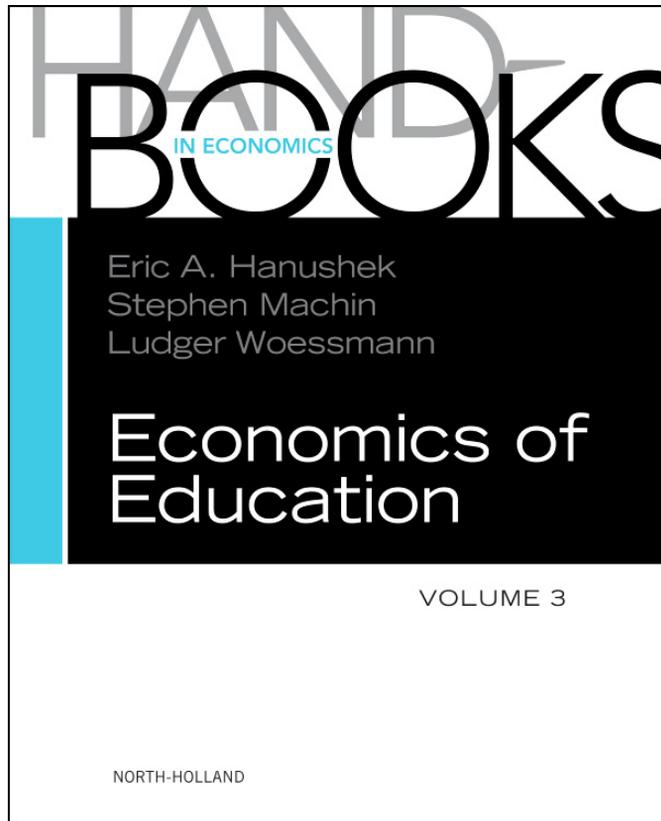


**Provided for non-commercial research and educational use only.
Not for reproduction, distribution or commercial use.**

This chapter was originally published in the book *Handbooks in Economics*, Vol. 3, published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues who know you, and providing a copy to your institution's administrator.



All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

<http://www.elsevier.com/locate/permissionusematerial>

From: David Figlio and Susanna Loeb, School Accountability. In Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, editor: *Handbooks in Economics*, Vol. 3, The Netherlands: North-Holland, 2011, pp. 383-421.

ISBN: 978-0-444-53429-3

© Copyright 2011 Elsevier B.V.
North-Holland

CHAPTER 8

School Accountability

David Figlio* and Susanna Loeb**

*Northwestern University, CESifo, and NBER

**Stanford University and NBER

Contents

1. Introduction	384
2. The Rationale for School-Based Accountability	386
3. The Nature of Accountability	388
3.1 The consequences of accountability	388
3.2 Scope and domains of accountability indicators	389
3.3 Measuring school performance	391
3.4 Exclusions	394
3.5 Subgroup identification	395
3.6 Time considered for rating schools	396
4. Accountability Might Not Improve School Performance	397
4.1 Improving measured, but not generalizable, achievement	397
4.2 Strategic behavior	399
4.3 Failure or inability to respond to incentives	401
5. Evidence on Student Outcomes	402
5.1 Differential effects of accountability	410
5.2 Early versus late adopters of accountability	411
5.3 Size and policy significance of the estimated effects	412
6. Accountability and Teacher Labor Markets	412
7. Directions for Future Research	416
References	417

Abstract

School accountability—the process of evaluating school performance on the basis of student performance measures—is increasingly prevalent around the world. In the United States, accountability has become a centerpiece of both Democratic and Republican federal administrations' education policies. This chapter reviews the theory of school-based accountability, describes variations across programs, and identifies key features influencing the effectiveness and possible unintended consequences of accountability policies. The chapter then summarizes the research literature on the effects of test-based accountability on students and teachers, concluding that the preponderance of evidence suggests positive effects of the accountability movement in the United States during the 1990s and early 2000s on student achievement, especially in math. The effects on teachers and on students' long-run outcomes are more difficult to judge. It is also clear that school personnel respond to accountability in both positive and negative ways, and that accountability systems run the risk of being counter-productive if not carefully thought out and monitored.

JEL classification: I21 I28 L15

Keywords

Accountability
Achievement
Education
NCLB

1. INTRODUCTION

School accountability—the process of evaluating school performance on the basis of student performance measures—is increasingly prevalent around the world. In the United States, accountability measures have become a centerpiece of both Democratic and Republican federal administrations' education policies, following the movement by individual states to introduce accountability systems in the 1990s. Centralized reporting of school-wide examination scores has occurred for over two decades in the United Kingdom (Burgess et al., 2005) and in Chile (Mizala, Romaguera and Urquiola, 2007). Most Western European and Latin American countries have had national assessment systems and some semblance of reporting for over a decade, and a new system is being unveiled at the time of writing in Australia.

Accountability in education is a broad concept that could be addressed in many ways, such as using political processes to assure democratic accountability, introducing market-based reforms to increase accountability to parents and children, or developing peer-based accountability systems to increase the professional accountability of teachers (and now, especially following the “Race to the Top” initiative of President Obama, using similar tools to evaluate, reward, and sanction individual teachers as well). The most commonly considered definition of accountability involves using administrative data-based mechanisms aimed at increasing student achievement.

We focus in this chapter on accountability systems that generate explicit or implicit rewards and/or sanctions to schools on the basis of aggregate student performance on standardized tests. We concentrate on accountability systems in which the school is the unit of analysis, rather than systems that demand higher standards of students or those that evaluate and compensate teachers based on their students' performance. The rewards and sanctions associated with accountability systems could be explicit, such as bonuses for educators in schools considered to be excellent or threats of restructuring or closing low-performing schools, and they could also be implicit—operating less through direct action by central decision-makers and more through community pressure on schools to improve. Thus, school accountability incentives can work through direct government action or through the provision of information. Accountability ratings help community stakeholders observe school performance. For example, school accountability ratings may affect the housing market in a community (Figlio and Lucas, 2004) and could

influence private donations to schools (Figlio and Kenny, 2009). Figlio and Ladd (2007), in a previous survey of the school accountability literature, lay out many of the issues regarding school accountability; the present chapter summarizes these concepts and provides more detailed evidence on the effects of school accountability policies and programs.

The school accountability systems that we consider operate primarily within the traditional public school system and are based in large measure on student testing (Elmore, Abelman, and Fuhrman, 1996; Clotfelter and Ladd, 1996; Carnoy and Loeb, 2002; Hanushek and Raymond, 2003). The most famous of these systems is the federal No Child Left Behind Act (NCLB), which became law in the United States in 2002. NCLB requires states to test students in reading and mathematics in grades three through eight, as well as in one high school grade. NCLB also requires science testing in at least one grade per traditional school level.¹ In addition, it requires states to determine what it means to be proficient on the state assessments and to evaluate schools based on whether their students, in aggregate and by subgroup, are progressing adequately toward an ultimate goal of 100% proficiency by 2014. This law was preceded in the United States by the Clinton Administration's 1994 Goals 2000: Educate America Act, though NCLB is more focused on school accountability than on standards, and exerts a stronger federal role in education policy than did previous accountability laws. Many U.S. states also had accountability initiatives in place before NCLB and even before the Goals 2000: Educate America Act, partially because the U.S. impetus for accountability emerged from a 1989 meeting between President George H.W. Bush and the set of state governors. Prior to NCLB, 45 states published report cards on schools and 27 rated schools or identified low-performing ones (Education Week, 2001). The state-based accountability movement pre-NCLB was strongest in the southern United States.²

The United States has not been alone in the introduction of school accountability; as noted above, English schools' performance has been reported since 1988. The most-developed accountability systems operate in the United States (both federally and at the individual state level), England, and Chile, and these are the systems on which the overwhelming majority of academic research has been based. Other countries vary in the degree to which they assess students, and whether they publicly report scores at the school level. In Latin America, for instance, scores are publicly reported in Brazil, Chile, Colombia, and Mexico (in some regions), and most

¹ Science testing is not, however, currently used for accountability purposes in every state.

² The concentration of early accountability efforts in the south was motivated by southern governors' desire to foster economic development; the fact that for historical reasons state governments in the south typically had more authority over education finance and governance than in other parts of the country and hence were in a position to impose accountability; and that teachers unions, which might have opposed accountability programs, were not a major factor in most southern states.

governments record school-level reports for internal purposes (Vegas and Petrow, 2008). Countries vary considerably in terms of the quality and fidelity of assessment practices: Brazil, Chile, Colombia, and Mexico rank at the top in Latin America in their present capacity for assessment (Ferrer, 2006). Other countries, such as Costa Rica, Cuba, Guatemala, and Panama, though they regularly test students and measure aggregate scores at the school level for internal purposes, rank much lower along the same metrics. Nonetheless, it is clear that many countries are developing capacities for conducting accountability systems, and the methodological and conceptual issues touched upon in this chapter will be important in any system design and implementation.

2. THE RATIONALE FOR SCHOOL-BASED ACCOUNTABILITY

The current school-based accountability movement emerged out of a desire, particularly seen in the United States and the United Kingdom beginning in the 1980s in the Reagan and Thatcher eras, to measure performance in the public and nonprofit sectors (Figlio and Kenny, 2009). In the United States, this movement aligned well with the broader standards-based reform movement both in terms of intent and substantive areas (O'Day and Smith, 1993). The objective of standards-based reform is to identify a set of clear, measurable, and ambitious performance standards for students across a number of core subject areas, to align curriculum to these standards, and to expect students to meet these high standards. A central component of standards-based reform is the assessment of students to ensure that they are meeting the expectations set out for them, to identify the schools that have students who are relatively successful (or unsuccessful) in meeting these expectations, and to encourage schools to improve student outcomes.

Accountability, in the context of standards-based reform, is part of a broader integrated policy package, providing incentives for students, teachers, schools, or districts to perform. The principal-agent problem, well-known to economists, provides a rationale for accountability: if stakeholders—be they parents, local firms, or policy makers—have difficulty monitoring the activities of schools, then educators might behave in a manner contrary to the interests of these stakeholders. In such a case, it would follow that more effective monitoring of educators could result in improved student outcomes.

The information content in school accountability systems can provide a powerful mechanism for overcoming the principal-agent problem. Assessing schools against the common metric of standardized student test scores provides policy makers and members of the general public with independent information regarding how well schools and school districts (and potentially teachers) are doing in comparison to their peers or to outside performance standards. Measuring and reporting school performance and attaching positive and negative consequences to meeting or failing to meet

performance objectives provides incentives that encourage educators to concentrate on the subjects and materials that are being measured and to potentially alter the methods through which they educate students.³ The measurement and reporting of a school's progress allows policy makers to assess how successful a school has been in meeting the state's achievement goals.

The school is not the only level at which accountability could be targeted. Some policy makers favor accountability for individual teachers—through, for example, merit or performance-based pay, as advocated by President Obama and others—rather than for schools; and some researchers have found evidence indicating that performance incentives for teachers can be beneficial for student outcomes (e.g., [Lavy, 2007](#); [Figlio and Kenny, 2007](#)). Others view accountability at the school level as preferable both because it promotes collaboration among teachers and because schools have more opportunities than do individual teachers to enact the types of changes in resource allocation and practices that may be needed to raise student achievement ([Ladd, 2001](#)). Exclusive accountability at the school district level, instead of at the school or teacher levels, could mask the substantial heterogeneity in school performance observed across schools within a district. This is particularly the case in the larger school districts with dozens or hundreds of schools, though district level accountability has the benefit of allowing the reallocation of resources across schools in response to accountability incentives.

