying an On-Track Indicator for High School Graduation: Adapting the Consortium on Chicago School Research Indicator for Five Texas Districts*, Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest, REL 2011–No. 100, January 2011. As of February 16, 2011:
http://ies.ed.gov/ncee/edlabs/projects/project.asp?ProjectID=264

Hill, Richard, and Charles DePascale, *Determining the Reliability of School Scores*, Dover, N.H.: National Center for the Improvement of Educational Assessment, November 2002. As of February 2, 2011:
http://www.nciea.org/publications/CCSSO02_Reliability_RHCD03.pdf

Ho, Bonnie, "Few States Survey Teachers on School Climate," *Education Week*, September 24, 2008. As of February 16, 2011:
http://www.edweek.org/rc/articles/2008/09/24/sow0924.h27.html

Holland, Paul W., "Two Measures of Change in the Gaps Between the CDFs of Test-Score Distributions," *Journal of Educational and Behavioral Statistics*, Vol. 27, No. 1, Spring 2002, pp. 3–17. As of February 16, 2011:
http://www.jstor.org/stable/3648143

Holtzapple, Elizabeth, director of research and evaluation, Cincinnati Public Schools, telephone interview, October 7, 2010.

Huber, Stephan, Hunter Moorman, and Beatriz Pont, "The English Approach to System Leadership," in Beatriz Pont, Deborah Nusche, and David Hopkins, eds., *Improving School Leadership*, Volume 2: *Case Studies on System Leadership*, Paris: OECD Publishing, 2008, pp. 111–152. As of February 17, 2011:
http://www.oecd-ilibrary.org/education/improving-school-leadership_9789264039551-en

Jeong, Dong Wook, "Student Participation and Performance on Advanced Placement Exams: Do State-Sponsored Incentives Make a Difference?" *Educational Evaluation and Policy Analysis*, Vol. 31, No. 4, December 2009, pp. 346–366.

Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, Washington, D.C.: American Educational Research Association, 1999.

Kane, Michael T., "Validation," in Robert L. Brennan, ed., *Educational Measurement*, 4th ed., Westport, Conn.: Praeger Publishers, 2006, pp. 17–64.

Kaplan, Robert S., and David P. Norton, "The Balanced Scorecard: Measures That Drive Performance," *Harvard Business Review*, Vol. 70, January–February 1992, pp. 64–72.

Koretz, Daniel M., *Measuring Up: What Educational Testing Really Tells Us*, Cambridge, Mass.: Harvard University Press, 2008.

Lankford, Hamilton, Donald Boyd, Susanna Loeb, and James Wyckoff, *Measuring Test Measurement Error: A General Approach and Policy Implications*, paper presented at the Association of Public Policy Analysis and Management Fall Conference, Boston, Mass., November 6, 2010.

Leithwood, Kenneth, Rosanne Steinbach, and Doris Jantzi, "School Leadership and Teachers' Motivation to Implement Accountability Policies," *Educational Administration Quarterly*, Vol. 38, No. 1, February 2002, pp. 94–119.

Linn, Robert L., "Assessments and Accountability," *Educational Researcher*, Vol. 29, No. 2, March 2000, pp. 4–16.

———, "Accountability: Responsibility and Reasonable Expectations," *Educational Researcher*, Vol. 32, No. 7, October 2003, pp. 3–13.

Mac Iver, Martha Abele, *Gradual Disengagement: A Portrait of the 2008–09 Dropouts in the Baltimore City Schools*, Baltimore, Md.: Baltimore Education Research Consortium, August 2010. As of February 17, 2011: http://baltimore-berc.org/pdfs/Gradual%20Disengagement%20final.pdf

Massachusetts Department of Elementary and Secondary Education, "2009–10 Early Warning Indicator Index," last updated March 1, 2010. As of February 8, 2011: http://www.doe.mass.edu/dropout/fy10earlyindicator.pdf

