Framework for an Effective Assessment and Accountability Program: The Philadelphia Example

Andrew C. Porter—University of Wisconsin–Madison
andyp@education.wisc.edu
(608) 263-4200
Wisconsin Center for Education Research
1025 West Johnson Street
Madison, WI 53706

Andrew Porter is Anderson-Bascom Professor of Education and Director of the Wisconsin Center for Education Research at the University of Wisconsin–Madison. He has published widely on psychometrics, student assessment, education indicators, and research on teaching. His current work focuses on curriculum policies and their effects on opportunity to learn. He is an elected member and former officer of the National Academy of Education, lifetime national associate of the National Academies, and Immediate Past-President of the American Educational Research Association.

Mitchell D. Chester—Ohio Department of Education
Mitchell.Chester@ode.state.oh.us
(614) 466-3175
Ohio Department of Education
25 South Front Street, Mail Stop 702
Columbus, OH 43215

Mitchell Chester is Assistant Superintendent for Policy Development with the Ohio Department of Education. His research interests include large-scale assessment, accountability systems, school improvement initiatives, and urban education. Chester is the co-author, along with Andrew Porter, of a paper on accountability that was published in the 2002 volume of *Brookings Papers on Education Policy*. Chester authored an article on the use of multiple measures for high-stakes student decisions that will be published in 2003 in *Educational Measurement: Issues and Practices*.

Michael D. Schlesinger—School District of Philadelphia
Mschlesi@phila.k12.pa.us
(215) 299-7906
130 Kenilworth Road
Merion, PA 19066

Michael Schlesinger currently is Director of the Office of Student and School Progress in the School District of Philadelphia. He has worked for over 30 years in the district, holding a variety of positions including teacher, research associate, evaluation specialist, and assistant to the deputy superintendent. He earned undergraduate and master's degrees from the University of Pennsylvania and a doctorate from Temple University. His research interests include state and district accountability systems and school use of achievement data. Schlesinger also teaches courses in program evaluation and educational measurement at two local universities.

Abstract

The purpose of this article is to put in the hands of researchers, practitioners, and policymakers a powerful framework for building and studying the effects of high-quality assessment and accountability programs. The framework is illustrated through a description and analysis of the assessment and accountability program in the School District of Philadelphia.

Executive Summary

The purpose of this article is to put in the hands of researchers, practitioners, and policymakers a powerful framework for building and studying the effects of high-quality assessment and accountability programs. The framework can be used to critique existing programs and analyze how they might best be strengthened, as well as to build high-quality new programs. The framework can also help make sense out of the research literature on the effects of high-stakes testing. What the framework makes clear is that assessment and accountability programs can and do differ in important ways. The framework explains why some programs can be expected to yield negative effects and others positive effects. Knowing how to build high-quality assessment and accountability programs has become a matter of greater urgency with the passage of the No Child Left Behind Act of 2001.

The most important statement for guiding the design of an assessment and accountability program is *Standards for Educational and Psychological Testing*, published in 1999 by the American Educational Research Association (AERA). Earlier versions of these standards have been in place for some time and have become the legal and industry standards for test development and use. Because the standards cover much more than student achievement testing and accountability, in 2000 AERA developed a separate position statement concerning high-stakes testing in preK-12 education. The AERA position articulates 12 conditions that every high-stakes testing program should meet.

Essentially, the standards and the AERA position statement identify the following three criteria that any effective assessment and accountability program should meet:

1. *The assessment and accountability program should provide a good target for student and school effort.* If assessment and accountability can focus effort, they must focus effort in constructive and coherent directions. In light of federal and state accountability requirements, it is essential that district and state accountability efforts present a coherent set of targets to guide school efforts. These expectations should focus educators and students on valued outcomes.
2. *The assessment and accountability program should be symmetrical.* To produce high levels of student achievement, students and schools must work together. No school is so good that it can be successful without students who are motivated and ready to learn. Similarly, even students who are motivated and ready to learn must be provided adequate opportunities to learn worthwhile content. Students from low-income families are especially dependent on school-based opportunities to learn. The assessment and accountability program should include stakes that schools and students share so that both have incentives to improve the same outcomes.
3. *The assessment and accountability program should be fair*. For students, fairness requires that schools provide an adequate opportunity to learn. For schools, fairness requires access to the resources needed to be successful. A fair assessment and accountability program relies on tests that are reliable and valid for the ways in which they are used. Additionally, any inferences about school effectiveness drawn from the program must be consistent and accurate.

We begin our article with a description of the framework's three parts: (a) setting coherent and good targets for instruction; (b) creating an assessment and accountability program that holds both schools and students accountable; and (c) creating an assessment

and accountability program that is fair. Next, we use the Philadelphia assessment and accountability program to illustrate how the framework can be used. The Philadelphia example reveals the complexities of putting together a system with the desired features. Our analysis of the effects of the Philadelphia assessment and accountability program through 1999–2000 leads us to conclude that the program as initially implemented demonstrated some positive effects. We provide a brief discussion of the period beginning with the 2000–01 school year to bring the reader up-to-date on the Philadelphia program. In this regard, it is important to note that many of the planned improvements to the program were not implemented in 2000–01 and 2001–02 due to financial constraints, leadership transition, and the uncertainty caused by the state takeover of the district in December 2001. We conclude the article with lessons learned, pointing toward future directions for improvement.

Education researchers and practitioners hold strong beliefs about the value of high-stakes testing. Some believe that high-stakes testing leads to a dumbed-down curriculum and unfair penalties for students and schools. Others believe equally strongly that, without high-stakes testing, many schools will continue to provide inadequate opportunities to learn for students, especially students from low-income families. We believe that a carefully crafted and continuously refined assessment and accountability program can lead to more effective schools and higher levels of student persistence and achievement on content critical for future success.

Framework for an Effective Assessment and Accountability Program: The Philadelphia Example[1]

Andrew C Porter—University of Wisconsin–Madison
Mitchell D. Chester—Ohio Department of Education
Michael D. Schlesinger—School District of Philadelphia

The purpose of this article is to put in the hands of researchers, practitioners, and policymakers a powerful framework for building and studying the effects of high-quality assessment and accountability programs.[2] The framework draws from the first author's experiences as chair of the technical advisory panels for the assessment and accountability programs in Philadelphia, Missouri, and Ohio, and member of the technical advisory panel in Kentucky; from the second author's experiences developing and implementing accountability systems in the School District of Philadelphia and the State of Ohio; and the third author's ongoing experiences developing and implementing the School District of Philadelphia accountability system. In addition, the framework is grounded in the test standards developed jointly by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) and by the AERA position statement on high-stakes testing (AERA, APA, & NCME, 1999; AERA, 2000).

The framework can be used to critique existing programs and analyze how they might best be strengthened, as well as to build high-quality new programs. The framework can also help make sense out of the research literature on the effects of high-stakes testing. What the framework makes clear is that assessment and accountability

[1] The research in this article was supported by the Wisconsin Center for Education Research, School of Education, University of Wisconsin–Madison, and by the School District of Philadelphia. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Wisconsin Center for Education Research, the Ohio Department of Education, or the School District of Philadelphia.
[2] What follows draws heavily upon Porter and Chester (2002). The framework remains largely unchanged, but the Philadelphia example is updated.

programs can and do differ in important ways. The framework explains why some programs can be expected to yield negative effects and others positive effects. Knowing how to build high-quality assessment and accountability programs has become a matter of greater urgency with the passage of the No Child Left Behind Act of 2001 (2002; reauthorizing the Elementary and Secondary Education Act).

We begin our article with a description of the framework's three parts: (a) setting coherent and good targets for instruction; (b) creating an assessment and accountability program that holds both schools and students accountable; and (c) creating an assessment and accountability program that is fair. Next, we use the Philadelphia assessment and accountability program to illustrate how the framework can be used. The Philadelphia example reveals the complexities of putting together a system with the desired features. Our analysis of the effects of the Philadelphia assessment and accountability program through 1999–2000 leads us to conclude that the program as initially implemented demonstrated some positive effects. We provide a brief discussion of the period beginning with the 2000–01 school year to bring the reader up-to-date on the Philadelphia program (Porter & Chester, 2002). In this regard, it is important to note that many of the planned improvements to the program were not implemented in 2000–01 and 2001–02 due to financial constraints, leadership transition, and the uncertainty caused by the state takeover of the district in December 2001. We conclude the article with lessons learned, pointing toward future directions for improvement.

A Framework for Building an Effective Assessment and Accountability System

The purpose of implementing an assessment and accountability program is to improve student learning of worthwhile content (Educational Testing Service, 2001;

Haertel, 1999). Current levels of achievement in most U.S. urban districts are unacceptably low. Average achievement test results conceal the fact that achievement levels of students of color are substantially lower than those of White students. Everyone agrees that improvements are urgently needed. One potentially important piece of comprehensive reform in urban education is an effective assessment and accountability program.

Assessment and accountability, by themselves, are unlikely to turn around the low levels of student achievement in urban settings (Linn, 1997, 2000). Supports must be put in place to enable students and schools to be successful. We argue that such supports must be an integral part of an effective assessment and accountability program. Nevertheless, high-stakes testing can be a powerful policy lever in a more comprehensive reform initiative (Popham, 2000).[3]

Education researchers and practitioners hold strong beliefs about the value of high-stakes testing. Some believe that high-stakes testing leads to a dumbed-down curriculum and unfair penalties for students and schools (McNeil, 2000; Shepard & Smith, 1989; Smith, 1991). Others believe equally strongly that, without high-stakes testing, many schools will continue to provide inadequate opportunities to learn for students, especially students from low-income families. We believe that a carefully crafted and continuously refined assessment and accountability program can lead to more

[3] Statements of the desired curriculum are usually given in the form of content standards and curriculum frameworks. But these documents, by themselves, have neither provided students and schools with clear targets for instruction nor exerted much influence on teacher content decisions (Porter, 1994). Assessment and accountability serve to focus the attention of schools (e.g., administrators, teachers, and parents) and students on the content to be learned. At the same time as the assessments are intended to focus attention on the targeted content, accountability (rewards and sanctions) is intended to give the target weight, stimulating greater effort on the part of schools and students.

effective schools and higher levels of student persistence and achievement on content critical for future success.

An effective assessment and accountability program has many components. Much more is involved than simply administering a test and adopting a policy that if students and schools do not achieve to a standard they will be punished. The most important statement for guiding the design of an assessment and accountability program is *Standards for Educational and Psychological Testing* (AERA et al., 1999). Earlier versions of these standards have been in place for some time and have become the legal and industry standards for test development and use (Heubert & Hauser, 1999). Because the standards cover much more than student achievement testing and accountability, in 2000 AERA developed a separate position statement concerning high-stakes testing in preK-12 education (AERA, 2000). The AERA position articulates 12 conditions that every high-stakes testing program should meet.[4]

Those opposed to high-stakes testing use the *Standards for Educational And Psychological Testing* and AERA's position statement to critique assessment and accountability programs and find them wanting. For them, the standards and position statement are valuable insofar as they can be used to help put a halt to high-stakes testing. We see the standards and position statement in a different light. For us, they provide the criteria and guidelines for building an effective assessment and accountability program that will strengthen instruction and improve student achievement.