School accountability systems have the potential benefits of aligning effort with stakeholders' goals and providing information for improvement; however, they are limited by the fact that they can only measure a small number of the dimensions that stakeholders value. [Rothstein, Jacobson, and Wilder \(2008\)](#) demonstrate that educational stakeholders value a wide range of outcomes including not just academic performance and educational attainment but also areas such as citizenship, work ethic, and critical thinking. But school accountability systems generally do not cover even the full set of valued academic outcomes, instead often focusing solely on reading and mathematics performance, and the nontest measures like graduation rates or attendance rates are also crude proxies for the behavioral and attainment outcomes that stakeholders value. By focusing attention on the set of outcomes that are easily measurable, school accountability systems may lead some valued outcomes to be treated as more important than other valued outcomes.

The limitations of the outcome measures notwithstanding, school accountability can be successful in attaining its objectives if stakeholders value the information embedded within the accountability systems. A long line of papers, including work by [Black \(1999\)](#) and the papers summarized in the entry on real estate values by Black and Machin in Chapter 10 of this book, demonstrate that aggregate test score results are

³ Such information may also facilitate improved monitoring by another important set of stakeholders in the education system, namely parents. Whether by complaining about poor performance or by threatening to withdraw their child from the school, parents could potentially use the publicly provided information on school performance to induce their children's schools to improve.

capitalized into real estate prices. Figlio and Lucas (2004) show that school accountability grades have major consequences for real estate valuation (even holding other measures of school effectiveness constant) demonstrating that the nature of the presentation of school accountability information is itself quite consequential, not only for parents of students but also for members of the general public in terms of their asset values. In recent work, Figlio and Kenny (2009) provide new evidence that suggests that school accountability measures influence voluntary contributions to public schools. Specifically, they find that schools that experience negative accountability information “shocks” lose financial support from parents and community members, while those that experience positive accountability information shocks gain financial support. These responses are particularly strong for schools serving minority students and lower-income families that might have lower levels of monitoring of schools than might other families, and are consistent with the findings from the psychology, charitable contribution, and marketing literatures, that stakeholders tend to wish to avoid “throwing good money after bad.” In sum, the weight of the available evidence indicates that stakeholders of many stripes care deeply about the outcomes of school accountability systems, and this suggests that educators are likely to wish to respond as well.

3. THE NATURE OF ACCOUNTABILITY

How a school accountability system is designed can have a significant impact on the nature of and the strength of the incentives that schools face to raise student achievement in the tested subjects. Moreover the design can affect which students receive the most attention.

3.1 The consequences of accountability

School accountability systems can take two different approaches with regard to the consequences of accountability. One possibility is to include explicit rewards and/or sanctions for performance that exceeds or does not meet expectations. Examples of positive consequences for schools and educators in these systems may include increased resources or autonomy to spend these resources at the school level; and bonuses for educators in successful schools. Some accountability systems offer rewards to schools that are either exceeding stated expectations or moving strongly in the direction of doing so. On the flip side, accountability systems also frequently include explicit sanctions for schools not meeting expectations. Examples of these sanctions include the withdrawal of autonomy; requiring local education agencies to provide additional schooling options—either school choice or supplemental services—to students in these schools; and outright school restructuring or closure. Several studies, including Hanushek and Raymond (2005) and Dee and Jacob (2009), specifically identify systems with this more “consequential” accountability and provide evidence that these

consequences appear to translate to improved student outcomes, suggesting that educators respond to the explicit consequential incentives.

School accountability systems may not need to have explicit consequences from central authorities to influence educator behavior, however. As mentioned above, central governments are only one of many monitors of school performance, and other performance monitors—parents and community members—may pack enough punch to influence educator behavior. The broader economics literature on the role of information on product quality (e.g., Figlio and Lucas, 2004; Jin and Leslie, 2003; Mathios, 2000) shows how strong information disclosure can be in influencing markets, and it is realistic to expect that a major source of consequences of school accountability would be community and local pressure provoked by increased accessibility of information. Black (1999) shows that school test scores are capitalized in housing prices, and Figlio and Lucas (2004) demonstrate that housing markets react even more to the information embedded in school accountability systems. The findings of these two studies have since been replicated in numerous settings in North America and Europe. Figlio and Kenny (2009) also show that parents and community members withhold financial support from schools that central governments say are performing poorly and offer more financial support to those that central governments say are performing well. This financial pressure, coupled with the other informal pressure that surely accompanies it, strongly suggests that even absent formal consequences of accountability, accountability systems may be effective in influencing educator behavior.

This last point is particularly important because it implies that accountability systems have the potential to be effective even when the threatened sanctions associated with poor performance are not viewed as credible. Accountability systems that set standards such that a massive fraction of schools would likely fail may be perceived as incredible by educators who do not believe that central authorities would shut down schools or fire educators on a grand scale. Indeed, there exist very few examples of large-scale implementation of the more draconian elements of some school accountability systems. Economic theory would indicate that educators, when faced with an incredible threat, would not react to those threats. But, if the less draconian consequences of school accountability systems, including those that come through community pressure as a result of reporting systems, are sufficient to generate educator responses, then it may be that the severe but less credible threats associated with an accountability system are unnecessary for generating educator responses. Of course, since there have been so few instances of large-scale implementation of these severe threats, there is little data to shed light on the degree to which more credible serious threats might impact educator behavior and school outcomes.

3.2 Scope and domains of accountability indicators

School accountability systems differ in the number and types of tests, or other performance indicators, they include. Central governments face important tradeoffs when determining how broad-based to make their accountability systems. In particular, systems that align

accountability with a smaller set of outcomes tend to narrow the scope of the education provided to students. On the other hand, a broad set of outcomes is more difficult to measure reliably and may blur the focus of school and district personnel.

School accountability systems are intended to provide incentives for schools to generate higher performance in academic subjects, and indeed, schools appear to pay attention to the subject matter on which the tests are based. The available evidence strongly supports the conclusion that schools tend to concentrate their attention on the subjects tested and on the grades that have high stakes tests (Deere and Strayer, 2001; Ladd and Zelli 2002; Stecher et al., 2000). Other studies (e.g., Hamilton et al., 2005; Jones et al., 1999; Koretz and Hamilton, 2003; Linn, 2000; Stecher et al., 1998; and Stecher et al., 2000) show that teachers and schools tend to narrow the curriculum and shift their instructional emphasis from nontested to tested subjects, while earlier work by Shepard and Dougherty (1991) and Romberg et al. (1989) suggest that teachers focus more on tested content areas within specific subjects. In related work, Chakrabarti (2005) presents evidence that schools may concentrate their energies on the most easily-improved areas of instruction, rather than on subjects across the board.

This evidence on the narrowing of the curriculum in response to accountability implies that governments intent on school improvements along a wide variety of dimensions may wish to include a large number of subjects in the accountability system. However, increasing the scope of testing is costly, both in terms of financial costs and in terms of either the opportunity cost of foregone instructional time instead devoted to testing or the reliability of the test measures generated.

The scope and domain of accountability is also limited by the technology available to assess students' progress. Some subjects are simply more challenging to assess than are others. Given the well-established tendency of educators to focus their attention on the material most likely to be covered on the assessment, a behavior known commonly as "teaching to the test," it seems likely that educators will concentrate on the assessed components of difficult-to-assess subjects. Such a pattern of behavior could lead to attention redirected from the desired, but difficult to measure, knowledge and skills in favor of the less desirable, but easier to measure, aspects of a subject. A recent National Research Council panel (Wilson and Bertenthal, 2006) warns of this potential with regard to science assessment, as the members note how challenging it is to design a science assessment to tests students' scientific inquiry skills. It is therefore important to carefully consider the specific nature of assessments administered to students when deciding how broadly to base an accountability system.

There exists considerable heterogeneity across U.S. states in the substantive breadth of the state accountability system. While most states assess schools principally on the basis of reading, mathematics (and sometimes writing), some states administer a much farther-reaching set of tests for school accountability. Virginia, for example, tests students in more subjects, including science, U.S. and Virginia history, and social studies,

and reports on end-of-course examinations in high school. Kentucky's core content areas identified for the basis of school accountability further include fine arts and humanities. Nebraska assesses schools on a narrower set of substantive areas, but at a greater depth, including portfolio reviews as part of the assessment.⁴

In principle, accountability systems could be expanded to incorporate other measures of school performance besides student performance on standardized tests. [Hanushek and Raymond \(2003\)](#) construct a hierarchy of nontest indicators of school performance, ranked on the basis of their relevance and likely alignment to objective measures of school progress. For instance, they argue that certain measures such as the drop-out rate, graduation rate, number of students in advanced courses, percent of students passing end-of-course exams, retention rate, student mobility, and suspension rate, are relatively closely related to student achievement. If schools are trading off these outcomes in order to increase measured and incentivized outcomes, accountability may be counter-productive (though [Carnoy, Loeb and Smith, 2001](#), find no evidence of this tradeoff in Texas). Other variables, however, such as college entrance exam scores, course offerings, number of computers, number of noncredentialed teachers, parental satisfaction, school crime rate, principal mobility, or teacher mobility, are only weakly related to student achievement. To the extent that the goal of the accountability system is to increase student achievement, therefore, some of these measures would be more appropriate than others as elements of an accountability program. Some of these factors are already incorporated into many accountability systems. For example, graduation rates are part of NCLB's assessment of high schools, as well as those in numerous states. One shortcoming of this broader set of outcomes, however, is that they are more easily manipulated by school officials than achievement tests.

3.3 Measuring school performance

There have been two main approaches used to measure school performance on the basis of test scores. In "status" measures, a school's performance is judged based on levels of performance, such as the fraction of students attaining a given proficiency level or the average test score in the school. "Growth" measures, often called "gain scores" or "value-added" measures, evaluate schools on the degree to which their students improve in their test performance from one year to the next, or from fall to spring of a given school year. Growth measures can be technical complicated, and a thorough discussion of the issues is beyond the scope of this chapter. The simplest of these measures averages year-to-year or fall-to-spring changes in test scores across all students in a school while more complicated measures regression-adjust test score changes for various student characteristics or take into account the variance in observed test score changes. The No Child Left Behind law in the United States is currently based on a status model of evaluating schools,

⁴ At the time of writing, however, Nebraska was phasing out the portfolio review component of school accountability.

though the U.S. Department of Education has granted some states the ability to evaluate schools using a growth model. Proposals for reauthorization of the NCLB law being considered at the time of this writing would further expand the use of growth models in federal accountability evaluations in the United States.

The two types of approaches—status and growth—measure different outcomes and tend to generate different objectives and incentives for schools. Status-based systems that focus on the percent of students who achieve at proficient levels seek to encourage schools to raise performance at least to that level (Krieg, 2008; Neal and Schanzenbach, 2010). This approach is appealing to many policy makers because it sets the same target for all groups of students and because it encourages schools to focus attention on the set of low performing students who in the past may have received little attention. Status-based systems also have the advantage of being transparent.