Matthews, Peter, Hunter Moorman, and Deborah Nusche, "Building Leadership Capacity for System Improvement in Victoria, Australia," in Beatriz Pont, Deborah Nusche, and David Hopkins, eds., *Improving School Leadership*, Volume 2: *Case Studies on System Leadership*, Paris: OECD Publishing, 2008, pp. 179–213. As of February 17, 2011: http://www.oecd-ilibrary.org/education/improving-school-leadership_9789264039551-en

Mayer, Daniel P., John E. Mullens, and Mary T. Moore, *Monitoring School Quality: An Indicators Report*, Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, NCES 2001-030, 2000. As of February 17, 2011: http://purl.access.gpo.gov/GPO/LPS10877

McAdams, Donald R., Michelle Wisdom, Sarah Glover, and Anne McClellan, *Urban School District Accountability Systems*, Houston, Tex.: Center for Reform of School Systems, December 2003. As of February 17, 2011: http://www.ecs.org/html/educationissues/accountability/mcadams_report.pdf

McCaffrey, Daniel F., Daniel Koretz, J. R. Lockwood, and Laura S. Hamilton, *Evaluating Value-Added Models for Teacher Accountability*, Santa Monica, Calif.: RAND Corporation, MG-158-EDU, 2004. As of February 17, 2011: http://www.rand.org/pubs/monographs/MG158.html

McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly, "The Intertemporal Variability of Teacher Effect Estimates," *Education Finance and Policy*, Vol. 4, No. 4, Fall 2009, pp. 572–606.

McGabe, Libby, and Jonathan Cohen, *State Department of Education School Climate–Related Policy: A Summary*, September 2006. As of February 17, 2011:
http://www.schoolclimate.org/climate/documents/policyscan.pdf

Means, Barbara, *Implementing Data-Informed Decision Making in Schools: Teacher Access, Supports and Use*, Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation and Policy Development, January 2009. As of February 17, 2011:
http://www.espsolutionsgroup.com/news/data-informed-decision.pdf

Mosteller, Frederick, "The Tennessee Study of Class Size in the Early School Grades," *Future of Children*, Vol. 5, No. 2, Summer–Fall 1995, pp. 113–127.

National Center for Education Statistics, "National Education Data Model," undated flyer. As of February 17, 2011:
http://nces.sifinfo.org/datamodel/files/NEDM_FLYER.pdf

New York City Department of Education, "NYC School Survey," undated web page. As of February 8, 2011:
http://schools.nyc.gov/Accountability/tools/survey/default.htm

Osher, David, vice president, American Institutes for Research, telephone interview, October 4, 2010.

Osher, D., J. Sprague, R. P. Weissberg, J. Axelrod, S. Keenan, K. Kendziora, and J. E. Zins, "A Comprehensive Approach to Promoting Social, Emotional, and Academic Growth in Contemporary Schools," in Alex Thomas and Jeff Grimes, eds., *Best Practices in School Psychology, V*, Vol. 4, Bethesda, Md.: National Association of School Psychologists, 2008, pp. 1263–1278.

Perie, Marianne, Scott Marion, Brian Gong, and Judy Wurtzel, *The Role of Interim Assessments in a Comprehensive Assessment System: A Policy Brief*, Dover, N.H.: National Center for the Improvement of Educational Assessment, 2007. As of February 17, 2011:
http://www.achieve.org/files/TheRoleofInterimAssessments.pdf

Pinkus, Lyndsay M., *Moving Beyond AYP: High School Performance Indicators*, Washington, D.C.: Alliance for Excellent Education, June 2009. As of February 17, 2011:
http://www.all4ed.org/files/SPIMovingBeyondAYP.pdf

Public Law 89-10, Elementary and Secondary Education Act, April 11, 1965.

Public Law 107-110, No Child Left Behind Act of 2001, January 8, 2002. As of February 16, 2011:
http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=107_cong_public_laws&docid=f:publ110.107.pdf

Reback, Randall, Jonah E. Rockoff, and Heather L. Schwartz, *Under Pressure: Job Security, Resource Allocation, and Productivity in Schools Under NCLB*, Cambridge, Mass.: National Bureau of Economic Research, Working Paper 16745, January 2011. As of February 17, 2011:
http://papers.nber.org/papers/16745

Rosenthal, Leslie, "Do School Inspections Improve School Quality? Ofsted Inspections and School Examination Results in the UK," *Economics of Education Review*, Vol. 23, No. 2, April 2004, pp. 143–151.