[4] The 12 conditions are (a) protection against high-stakes decisions based on a single test; (b) adequate resources and opportunity to learn; (c) validation for each separate intended use; (d) full disclosure of likely negative consequences of the program; (e) alignment between the tests and the curriculum; (f) validity of passing scores and achievement levels; (g) opportunities for meaningful remediation for students who fail high-stakes tests; (h) appropriate attention to language differences among students; (i) appropriate attention to students with disabilities; (j) careful attention to explicit rules for determining which students are to be

Essentially, the standards and the AERA position statement identify the following three criteria that any effective assessment and accountability program should meet (Porter, 2000):

1. *The assessment and accountability program should provide a good target for student and school effort.* If assessment and accountability can focus effort, as we have argued above, then they must focus effort in constructive and coherent directions. In light of federal and state accountability requirements,[5] it is essential that district and state accountability efforts present a coherent set of targets to guide school efforts. These expectations should focus educators and students on valued outcomes.
2. *The assessment and accountability program should be symmetrical.* To produce high levels of student achievement, students and schools must work together. No school is so good that it can be successful without students who are motivated and ready to learn. Similarly, even students who are motivated and ready to learn must be provided adequate opportunities to learn worthwhile content. Students from low-income families are especially dependent on school-based opportunities to learn.[6] The assessment and accountability program should include stakes that schools and students share so that both have incentives to improve the same outcomes.
3. *The assessment and accountability program should be fair.* For students, fairness requires that schools provide an adequate opportunity to learn. For schools, fairness requires access to the resources needed to be successful. A fair assessment and

---

tested; (k) sufficient reliability for each intended use; and (l) ongoing evaluation of intended and unintended effects of high-stakes testing.

[5] The No Child Left Behind Act of 2001 (2002) is the primary source of federal accountability requirements. Goertz, Duffy, and Carlson Le Floch (2001) provides an analysis of state accountability requirements.

accountability program relies on tests that are reliable and valid for the ways in which they are used. Additionally, any inferences about school effectiveness drawn from the program must be consistent and accurate.

Setting Good Targets

Those who argue against high-stakes testing believe that it will lead to a dumbed-down and narrow enacted curriculum. Of course, in many urban schools, the curriculum is already dumbed down and narrow. Assessment and accountability can set a new and better target for instruction (National Research Council, 2001; Webb, 1997).

Carefully constructed content standards are the first step. These content standards should call for a balanced curriculum, emphasizing students' mastery of key concepts and ideas and the ability to use those concepts and ideas to reason, communicate, and solve novel problems. Students should also master a foundation of facts and skills. The tests used in an assessment and accountability system must be aligned with these ambitious content standards.

There are a number of key dimensions to keep in mind in setting good targets:

- A range of academic subjects should be tested. If, for example, only reading and mathematics are tested, the result may be to push school and student effort toward those subjects and away from other subjects, such as science and social studies.
- The whole school should accept appropriate responsibility for student success or failure. If only selected grades are tested (e.g., 4^{th}, 8^{th}, and 12^{th}), an uneven assessment and accountability system may result, with undue pressure felt at grade levels that are tested and undue freedom felt at those that are not.

[6] Roderick and Engel (2001), in their analysis of accountability in Chicago Public Schools, conclude that the most powerful results occurred in the schools in which educators demonstrated to students their own

- Some content standards are not easily assessed in an on-demand paper-and-pencil test. If these standards are important, alternative assessment procedures should be put in place. Similarly, it has proven more difficult and expensive to assess students' ability to reason, communicate, and use their knowledge than to assess their mastery of facts and skills. A test should be as balanced on these dimensions as the desired curriculum (Baker, 1997).
- As we noted above, achievement is lower for some groups than others. Setting a good target involves accountability for results disaggregated by groups that have too often not been well served by schools.[7]
- A good assessment and accountability program focuses attention on the domain of content desired, as opposed to the specific sample tested. If the same form of a test is used repeatedly, undoubtedly the sample of content represented in that test form will become the target and lead to an inappropriate narrowing of the enacted curriculum (Shepard, 1990). Ideally, each time the test is administered, a different but parallel form should be used.[8]
- Holding schools and students accountable requires specifying standards for performance. For example, if test scores are to be used as the basis for promotion from one grade to the next or graduation from high school, minimum levels of satisfactory performance must be established. Similarly, if schools are to receive

---

investment in student success.

[7] Section 1111(b)(2)(C)(v)(II) of the No Child Left Behind Act of 2001 (2002) includes a core emphasis on subgroup performance. Strong aggregate school and district performance is not sufficient if any of the identified subgroups—racial/ethnic groups, students with disabilities, students with limited English proficiency, and economically disadvantaged students—are not meeting reading and mathematics standards.

[8] At the same time, since it is the test and the accountability associated with testing results that get the attention of educators, students, and parents, it is important to communicate the content being tested. Some

rewards for producing high levels of student achievement, qualifying levels of achievement must be identified (Jaeger, Mullis, Bourque, & Shakrani, 1996; Linn, Koretz, Baker, & Burstein, 1991).[9] Standards that are set too low will be too easily met, and standards that are set too high will be dismissed as unattainable; neither will have a positive impact. We do not argue against setting high standards in the long run, but we do argue against setting unreasonably high standards in the near term. Near-term standards should be set so that they put pressure on schools and students to exceed current levels of achievement. They should then be revisited periodically to see if they need to be revised and set higher.

- Setting targets raises the question, what should be the size and nature of rewards and sanctions for students and schools (Clodfelter & Ladd, 1996; Richards & Sheu, 1992)? We do not know the answer to this question. For students, promotion and retention represent large rewards and sanctions; using test performance as a component of grades is smaller. For schools, reconstitution is a large sanction. For teachers, a modest increase in salary or a bonus is a small reward. The goal is to set

---

representative fraction of items might be released to the public, following each test administration, as a mechanism for communicating the desired content.

[9] There are many different approaches to setting standards for testing programs (e.g., bookmark, known groups, Angolf, Jaeger-Mills). There are different approaches to setting targets for school accountability, as well. One approach is to set a long-term target that all schools are to meet or exceed, determine each school's present level of achievement (i.e., its baseline), then take the distance between each school's baseline and the target and divide it into, say, 12 equal amounts (if each school is to reach the long-term target in 12 years). Because all schools have the same long-range target but different baselines, the lowest performing schools must progress at the fastest rate.

Another strategy for setting targets is to expect the same rate of growth for each school. This approach institutionalizes differences in school achievement, however, since low-performing schools are not expected to close the gap between their achievement levels and those of higher performing schools. Other approaches use (a) criterion-referenced standards (e.g., Virginia requires that at least 70% of students in each school meet the mastery level); (b) norm-referenced standards (e.g., Pennsylvania's Empowerment Act requires that no more than 50% of a district's students score in the bottom quartile of statewide performance); or (c) a combination of norm- and criterion-referenced standards (e.g., California generally requires gains of schools, but those performing above a mastery threshold may receive rewards regardless of their rates of progress).

rewards and sanctions that are sufficient to focus attention on desired content and increase effort.

The federal role in accountability for school performance began in earnest with the 1994 reauthorization of the Elementary and Secondary Education Act (ESEA). The 2001 reauthorization of ESEA, the No Child Left Behind Act (2002), substantially elevated the role of the federal government in prescribing state accountability requirements.[10] Where accountability programs existed before the reauthorization, the new federal requirements have challenged states and districts to merge federal, state, and local requirements in ways that result in coherent expectations and clear signals about what constitutes effective school performance.

Creating an Assessment and Accountability Program That Is Symmetrical

A surprisingly large number of assessment and accountability programs are not symmetrical, either holding schools accountable without holding students accountable or holding students accountable without holding schools accountable. Such approaches are incomplete. If there are no stakes for students, students may not try their best on the test (especially at the upper grades). Even worse, students may not try their best in school. Either way, schools are left in the difficult situation of having to raise the achievement of potentially unmotivated students. Similarly, if students are held accountable, they have the right to expect schools that are doing their best to meet student needs. School accountability may help.[11]

---

[10] The No Child Left Behind Act extended federal requirements to include the implementation of unitary state accountability systems for all public elementary and secondary schools, regardless of Title I funding. The act prescribed in substantial detail a number of design features for setting reading and mathematics achievement targets.

[11] For school accountability, a value-added criterion is appropriate (Meyer, 1996). Given the students with whom a school is working, what value does the students' school experience add to their levels of achievement? Ideally, school value added to student achievement is estimated based on longitudinal data,

Other factors define symmetry, including (a) the degree to which teachers of tested and untested grades share responsibility for student gains, (b) the degree to which schools, teachers, and students own the accountability program results, and (c) the degree to which evaluation criteria for teachers are based on the same data used to set evaluation criteria for students. Our framework calls for symmetry not only of school and student accountability, but also of state and district accountability. If a state holds a district accountable for student performance on state tests, then those state tests should be a part of the district accountability program.

Creating an Assessment and Accountability Program That Is Fair

The key ideas behind a fair assessment and accountability program are opportunity to learn, adequate resources, and the reliability and validity of measures on which inferences are made (Heubert & Hauser, 1999). If students are to be held accountable, then schools must provide them with an adequate opportunity to learn the content for which they are being held accountable (Porter, 1993, 1995). If schools are to be held accountable, then the state and district must provide those schools with the resources they need to be successful. Unfortunately, determining whether a student has an adequate opportunity to learn is not straightforward. Neither is it straightforward to pinpoint exactly what resources are necessary for a school to be successful. Just because opportunity to learn and adequate resources are complicated concepts and difficult to measure in no way reduces their importance, however.

---

with each student serving as his or her own control. This approach requires testing at multiple grade levels, a cost in money and burden. Sometimes value added is estimated based on cross-cohort comparisons (e.g., repeated testing at a single grade level). The assumption here is that each cohort of students at a grade level is a fair comparison group for any other cohort of students at that same grade level. Whatever way value added is estimated, according to our framework the value added should be based on the same tests that are used for student accountability. In short, there needs to be alignment between the school and student accountability programs for them to work in concert.

One way to address opportunity to learn is to hold schools accountable for a period of time before implementing the student accountability component of the program. The idea is that school accountability will bring the enacted curriculum into alignment with the district content standards and assessments; and it is this alignment that is crucial to providing students an adequate opportunity to learn. Of course, instruction must not only be aligned with the content of the standards and assessments; it must be effective as well. Similarly, student accountability can be phased in by setting interim targets for achievement that are beyond current levels but short of levels ultimately desired. If schools are to be increasingly successful, they must be provided qualified staff, instructional materials including technology, high-quality teacher professional development, and reasonable class sizes.

In addition to providing adequate resources for schools and opportunity to learn for students, an assessment and accountability program that is fair makes decisions based on reliable and valid information (Messick, 1993). Actually, reliability is probably not the most useful concept here; the real issue is one of decision consistency. For example, if a student's score falls below a criterion cut point, what is the probability that the student's true achievement level falls below that cut point? Validity is more complicated. If a student is correctly classified as failing the standard and is therefore retained, is retention in the student's best interest in the long run? For example, will the student receive a better and more appropriate educational experience if retained than if promoted? Alternatively, a decision about promotion or retention can be based on a standard of mastery, which would require, for example, that a student demonstrate

mastery of the desired curriculum for kindergarten through fourth grade in order to be promoted to the fifth grade.