The goal of the growth model approach is to encourage schools to improve the performance of their students independently of the absolute level of that achievement. Such an approach is appealing to many people because of its perceived fairness. It explicitly takes into account the fact that where students end up is heavily dependent on where they start and the fact that the starting points tend to be highly correlated with family background characteristics. At the same time, the use of the growth model approach may raise political concerns, both because the public may find the approach less transparent than the status approach and because some see it as a way of letting schools with low average performance off the hook.

Systems using status and growth models generate different incentives in part because they lead to different rankings of schools. Many schools deemed ineffective based on their aggregate performance levels may actually have quite high “value added” and vice versa (Clotfelter and Ladd, 1996; Ladd and Walsh, 2002; Kane and Staiger, 2002; and Stiefel et al., 2005). Some accountability systems (e.g., North Carolina’s) encourage both high levels of performance and high test score growth, by including both levels and gains in the index of success for schools under the accountability system (Ladd and Zelli, 2002).

The status and growth approaches send different signals to schools about which students deserve more attention. Under a status-based system designed to encourage schools to raise student performance to some threshold level, the position of the threshold matters. A challenging performance threshold—one that would be consistent with the high aspirations of the standards-based reform movement, for example—would provide incentives for schools to focus attention on a larger group of students than would be the case with a lower threshold. Evaluating schools on the basis of “value added,” by contrast, provides incentive for schools to distribute their effort more broadly across the entire student body. In such a system, however, schools may have an incentive to focus attention on the more advantaged students if the test score gains of those students are easier to increase, bringing up the average gains for the school (Ladd and Walsh, 2002; and Richards and Sheu, 1992).

Under either approach, random errors in the measurement of student performance can generate inconsistent rankings of schools over time—a factor that weakens incentives for improvement. The implications of measurement error are especially strong for small schools because the smaller the number of students in the school, the larger the school-wide average measurement error, and hence the less consistent the school's ranking is likely to be from one year to the next. Schools deemed to be improving at one point in time are often found to be declining the next year due to measurement error (Kane and Staiger, 2002). The problem of measurement error is exacerbated when schools are rated based on the growth model because it requires test scores for two years instead of one and both of these scores are measured with error. The danger is that personnel in such schools may receive such inconsistent signals from one year to the next that they have little incentive to respond in a constructive way. The policy relevant issue with measurement error is not whether there is any measurement error, there always is. The issue is whether this error is large enough to mask the signal that drives the incentives in the accountability system.

Neither the status nor the growth approach to measuring school performance perfectly captures school efficiency—the effectiveness with which schools use their resources to maximize student outcomes, given the students they serve. According to the “education production function” model, student achievement is determined by the characteristics of the student and his or her classmates, the school's resources (including the quantity and qualifications of the teachers), and the efficiency with which those resources are used. Because efficiency cannot be observed directly, it must be inferred from statistical analysis that controls both for the resources available to the school and the characteristics of the students being served (Stiefel et al., 2005). If the goal of an accountability system is to induce schools to use the resources they have more effectively, then, in principle, schools should be rated on their efficiency, not simply on the level or growth of their students' achievement. The problem is that the data requirements for such efficiency measures are often daunting and the statistical techniques can be complex (Ladd and Walsh, 2002; and Stiefel et al., 2005).

In contrast to a measure of school efficiency, the status and growth measures provide information on whether or not schools are meeting expectations for either the level of achievement or the growth in achievement with no attention to what accounts for that performance. Although inefficient use of available resources may be one reason for poor performance, another could well be that the resources available to the school are insufficient for the school to meet the accountability standard given the profile of the students in the school. In the latter case, it is neither fair nor likely to be productive for state or federal policy makers to hold the teachers or other school personnel responsible for the poor performance of the school's students (Ladd and Walsh, 2002). Thus, accountability and the financing of schools are closely intertwined.

3.4 Exclusions

Designers of accountability systems must also determine which students should be counted when evaluating student learning. It seems at first glance to be obvious that all students should be credited to a school—especially when accountability laws have names such as No Child Left Behind. But universal inclusion raises important questions about fairness and attribution. For instance, should a school be held responsible for the performance of a student who just arrived at the school a week prior or even a month prior to the test administration? Should schools be held responsible for students for whom testing is more challenging, or potentially less reliable, such as students with disabilities? The fact that these questions have no immediately obvious answer is evident in Florida's treatment of mobile students in successive iterations of its accountability system. When Florida introduced its system in 1999, the state included both those students who spent the full academic year and those who were recent in-migrants in its calculations of school grades. The following year the state amended its policies to include only students who had spent the full academic year up to testing in the school. These rule changes influenced the sets of schools identified as low- or high-quality (Figlio and Lucas, 2004). At the federal level, NCLB counts only those students who had spent the full year in the school toward school proficiency goals but still includes all students in the calculations of average proficiency rates for the purposes of public reporting. Transient students count for school *district* accountability under NCLB.

NCLB mandates that students with disabilities and English Language Learners be included in a school's aggregate proficiency counts, and these groups are specifically identified as separate subgroups of interest in the federal law. States, on the other hand, in implementing their own accountability systems, have diverging treatments of these students. Virginia, for instance, with an accountability system that predates NCLB, chose to include English Language Learners and students with disabilities in its calculations of school ratings, while Florida excludes all students with disabilities, even those who take tests, from the school-level aggregates used to measure performance. Florida schools, therefore, are subject to two different accountability treatments of these groups of students, one from the state level and one from the national level.

Policy makers face clear tradeoffs with respect to the treatment of these special populations. On the one hand, schools with large fractions of mobile and disabled students in many cases have a legitimate argument that holding them accountable for the academic achievement of such challenging-to-educate students puts them at an unfair disadvantage relative to other schools with fewer disabled students. On the other hand, excluding students on the basis of *classification* provides schools with less incentive to support these students as well as an incentive to selectively reclassify or move students in order to look better against performance metrics. The evidence is quite clear that schools have responded to accountability pressures by reclassifying low-performing students as students with disabilities (see, for example, Cullen and Reback, 2006; Deere

and Strayer, 2001; Figlio and Getzler, 2007; and Jacob, 2005). Thus, while the incentives to reclassify are small under NCLB, such incentives may still exist under state policy.⁵ NCLB (and other accountability systems that hold schools accountable for a disabled subgroup explicitly) also may provide incentives for schools to identify as disabled more marginal students so that the average proficiency rates amongst the school's disabled population increases. These incentives are all smaller under growth model accountability scenarios than under status model scenarios. Moreover, reclassification may not be detrimental; one can interpret the incentive effects of accountability with regard to disabled students as potentially providing incentives to correctly identify students as disabled, rather than to over-classify disability rates. Bokhari and Schneider (2009) find that accountability leads to increased prescription of psychostimulants, implying physician involvement in the identification of disabilities.

Accountability-based incentives for identifying students as special needs interact with other incentives embedded within the school finance system of a jurisdiction. Some U.S. states compensate school districts for disabled students on the basis of predicted—rather than actual—disability caseloads. In these cases, a district that reclassifies a student as disabled to avoid having that student counted for the purposes of accountability will generate higher costs for the district because it is responsible for the full costs of providing special services for the student. In places that compensate school districts for the extra costs of educating students who are specifically classified as disabled, in contrast, the finance system will exacerbate any incentives to over-classify students provided by the accountability system, and will also increase the incentives to correctly classify students in need of special education services. It is currently impossible to determine for certain the “correct” level of disability classification.

3.5 Subgroup identification

One purpose of accountability systems may be to focus attention on traditionally underperforming groups of students. Policy makers interested in doing so may explicitly require that schools meet certain performance targets for individual subgroups of students within a school's population. This focus on subgroups is central to federal accountability policy in the United States, as NCLB holds schools accountable not only for the performance of the full student body, but also for the performance of subgroups of students defined by their race, income, and disability status. Because of the small size of many subgroups, this subgroup requirement exacerbates the problems of measurement error highlighted by Kane and Staiger (2002). Nonetheless, NCLB requires such disaggregation on the grounds that it provides incentives for schools to pay attention to members of each subgroup and thereby prevents schools from leaving particular groups of children behind.

⁵ Consistent with this conclusion, during the during the 1990s, Texas and North Carolina, both of which had highly touted state accountability systems, excluded increasingly large number of students from the NAEP tests, thereby biasing upward the observed gains in NAEP scores in those states (Amrein and Berliner, 2002; Braun, 2004).

States have the authority under NCLB to determine the minimum size of subgroups that are separately measured and reported, and states have set very different thresholds. Thresholds vary from five students in Maryland to as many as 50 to 200 students depending on the size of the school in Texas. Setting the size of the subgroup thresholds involves a clear policy tradeoff. On the one hand, a higher threshold increases the accuracy with which school performance is measured. On the other hand, a higher threshold means that large segments of a school's population could fall under the radar screen, an outcome that would be inconsistent with the goals of NCLB. A recent and highly publicized Associated Press analysis, for example, reported that 1.9 million students are not counted under their racial and ethnic subgroups, including more than one-third of Asian students and nearly half of Native American students (Bass, Dizon and Feller, 2006). A potential alternative to the subgroup requirement would be to focus special attention on the segment of the school's students that performed at a low level in the previous year and to track that group's growth. This segment would likely include a large fraction of the economically disadvantaged and racial minority students, and so might capture the spirit of the NCLB law without exacerbating the problem of measurement error.

The identification of subgroups, and the attendant issue of the size requirements for subgroup identification, influences the likelihood that a school will meet all of the annual yearly progress (AYP) criteria. When the subgroup size thresholds are low, the more racially heterogeneous schools will have more measured subgroups and will face greater risks of low accountability ratings compared to more homogeneous schools because any negative random error in any single subgroup is sufficient to lead to a negative rating for a school. Using national data, Stullich et. al. (2006) show that among schools that missed AYP in 2004, 23% missed because of the failure of a single subgroup and 18% missed because of insufficient achievement of two or more subgroups. Given the correlation between subgroups (e.g., those based on race and free lunch eligibility), one can reasonably assume that subgroup size requirements were responsible for anywhere from one-fifth to one-third of the failure to make AYP among the schools that missed it.

3.6 Time considered for rating schools

A final design issue is the relevant time period for accountability. Kane and Staiger (2002) demonstrate, both conceptually and with data, that substituting multiyear moving averages for year-by-year analysis considerably reduces the instability of the measures of school performance over time, and thereby provides schools with more consistent incentives to raise student performance. Accountability systems based on a single year of data (or growth from one year to the next), as is largely the case in both NCLB and state accountability systems, are far more likely to misjudge the performance of schools. Increasing the time period over which schools are evaluated reduces the measurement error and the incorrect classification of schools, though also requires more years of data to spot indications of improvement or decline.

Figlio (2004) simulates how permitting accountability to apply to periods longer than a year affects the set of schools likely to be sanctioned under the NCLB Act. His simulations show that the fraction of schools sanctioned under the year-by-year system is approximately 20% higher than it would be under a system based on a three-year time period. In addition, he demonstrates that shifting to the longer time period reduces the rate at which schools are likely to be sanctioned more for racially heterogeneous schools with multiple subgroups than for other schools. Thus, extending the time period to three years reduces the random variation within subgroups and allows for a more accurate picture of trends in student performance within a student category.

4. ACCOUNTABILITY MIGHT NOT IMPROVE SCHOOL PERFORMANCE

The preceding discussion predicts that school accountability programs will increase student achievement, although the magnitude of the predicted effects for particular groups of students or types of schools may well differ depending on how the system is designed. For several reasons, however, school accountability systems might not generate higher achievement.