Rothstein, Jesse, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics*, Vol. 125, No. 1, February 2010, pp. 175–214.

———, "Review of 'Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project,'" Boulder, Colo.: National Education Policy Center, January 13, 2011. As of February 17, 2011:
http://nepc.colorado.edu/thinktank/review-learning-about-teaching

Schochet, Peter Z., and Hanley S. Chiang, *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains*, Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, NCEE 2010-4004, July 2010. As of February 17, 2011:
http://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf

Scott, Caitlin, *Mining the Opportunities in "Differentiated Accountability": Lessons from the No Child Left Behind Pilots in Four States*, Washington, D.C.: Center on Education Policy, August 2009. As of February 16, 2011:
http://www.eric.ed.gov/PDFS/ED506742.pdf

Sparks, Sarah D., "Experts Begin to Identify Nonacademic Skills Key to Success," *Education Week*, December 23, 2010.

Stecher, Brian M., "Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practice," in Laura S. Hamilton, Brian M. Stecher, and Stephen P. Klein, *Making Sense of Test-Based Accountability in Education*, Santa Monica, Calif.: RAND Corporation, MR-1554-EDU, 2002, pp. 79–100. As of February 21, 2011:
http://www.rand.org/pubs/monograph_reports/MR1554.html

Stecher, Brian M., Scott Epstein, Laura S. Hamilton, Julie A. Marsh, Abby Robyn, Jennifer Sloan McCombs, Jennifer Russell, and Scott Naftel, *Pain and Gain: Implementing No Child Left Behind in Three States, 2004–2006*, Santa Monica, Calif.: RAND Corporation, MG-784-NSF, 2008. As of February 17, 2011:
http://www.rand.org/pubs/monographs/MG784.html

Vaznis, James, "Thousands Called Dropout Risks," *Boston Globe*, November 29, 2010, p. A1. As of February 17, 2011:
http://www.boston.com/news/education/k_12/articles/2010/11/29/thousands_called_dropout_risks/

Victoria Department of Education and Early Childhood Development, *School Review Guidelines 2010*, Melbourne, October 2009. As of February 17, 2011:
http://www.eduweb.vic.gov.au/edulibrary/public/account/operate/2010_School_Review_Guidelines.pdf

Victoria Department of Education and Training, "Blueprint for Government Schools," last updated January 4, 2006. As of February 17, 2011:
http://pandora.nla.gov.au/pan/20690/20060620-0000/www.sofweb.vic.edu.au/blueprint/default.html

Votta, Peg, research analyst, Rhode Island Department of Elementary and Secondary Education, telephone interview, November 8, 2010.

Wayman, Jeffrey C., Vincent Cho, and Shana Shaw, *First-Year Results from an Efficacy Study of the Acuity Data System*, Austin, Tex.: University of Texas at Austin, November 1, 2009. As of February 17, 2011:
http://edadmin.edb.utexas.edu/datause/papers/Wayman_Cho_Shaw_Acuity_Study_Year_One.pdf

Wayman, Jeffrey, and Sam Stringfield, "Data Use for School Improvement: School Practices and Research Perspectives," *American Journal of Education*, Vol. 112, 2006, pp. 463–468.

Wöbmann, Ludger, Elke Lüdemann, Gabriela Schütz, and Martin R. West, *School Accountability, Autonomy, Choice, and the Level of Student Achievement: International Evidence from PISA 2003*, Paris: OECD, OECD Education Working Paper 13, 2007.

Wong, Manyee, Thomas D. Cook, and Peter M. Steiner, *No Child Left Behind: An Interim Evaluation of Its Effects on Learning Using Two Interrupted Time Series Each with Its Own Non-Equivalent Comparison Series*, Evanston, Ill.: Northwestern University, WP-09-11, 2009. As of February 17, 2011:
http://www.northwestern.edu/ipr/publications/papers/2009/wp0911.pdf