When high-stakes decisions are made about students, fairness requires that the decisions not be based on a single test administration or even a single test. Giving students multiple opportunities to demonstrate that they meet a standard improves the probability that a student whose true performance meets the standard will be judged as having met the standard. Providing multiple opportunities to meet the standard also creates a bias in favor of success. That is, the proportion of students certified as meeting the standard even though they did not will exceed the proportion of students certified as failing to meet the standard even though they did. For high-stakes promotion and graduation decisions, the bias should probably be in the positive direction.[12]

For schools, decision consistency is the probability that a school judged to exceed a standard and merit rewards has truly met the standard. Often, a performance index is created against which the level of school performance is tracked over time. Small schools have larger standard errors on such a school performance index, and thus lower decision consistency, than large schools (see, e.g., Kane & Staiger, 2002). This pattern is

[12] When deciding whether to promote students from one grade to the next, the decision might be based on their achievement in several academic subjects. There are two types of standards, conjunctive and compensatory. Under the *conjunctive standard*, a student must meet a separate standard in each academic subject. The reliability of a decision based on this standard is a function of the least reliable of the subject-matter tests. Under the *compensatory standard*, a student's scores are added together to create a total achievement score, and thus high performance in one subject can offset lower performance in another. From the point of view of decision consistency, the compensatory standard is better, since the reliability of the total score is invariably higher than the reliability of the least reliable separate subject-matter test.

An assessment and accountability program that is fair to students will have additional features. Students who are at risk of failing a standard, such as promotion from Grade 4 to 5, should be identified in the earlier grades. Those on a low trajectory should be given an early warning and receive appropriate support. Students with special needs should receive appropriate accommodations (Koretz & Hamilton, 2000; Phillips, 1994; Thurlow, Elliott, & Ysseldyke, 1998). For example, English language learners should be assessed in the language of their instruction. Students with disabilities should be provided accommodations that are construct irrelevant; that is, the accommodation should not change the construct being assessed but rather should allow the student to demonstrate his or her accomplishment on the construct.

due to the fact that the cohort of students in a given year may not be representative of cohorts of students at the same grade in other years. Sometimes, standards are based on averages across multiple years of data to improve the decision consistency for schools.[13]

Finally, to ensure that an assessment and accountability program is as fair as possible, a district must be committed to evaluating the program over time. One part of this evaluation should address consequential validity. Is the program having the intended effects? Is the curriculum getting better over time and coming increasingly into alignment with district content standards? Are decisions about students leading to increased levels of achievement across all subgroups of students? Evaluation of the assessment and accountability program should also take its impact into account. Are acceptable numbers of students being promoted from grade to grade and ultimately graduating from high school? Are reasonable numbers of schools reaching their targets and receiving rewards? Ongoing evaluation of an assessment and accountability program can guard against unintended negative effects and identify ways in which the program can be made more effective.

The Philadelphia School District

This article chronicles accountability in the School District of Philadelphia beginning with the 1995–96 school year. This time frame encompasses two distinct periods—the tenure of Superintendent David W. Hornbeck, through the 1999–2000 school year, and the 2 school years following. Under Hornbeck's leadership, the district

[13] An assessment and accountability program that is fair to schools requires still other features. For example, there should be incentives for schools to test all of their students; a school should not have a higher probability of receiving rewards by not testing its weakest students. Neither should teachers score their own students' tests if the test scores are to be used for school accountability; the pressure for bias may simply be too great. More generally, a district must guard against cheating. Tests must be secure, and the administration protocol must be faithfully implemented.

employed a multifaceted approach to school improvement, a cornerstone of which was the implementation of a comprehensive assessment and accountability program.[14]

We characterize the period of 1995–96 through 1999–2000 as having a focus on implementing and refining a coherent approach to assessment and accountability. In contrast, the 2000–01 and 2001–02 school years were a period of transition, during which this focus was absent. Growing budget deficits, a looming state takeover that was eventually initiated in late 2001, and concerns about test burden contributed to transitory district leadership and lack of commitment to maintaining the assessment and accountability system as begun under Hornbeck. The state takeover resulted in the disbanding of the Philadelphia Board of Education and the forming of the School Reform Commission. Effective with the 2002–03 school year, the School Reform Commission selected Paul Vallas, former Chief Executive Officer of the Chicago Public Schools, to head the School District of Philadelphia. This article chronicles the time up to, but not including, Vallas's tenure.

During the period 1996–2000, the Philadelphia School District served approximately 210,000 students, making it the seventh largest school district in the nation. Eighty percent of the students served were of color, with 65% African American, 11% Hispanic, and 4% Asian American. Almost 80% of the students were from low-

[14] Superintendent David W. Hornbeck's program for improving student learning, entitled *Children Achieving*, employed 10 reform components: (a) establishing academic content standards that signal high expectations for student performance; (b) creating an assessment and accountability program that measures performance against the standards; (c) vesting schools with greater decision-making authority; (d) implementing effective professional development for staff; (e) ensuring that students are ready for school; (f) providing community supports and services; (g) providing up-to-date technology and instructional resources; (h) engaging the public; and (i) ensuring adequate resources. The 10th component required that all of the other 9 components be implemented; that is, the components were not a menu from which to choose. In this article, we focus on the second component—the role of assessment and accountability. It should be noted, however, that substantial commitment and resources were devoted to each of the other components of the *Children Achieving* program.

income families. There were 259 schools organized into 22 clusters, with each cluster consisting of a high school and its feeder schools (about 8–16 schools per cluster). Just over 10% of the students were disabled. Student persistence was a serious challenge, with the district estimating that it failed to adequately educate 75% of its students by the time they were due to graduate, with half actually dropping out.[15]

By the beginning of the 2002–03 school year, the school district looked quite different. Schools were no longer organized into clusters, but rather into nine regions, with each region serving from 22 to 48 schools. Enrollment in public schools had dropped to about 194,000 students. The charter school movement that began in the late 1990s had given rise to 46 charter schools serving approximately 20,000 students.

Under Superintendent Hornbeck, Philadelphia created an accountability system designed to improve (a) student reading, mathematics, and science achievement; (b) attendance of staff and students; (c) student promotion rates; and (d) student persistence rates. The baseline for the system was set in 1995–96. The long-term performance target was that within one student generation—12 years—95% of Philadelphia students would be proficient on citywide measures of district academic standards, would graduate from high school, and would be prepared for further education or workplace employment.

The performance index was the composite measure that summarized (a) a school's scores in reading, math, and science; (b) the school's promotion rate (in Grades K-8) or persistence rate (in high school); and (c) the attendance of students and staff. When calculating the achievement components of each school's performance index, the proportion of students scoring at each performance level (*advanced*, *proficient*, *basic*, and

[15] Remarks delivered by then-Superintendent David W. Hornbeck on September 1, 1998, during a School District of Philadelphia management convocation. Hornbeck stated that "50 percent don't graduate and

below basic) was multiplied by a weighting factor. Because the 12-year goal was to have most students achieve at proficient levels, the *proficient* performance level was weighted 1.0; the *advanced* level, 1.2; and the *basic* level, 0.8. The *below basic* level was divided into three performance bands to be sure that schools were credited for progress at the lower end of the achievement scale. These bands were assigned weights of 0.6, 0.4, and 0.2. Finally, the proportion of students who were *untested* was weighted zero.

Schools had two improvement targets. The first was to close the gap between their baseline and the 12-year goal of a performance index score of 95. Although many permutations will result in a performance index score of 95, schools attaining that score had to have roughly three fourths of their students performing at *proficient* or higher levels across the index components. Schools that improved over the first 2-year period at a rate that, if sustained, would have resulted in attainment of the 12-year goal met the target for that first period. Every 2 years, a new baseline was established, and a new 2-year improvement target calculated. In the second 2-year accountability cycle, for example, schools had to gain one fifth of the distance to the long-range goal, which was then 10 years distant.

The second improvement target was to reduce the proportion of low-scoring and untested students. To meet this target, every 2 years schools were to reduce by 10 points the proportion of students who scored at *below basic* levels or were *untested*. Schools for which the proportion was below 30% were to reduce this proportion by one third. This second target was an incentive for schools to attend to the achievement of all students, not simply those in the middle and upper achievement ranges.

50% of those who do cannot perform 12th-grade work."

After the 1999–2000 school year, the district discontinued using the performance index. Interim district leadership during 2000–01 and 2001–02 was ambivalent about continued investment in the accountability system that began in 1995–96. The remainder of this article distinguishes between (a) the initial period of 1995–96 through 1999–2000 and (b) the two subsequent school years, 2000–01 and 2001–02. The period through 1999–2000 was marked by a systematic effort to incorporate consistent but evolving assessment and accountability initiatives into a comprehensive district approach to improve student achievement. During the 2000–01 and 2001–02 years, the assessment and accountability system devolved into a patchwork of initiatives designed to address the increased state oversight role, growing fiscal concerns, concern about test burden, and the absence of a comprehensive plan for school reform.

Setting a Good Target

Setting a good target is the art of successive approximations. Maintaining the vitality and relevance of the accountability system requires a willingness to learn from experience and make refinements along the way. In this section, we draw on Philadelphia's experience to illustrate the adjustments that were made to improve and refine the target.

1995–96 Through 1999–2000

Multiple measures. Philadelphia's accountability system incorporated both direct and indirect measures of student achievement, as well as noncognitive factors that facilitate student achievement. Through the first 4 years, achievement was measured directly through the Stanford Achievement Test, Ninth Edition (SAT-9). Reading, mathematics, and science achievement were assessed through both multiple-choice and

open-ended formats. Student promotion rates in Grades 1 through 8 and high school persistence rates constituted the indirect measures of achievement. Student and staff attendance were particularly important to Philadelphia because the rates had been low historically, even in comparison to those of other large city districts.

From the start, Philadelphia realized that refinements in the performance index would be necessary to maintain the effectiveness of the accountability system. In particular, the district wanted to reduce its reliance on the SAT-9 by incorporating assessments that provided better instructional targets for teachers and schools.[16] The new assessments were to be systematically aligned with the district curriculum standards and frameworks, require strong communication and problem-solving skills, and reflect the cultural, racial, and ethnic diversity that characterizes Philadelphia's schools.

The first adjustment to the accountability system involved customization of the SAT-9. Working with the test publisher, Harcourt Educational Measurement, the district developed and field-tested reading, mathematics, and science items that were culturally representative. On each form of the SAT-9, the district replaced off-the-shelf items with custom open-ended and multiple-choice items, using appropriate equating methodologies to ensure comparability of scores for the new and old test forms.

An additional refinement was made to the accountability system after the first cycle. For the first 2 years that the performance index was in place, high school persistence was simply a measure of the proportion of first-time 9th graders who graduated within 4 years. In the 3rd year, persistence was expanded to include calculations of the promotion rates of first-time 9th graders to 10th grade and the

proportion of students who graduated in 5 or 6 years. The promotion rates from 9th to 10th grade were included to highlight the high failure rate that occurs in 9th grade and to provide an incentive for schools to focus on 9th-grade success. The 5- and 6-year graduation rates were included to reward schools that successfully engaged nongraduating students beyond 4 years. On-time, 4-year graduation was weighted more heavily in the performance index than 9th-grade promotion or the 5- and 6-year graduation rates.