4.1 Improving measured, but not generalizable, achievement

Monitoring provides incentives for those being monitored to appear as effective as possible against the metric being assessed. It is certainly possible, therefore, that educators could teach very narrowly to the specific material covered on the tests, and little or no generalizable learning outside of that covered on the test would take place (Koretz and Barron, 1998). This restriction on the domains of learning may not be a concern if the tests that come with high stakes for schools cover a wide range of material considered important by society; in fact, this “teaching to the test” may be desirable. On the other hand if the high-stakes tests reflect only a subset of the knowledge and skills desired by stakeholders, then teaching to the test could have negative consequences for students. Furthermore, educators can go further than teaching to the test, and teach test-taking strategies with little long-term benefit for students or even engage in outright cheating to appear better on the accountability examinations. For example, Jacob and Levitt (2003) show that a small fraction of Chicago teachers responded to accountability pressures in that city by fraudulently completing student examinations in an attempt to improve observed student outcomes.

A popular approach for determining the extent to which an accountability policy has resulted in generalized learning involves seeing whether gains observed on high-stakes tests are also observed using low-stakes tests with no particular consequences for schools or students. A natural test for that purpose is the National Assessment of Educational Progress (NAEP), which has been administered to a nationally representative random

sample of students since the early 1970s and to representative samples of students in grades four and eight in most states since the 1990s. Because of its high profile and national scope, the NAEP has been widely employed in the studies described later in this chapter assessing the effects of accountability on student learning.

Observing a low-stakes test has considerable appeal, but it has at least two downsides. One downside is that students may not take a low-stakes test sufficiently seriously to do their best work. That said, this lack of effort would mainly introduce measurement error into the dependent variable of analyses of the effects of accountability on student outcomes and one might expect that a main consequence of using low-stakes tests to evaluate the effects of accountability is the imprecision associated with the estimates. Unless student effort differs from one administration of the low-stakes test to the next, changes in performance on the low-stakes test should provide a reasonable estimate of gains in student learning. A second downside to using low-stakes tests is that the high-stakes tests are often aligned to the standards valued by policy makers, while the low-stakes tests are not as aligned. Hence, findings of smaller effects of accountability when low-stakes tests are used to measure performance may simply reflect differences in the degree to which the two types of tests reflect the material that policy makers want to see covered.

The accountability experience in Texas illustrates the importance of this distinction between performance on high- and low-stakes tests. After a series of education reforms starting in the early 1980s, Texas introduced in 1990 a criterion-referenced testing program called the Texas Assessment of Academic Skills (TAAS) that was designed to shift the focus from minimum skills to higher-order thinking skills (Haney, 2000). By 1994, tests were being administered annually to all students in grades 3–8 and students had to pass a 10th-grade test to graduate. The state then used passing rates on the TAAS, along with dropout rates and student attendance rates, to hold individual schools accountable for their students' performance. Schools were held accountable not only for the overall pass rate in the school but also for the pass rates of four student subgroups: African Americans, Hispanics, whites, and economically disadvantaged students. Between 1994 and 1998, TAAS scores in both math and reading increased quite dramatically, suggesting that the accountability program had a large and positive impact on student achievement. Klein et al. (2000), however, showed that the large gains on TAAS did not translate into comparably large gains in the lower-stakes Texas NAEP scores. In general, the gains in NAEP scores were about a third the size of the gains in TAAS scores, though still meaningfully positive.

Further, the TAAS and NAEP results generate conflicting stories about how accountability affected racial achievement gaps in Texas. In particular, the gaps between blacks and whites based on the TAAS scores in fourth-grade reading and math and eighth-grade math decreased significantly between 1994 and 1998, while the

comparable gaps based on the NAEP increased slightly (Klein et al., 2000). Similar patterns also emerge for Hispanics. Klein et al. speculate that the reasons for the differing patterns for TAAS and NAEP results is that Texas teachers may be teaching very narrowly to the TAAS and that the schools serving minority students may be doing so even more than other schools.

Additional evidence on whether the transferability of knowledge from high-stakes to low-stakes tests emerges from Jacob's 2005 study of accountability in Chicago. Jacob compared achievement gains for fourth and eighth graders in math as measured by scores on the district's high stakes test to those on a comparable, but low-stakes, test administered by the state of Illinois. Those comparisons show that gains for eighth graders generalized to the state test but that those for fourth graders did not.⁶ In Florida, Figlio and Rouse (2006) find consistently smaller estimated effects of accountability on low-stakes tests than they do using high-stakes tests.

4.2 Strategic behavior

Teaching to the test is not the only mechanism through which schools might alter their behaviors in response to the incentives embedded within accountability systems. There exists considerable evidence that schools engage in strategies that artificially improve test scores by changing the group of students subject to the test. The most widely studied behavior of this type is the selective assignment of students to special education programs. As mentioned above, many studies show that schools tend to classify low-achievers as learning disabled in the context of accountability systems. Though there may be some debate about whether the greater rates of classification are undesirable in all cases, nonetheless, they highlight the possibility that schools are manipulating the testing pool specifically to inflate measured school performance. These decisions may have spillover consequences outside of education: Bokhari and Schneider's (2009) finding that school accountability policies enhances the use of psychostimulants suggests that there are health consequences of education policies. Likewise, Figlio's (2006) finding that some Florida schools changed their discipline and suspension patterns around the time of the testing in ways consistent with the goal of improving test-takers' average scores reinforces the concern that schools might engage in artificial improvements of student test performance, with possible significant ramifications for students involved.

Schools may engage in other types of strategic behavior that affect student performance. For example, Figlio and Winicki (2005) demonstrate that schools change their meals programs at the time of the tests in an apparent attempt to raise performance on high-stakes examinations, while Anderson and Butcher (2006) find that schools subject

⁶ Data were not available for a comparable analysis of reading scores.

to accountability pressure are more apt than other schools to sell soft drinks and snacks through vending machines.⁷ Finally, [Boyd et al. \(2008\)](#) illustrate how high-stakes testing in certain grades in New York altered which teacher taught in particular grades and schools.

Many of these behaviors are less likely to occur in growth model accountability systems than in status-based systems. The reason is that in the growth approach, the manipulative behavior that increases student achievement in one year would make it more difficult for the school to attain accountability goals the following year. No such tradeoff arises in status-based accountability systems. Indeed, [Rouse et al. \(2007\)](#) document a series of significant substantive responses with regard to instructional policies and practices as a consequence of Florida's school accountability system that assesses schools on the basis of student achievement growth.

In a federal system with multiple levels of accountability decision-making, states (or other subfederal units) may themselves respond strategically to federally imposed accountability pressures in ways antithetical to higher achievement. For example, at the same time that NCLB delegates to them the task of defining proficiency standards, it imposes penalties on schools and districts that fail to make adequate progress toward those standards. Consequently, states have incentives to set low proficiency levels. [Peterson and Hess \(2006\)](#) document the low level of concordance between their students' progress toward state-defined proficiency and their performance on the NAEP. States such as South Carolina that set very high standards for their students find themselves with large fractions of schools deemed in need of improvement, while states such as Texas that set low standards have few such schools. This interaction between state-set standards and the likelihood that their schools will face sanctions has the potential to lead, in Peterson and Hess's words, to a "race to the bottom" in terms of setting proficiency standards.

Another potentially adverse effect on achievement works through funding provisions. Under NCLB, districts that are sanctioned for low performance are required to use their federal Title I grants to pay for privately-provide supplemental services and for transportation for students who choose to opt out of failing schools. [Figlio \(2003\)](#) shows that the districts with the highest fractions of minority and low-income students are likely to lose the most Title I funding under this provision. Unless the district or the state replaces that funding with other revenue, NCLB could reduce the instructional resources available to students in those districts, which potentially could have adverse effects on student achievement.

⁷ [Anderson, Butcher, and Schanzenbach \(2009\)](#) and [Yin \(2009\)](#) even find that school accountability systems have contributed to childhood obesity, though there is little evidence that the principal pathway through which this is happening is changes in school nutrition or food served in vending machines.

4.3 Failure or inability to respond to incentives

While the evidence mentioned above indicates that many school administrators and teachers are highly responsive to incentives,⁸ some educators might not react to incentives—whether those incentives are in the form of bonuses, positive recognition, or negative sanctions—by changing their behavior in ways consistent with the goals of the accountability system. External incentives may be too small, for example, to override the professional judgments of teachers and school administrators, and if so, one may not see substantial changes in educator behavior as a result of the accountability system. Additionally, as [Frey \(2000\)](#) points out, if the extrinsic incentives associated with the accountability system crowd out the intrinsic motivations that attracted educators into teaching, one might observe stagnant or decreasing performance by students in a new accountability regime.

Schools personnel also might fail to incentives embedded within accountability systems if they lack the capacity to respond in ways desired by state or federal policy makers. Some schools may have insufficient resources to effect serious change in student outcomes, while others may lack the leadership required for significant change. Teachers may lack the necessary skills and knowledge to meet the expectations of an accountability system that requires rates of improvement far larger than historical experience has shown to be feasible, as is required under the initial iteration of NCLB. Thus, it could be that one of the major assumptions underlying standalone accountability programs—namely that teachers and schools are underperforming because of insufficient monitoring of their behavior—is incorrect. If school resources must be at a certain level to bring about positive performance improvements, or if principals and teachers have sufficient resources but lack the specific policy and practice knowledge necessary to implement highly successful instructional policy and practice changes, then accountability might not lead to meaningful improvements in student outcomes.

The lack of potential responsiveness to accountability could be exacerbated by the fact that accountability systems generally concentrate on shorter-term achievement improvements while many of the policies and practices that schools may wish to implement can take longer to bring to fruition. School accountability may solve one principal-agent problem by introducing a new one—educators may eschew the types of policies that might yield large-scale long-term success in favor of those that would be less successful in the longer term but might generate bigger boosts today.

Thus, despite the theoretical prediction that school accountability systems will improve student achievement—at least for certain segments of the school population—

⁸ Recall that the responses to accountability are often substantive as well as potentially considered to be “gaming.” [Rouse et al. \(2007\)](#) demonstrate that Florida schools engaged in a series of substantive changes in instructional policies and practices as a result of accountability incentives. [Ladd and Zelli \(2002\)](#), in a survey of elementary school principals in North Carolina, also find evidence that principal behaviors changed in line with state goals in the wake of accountability.

such gains are not a foregone conclusion. In some cases schools may focus on test scores to the exclusion of transferable knowledge or may end up with less funding for instruction. Potentially most important, schools may lack the knowledge and capacity to produce significant gains in student achievement.

5. EVIDENCE ON STUDENT OUTCOMES

Measuring the effects of test-based accountability systems on student achievement is not a simple task. When such systems are part of a larger standards based reform effort, it is difficult to separate the effects of the accountability system from those of other components of the reform package. In addition, researchers face the challenge of finding appropriate control groups to determine what would have happened to student achievement in the absence of the accountability system. In practice, researchers have used a variety of empirical strategies to address these challenges.

A few recent studies have tried to determine the achievement effects of NCLB. Given the difficulty of isolating the effect of NCLB from other concurrent changes, [Wong, Cook, and Steiner \(2009\)](#) use multiple approaches. First they compare the change in achievement in Catholic schools, not subject to NCLB, to the change in public schools since implementation. Second, they compare the growth in states with low proficiency standards and thus fewer schools failing to meet NCLB goals to those with higher proficiency standards. In both cases, they find positive effects of the accountability provisions on student achievement in the fourth and eighth grades. Using similar data from NAEP, [Dee and Jacob \(2009\)](#) compare states that had school-level accountability systems prior to NCLB to those that did not and find greater achievement gains in math for both fourth and eighth grade students in states that did not have assessment-based accountability at the school level prior to NCLB. They do not find corresponding gains in reading achievement. [Cronin et al. \(2005\)](#) use longitudinal data on students just before (2001–2002) and just after (2003–2004) NCLB was first implemented to assess the extent to which students are learning more after NCLB. They find higher achievement post NCLB especially on the math exams, but they find lower achievement gains over the course of the year; findings that are difficult to reconcile. Finally, [Neal and Schanzenbach \(2010\)](#) use a different approach comparing students who took the same test right before and right after NCLB was passed and found substantially higher scores among students in the middle of the achievement distribution. [Table 8.1](#) summarizes the results of these four studies. Taken together, they provide some evidence that NCLB increased student test performance especially in math.

The short time-period since that federal law was implemented combined with the lack of variability in the law across states limits the conclusions that can be drawn from studies of NCLB. More compelling studies of how accountability affects student

Table 8.1 Studies of the Effects of NCLB on Student Achievement

Study	Data	Identification	Findings
Cronin et al. (2005)	Northwest Education Association longitudinal data	A comparison of achievement and student growth just prior (2001–2002) and just post (2003–2004) NCLB implementation.	0.05 and 0.01 standard deviations higher post NCLB fall scores in math and reading scores, respectively. 0.04 and .02 standard deviations lower post reform growth in math and reading respectively.
Neal and Schanzenbach, (2010)	Chicago public schools data	Compares students who took a high-stakes test under a new accountability system with students who took the same exam under low stakes in the year before the accountability system was implemented.	0.040, 0.073, 0.053, 0.093, 0.087, 0.080—5th grade reading gains between 2001 and 2002 for students in 3rd–8th decile of 3rd grade scores 0.080, 0.040, 0.060, 0.140, 0.107, 0.127. 0.080—5th grade math gains between 2001 and 2002 for students in 3rd–9th decile of 3rd grade scores
Dee and Jacob (2009)	Main NAEP state data: 1990–2007	A comparison of state-level achievement growth since NCLB implementation for states that did and did not have school accountability prior to NCLB.	0.23 and 0.10 standard deviation higher 4th and 8th grade math achievement by 2007, respectively. Estimates differ somewhat across specifications. No effect for 4th or 8th grade reading.
Wong, Cook and Steiner (2009)	Main NAEP state data and trend NAEP national data: 1990–2009	Comparison of (1) Catholic schools to public schools, (2) states with lower proficiency standard to states with higher proficiency standards, and (3) states with accountability systems that included school sanctions prior to NCLB and those that did not.	0.34 and 0.24 standard deviations higher gains for public school students than Catholic school students post-NCLB in 4th and 8th grade math, respectively. No effect for reading. 0.26 and 0.19 standard deviations higher gains for states with higher proficiency standards for 4th and 8th grade math. No effect for reading. 0.11 standard deviations higher 4th grade reading achievement for states with both higher proficiency standards and school sanctions.

achievement are based on the state and local accountability systems that preceded NCLB. This research includes district or state specific-studies as well as cross-state studies that measure achievement using the NAEP data. Researchers conducting district and state-specific studies have used a combination of state or district-wide trends in achievement along with trends or patterns in school and student level achievement in other comparable districts or states to sort out how the specific accountability system in that district or state affected student achievement. The main advantage of district and state studies is that the analysis is firmly focused on a specific, well-defined accountability system. Some of the studies, especially those for particular states, are hampered by the difficulty of predicting what would have happened to student achievement in the absence of the state's accountability system.⁹

Table 8.2 summarizes the results of state- or district-specific studies of the effects of accountability on student outcomes. The findings are, on average though not universally, positive. One set of papers describe trends in student achievement after districts or states implemented accountability measures. Richards and Sheu (1992) find positive trends in South Carolina following reform. Jacob (2005) also finds positive trends in both math and reading scores following accountability reforms in Chicago using a more sophisticated interrupted time series design, though these positive results are limited to the high-stakes test. Klein et al. (2000) compares high- and low-stakes tests in Texas following reform and similarly finds more positive results on the Texas state test than on the lower stakes NAEP, though even the low-stakes exam showed positive trends.

A second set of district-specific studies compares districts that implemented accountability reforms to other nearby jurisdictions. Ladd (1999) finds greater increases in pass rates in Dallas after the district implemented accountability than in other Texas districts. However, Smith and Mickelson (2000) find no difference in achievement trends between Charlotte-Mecklenburg and other North Carolina districts after accountability reforms.

The final set of district- or state-specific studies use variation in accountability pressures within a given system to identify effects. Figlio and Rouse (2006), Rouse et al. (2007), and Chiang (2007) all exploit discontinuities in school accountability grades and find positive effects of receiving low grades on student achievement gains with effects up to 0.20 standard deviations, though most between 0.05 and 0.10. Rockoff and Turner (2008) take a similar approach in New York City and find positive effects of accountability pressures associated with receiving a failing grade. Taken together the district- and state-specific studies, like the studies of NCLB, provide some evidence of a positive relationship between accountability and student achievement, though they are not universal in this conclusion.

⁹ However, some studies (e.g., Figlio and Rouse, 2006; Figlio, 2006) focus not on overall achievement but rather on how specific provisions of Florida's accountability system affect student achievement.

Table 8.2 State-Specific or District-Specific Studies of the Effects of Accountability on Student Outcomes

Study	Data	Identification	Findings
Richards and Sheu (1992)	South Carolina data	Simple trends in student achievement and attendance following implementation of accountability	No comparison group. Upward trend in test performance but no change in attendance.
Ladd (1999)	Panel data for schools in large Texas cities from 1990–91 to 1994–95	Compares Dallas student outcomes to the outcomes of students in other districts after Dallas implements accountability system	After one year of program implementation Dallas pass rates increased by 15.5, 16.8 and 12.1 percentage points one, two and three years post implementation. Consistently positive effects for Hispanic and white 7th graders, but none for black students. The study does not provide information on pooled standard deviations for pass rates to compute effect sizes.
Klein et al. (2000)	Trends in State NAEP scores for Texas and in Texas state test scores (TAAS) separately for white, black and Latino students	Compares trends between 1992 and 1998 on NAEP and 1994 and 1998 on TAAS	0.13–0.15 increase on NAEP in 4th grade math compared to 0.31–0.49 increase in TAAS.
Smith and Mickelson (2000)	District average student outcomes from three North Carolina districts	Compares district average student outcomes in Charlotte–Mecklenburg after accountability implementation to those in two other North Carolina districts	No evidence of achievement effects.
Jacob (2005)	Student-level data from Chicago with low-stakes and high-stakes exam scores	Interrupted time series design	0.35 and 0.25 standard deviation increase on high-stakes in math and reading four years post reform. No effect on low-stakes exam.

Continued

Table 8.2 State-Specific or District-Specific Studies of the Effects of Accountability on Student Outcomes—cont'd

Study	Data	Identification	Findings
Figlio and Rouse (2006)	Student-level test scores from a subset of Florida districts	Compares schools with failing grade after change in accountability formula	0.20—effect of receiving an “F” in math on gains on the high-stakes test in the high-stakes grade (0.10 for all grades). 0.06 for low-stakes test in math for all grades.
Rouse et al. (2007)	Florida administrative data combined with a survey of all public schools	Uses regression discontinuity to estimate the effect of receiving an “F” grade on student achievement and school policy	0.099, 0.141, 0.069, 0.076—effect of receiving an “F” on high-stakes reading and math and low-stakes reading and math. 0.140, 0.212, 0.074, 0.122—effect one year later, respectively
Chiang (2007)	Florida administrative data	Regression discontinuity exploiting Florida’s criteria for identifying schools that receive sanctions	0.11, 0.12—effect of attending F-graded school on reading and math after one year. 0.00 and 0.12 in year 2. 0.03 and 0.08 in year 3
Rockoff and Turner (2008)	New York City school-level data	Discontinuities in grade assignment across schools	0.10 and 0.05—effect of receiving an “F” relative to a “D” in math and ELA.

A final set of U.S. studies has sought to measure the effects of accountability by comparing achievement trends across states prior to NCLB. [Table 8.3](#) summarizes these results. The cross-state studies make use of variation across states in the nature or timing of accountability systems. Although the conclusions are sensitive to how accountability policies are defined as well as to methodological considerations such as the determination of control groups, the findings of cross-state studies are likely to be less idiosyncratic and more generalizable than those that emerge from the analysis of a specific program. The earliest studies in this group compare states that implemented minimum competency exams or graduation exams. [Fredericksen \(1994\)](#) finds increases in test performance particularly for nine-year-olds following the implementation minimum competency exams, while [Jacob \(2001\)](#) finds increases in math and reading gains in states that implemented graduation exams.

The 1990s saw substantial increases in state accountability systems and a series of studies exploit this variation to estimate the effects of accountability ([Amrein and Berliner, 2002, 2003](#); [Carnoy and Loeb, 2002, 2005](#); [Rosenshine, 2003](#); [Braun, 2004](#); [Hanushek and Raymond, 2005](#). [Hanushek and Raymond, 2005](#)) find that the introduction of an accountability system with consequences for schools during the 1990s raised eighth grade student test scores on the NAEP by about 3.2 scale points¹⁰. The study does not distinguish between effects on reading and on math. The effect is about a fifth of the 16.2 point standard deviation of average eighth grade scores across states, but would be a far smaller fraction of the deviation across individual students, which is the way effect sizes are more commonly measured in the education literature. Thus the effect is modest at best. This conclusion is similar to that reported by [Lee \(2006\)](#) based on his meta-analysis of 12 cross-state studies completed between 1994 and 2004.

In one of the first careful cross-state studies of accountability, [Carnoy and Loeb \(2002\)](#) show that the relationship between the strength of a state accountability system and student performance on NAEP math is stronger at the basic level than at the proficient level. Given that NCLB calls for performance at the proficient rather than the basic level, this finding suggests that even the strongest current state-level accountability systems may have little success in raising students to levels required under NCLB—except to the extent that states maintain proficiency levels far below the NAEP standards. Consistent with that conclusion, but inconsistent with findings by [Hanushek and Raymond \(2005\)](#), other studies indicate that even though high-stakes tests may be associated with gains in math scores at the fourth-grade level, they may not be associated with gains as students progress from fourth grade to eighth grade and, hence, as the students confront more challenging material.

¹⁰ Even this study is not free from criticism. The study identifies the effects of accountability systems by making use of the variation in their time of introduction. The choice of specific starting dates for some of the states, including key states such Florida and North Carolina, raises some cause for concern. Hence, replications of this study would be useful.