Further adjustments to the accountability system had been anticipated. Beginning in 1998–99, the district developed new assessments aligned to district academic content standards. Proficiency exams at the middle and high school levels in English, mathematics, and science (e.g., eighth-grade English, Algebra 1, Living Environments) utilized multiple-choice, short-answer, and extended constructed response formats to assess student achievement on district content standards. A system of reading and mathematics assessments in kindergarten through Grade 3 incorporated teacher observations, curriculum-embedded assessments, and on-demand tasks to evaluate student progress and diagnose student strengths and weaknesses. The exams were being phased in, with the first operational administrations occurring during the 2000–01 school year. Beginning with the 2001–02 school year, the district had planned to use proficiency exam results in the performance index, thus reducing the reliance on the SAT-9 and the Pennsylvania System of School Assessment (PSSA) administered by the state.

Accounting for all enrolled students. From the start, the Philadelphia School District took a strong philosophical stand that schools were responsible for the

[16] Among the features of tests that perform as powerful incentives are that they define achievement relative to an external standard, they are organized by discipline and keyed to the content of specific courses, and

achievement of all students and that, with very few exceptions, all students could reach high levels of achievement in core subjects. This philosophy was operationalized by incorporating untested students into the performance index score, by exempting few students from the performance index calculation, and by instituting testing accommodations that were meant to maximize student access to the assessments. Many large-scale testing programs fail to take differences in participation and exemption rates into account when reporting results. Philadelphia went to great lengths to account for all enrolled students in its program.

The *untested* category was included to provide an incentive to focus on all students. The proportion of students who were untested was weighted zero when calculating the achievement components of the performance index. Since the percentage of students who took the tests varied considerably from school to school, the school district did not want schools that were reaching out to test all of their students to appear to have lower scores than schools that tested only their more successful students. No matter how poorly students performed on the assessment, a school received more credit if they were tested than if they were not.

Only 4% of Philadelphia students—the most severely disabled and English language learners who lacked proficiency in both English and their native language—were excluded from testing. To achieve this level of inclusion, the school district permitted a range of test accommodations—for example, the provision of extra time, one-on-one testing in shorter time segments, and real-time translation of vocabulary on the mathematics and science sections.

---

they assess a major portion of what students are studying and are expected to know (Jacobs, 2001).

For students who received instruction in their native languages, Philadelphia used native language assessments. When possible, the school district purchased existing assessments in the two languages for which bilingual programs existed—Spanish and Chinese. When existing native language assessments were not available, tests were translated or native language versions were developed in Spanish and Chinese in parallel with the English language version.

Performance goal and improvement targets. When the accountability system was initially implemented, the SAT-9 performance levels served as the benchmarks for setting the 12-year goal and the 2-year improvement targets. The SAT-9 performance levels had features that supported the district's goal of having Philadelphia students achieve at levels that would allow them to compete with high school graduates anywhere: The performance levels were based on academic content standards developed with the input of national professional organizations; they represented achievement based on judgments about how students should perform; and the distribution of students from the norming sample across SAT-9 performance levels paralleled the distribution of students nationally across the National Assessment of Educational Progress performance levels.

The SAT-9 performance standards probably received more attention from teachers and administrators than other components of the performance index. Although student and staff attendance, Grade 1–8 promotion rates, and high school persistence rates contributed to the performance index score, the test score components were 60% of the total score. Further, most schools attained higher scores on the nontest components, and therefore there was less room for improvement on these components than on the test components.

In the first 2-year accountability cycle (1996–97 and 1997–98), 56% of Philadelphia's schools exceeded both their accountability targets (i.e., their 2-year improvement targets towards the 12-year goal of attaining a performance index score of 95 and the required reduction in the proportion of low-scoring and untested students). Substantial performance gains were also realized in the subsequent 2 years, but the rate of progress slowed: Only 20% of schools exceeded both accountability targets in the second 2-year accountability cycle.

The credibility and value of an accountability system are tied to educator and public perception of the reasonableness and appropriateness of the improvement targets. As the district gained more experience with its targets, it became apparent that the SAT-9 performance levels that drove them were not realistic goals at the time for many Philadelphia schools. High schools, as well as many elementary and middle schools with very low baselines, were not succeeding in meeting their targets—often despite making substantial, consistent progress. Only 2 out of 22 comprehensive high schools earned rewards in the first 4 years, even though many had achieved strong and continuous gains. In part, this was an artifact of keying the performance levels to the SAT-9. Nationally, the *basic*, *proficient*, and *advanced* levels were much more difficult to attain at the higher grades than they were at the lower grades, particularly in mathematics and science. This phenomenon contributed to much lower baseline performance index scores in high schools than in other schools and thus to much higher targets for improvement.

The Philadelphia School District planned to utilize at least two strategies to set new improvement targets. First, the district was moving away from the use of the SAT-9, replacing it with the state PSSA and district-developed assessments. Second, the district

was studying gains achieved by schools over the first 4 years to help determine levels of improvement that were possible.

Promoting the learning of worthwhile content. Setting good targets requires that the academic components of the accountability system represent worthwhile content. This principle has several dimensions, including the range of content areas assessed, the alignment of the assessments with curriculum standards, and the cognitive demands (including students' mastery of key concepts and ideas and their ability to use those concepts and ideas to reason, communicate, and solve novel problems).

The initial adoption of the SAT-9 was based in part on content considerations. SAT-9 reporting was based on rigorous performance standards, and the test assessed science as well as reading and mathematics. In addition, the test publisher permitted the Philadelphia School District to customize the assessments with items that were representative of the cultural backgrounds of its students. The district maximized the assessment's promotion of application of skills, problem solving, and communication by including the open-ended as well as the multiple-choice components of the SAT-9. As stated earlier, the district developed custom assessments that would have reduced its reliance on the SAT-9. These newer assessments included proficiency exams in middle and high school and a K-3 literacy and mathematics assessment system. The custom assessments were systematically aligned with district standards and curriculum. They emphasized higher order processing, including application of skills to solve problems, communication of understandings, and explication of students' reasoning. Such customized high-quality and high-stakes assessments do not come without a price,

however. Philadelphia spent approximately $7.5 million on the development of customized assessments during the 1998–99 and 2000–01 school years.

Rewards and sanctions. The accountability system's leverage was maximized through the application of rewards and sanctions. The rewards and sanctions were designed to motivate schools to focus on those outcomes that were most directly related to achievement and over which they had control. The first accountability cycle (school years 1996–97 and 1997–98) gave educators a concrete grounding in the reality of rewards and sanctions. Of the 259 district schools, 145 received a total of $11.5 million in reward funding for exceeding their 2-year targets; 13 schools with declining performance participated in a quality review process that resulted in the identification of improvement initiatives and the establishment of timelines and benchmarks to be met; and 2 schools were reconstituted (an arbitrator subsequently reversed the reconstitution decision on procedural grounds, although he validated the legality of reconstitution).

Rewards and sanctions affected schools in different—and interesting—ways. Schools that exceeded their targets and earned rewards were quite proud of their accomplishment. They boasted of the achievement in their communications and prominently displayed the banners they were awarded. Philadelphia required that reward dollars, which averaged almost $80,000 per successful school, be reinvested in the school program; they could not be used for salary bonuses. Schools used the dollars to celebrate their success, support new professional development activities, fund new positions, and purchase new equipment and materials.

For some schools, avoiding sanctions was a stronger motivator than exceeding the targets and earning rewards. In schools that came close to meeting their 2-year target

after 1 year, the faculty rallied around a goal of exceeding the target. Schools that were still far from their 2-year target after 1 year were likely to see their goal as avoiding designation as a low-progress school.

2000–01 and 2001–02

Beginning with the 2000–01 school year, the Philadelphia district had planned to add end-of-course proficiency exams to the performance index calculation and to replace some of the SAT-9 reading and mathematics testing with the state PSSA in those subjects. The plan to incorporate PSSA test scores responded in part to Pennsylvania legislation that had raised the stakes for Pennsylvania districts in which more than half of students scored in the bottom quartile on the PSSA. The decision also reflected a desire to reduce the amount of testing because in some grades reading and mathematics were being assessed through both the SAT-9 and the PSSA. As mentioned earlier, however, use of the performance index was discontinued after 1999–2000.

Including PSSA results in the school district accountability system would have permitted Philadelphia to achieve a degree of alignment among the criteria for judging school effectiveness. An interesting sidebar to these overlapping district and state accountability systems is the coherence, or lack thereof, that sometimes resulted. For example, because there was a strong correlation between school and student results on the SAT-9 and PSSA, most schools that realized gains on one assessment made gains on the other. Occasionally, however, this was not the case.

The contrast between the conclusions of the district and state accountability programs was sometimes dramatic. These inconsistencies happened because the district program was based on school gains, whereas the state program was based on school

levels of achievement. For example, in the fall of 2000, 53 Philadelphia schools that received rewards from the district were identified by the state as low performers. Similar inconsistencies have been observed in the state's accountability program, sending contradictory messages to schools. In December 2002, for example, 64 Philadelphia schools were awarded almost $3 million for increases on the state exam through the' School Performance Funding Program of the Pennsylvania Department of Education (PDE). Approximately 1 month later, these same schools were placed on PDE's list of schools that had failed to achieve adequate yearly progress.

Improving the Symmetry

Symmetry has multiple dimensions. Although it primarily concerns the relationship between school and student accountability, it also encompasses the relationship between teachers of grades that are tested and teachers of grades that are not, and the degree to which each group feels responsible for student performance.

1995–96 Through 1999–2000

In the early years of Philadelphia's accountability system, it became apparent that not all educators perceived that they had equal responsibility for school performance. In the first 2 years, the achievement components of the performance index consisted of test scores solely from Grades 4, 8, and 11. Schools placed inordinate emphasis on programs and staff in tested grades, often at the expense of attending to other grades. Beginning with the 3rd year of accountability, the school district expanded the grades that contributed to the achievement components of the performance index. By the 1998–1999 school year, results from at least two grades per elementary, middle, and high school level were included in the performance index.

Initially, Philadelphia's accountability system held schools accountable; accountability for students began with the 4th year of the program. This sequence placed the initial responsibility on schools to improve. Educators unaccustomed to being held accountable for student achievement frequently lamented the fact that students did not have as concrete a stake in the outcomes. Yet these initial years of accountability encouraged educators to rethink what they were teaching, to whom, and how, before individual student stakes were implemented.

Beginning with the 1999–2000 school year, promotion from fourth to fifth grade required students to (a) earn a passing mark from their teachers in reading/language arts, mathematics, science, and social studies; (b) successfully complete a multidisciplinary project; and (c) achieve a modest passing score in reading and mathematics on the SAT-9 or on a "second-chance" assessment that the school district developed. Beginning in the 2001–02 school year, the district planned to use a combination of teacher marks and test score achievement in reading, mathematics, and science to determine eighth-grade promotion and high school graduation.

The district planned to implement new promotion and graduation standards, and would have modified the performance index to incorporate the new student requirements. Promotion and graduation rates would have continued to be components of the accountability system. Proficiency exam scores would have contributed to promotion and graduation decisions and would have been incorporated in the achievement components of the performance index. Had these developments occurred, both educators and students would have had a stake in common measures, therefore providing the symmetry needed for students and teachers to work toward common academic goals.

Another dimension of symmetry concerns the overlap between the criteria by which individual teachers are evaluated and the criteria by which student progress and student achievement are assessed. The Philadelphia district had been developing a pay-for-performance evaluation system that would have rewarded teachers for demonstrating the knowledge and skills associated with effective teaching. To the extent that the criteria by which individual teachers were judged would have contributed to school improvement and student attainment of promotion and graduation requirements, the accountability system would have promoted coherent efforts by schools, teachers, and students.

2000–01 and 2001–02

In the spring of 2002, the test score requirement for Grade 4 promotion was dropped because of the district's inability to provide the necessary funding for supports, such as transition services and extended time, to students in danger of being retained (the other requirements—pass the four major subjects and complete an acceptable project—were kept.) In addition, a moratorium was placed on the requirements for promotion to high school and for graduation as the district went through the transition that led to state takeover.

Under the Vallas administration, a new promotion policy is being planned and will be implemented in Grades 3 and 8 for the 2002–03 school year. The policy is similar to the one the district previously used for Grade 4 but differs in one respect: students whose results fall in a range just below the test score requirement will be able to meet this requirement if they pass muster on five other indicators: (a) grades in reading and math

(A or B), (b) average daily attendance, (c) disciplinary record, (d) homework completion, and (e) teacher-made tests.

Developing a Fair System

Fairness has both programmatic and technical dimensions. Programmatic dimensions concern the opportunities students receive to learn and demonstrate what they know and the resources schools receive to be successful. Technical dimensions relate to the quality of the instruments used and the inferences drawn about what students know and can do and about the status and progress of schools (McLaughlin & Shepard, 1995).

1995–96 Through 1999–2000

Supports for students. One of the most fundamental fairness considerations is whether students have had the opportunity to attain the performance levels for which they are being held accountable. In Philadelphia, adequate opportunities to learn were fostered by implementing the school accountability component before the student accountability component. Individual students were expected to meet more rigorous promotion and graduation requirements only after educators had been held accountable for improving student learning. Because school accountability was implemented first, schools had 4 years to improve instruction and focus on test score measures of school performance that also serve as individual student measures.

The implementation of higher credit accumulation requirements for high school graduation was also improving student opportunity to learn. Still, problems persisted. Some schools had not secured instructors to teach world language classes—yet students had to earn two world language credits to graduate. In an effort to maintain the new graduation standards while not unfairly penalizing students, the school district was

exploring a range of means of securing sufficient world language teachers, including contracting with community colleges and private language instructors.

Another way in which the district fostered adequate opportunities to learn was by phasing in the targets for student achievement. For example, Philadelphia initially held fourth graders responsible for meeting modest test score requirements, while making plans to raise the passing standards in later years. Initially, eighth graders were required to achieve passing grades in English, mathematics, science, and social studies; in future years, test score requirements were to be added. By setting lower initial standards and raising them over time, and by introducing new requirements over time, the district allowed the gradual unfolding of opportunities to learn.

To further improve students' opportunities to learn, the district was identifying students who were at risk of failing to meet the standards at least 1 year before the promotion grades and 3 years before graduation. The district provided students with their first opportunity to meet the test score requirements for promotion and graduation in Grades 3, 7, and 10 and identified students with low grades at the same time. Extended time, intensive instruction, and summer programs were provided to students who were at risk of failing or who had failed the promotion and graduation standards. Philadelphia was also phasing in a kindergarten through Grade 3 literacy and mathematics assessment system designed to identify students who needed additional help. The K-3 assessments were intended to strengthen and align instruction with the requirements for promotion to fifth grade.

Supports for schools. From 1995–95 through 1999–2000, the Philadelphia School District experienced a teacher shortage that was not equally distributed. Certain subject

areas—for example, mathematics, science, and world languages—were harder to staff than others. Also, some schools were viewed as less desirable placements than others. As a result, for example, many middle schools that served poor, minority populations had chronic teacher shortages that they filled with emergency-credentialed or substitute teachers.

The school district had to be diligent in seeking ways to provide adequate support for schools. Within the context of district underfunding, low salaries compared to those in surrounding communities, often-neglected facilities, and large class sizes, there is the danger that imposing accountability for results would exacerbate low teacher and principal morale and commitment. When educators perceive that accountability for results is not balanced by support to achieve targets (e.g., full staffing), the system may foster cynicism rather than motivation.

Accuracy of data. The Philadelphia School District made a number of adjustments to ensure that the data on which decisions were made were reliable and that the decisions themselves were consistent with the decision criteria. At the school level, achievement results from additional grades and tests were added to the performance index to increase the probability that score increases and decreases reflected real improvement or decline, and were not due to cohort effects or measurement error.

To ensure the accuracy of inferences made about the gains and losses of schools, the district calculated the standard error of the performance index score for each school and then used these to calculate standard errors for performance index change scores. The results revealed that the standard error of the change score was never larger than 1.5 performance index points. The standard error analyses were used to create a confidence

band against which the district judged school growth. Schools were not sanctioned when an observed decline was smaller than could confidently be inferred to be a "true" decline.

At the student level, the school district adopted two approaches to maximize decision consistency: multiple measures and a compensatory scoring model. Wherever test scores were a requirement for promotion and graduation, Philadelphia provided students with multiple opportunities and employed multiple assessments to determine whether students met the standard. To be promoted to fourth grade, for example, students were required to achieve a passing score in reading and math on the SAT-9. As a second chance to meet the test score requirement, students were given an individually administered reading test and/or a small group–administered math test.

Test score requirements for promotion and graduation initially were to use a conjunctive approach, under which students had to meet the passing standard on each assessment. The district later adopted a compensatory approach, which allowed stronger performance in one subject to offset weaker performance in another and increased the reliability of the results.

The use of multiple measures resulted in high levels of decision consistency. Fourth graders were not retained for failing to meet the test score requirement unless they failed to meet standards on one administration of the SAT-9 *and* on two administrations of the second-chance test (one administration in June, before the close of school, and a second administration, using a different test form, at the end of the summer). On the fourth-grade second-chance reading assessment, the 1999–2000 reliability index was .91.

2000–01 and 2001–02

As identified earlier, the Philadelphia district postponed the test score requirement for promotion to fifth grade out of concern that adequate opportunities to learn were not provided to students who failed the test score standards. In addition, the challenges of staffing Philadelphia schools with qualified teachers were heightened under the uncertainty that interim leadership, growing deficits, and the threat of state takeover brought. The district secured the services of private-sector recruitment firms to assist with the effort to fully staff schools—an effort that continued to come up short of its goal. Against this backdrop, the district suspended the use of the performance index system and test standards for promotion and graduation.

Complexities and Interdependencies

Creating an effective assessment and accountability system for the Philadelphia Public Schools was not an easy task. The district's approach was to wade right in, doing the best that it could at the time, but with a commitment to continually improve the system. Perhaps not surprisingly, each new feature added to the system and each change had a domino effect, requiring rethinking and adjustments throughout the system. The following are a few illustrations of these complexities and interdependencies.

Increasing high school graduation requirements. At the time of the state takeover, the district was in the process of putting in place increased credit requirements for high school graduation and proficiency exams for many of the core academic courses that students were required to take. To graduate, students were previously required to obtain 21.5 credits, with 4 in English, 3 in math, 3 in science, 3 in social studies, 1.5 in health and physical education, 2 in arts and humanities, and 5 electives. The plan was to

increase the requirement by June 2002 to 23.5 credits, with 4 in English, 4 in math, 4 in science, 3 in social studies, 2 in world languages, 1.5 in health, 2 in art, and 3 electives. One sixth of the final grade in a course was to be based on the new citywide proficiency exams in English, math, and science. At first glance, these planned increases in high school graduation requirements seem straightforward, although perhaps difficult to meet. Unfortunately, implementation proved to be anything but straightforward.

First, the 23.5 credits created a very tight schedule for students, who typically earned only 6 credits per year. This problem led the district to offer block scheduling, which enabled students to complete a course in one semester rather than two. Block scheduling allowed students to complete 8 credits per year, thus introducing greater flexibility.

Second, the proficiency exams were originally thought of as end-of-course exams that would count as one sixth of the final grade in a course. But using the proficiency exams as part of the final grade required a quick turnaround time for scoring and reporting back to teachers. It soon became apparent that the proficiency exams could not be end-of-course exams; rather, they would need to be given before the end of the course if the results were to be available in time to be incorporated into course grades.

To shorten the scoring turnaround time, the district planned to revise the proficiency exams, reducing their length and adopting a largely multiple-choice format. At one point, consideration was given to using only the multiple-choice portion of the test for teacher grades. That approach was rejected as setting a poor target for students and teachers, but that decision meant that students' constructed response answers would have to be scored either by the contractor (in time to return scores to teachers before the end of

the year) or by the teachers. If the latter option had been selected, students' constructed response answers would have been copied, with the originals taken away for external scoring and the copies left for the teacher to score and use as one sixth of the students' grades.

External scoring helps to ensure lack of bias when exam scores are used for high stakes promotion and graduation decisions. This option drives an increase in test cost, however, as contractor scoring services must be secured. Teacher scoring, in turn, requires professional development for teachers to learn to score the exams, which may in fact turn out to be a very powerful and positive intervention. Professional development also drives cost, however, as the collective bargaining agreement in Philadelphia limits administrative control over teacher time. Ultimately, the district decided to secure external contractor scoring of the exams.

A third complication arose from the need to administer the exams approximately 6 weeks before the end of the course. Block-scheduled courses move at twice the pace of courses completed over two semesters. Since the proficiency exams were to be used as a part of the school performance index, how would results from block-scheduled courses be adjusted to be comparable to results from courses completed over two semesters? Clearly, less instruction would have been completed in a block-scheduled course at the time of administration.

Finally, questions came up about how the district should implement the requirement that proficiency exam scores count as one sixth of the teacher's course grade. Ultimately, it was decided to provide information on how each student scored relative to the district and school distribution, and within that context let each teacher decide how to

implement the one-sixth requirement. Obviously, this resulted in differences among teachers in how much the performance examination scores actually counted. Because the high school exams were discontinued after the 2000–01 school year, the district conducted no systematic examination of how teachers used the proficiency exam results when assigning student grades.

Creating a new school performance index. The district had planned to field a new performance index in 2000–01, using the 1999–2000 school year as a baseline. Although the new index was completed, and included PSSA results and new targets, it was ultimately not implemented. We examine the reasons for this outcome at the end of this section. First, however, we focus on the complexities the district encountered in developing the new performance index.

As reported previously, the district had decided to substitute state assessment scores for SAT-9 scores as part of its school accountability program to reduce the total amount of testing time for students and to better integrate the assessment and accountability programs of the district with those of the state. This decision required the development of a new school performance index, however, since state testing was replacing SAT-9 testing at various levels and in various subjects. Creating a new school performance index, in turn, required setting new school targets. All of these changes were sure to create confusion. At the same time, they created the opportunity to improve the initial performance index.

The initial school performance index was a function of students' proficiency levels on the SAT-9, as well as student and staff attendance and student promotion and persistence rates. The approach appeared to keep teachers and administrators focused on

the lower end of the student achievement distribution. Further, the index was not easy for educators and lay people to understand.

The new index was to be based on students' average scale scores across the various tests involved. The new index would have incorporated several advantages over the initial index:

- The index was to have been a function of changes in each student's achievement as measured by scale score, not simply changes in the performance level distribution.
- Calculating the index and determining its statistical properties would have been more straightforward and, it was hoped, would have made the index easier to explain and understand.
- Changing the metric for the performance index would have made clear that the index really was new. In contrast, an index that looked similar to the first index might have been confusing, inviting inappropriate comparisons over time.
- A new index would have allowed the district to move away from reliance on the SAT-9 proficiency levels, which were inappropriately high, especially in the upper grades and in math and science.