Table 8.3 Cross-state Studies of the Effects of Accountability on Student Outcomes

Study	Data	Identification	Findings
Fredericksen (1994)	Long-term NAEP from 1978 and 1986	Difference-in-difference analysis of students in states that did and did not implement Minimum Competency Tests	<p>0.22, 0.13—difference in gains between high-stakes and low-stakes states on 9-year-old math routine and nonroutine items</p> <hr/> <p>0.08, 0.12—difference in gains between high-stakes and low-stakes states on 13-year-old math routine and nonroutine items</p> <hr/> <p>0.02, 0.05—difference in gains between high-stakes and low-stakes states on 17-year-old math routine and nonroutine items</p>
Jacob (2001)	National Educational Longitudinal Survey (NELS)	Models probability of dropping out of high school as a function of the state requiring graduation exams	0.04 and 0.001—effect of graduation exam on math and reading gains between 1998 (8th grade) and 1992 (12th grade)
Amrein and Berliner (2002)	Trend data on 18 states that implemented accountability programs	Simple trends in student achievement and attendance following implementation of accountability	No comparison group
Carnoy and Loeb (2002)	State NAEP data from 1996–2000	Rated state accountability policies on a scale from 0 to 5 and modeled test score growth as a function of accountability strength	<p>0.78, 0.95, and 1.05; 0.80, 1.14, and 0.93—for a two-step move in accountability and the % white, black, and Hispanic students attaining basic skills and then proficiency on 8th grade math</p> <hr/> <p>0.10, 0.77, 0.54—relationship between a two-step move in accountability index and the % white, black, and Hispanic students attaining basic skills on 4th grade math assessment</p>

Amrein and Berliner (2003)	State NAEP data from 1994–2000	Compare achievement gains between states that have and do not have high-stakes tests	<p>1.2—effect of testing on 1996–2000 4th grade math gains</p> <hr/> <p>No statistically significant effect on 1996–2000 8th grade math (positive and significant when all states are included)</p> <hr/> <p>No statistically significant effect on 1994–1998 4th grade reading</p>
Rosenshine (2003)	State NAEP data from 1994–2000	Compares achievement gains of states with and without high stakes testing	<p>0.35 and 0.79—effect of high-stakes tests on 4th and 8th grade math gains between 1996 and 2000</p> <hr/> <p>0.61—effect of high-stakes tests on 4th grade reading gains between 1994 and 1998</p>
Braun (2004)	State NAEP math assessments from 1992–2000	Compares the achievement gains and cohort gains in states with high-stakes tests to other states	<p>0.96 and 0.81—difference in 4th grade and 8th grade math gains between high- and low-stakes states (1992–2000)</p> <hr/> <p>–0.67 and –0.31—difference in cohort gains (1992/96 4th grade to 1996/2000 8th grade math) between high- and low- stakes states</p>
Hanushek and Raymond (2005)	State NAEP data from 1992–2002 combined with the timing of accountability	Compares cohort gains using 4th and 8th grade NAEP scores four years apart between states with and without accountability policies	0.22, 0.21, 0.09, 0.54—effect of states attaching consequences to school performance on NAEP gains in math and reading (overall, white, black, and Hispanic students)

Though no one approach or study is flawless and many inconsistencies remain, taken as a whole, the body of research on implemented programs suggests that school accountability improves average student performance in affected schools, at least in general. Experimental evaluations of test score reporting, such as [Andrabi et al.'s \(2009\)](#) new results from Pakistan, also support the notion that accountability can boost student outcomes.

While, in general, the findings of the available studies indicate achievement growth in schools subject to accountability pressure, the estimated positive achievement effects of accountability systems emerge far more clearly and frequently for mathematics than for reading. This pattern is particularly clear when the outcome measure is based on a national test, such as NAEP, but it also emerges in some of the district or state level studies such as [Figlio and Rouse \(2006\)](#). In part this pattern reflects the fact that some authors report results only for math, although that is presumably because of the smaller effects for reading. The larger effects for math are intuitively plausible and are consistent with findings from other policy interventions such as voucher programs ([Zimmer and Bettinger, 2008](#)) and tax and expenditure limitations ([Downes and Figlio, 1998](#)). Compared to reading skills, math skills are more likely to be learned in the classroom, the curriculum is well-defined and sequenced, and there is less opportunity for parents to substitute for what goes on the classroom ([Cronin et al., 2005](#), p. 58).

One exception to this finding of larger effects for math emerges from [Jacob's 2005](#) study of accountability in Chicago, where the positive effects for low performing students were somewhat stronger in reading than in math. This finding, however, is based on results from the district's high-stakes test rather than from a low-stakes test, and may well reflect the particular characteristics of Chicago's accountability system.

Several studies have documented that school accountability systems have had long-lasting effects on student test scores, even after the students have left the schools directly affected by accountability pressure. [Rouse et al. \(2007\)](#) and [Chiang \(2007\)](#) both show that student test scores, in mathematics and to a lesser degree in reading, are persistently higher for several years following a student's departure from an affected public school. This evidence provides support for the notion that the estimated test score responses to school accountability pressure, at least in Florida, are genuine.

5.1 Differential effects of accountability

The studies described above generate mixed results by racial group, with at least one study ([Carnoy and Loeb, 2002](#)) for the late 1990s finding larger effect sizes on passing rates at the basic level on NAEP for black and Hispanic students than for white students. Other studies with different outcome measures find different patterns. In particular, [Hanushek and Raymond \(2005\)](#) find essentially no effects of accountability on the eighth grade achievement of black students, but positive effects for Hispanic students, patterns that are consistent with early findings by racial group for seventh graders

in Dallas (Ladd, 1999). Effects of accountability on racial achievement gaps are similarly mixed. The Hanushek and Raymond study finds that state accountability systems may have reduced the gap for Hispanics but raised it for blacks. The two recent national studies find little effect of NCLB on racially defined achievement gaps.

Some evidence from the district or state-specific studies suggests that the schools at the bottom of the performance distributions exhibit the greatest gains under an accountability system. This conclusion emerges from both Chicago (Jacob, 2005) and Florida (Figlio and Rouse, 2006). Working in the other direction is the finding from Cronin et al.'s (2005) national study that the effects of high stakes are greater for the higher scoring students. That said, there exists an emerging consensus that students whose scores are the most consequential for school accountability are those who gain the most, indicating that schools concentrate their energies on marginal students. Prominent examples of this evidence include Neal and Schanzenbach (2010), Reback (2008), and Krieg (2008) in the United States and Burgess et al. (2005), West and Penell (2000), and Wiggins and Tymms (2002) in England.

The effects of accountability on students may also interact with other state policies. In particular, the theory of action behind accountability reforms is that school personnel will adjust their behavior to increase student achievement on the incentivized tests. However if local actors have little control over the education (e.g., over budget allocation or curriculum choice), they will be less able to respond to the new incentives. Loeb and Strunk (2007) using an approach similar to Carnoy and Loeb (2002) find far greater positive effects of accountability in states with greater local autonomy. This finding is consistent with more recent cross-national studies. A series of studies making use of international assessments—PISA and TIMSS—finds that countries in which schools have more autonomy experience improved test performance in the cases in which there are mandated external school exit examinations (Fuchs and Woessmann, 2007; Woessmann, 2003, 2005, 2007). Not surprisingly, school-level accountability incentives are more salient when schools have the discretion to respond.

5.2 Early versus late adopters of accountability

The evidence suggests that accountability systems generated larger effects on achievement in the late 1990s than in the early 1990s, although Carnoy and Loeb (2005) suggest that their effectiveness may now be declining. The larger estimated effects in the late 1990s relative to the early 1990s are consistent with the observation that the programs introduced in the late 1990s were typically more ambitious than those introduced earlier in the decade. The possibility that the size of the effects is now declining suggests either that accountability generates decreasing marginal returns over time within a state, or that the early state adopters were the most likely

to benefit from accountability given their low initial test scores. A related potential explanation is that the early adopters were also more likely than later adopters—primarily those who introduced accountability in response to federal legislation—to embed their accountability systems in comprehensive standards based reform packages that included other elements such as additional funding or professional development for teachers. Although some of the studies control for certain elements of comprehensive reforms such as changes in funding, no study controls fully for all the components such as the development of organizational capacity, and investments in the capacity of teachers. In any case, both the recent decline and these possible explanations should be viewed as speculative at this time. More research would be useful.

5.3 Size and policy significance of the estimated effects

Tables 8.1 through 8.3 give an indication of the estimated size of the accountability effects. [Dee and Jacob \(2009\)](#) and [Wong, Cook, and Steiner \(2009\)](#) estimate effect sizes of up to 0.34 standard deviations for NCLB. Most of the other estimated effects range from no effect to up to 0.20 standard deviations. Judging the policy relevance and size of these effects is not easy. Achievement gaps between racial and ethnic groups and across income groups can be far larger than these gains, making the effects look small. On the other hand, these effects sizes are, in many cases, as great as a full standard deviation in teacher effectiveness as currently measured by value-added techniques (see [Hanushek and Rivkin, 2010](#)).

6. ACCOUNTABILITY AND TEACHER LABOR MARKETS

School reforms affect school personnel as well as students. Assessment-based accountability likely led to substantial changes in teachers' and principals' work lives, including increased scrutiny in the classroom, a more intense focus on student performance, and direct consequences for school funding and management. These changes, in turn, may affect career decisions about whether to join the profession, where to work, and, once working, whether to transfer to another school or to leave the profession. Likewise, accountability reforms may help administrators identify and replace ineffective teachers and principals.

As in any profession, turnover of personnel can be both beneficial and harmful. Turnover can be costly because recruitment and hiring takes time and resources away from a focus on instruction. Moreover as teachers leave they take with them specific human capital—an understanding of the school's instructional program, students, and community. New teachers need time and resources to develop these understandings. However, if the less-effective teachers leave and more effective teachers replace them, then turnover can benefit schools.

Accountability reforms can affect turnover and the cost of turnover through each of these mechanisms. They may simply either increase or reduce overall mobility. If teachers leave more or less as a result of accountability, the recruitment and hiring costs may increase or decrease. If accountability brings with it more school-specific human capital, then the cost of losing this knowledge may increase; while, if accountability leads to more similarity in the needed knowledge across schools then the cost of turnover may decrease. Finally, if accountability changes the composition of leavers, either encouraging more or less effective teachers to leave, then even the same level of turnover may be either more or less detrimental than it has been.

There are reasons to believe that accountability could change which teachers stay and which leave; and similarly, that it could affect who enters teaching. First, assessment-based accountability provides information on student performance that teachers can use to assess their own effectiveness and school leaders can use to assess their teachers' effectiveness. Teachers may be more likely to stay if they see themselves as benefiting their students. Moreover, with this information school leaders may put greater effort into keeping their best teachers and encouraging their less effective teachers to leave. Second, accountability reforms create pressure for school leaders to improve the achievement in their schools. This pressure may in turn lead these leaders to do more than they have done in the past to keep their best teachers and encourage their less effective teachers to leave. It may also lead these leaders to work harder to recruit more effective teachers. Finally, accountability may change who enters and who stays by changing the appeal of teaching differentially for more and less effective teachers. For example, if more effective teachers like the emphasis on test performance more than less effective teachers, then they might be relatively more likely to enter and stay than their less effective counterparts. The reverse could also be true.

While there are many reasons to believe that accountability policies could affect the teacher workforce, the research on the effects of accountability is relatively sparse. Ideally, we would be able to answer the following questions: Has accountability changed who enters teaching? Has accountability changed the mobility and attrition of teachers? Has accountability changed the mobility and attrition of teachers differentially across schools? Has accountability changed the mobility and attrition of teachers differentially for more or less effective teachers? We cannot answer any of these questions definitively but the extant research provides suggestive evidence.