In the initial index, the various components, with a couple of exceptions, were to be weighted equally. In fact, they were not. In the new index, each component would have been standardized, with a common mean and variance in the baseline year. This would have resulted in equal weighting or, for example, weighting student and staff attendance .5, as the policy required.

District simulations revealed that the new and old performance indexes were correlated nearly 1.0, and correlations of school gains were a surprisingly high .8. This

meant that schools would have been rank-ordered on the new performance index essentially as they were on the initial index, unless, of course, some schools became more productive than they were previously.

The district carefully thought about how to set targets for school gains on the new index. Three approaches were considered. The first was to translate standards set on the old index into the metric of the new. The second was to regress student performance against year of performance for each school, setting standards in terms of some function of the slopes of these school regressions (e.g., a school's target could be the average slope). A third was to take the slopes and intercepts from the school-by-school regressions, regressing slope on intercept; each school's target would be set as the predicted slope given their intercept.

School targets from 1995–96 through 1999–2000 set the highest standards for the lowest performing schools. This approach reflected a commitment to bringing all students in all schools to the same high standard. Setting each school's target as the average of the rates of gain across schools would have given each school the same target. This approach might have seemed fair to the lowest scoring schools, but it would have been less fair to the highest scoring schools. Further, it would have stepped away from the commitment to bring all students and schools to the same high standard.

One way to take advantage of the strengths of these approaches would be to allow different schools to have different targets in the short run, but, at the same time, to select targets that schools have demonstrated are possible to reach. For schools distant from the mean slope and intercept, the predicted slopes would have large standard errors. Some predicted values might be quite inappropriate.

After careful study, the district decided to use this last approach. The index and the goals were constructed in such a way as to require low-achieving schools to improve more than higher achieving schools, although the growth targets were to be set based on gains that schools with similar achievement levels had attained in the past. An equation for deriving new performance index values from the old index was developed for each type of school (elementary, K-8, etc.). The equations were applied to all of the performance index values earned by schools from 1995–96 through 1999–2000. This provided new performance index values for every school for every year. The historic growth trajectory of each school through 1999–2000 was calculated by regressing student performance against year of performance for each school. This growth trajectory yielded a starting value estimate of 1995–96 and an annual historic growth estimate for each school.

From these data, the district calculated the rate of growth that could be expected as a function of the starting value of a school, based on the past performance of schools. This relationship was to have become the basis of the goals set for subsequent years. Each school's 1999–2000 index score was treated as its new starting value. The predicted rate of growth for schools with that starting value (based on 1995–96 through 1999–2000 gains) was to be set as the annual goal. Thus, each school's goal was to be related to the improvement that schools with similar starting points had demonstrated was possible.

As previously noted, this revised performance index, which included PSSA results and new targets, has not been implemented to date. Four factors delayed and then halted the implementation of the new index. First, the development of the new index and its incorporation into district data systems took longer than initially anticipated. Second, a

volatile mix of governance, financial, and leadership issues was coming to the fore, beginning with the 2000–01 school year, which culminated in the takeover of the district by the Commonwealth of Pennsylvania in December 2001. Third, with the passage of the No Child Left Behind Act, there arose the possibility that the district's new index might be inconsistent with the federal act's accountability requirements, and ultimately with requirements adopted by Pennsylvania to comply with the new federal requirements. Finally, interim district leadership during 2000–01 and 2001–02 was ambivalent about continued investment in the assessment and accountability program. These factors created an environment that derailed the introduction of a revised accountability system.

Minimizing testing burden. One of the high-profile issues in any assessment and accountability program concerns amount of testing time. Valid and reliable tests take time to complete. Offering students multiple opportunities to demonstrate that they have reached a standard also takes testing time. The best estimates of school value added to student achievement require longitudinal data on students, the generation of which, in turn, requires frequent testing.

The Philadelphia School District struggled mightily with the conflict between creating good data for accountability decisions and keeping test burden under control. First, as noted earlier, the district replaced some district testing with state tests to eliminate duplicate testing in the same grade. Second, the district used cross-cohort data, rather than longitudinal data, to estimate school accountability. Third, testing time limits were reduced; unfortunately, there was evidence that, to some extent, tests became difficult to complete in the time allotted. Fourth, the district moved toward less reliance on extended response performance assessments and greater reliance on multiple-choice

formats, which allowed testing a wider range of content in a shorter period of time. But some student achievement is hard to assess with a multiple-choice format. Fifth, the number of pilot items included in each test form was kept to a minimum. Creating equivalent forms of tests for future administration requires embedding pilot items in each assessment. The more pilot items embedded, the longer it takes for students to complete the test. To some extent, this problem can be addressed by using rotated forms. Nevertheless, fewer pilot items were embedded than would have been ideal, which meant fewer items were released to the public because fewer replacement items were available.

The many possible uses of test results—e.g., benchmarking school and district performance against national performance, promoting instructional improvement, informing student placement decisions, judging school progress—cannot be met by one, or even two, tests. The reduction of test burden forces district officials to clarify the purposes and prioritize the uses of the testing program.

During the 2000–01 and 2001–02 school years, the district dramatically altered its assessment program, due in part to funding and test burden considerations. Also contributing to the assessment changes were the lack of focus on assessment and accountability under interim leadership, and a transition period that was fraught with funding and governance concerns. Initially, SAT-9 testing was reduced, with the elimination in 2000–01 of the reading and math tests in Grades 8 and 11, two grade levels already tested with the PSSA. Then, in 2001–02, the Grade 4 reading and math tests were eliminated, along with the open-ended items from the tests for Grades 3, 7, and 10, the three remaining SAT-9–tested grades. Some of these changes were enacted to ease the test burden on Grades 8 and 11, which had been "double-tested" with the SAT-9 and the

PSSA. Other changes were carried out to cut district expenditures; the decision to abandon open-ended items saved the district approximately $1 million annually.

Similarly, the district placed on hold the development and administration of proficiency exams. The high school exams were not administered after the 2000–01 school year. Although the Grade 8 proficiency exams were administered systemwide in 2001–02, there are no plans to administer these exams again. The decision to make proficiency exam results count as a portion of a student's report card grade was not implemented in a systematic and robust way; teachers were provided some guidance but ultimately had a great deal of latitude in determining how to integrate exam results into final grades. As with the SAT-9, the primary reasons for moving away from the proficiency exams were to decrease cost and testing time.

The changes since 1999–2000 likely diminish the quality of Philadelphia's assessment and accountability program. Reduction in tested grades and elimination of open-ended items compromise the assessment targets, encourage curricular narrowing, and diminish incentives for schoolwide ownership of the accountability results. Discontinuation of the development and implementation of proficiency exams represents a lost opportunity to leverage curricular and instructional improvement in middle and high school grades by aligning assessments to course-by-course expectations.

Setting standards for high school graduation and grade-to-grade promotion. As noted earlier, graduation and promotion requirements were scaled back after 1999–2000. With increasing doubt as to whether students were receiving adequate opportunity to learn and whether schools had adequate resources to support students' opportunities, the district decided to back off from the planned phase-in of promotion and graduation

requirements. Before making this decision, however, the district invested much time and thought into the setting of standards.

District simulations of the possible impact of the new standards were sobering. For example, by using historical test data to anticipate the 2001–02 school year results, the district estimated that, were there not an opportunity for students to take a "second-chance" test, less than 50% of fourth-grade students would have met the standard for promotion to fifth, less than 25% of eighth-grade students would have met the standard for promotion to ninth, and less than 10% of seniors would have met the standard for graduation from high school. Not surprisingly, the impact differed by ethnic group, with African Americans and Hispanics the hardest hit. For example, slightly less than 15% of African American or Hispanic students would have met the standard for promotion to ninth grade, and slightly less than 4% would have met the standard for high school graduation.

Clearly, these would have been unacceptable results. The district was caught between wanting to maintain a commitment to high standards, on the one hand, and realizing that its plans would have resulted in unacceptable outcomes, on the other. The district did not want to be seen as watering down its standards. At the same time, the district recognized that students should not be inappropriately penalized by unreasonably high standards. Simulations suggested alternative courses of action, leading the district to consider interim standards.

One alternative would have been to require students to meet the *below basic 3* standard (the highest of the three *below basic* score bands), rather than the *basic* standard on the SAT-9. Another possibility would have been to move from requiring that students

meet the standard in each separate subject (the conjunctive standard) to requiring that students meet the standard in an average across subjects (the compensatory standard). The district explored two approaches to operationalizing the compensatory standard. One was to translate student scores into performance levels and then average across subjects. The other was to translate student performance into normal curve equivalents and then average them. Students who were not tested would have received a zero under either approach. Simulations revealed that a higher percentage of students would have met the standards if (a) they were set at *below basic 3*, (b) they were based on the compensatory method, and (c) they used normal curve equivalents.

In 2000–01, Grade 4 promotion requirements were applied in the same manner as in 1999–2000. To be promoted, students needed to achieve a test score of *below basic 3* on the SAT-9 in reading and math, pass the four major subjects, and acceptably complete a multidisciplinary project. Students who did not meet the test score requirement were given another opportunity to do so by passing a second-chance test, which was administered in the first part of June 2001 and at the end of summer programs in August.

For the 2001–02 school year, the Grade 4 test score requirement was to be raised to *basic*. The board of education approved a compensatory model for the test score requirement; this model stipulated that a student would be considered promoted if (a) the average of the reading and math test scores exceeded the *basic* level and (b) neither score was lower than *below basic 3*. Thus, the promotion requirements for 2001–02 were to include more rigorous test score requirements while keeping the other requirements—pass four major subjects and acceptably complete a multidisciplinary project.

The district was considering initially setting higher standards in elementary schools than in middle and high schools. This approach would (a) reflect the fact that elementary schools were achieving higher levels of performance and (b) recognize that reforms need to be phased in, so that students will have benefited from a high-quality elementary, middle, and high school experience before they are required to meet demanding high school graduation standards. Clearly, the interim standards were to be more demanding than previous standards. Thus, the district would have maintained its commitment to increasing standards, but it would have been implementing the increase in such a way as to allow time for the system to come into alignment instructionally and to ensure that the system did not unfairly penalize students.

Evidence of Effects and Prognosis for Continued Impact

This article presents a framework for designing and implementing high-quality assessment and accountability programs. Such programs should lead to improved instruction and better student achievement. Estimating the impact of the accountability system in Philadelphia is difficult, however. As will become clear, the results are not certain, nor is their interpretation. A controlled experiment was not done; no comparison groups were in place. Further, the program was interrupted after the 1999–2000 school year. Still, judgments must be made. Overall, we conclude that the effects of Philadelphia's assessment and accountability program through 1999–2000 were positive for both student persistence and student achievement. Before presenting our analysis, however, it is important to consider three aspects of estimating effects that make our evaluation difficult.

First, one must estimate what the true changes are in student persistence and achievement since the accountability program was initiated. The SAT-9 achievement test is not only an indicator of student achievement, but also an integral part of the assessment and accountability intervention. As will be seen, there could be spurious gains on the SAT-9, especially since the same form was used repeatedly, thus setting a narrower target than might be ideal. Given a good record-keeping system, such as the one in Philadelphia, statistics on persistence are less susceptible to inflation. Although not proof alone of the validity of Philadelphia's record-keeping system, it is interesting to note that the 4-year, on-time graduation rate for students who should have graduated in June 2002 declined when compared to the 2001 rate. This cohort of first-time ninth graders, who began high school in 1998–99, was the first cohort of students subject to the increased credit requirements, and it was hypothesized that the graduation rate for these students might decline given the more rigorous requirements.

Second, assessing the effects of the program requires deciding what is included as a part of the program and what is not. For example, charter schools were initiated in the district at roughly the same time as the assessment and accountability program. We see these two initiatives as separate. A look at the results for the charter schools from 1995–96 through 1999–2000 suggests that the schools were (a) too few in number to affect district averages and (b) too variable in their performance to have a large average effect on their own. On the other hand, given that fairness is one of our three criteria for a good assessment and accountability program, we do include as part of the program such supports to students as a summer school experience for those who score low on the SAT-9 and such supports to schools as reduced class sizes for students who are retained.

Third, the Philadelphia assessment and accountability program was interrupted, and the commitment to new assessments more closely aligned to district curriculum discontinued.

On the near horizon are a number of important challenges. As the Vallas administration takes over after 2 years of transition, the credibility of the efforts going forward will be measured against the district's recent history with assessment and accountability. The district will need to consider federal and state requirements when setting new targets and combine these with locally valued outcomes in a manner that results in clear and coherent signals. The implementation of new student-level accountability requirements will force the district to revisit issues of standard setting and combining multiple measures to arrive at fair and consistent decisions. With funding a continuing issue, an ongoing concern will be whether schools have the resources, including qualified staff, to provide students with the opportunity to achieve the district standards.

Gains, but slowing improvement. The purpose of implementing an assessment and accountability program is to improve student learning of worthwhile content. Five years of Philadelphia SAT-9 data (from the 1995–96 school year baseline scores through the 1999–2000 school year results) reveal improved test scores at the same time that increased numbers of students were participating in the testing program (Table 1). For example, in fourth-grade reading, the percentage of students scoring at the *basic* level or above increased from 43.5% to 59.6%. In fourth-grade mathematics, the percentage increased from 34.9% to 51.5%. Advances toward the long-term goal of having most students achieve at *proficient* levels were also realized. From the baseline year through

1999–2000, the percentage of students achieving at *proficient* levels districtwide improved from 7.7% to 12.2%. On the other hand, although the rate of improvement was substantial, fewer than one in eight students had achieved *proficient* performance by 1999–2000. SAT-9 testing was discontinued in 2000–01 so no SAT-9 results are available beyond 1999–2000.

Table 1

Proportion of Philadelphia Students Scoring* Basic *or Higher on the Stanford Achievement Test, Ninth Edition

Reading

Grade	1995–96	1996–97	1997–98	1998–99	1999–2000
4	43.5%	51.3%	56.9%	57.6%	59.6%
8	48.7%	55.0%	59.6%	62.5%	60.5%
11	25.9%	34.9%	34.1%	37.2%	38.7%

Mathematics

Grade	1995–96	1996–97	1997–98	1998–99	1999–2000
4	34.9%	44.4%	47.7%	50.0%	51.5%
8	21.1%	24.5%	31.6%	30.7%	28.4%
11	12.0%	14.8%	16.5%	15.8%	17.2%

Authors' Calculations

Over the same period, assessment participation rates increased substantially, from 71.5% to 88.1%. The strongest increases in participation occurred in high schools (from 51% to 78%) and among special education students (from 31% to 70%) and English language learners (from 56% to 85%). This result is significant since in large-scale testing programs, increases in the proportion of participating students are typically associated with declines in achievement.

From 1995–96 to 1999–2000, on-time, 4-year graduation rates increased from 49% to 57%, and promotion rates improved from 85% to 93%. Both staff and student attendance also improved over this period. For example, the proportion of staff attending at least 95% of the time increased from 54% to 63%.

Some may question whether the SAT-9 gains represent real achievement gains. Unfortunately, these results are on repeated uses of the same form of the SAT-9, so there probably is some inflation of gains that would not be found on a new form (Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998). Koretz, McCaffrey, and Hamilton (2001) concluded that validating gains under high-stakes conditions is not currently possible, though, until methodologies are developed, they recommend examination of changes on external measures. Fortunately, in Philadelphia, the state test presents just such an opportunity.

As seen in Table 2, Philadelphia's advances on the SAT-9 were also seen on the state test scale scores, with mean 1300 and standard deviations 200. District- and school-level performance generally improved on the PSSA from 1995–96 through 1999–2000, despite the fact that the test was not being used during that period for student or school accountability. Districtwide reading and mathematics scale scores improved in each grade except 11, where the participation rate rose dramatically over time. The PSSA reading and mathematics scores improved over the 4 years in 9 out of every 10 Philadelphia schools. Generally, PSSA scores continued to improve for the years 2000–01 and 2001–02, though not as clearly in math as in reading. Of course, the PSSA results were high-stakes in those years due to state sanctions.

Table 2

Mean Philadelphia Scale Score on the PSSA

Reading

Grade	1995–96	1996–97	1997–98	1998–99	1999–2000	2000–01	2001–02
5	1090	1110	1090	1120	1140	1140	1150
8	1080	1140	1120	1130	1120	1130	1140
11	1160	1140	1140	1140	1130	1180	1170

Mathematics

Grade	1995–96	1996–97	1997–98	1998–99	1999–2000	2000–01	2001–02
5	1110	1130	1140	1140	1140	1150	1150
8	1070	1110	1120	1120	1130	1150	1170
11	1170	1130	1120	1140	1160	1190	1180

Authors' Calculations

To what should the increase in student persistence and achievement be attributed? Would the same level of improvement have been achieved without implementation of the assessment and accountability program? We find no alternative explanation for the gains that is as compelling as the district's aggressive and ambitious assessment and accountability program, together with its capacity-building initiatives, including targeted summer school and reduced class size. At the same time, Corcoran and Christman (2002), in a separate study of Philadelphia school reform, conclude that although progress has been made, "the high hopes that greeted this ambitious reform were not fulfilled" (p. 37).

Two factors are important to watch over the next few years: (a) the need to turn early achievement gains into continuous program improvement and (b) the limitations of low funding levels. Both are related to the fact that although gains were achieved in each of the first 4 years of Philadelphia's accountability program, most of the gains occurred

early, and the rate of improvement slowed in successive years. Both are also related to the changes in district policies.

T Strong early gains followed by slowing improvement is a pattern that has been observed in other large-scale testing programs. In the early years of the program, schools learn to signal the importance of the tests to teachers, students, and parents, create optimal testing environments, and ensure that students are familiar with test formats. Once these factors are optimized, the hard work of upgrading curriculum and improving instructional programs begins.

There is evidence that curricular and instructional improvement is occurring in the Philadelphia district where it was not happening before, although program improvement is not consistent across the district. In many schools, students are being assigned work that requires them to communicate their understandings orally and in writing and to apply their understandings to solve real-world problems. Teachers within and across grades are beginning to collaborate to develop curriculum and design instructional programs. The schools where this work is happening systematically are the exception, however, and not the rule. These findings of changes in instructional practices are based on a number of studies, some conducted by the district, but most conducted by researchers outside the district (see, e.g., Chester, Orr, & Christman, 2001; Consortium for Policy Research in Education, Research for Action, & OMG Center for Collaborative Learning, 1998). Unfortunately, there are many holes in the set of studies, leaving us less knowledgeable about assessment and accountability's effects on instructional practices than we would like.

The fact that Philadelphia's 2000–01 and 2001–02 PSSA test scores continued to rise in most grades and subjects (Table 2) provides promise that the reforms begun in the mid-1990s continue to have an impact on student achievement. According to a Standard and Poor's (n.d.) analysis, Philadelphia schools exhibited consistent and significant gains on state math and reading tests—gains that exceeded the statewide improvement in achievement—over the academic years 1996–97 through 2000–01.

The challenges of coherent capacity building. One impediment to policy coherence in Philadelphia has been the range of initiatives implemented by the central office. Although the stream of policy initiatives may have emanated from a framework that was coherent to its architects, it is rare to find a school that implemented them coherently. The pace of change was frenetic, and the quantity of initiatives that were introduced (e.g., small learning communities, new curriculum standards, new curriculum frameworks, comprehensive support process, school improvement planning, multidisciplinary project requirements, service learning project requirements, adaptive instruction) was too large for substantial organizational learning to occur. Few schools were able to incorporate the many new district initiatives in a manner that resulted in increased organizational capacity. Most schools struggled to comply with the initiatives. The changes in administration and governance since 1999-2000 have not only slowed the pace of reform, but have also probably exacerbated the lack of coherence among district initiatives.

Funding and governance. Another factor that bears watching is the intersection of accountability, inadequate funding, and state oversight. Low teacher and administrator salaries, high turnover of teachers and principals, old and poorly maintained facilities,

and large schools and classes are artifacts of the Philadelphia School District's underfunding. The perception that accountability exists without matched support lowers morale and commitment. This is a conundrum that the district has long recognized: Philadelphia needs to show results before it can expect additional state support, but without additional state support, there is a ceiling on the amount of improvement the district is able to attain.

After the state takeover in late 2001, the School Reform Commission enacted a series of dramatic changes in the district, including privatizing the management of 45 low-performing schools. The new index scores and targets were included in each Education Management Organization (EMO) contract, with the proviso that once the No Child Left Behind accountability provisions were operationalized, the index and targets would be revised, and the EMOs would be held accountable for these new targets.

Summary and Conclusions

Student achievement in Philadelphia was low in 1995–96. Admittedly, large percentages of Philadelphia students came from low-income families; some might expect their achievement to be low. At the same time, it was clear that the opportunities to learn available to Philadelphia students could have been better. The content of instruction provided to students could have been more ambitious, focused on key ideas in the academic subjects and balanced between communicating, reasoning, and problem solving, on the one hand, and mastery of facts and basic skills, on the other. Teachers could have accepted more responsibility for student achievement. And students could have tried harder. In short, the Philadelphia school system was very much in need of

reform when Superintendent David Hornbeck took the reins, launching the assessment and accountability program.

Many argue that a high-stakes assessment and accountability program dumbs down the curriculum. This can be true, of course, and has, in some cases, been documented (Smith, 1991). But an assessment and accountability program that pursues the three goals of setting a good target, making accountability symmetrical for schools and students, and achieving fairness in implementation is unlikely to dumb down the curriculum. Further, one must take into account a curriculum's starting point when initiating and evaluating an assessment and accountability program. In the case of Philadelphia, improvement clearly was needed. Doing nothing was not an option.

Some criticize high-stakes assessment and accountability for extinguishing students' intrinsic motivations to learn. This outcome also seems possible. On the other hand, if intrinsic motivation is already low, perhaps a little extrinsic motivation can help. The introduction of the accountability system jarred the prevailing culture and created opportunities for the board of education and educators to rethink the manner in which resources were allocated, the use of time, and student grouping practices. Based on the evidence, we conclude that the assessment and accountability program in Philadelphia had positive effects on the quality of instruction offered to students, and in turn, on the persistence and achievement of students.