Interview and survey research suggests that teachers feel pressure to deliver high student test scores (Barksdale-Ladd and Thomas, 2000; Hoffman, Assaf, and Paris, 2001). In addition, many teachers indicate that they view the high-stakes tests as an imposition on their professional autonomy, an invasion into their classrooms, a message that the state views them as incompetent, and a hindrance to professional creativity (Luna and Turner, 2001). However, teachers value cohesive, supportive work environments that acknowledge their efforts to promote student achievement (Johnson and

Birkeland, 2002; Luna and Turner, 2001; Heneman, 1998). Therefore, reforms, to the extent that they positively or negatively influence these aspects of the work place, will likely influence migration and attrition decisions. Disagreement with reforms is not one of the main reasons that teachers choose to leave. For example, analysis of a national survey of teachers in 2000, the Schools and Staffing Surveys, shows that less than 10% of teachers who leave indicate that disagreement with reforms was very important in their decision to leave, far below, for example the importance they place on salary for their attrition decisions. In addition, the proportion of teachers who indicate that reforms were important for them was no higher in states with stronger accountability systems (Loeb and Cuhna, 2007).

During the 1990s states varied substantially in the strength of their accountability systems. Just as Carnoy and Loeb (2002) and Hanushek and Raymond (2005) used this variation to try to identify the effects of accountability on student test performance, Loeb and Cuhna (2007) used it to try to identify the effects of accountability on teacher attrition. This cross-state analysis is constrained by the availability of national data. Yearly surveys of turnover, spanning the reform years, would be ideal; however, the only nationwide survey of teachers and turnover rates is the U.S. Department of Education's Schools and Staffing Surveys (SASS)—a nationally representative, random survey of U.S. districts, schools, and teachers—and its companion, the Teacher Follow-Up Survey (TFS). This study uses the 1993–94 and 1999–00 waves of SASS. In the year following each wave, sampled schools were recontacted to determine whether SASS-surveyed teachers had moved to a different school or left the teaching profession. A random sample of these “movers,” “leavers,” and “stayers” were administered the TFS. Unlike the similar studies of student test performance, this study finds no difference in turnover related to the introduction of state accountability system. The data used in this study are not ideal because they are based on survey responses instead of work history files, because there are only 50 states so the state-level variation in accountability has low power, and because there are a relatively low number of teachers in each state. However, it is the only study that we know that looks at the overall effect of accountability instead of the relative effect of accountability on one set of teachers in comparison to another.

Empirical research does provide evidence that accountability affects different groups of teachers differently. For example attrition appears higher in schools that are designated as low performing. Feng, Figlio, and Sass (2009) provide the most convincing evidence of this in a recent study of teachers in Florida. They exploit a rule change in Florida's school accountability system in the summer of 2002, and employ a similar identification strategy to that used by Rouse et al. (2007), Chiang (2007), and Figlio and Kenny (2009) in different contexts. Florida had graded every school in the state on a scale from “A” to “F” since the summer of 1999, based on proficiency rates in reading, writing, and mathematics. In 2002, the state changed its grading system to

both recalibrate the acceptable student proficiency levels for the purposes of school accountability and to introduce student-level changes as an important determinant of school grades. Using student-level microdata to calculate the school grades that would have occurred absent this change, they show that over half of all schools in the state experienced an accountability “shock” due to this grading change. Some schools were shocked downward to receive a grade of F, which no school in the state had received the prior year of grading. They find that schools that experienced positive shocks showed a decrease in attrition (both movement to other schools in Florida and exit from the Florida public school system), while schools that experienced a negative shock saw an increase in attrition.

This recent study mirrors earlier results in North Carolina. [Clotfelter, Ladd, Vigdor, and Diaz \(2004\)](#) used the introduction of the state level accountability system in the 1996–97 academic year to assess the differential affect of accountability on low-performing and high-performing schools. In this system, students in kindergarten through 8th grade were tested each year and, using a combination of the average level of student achievement and the yearly change in average test scores, schools were ranked as “exemplary,” “no recognition,” or “low-performing.” Low-performing schools fail to meet both the state-mandated standard for growth in test scores and have more than 50% of their students performing below grade level. Exemplary schools meet both of these requirements and teachers in those schools are rewarded with a bonus of \$1500. The paper finds that turnover increased in low-performing schools post-reform. For a typical teacher with 10 years of experience working in low-performing schools prior to the reform, the probability of leaving the school was approximately 17.6%. After the reform this increased to 19.1%. This 1.5 percentage point increase compares to a 0.5 percentage point increase for teachers who were not in low-performing schools. For new teachers, the change was 5.1 percentage points for low-performing schools and 0.8 percentage points for those in other schools. The increase in the probability of leaving was even greater for those low-performing schools labeled as such by the state. Following reform, low-performing schools saw a substantially greater increase in the turnover rate of their teachers than did higher performing schools.

A third study also provides support for the hypothesis that attrition is disproportional in low-performing schools as a result of reform. [Sims \(2009\)](#) finds that schools in California with subgroups large enough to qualify them for subgroup-based assessment are more likely to fail to meet annual yearly progress goals and are also more likely to see increased teacher attrition than similar schools with slightly smaller subgroups.

While these studies indicate that attrition increases in low-performing schools, this increase may not be detrimental if less effective teachers leave. Recent evidence across states shows that while both highly effective and less effective teachers leave schools, on average, less-effective teachers are more likely to do so ([Boyd, Grossman, Lankford,](#)

Loeb, and Wyckoff, forthcoming; Goldhaber, Gross, and Player, 2007; Hanushek, Kain, O'Brien, and Rivkin, 2005). Only the Feng, Figlio, and Sass paper has directly addressed the effects of accountability on the differential attrition of low-performing teachers. They find little clear relationship between the quality of teachers leaving and accountability pressures in schools that experienced positive shocks to their accountability grade. However, schools that experienced negative shocks, on average, loose more of their more effective teachers. Prior to the accountability system change, the average quality of those who left these schools was lower than the average quality of the stayers, in keeping with the work in other states. In contrast, after the negative shock, the average quality of leavers tended to be higher than that of stayers. In particular, these negatively shocked schools tended to lose more effective teachers to other schools in the same district. This result is particularly remarkable given the findings of an earlier paper, Rouse et al. (2007), which reports that downward-shocked schools experience larger test score gains the next year.

The research to date suggests that accountability has not dramatically changed the career choices of teachers overall, but that it has likely increased attrition in schools classified as failing relative to other schools. While increased attrition is not necessarily bad if the least effective teachers leave, the evidence suggests that it is not the least effective teachers who are leaving these schools. These results provide a warning of the potential difficulties of maintaining a stable high-quality workforce in schools classified as failing. However, the results are also not necessarily condemning of assessment-based accountability. Even in Florida, where some highly effective teachers left schools that received lower than expected scores, student outcomes actually improved, likely the result of school-level reforms (see Rouse et al., 2007). In New York State, Boyd et al. (2008) found that attrition did not increase more in grades with state-level standardized tests than in grades without these tests. In fact, attrition dropped in the tested grades and new teachers to the tested grade were on average more qualified than teachers in other grades. Even if testing is not necessarily appealing to teachers, schools were able to compensate teachers enough to increase the retention when needed. These results indicate that teacher compensation policies deployed in tandem with school accountability policies may influence the labor market implications of school accountability.

7. DIRECTIONS FOR FUTURE RESEARCH

In this chapter, we have identified design issues in developing test-based accountability systems for schools. We also have briefly described the benefits and costs of various choices inherent in the design of such systems. It is clear that there is no one ideal accountability system. The optimal system for one context and one set of policy goals is unlikely to be the optimal system for another context and another set of policy goals. Nonetheless, the research literature makes clear that these policy decisions have considerable consequences

for the distribution of student learning, for teacher labor markets, and for housing markets. As a result, these policy decisions should be made very carefully.

Extant research tells us quite a bit about the intended and unintended consequences of accountability systems, particularly those implemented in the United States since the early 1990s. Yet, there exist a number of important directions for future research. At the time of writing, numerous states and localities have been experimenting with expansion of teacher accountability. It will be important to gauge the degree to which school accountability and individual teacher accountability programs jointly affect teacher performance and decisions. More generally, there exists very little information to date on the effects of accountability programs on teacher and principal labor markets, and it will be important to observe whether results seen in one setting are replicated in other settings. The data on accountability and education labor markets that do exist study the effects of accountability on the job decisions of incumbent teachers but do not speak to the question of whether accountability systems attract a different type of potential teacher than previously occurred.

In addition, and, perhaps most importantly, the research to date tells us relatively little about the ways in which school accountability affects outcomes other than the most easily-measured test scores. There have been very few attempts to explore the impacts of accountability on higher education or labor market outcomes, which would provide a longer-term view of whether school accountability programs achieve their goals of developing a better-educated workforce. Likewise, there have been a few studies linking school accountability to proximate health outcomes such as obesity; and we know of no attempts to link school accountability to measures of nonacademic outcomes such as civic engagement, voter participation, or crime. As school accountability systems mature, it should become more feasible to study the effects of test-based accountability on these long-term outcomes; outcomes which motivate much of the reforms.

REFERENCES

- Amrein-Beardley, A., Berliner, D., 2002. High stakes testing, uncertainty and student learning. *Educ. Policy Anal. Arch.* 10 (18).
- Amrein-Beardley, A., Berliner, D., 2003. Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Responses to Rosenshine. *Educ. Policy Anal. Arch.* 11 (25).
- Anderson, P.M., Butcher, K.F., 2006. Reading, writing, and refreshments: Are school finances contributing to children's obesity? *J. Hum. Res.* 41 (3), 467–494.
- Anderson, P.M., Butcher, K.F., Schanzenbach, D.W., 2009. The effect of school accountability policies on children's health. Working Paper. http://www.bus.lsu.edu/mcmillin/seminars/anderson_accountability.pdf.
- Andrabi, T., Das, J., Khwaja, A.I., 2009. Report cards: The impact of providing school and child test scores on educational markets. Bureau for Research and Economic Analysis of Development, BREAD Working Paper No. 226.