Although the results from the assessment and accountability program from 1995–96 through 1999–2000 are encouraging, the program was curtailed in 2000–01. The possibility remains that the district will restart the accountability system. The new CEO, Paul Vallas, demonstrated his commitment to accountability mechanisms during his

tenure in Chicago. If the assessment and accountability program is renewed, whether in its past format or a new configuration, we are concerned about capacity for instructional improvement. All too often, early gains in student achievement hit a wall after the first 3 or 4 years of reform. Once the most accessible inefficiencies have been corrected, will the assessment and accountability program in Philadelphia bring about continued improvements in the quality of instruction and student achievement? Will future resources be adequate to support continued improvements in instruction? Professional development, improved materials, summer programs, and reduced class size all cost substantial amounts of money. In short, will the assessment and accountability program be fair?

We have organized our analysis of the Philadelphia assessment and accountability program around the framework of setting a good target, achieving symmetry for students and schools, and fostering fairness. In many ways, the Philadelphia assessment and accountability program had significant strengths when judged against these three goals. Still, there were weaknesses, as well. Designing and implementing an assessment and accountability program for an urban school district require a commitment to continuous improvement. We are convinced that the perfect system can never be realized. We are equally convinced that useful systems can be put in place and that they can be improved over time. That is exactly what Philadelphia was doing through 1999–2000.

With regard to setting a good target, many positive outcomes of the Philadelphia assessment and accountability program can be cited. The district's movement away from the SAT-9 and toward greater reliance on proficiency exams aligned with district content standards was a case in point. The district's commitment to testing in multiple subjects

and multiple grades was another. The performance index gave schools incentives to test all students. Even for the SAT-9, the district had replaced some items with culturally inclusive items, thus tailoring the test to the Philadelphia context.

But there are weaknesses in the targets that were set as well. First, the same form of the SAT-9 was used repeatedly. This practice undoubtedly produced spurious gains in student achievement, though some of the gains were real, as well. Second, although the district struggled to set realistic targets for student achievement and school improvement, as of 1999–2000 the targets were set too high. To the district's credit, it was working hard to revise these standards. Third, the emphasis on moving students out of the *below basic* category on the SAT-9 may have taken attention away from improving student achievement at higher levels. Fourth, there is some reason to fear that testing only in selected grades created difficulties in getting the whole school, all teachers at all grades, to accept responsibility for student achievement. Fifth, as state and federal accountability requirements, including performance targets, gain salience, there is a need to incorporate them in a way that results in clear signals and expectations.

With regard to the goal of symmetry, Philadelphia is to be congratulated for holding schools accountable first, and then phasing in student accountability to make the program symmetrical. By 1999–2000, after 4 years of school-level accountability that was linked to SAT-9 performance, student accountability also linked to SAT-9 performance was phased in at Grade 4. Since the 2000–01 school year, however, accountability has been most salient for schools, with student-level requirements put on hold and the state takeover resulting in the assignment of 45 low-performing schools to education management organizations.

Similarly, with regard to the goal of fairness, there were both pluses and minuses in the Philadelphia assessment and accountability program. First, the district was committed to providing the support it believed necessary for schools to succeed with students. When the support was lacking, the district pulled back on the accountability expectations. Second, the district's decision to phase in student accountability after first instituting school accountability likely improved opportunities to learn before holding students accountable. Moreover, student accountability provided students multiple attempts to reach a standard on a particular test, as well as opportunities to take different "second-chance" tests. In the interim, students were required to attend summer school to improve their knowledge and skills. Finally, throughout the implementation process, the district expressed concern for the accuracy of its data and worked to bring accuracy to acceptable levels.

But there is more work to be done in creating fairness. The most important issue is making available the resources schools and teachers need to be successful. At the heart of the problem is attracting qualified teachers, who are in notoriously short supply, especially in urban settings. Philadelphia schools have lower salaries, larger classes, and older facilities than most suburban districts. In addition, limited funding and the climate created by the state takeover are likely to exacerbate the challenge of staffing Philadelphia classrooms with high-quality teachers.

Designing and implementing an assessment and accountability program often requires a careful balancing act. For example, tensions arise (a) between the need to have assessments over time that are carefully calibrated, one to another, so that trends can be monitored, and the need to improve the assessments, based on experience and shifting

content priorities; (b) between the desire to include multiple-choice items in assessments—because such tests take less time, cost less, are more reliable, and are easier to equate over time—and the recognition that multiple-choice items may not measure the same content dimensions as performance assessment tasks; (c) between the desire to use assessments to leverage changes in classroom practice and the reality of working with assessments, at least initially, that do not have the best psychometric properties; and (d) between the need to use an assessment on which students are motivated to perform their best and the difficulty of meeting the psychometric and opportunity-to-learn requirements for a high-stakes test for students.

The Philadelphia School District recognized these tensions and attempted to take reasonable and responsible positions on each. To evaluate the assessment and accountability program, in 1998–99 the school board created an external Accountability and Assessment Advisory Panel. Initially, the school board wanted a panel that would come in, evaluate the program, submit a report, and leave. Wisely, we believe, the board decided instead to put in place a panel that would work with it over time, with the goal of building and implementing the best assessment and accountability program possible. Fortunately, the panel and the school district formed a good working relationship. The first author of this article chaired the panel. The second and third authors were the main contacts in the district for the panel. We enjoyed working together and believe that our collaboration resulted in an assessment and accountability program that was benefiting the education of students in Philadelphia. The panel has not reconvened since the 2000–01 school year.

We end with one final observation: Turnover in district superintendents, although common, is debilitating. As one extreme, the Kansas City, Kansas, School District has had 20 superintendents in 30 years. This revolving-door style of leadership makes sustained school reform impossible.

The person who initiated the assessment and accountability program in Philadelphia, David Hornbeck, is no longer superintendent. Since his departure in 2000, the district replaced the superintendent with a chief executive officer and chief academic officer. An interim CEO was appointed and served until 2001, when the Pennsylvania initiated a takeover of the district, disbanded the school board, and appointed a school reform commission. The commission hired Paul Vallas as the new CEO, effective with the 2002–03 school year.

Vallas demonstrated his commitment to accountability and a hard-line approach to less successful schools during his tenure as CEO of the Chicago Public Schools (see, e.g., Hess, 2002). We can only hope that, as CEO of the Philadelphia School District, he will be able to successfully continue the difficult and challenging work of employing assessment and accountability to promote more effective instruction and improve the achievement of students.

References

American Educational Research Association. (2000). *AERA position statement: High-stakes testing in preK-12 education.* Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Baker, E. L. (1997). Model-based performance assessment. *Theory into Practice*, *36*(4), 247–254.

Chester, M. D., Orr, M., & Christman, J. (2001, April). *Consequential validity of Philadelphia's accountability system: Triangulating four years of multiple sources of evidence.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Clotfelter, C. T., & Ladd, H. F. (1996). Recognizing and rewarding success in public schools. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 23–63). Washington, DC: Brookings Institution.

Consortium for Policy Research in Education, Research for Action, & OMG Center for Collaborative Learning. (1998). *Children achieving: Philadelphia's education reform progress report series 1996–97*. Philadelphia: Author.

Corcoran, T., & Christman, J. B. (2002). *The limits and contradictions of systemic reform: The Philadelphia story*. Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Educational Testing Service. (2001). *Using assessments and accountability to raise student achievement*. Princeton, NJ: Author.

Goertz, M. E., Duffy, M. C., & Carlson Le Floch, K. (2001). *Assessment and accountability systems in the 50 states: 1999–2000* (CPRE Research Report Series RR-046). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, *18*(4), 5–9.

Hess, G. A., Jr. (2002). Accountability and support in Chicago: Consequences for students. In D. Ravitch (Ed.), *Brookings papers on education policy*. Washington, DC: Brookings Institution.

Heubert, J. P., & Hauser, R. M. (Eds.) (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Jacobs, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis, 23*(2), 99–121.

Jaeger, R. M., Mullis, I. V. S., Bourque, M. L., & Shakrani, S. (1996). Setting performance standards for performance assessments: Some fundamental issues, current practice, and technical dilemmas. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessments* (pp. 79–115). Washington, DC: U.S. Government Printing Office.

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 235–269). Washington, DC: Brookings Institution.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: RAND.

Koretz, D. M., & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND.

Koretz, D. M., & Hamilton, L. S. (2000). Assessment of students with disabilities in Kentucky: Inclusion, student performance, and validity. *Educational Evaluation and Policy Analysis, 22*(3), 255–272.

Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001, April). Toward a framework for validating gains under high-stakes conditions. In D. M. Koretz (Chair), *New work on the evaluation of high-stakes testing programs*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14–16.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Linn, R. L., Koretz, D., Baker, E. L., & Burstein, L. (1991). *The validity and credibility of the achievement levels for the 1990 National Assessment of Educational Progress in Mathematics* (CSE Technical Report No. 330). Los Angeles: University of California, Center for the Study of Excellence.

McLaughlin, M. W., & Shepard, L. A. with J. A. O'Day. (1995). *Improving education through standards-based reform: A report by the National Academy of Education*

Panel on Standards-Based Education Reform. Stanford, CA: National Academy of Education.

McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Phoenix, AZ: Oryx Press.

Meyer, R. H. (1996). Value-added indicators of school performance. In E. A. Hanushek & D. W. Jorgensen (Eds.), *Improving American schools: The role of incentives* (pp. 197–223). Washington, DC: National Academy Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment* (P. Pellegrino, N. Chudowsky, & R. Glaser, Eds.). Washington, DC: National Academy Press.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7*(2), 93–120.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders*. Needham, MA: Allyn & Bacon.

Porter, A. C. (1993). School delivery standards. *Educational Researcher, 22*(5), 24–30.

Porter, A. C. (1994). National standards and school improvement in the 1990s: Issues and promise. *American Journal of Education, 102*(4), 421–449.

Porter, A. C. (1995). The uses and misuses of opportunity-to-learn standards. *Educational Researcher, 24*(1), 21–27.

Porter, A. C. (2000). Doing high-stakes assessment right. *School Administrator, 11*(57), 28–31.

Porter, A. C., & Chester, M. (2002). Building a high-quality assessment and accountability program: The Philadelphia example. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 285–337). Washington, DC: Brookings Institution.

Richards, C. E., & Sheu, T. M. (1992). The South Carolina School Incentive Reward Program: A policy analysis. *Economics of Education Review, 11*(1), 71–86.

Roderick, M., & Engel, M. (2001). The grasshopper and the ant: Motivational responses of low-achieving students to high-stakes tests. *Educational Evaluation and Policy Analysis, 23*(3), 197–227.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice, 9*(3), 15–22.

Shepard, L. A., & Smith, M. L. (1989). *Flunking grades: Research and policies on retention*. Philadelphia: Falmer Press.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8–11.

Standard & Poor's. (n.d). *The greatest gains: Making consistent and significant improvements: A study of Pennsylvania schools and districts, 1997–2001*. Retrieved January 30, 2003 from http://www.ses.standardandpoors.com/pdf/pa_consistent_gains.pdf

Thurlow, M. L., Elliott, J. L., & Ysseldyke, J. E. (1998). *Testing students with disabilities: Practical strategies for complying with district and state requirements*. Thousand Oaks, CA: Corwin Press.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (National Institute for Science Education and Council of Chief State School Officers Research Monograph No. 6). Washington, DC: Council of Chief State School Officers.