- Barksdale-Ladd, M.A., Thomas, K.F., 2000. What's at stake in high-stakes testing: Teachers and parents speak out. *J. Teach. Educ.* 51 (5), 384–397.
- Bass, F., Dizon, N., Feller, B., 2006. Schools skirt “No Child Left Behind” rule. Associated Press, April 17.
- Black, S., 1999. Do better schools matter? Parental valuation of elementary education. *Q. J. Econ.* 114 (2), 577–599.
- Bokhari, F., Schneider, H., 2009. School accountability laws and the consumption of psychostimulants. Florida State University, Working paper.
- Boyd, D., Grossman, P., Lankford, H., Loeb, S., Wyckoff, J., forthcoming. Who leaves? Teacher attrition and student achievement. *Econ. Educ. Rev.*
- Boyd, D., Lankford, H., Loeb, S., Wyckoff, J., 2008. The impact of assessment and accountability on teacher recruitment and retention: Are there unintended consequences? *Public Financ. Rev.* 36, 88–111.
- Braun, H., 2004. Reconsidering the impact of high-stakes testing. *Educ. Policy Anal. Arch.* 12 (1).
- Burgess, S., Propper, C., Slater, H., Wilson, D., 2005. Who wins and who loses from school accountability? The distribution of educational gain in English secondary schools. University of Bristol, Working paper.
- Carnoy, M., Loeb, S., 2002. Does external accountability affect student outcomes? A cross-state analysis. *Educ. Eval. Policy Anal.* 24 (4), 305–331.
- Carnoy, M., Loeb, S., 2005. Revisiting external accountability effects on student outcomes: A cross-state analysis of NAEP reading and math results in the 1990s and early 2000s. Paper prepared for the Learning from Longitudinal Data in Education Conference. The Urban Institute.
- Carnoy, M., Loeb, S., Smith, T., 2001. Do higher scores in Texas make for better high school outcomes? Consortium for Policy Research in Education, CPRE Research Report no. RR-047.
- Chakrabarti, R., 2005. Do public schools facing vouchers behave strategically?. Harvard University, Working paper.
- Chiang, H., 2007. How accountability pressure on failing schools affects student achievement. Harvard University, Working paper.
- Clotfelter, C., Ladd, H., 1996. Recognizing and rewarding success in public schools. In: Ladd, H. (Ed.), *Holding Schools Accountable: Performance-Based Reform in Education*. Brookings Institution Press, pp. 23–64.
- Clotfelter, C., Ladd, H., Vigdor, J., Diaz, R., 2004. Do school accountability systems make it more difficult for low performing schools to attract and retain high quality teachers? *J. Policy Anal. Manag.* 23 (2), 251–271.
- Cronin, J., Kingsbury, G.G., McCall, M.S., Bowe, B., 2005. The impact of the No Child Left Behind Act on student achievement and growth, 2005 edition. Northwest Evaluation Association, Northwest Evaluation Association Technical Report.
- Cullen, J., Reback, R., 2006. Tinkering towards accolades: School gaming under a performance accountability system. In: Gronberg, T., Jansen, D. (Eds.), *Advances in Applied Microeconomics*. 14, Emerald Group Publishing Limited, pp. 1–34, *Improving School Accountability: Check-ups or Choice*.
- Dee, T., Jacob, B., 2009. The impact of No Child Left Behind on student achievement. National Bureau of Economic Research, NBER Working Paper No. 15531.
- Deere, D., Strayer, W., 2001. Putting schools to the test: School accountability, incentives and behavior. Texas A&M University, Working paper.
- Downes, T., Figlio, D., 1998. School finance reforms, tax limits, and student performance: Do reforms level-up or dumb down?. Department of Economics Tufts University, Discussion Papers Series 9805.
- Education Week, 2001. Quality counts annual report 2001. Education Week.
- Elmore, R., Abelman, C.H., Fuhrman, S.H., 1996. The new accountability in state education reform: From process to performance. In: Ladd, H.F. (Ed.), *Holding Schools Accountable: Performance-Based Reform in Education*. Brookings Institution Press, pp. 65–98.
- Feng, L., Figlio, D., Sass, T., 2009. School accountability and teacher mobility. Working paper. www.econ.wisc.edu/~scholz/Seminar/Figlio.pdf.
- Ferrer, G., 2006. Estado de Situación de los Sistemas Nacionales de Evaluación de Logros de Aprendizaje en América Latina. Partnership for Educational Revitalization in the Americas.

- Figlio, D., 2003. Fiscal implications of school accountability initiatives. *Tax Pol. Ec.* 17, 1–36.
- Figlio, D., 2004. Measuring school performance: Promise and pitfalls. In: Stiefel, L., Schwartz, A.E., Rubenstein, R., Zabel, J. (Eds.), *Measuring School Performance and Efficiency: Implications for Practice and Research*. Eye on Education, pp. 119–136.
- Figlio, D., 2006. Testing, crime and punishment. *J. Public Econ.* 90 (4–5), 837–851.
- Figlio, D., Getzler, L., 2007. Accountability, ability and disability: Gaming the system? In: Gronberg, T., Jansen, D. (Eds.), *Advances in Applied Microeconomics*. 14, Emerald Group Publishing Limited, pp. 35–49, *Improving School Accountability: Check-ups or Choice*.
- Figlio, D., Kenny, L., 2007. Individual teacher incentives and student performance. *J. Public Econ.* 91 (5–6), 901–914.
- Figlio, D., Kenny, L., 2009. Public sector performance measurement and stakeholder support. *J. Public Econ.* 93 (9–10), 1069–1077.
- Figlio, D., Ladd, H., 2007. School accountability and student achievement. In: Ladd, H., Fiske, E. (Eds.), *Handbook of Research on Education Finance and Policy*. Routledge.
- Figlio, D., Lucas, M., 2004. What's in a grade? School report cards and the housing market. *Am. Econ. Rev.* 94 (3), 591–604.
- Figlio, D., Rouse, C.E., 2006. Do accountability and voucher threats improve low-performing schools? *J. Public Econ.* 90 (1–2), 239–255.
- Figlio, D., Winicki, J., 2005. Food for thought: The effects of school accountability plans on school nutrition. *J. Public Econ.* 89 (2–3), 381–394.
- Fredericksen, N., 1994. The Influence of Minimum Competency Tests on Teaching and Learning. Educational Testing Services.
- Frey, B., 2000. Motivation and human behaviour. In: Taylor-Gooby, P. (Ed.), *Risk, Trust and Welfare*. Macmillan, pp. 31–50.
- Fuchs, T., Woessmann, L., 2007. What accounts for international differences in student performance? A re-examination using PISA data. In: Dustmann, C., Fitzenberger, B., Machin, S. (Eds.), *The Economics and Training of Education*. Physica-Verlag HD, pp. 209–240.
- Goldhaber, D., Gross, B., Player, D., 2007. Are public schools really losing their “best”? Assessing the career transitions of teachers and their implication for the quality of the teacher workforce. The Urban Institute Center for Analysis of Longitudinal Data in Education Research, Working Paper 12.
- Hamilton, L., Berends, M., Stechter, B., 2005. Teachers' Responses to Standards-Based Accountability. RAND.
- Haney, W., 2000. The myth of the Texas miracle in education. *Educ. Policy Anal. Arch.* 8 (41).
- Hanushek, E.A., Kain, J., O'Brien, D., Rivkin, S.G., 2005. The market for teacher quality. National Bureau of Economic Research, Technical report.
- Hanushek, E.A., Raymond, M., 2003. Lessons about the design of state accountability systems. In: Peterson, P.E., West, M.R. (Eds.), *No Child Left Behind?: The Politics and Practice of School Accountability*. Brookings Institution, pp. 127–151.
- Hanushek, E.A., Raymond, M., 2005. Does school accountability lead to improved school performance? *J. Policy Anal. Manag.* 24 (2), 297–329.
- Hanushek, E.A., Rivkin, S.G., 2010. Using value-added measures of teacher quality. National Center for Analysis of Longitudinal Data in Education Research, CALDER Brief 9.
- Heneman III, H.G., 1998. Assessment of the motivational reactions of teachers to a school-based performance award program. *J. Pers. Eval. Educ.* 12 (1), 43–59.
- Hoffman, J.V., Assaf, L.C., Paris, S.G., 2001. High-stakes testing in reading: Today in Texas, tomorrow? *Read. Teach.* 54 (5), 482–492.
- Jacob, B., 2001. Getting tough? The impact of high school graduation exams. *Educ. Eval. Policy Anal.* 23 (2), 99–121.
- Jacob, B.A., 2005. Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *J. Public Econ.* 89 (5–6), 761–796.
- Jacob, B.A., Levitt, S.D., 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Q. J. Econ.* 118 (3), 843–878.

- Jin, G.Z., Leslie, P., 2003. The effect of information on product quality: Evidence from restaurants hygiene grade cards. *Q. J. Econ.* 118 (2), 409–451.
- Johnson, S.M., Birkeland, S.E., 2002. Pursuing a “sense of success”: New teachers explain their career decisions. *Am. Educ. Res. J.* 40 (3), 581–617.
- Jones, G., Jones, B., Hardin, B., Chapman, L., Yarbrough, T., Davis, M., 1999. The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan* 81, 199–203.
- Kane, T., Staiger, D., 2002. Improving school accountability systems. National Bureau of Economic Research, NBER Working Paper, 8156.
- Klein, S., Hamilton, L., McCaffrey, D., Stecher, B., 2000. What do test scores in Texas tell us? *Educ. Policy Anal. Arch.* 9 (49).
- Koretz, D.M., Barron, S., 1998. The validity of gains on the Kentucky Instructional Results Information System (KIRIS). RAND Corporation, Working paper.
- Koretz, D.M., Hamilton, L.S., 2003. Teachers’ responses to high-stakes testing and the validity of gains: A pilot study. Center for Research on Evaluation, Standards, and Student Testing, CSE Technical Report 610.
- Krieg, J., 2008. Are students left behind? The distributional impacts of the No Child Left Behind Act. *Educ. Fin. Policy* 3 (2), 250–281.
- Ladd, H., 1999. The Dallas school accountability and incentive program: An evaluation of its impacts on student outcomes. *Econ. Educ. Rev.* 18, 1–16.
- Ladd, H., 2001. School-based educational accountability systems: The promise and the pitfalls. *Natl. Tax J.* 54 (2), 385–400.
- Ladd, H., Walsh, R., 2002. Implementing value-added measures of school effectiveness: Getting the incentives right. *Econ. Educ. Rev.* 21 (1), 1–17.
- Ladd, H., Zelli, F., 2002. School-based accountability in North Carolina: The responses of school principals. *Educ. Admin. Q.* 38 (4), 494–529.
- Lavy, V., 2007. Using performance-based pay to improve the quality of teachers. *Future Child* 17 (1), 87–109.
- Lee, J., 2006. in: Is test-driven external accountability effective? A meta-analysis of the evidence from cross-state causal-comparative and correlational studies, Paper presented at the annual meeting of the American Education Research Association, April.
- Linn, R.L., 2000. Assessments and accountability. *Educ. Researcher* 29 (2), 4–16.
- Loeb, S., Cuhna, J., 2007. Have assessment-based accountability reforms influenced the career decisions of teachers and principals? The Urban Institute, The Urban Institute Working Paper.
- Loeb, S., Strunk, K., 2007. Accountability and local control: Response to incentives with and without authority over resource allocation and generation. *Educ. Fin. Policy* 2 (1), 10–39.
- Luna, C., Turner, C.L., 2001. The impact of the MCAS: Teachers talk about high-stakes testing. *Engl. J.* 91 (1), 79–87.
- Mathios, A.D., 2000. The impact of mandatory disclosure laws on product choices: An analysis of the salad dressing market. *J. Law Econ.* 43 (2), 651–677.
- Mizala, A., Romaguera, P., Urquiola, M., 2007. Socioeconomic status or noise? Tradeoffs in the generation of school quality information. *J. Dev. Econ.* 84 (1), 61–75.
- Neal, D., Schanzenbach, D., 2010. Left behind by design: Proficiency counts and test-based accountability. *Rev. Econ. Stat.* 92 (2), 263–283.
- O’Day, J.A., Smith, M.S., 1993. Systemic reform and educational opportunity. In: Fuhrman, S. (Ed.), *Designing Coherent Education Policy: Improving the System*. Jossey Bass, pp. 250–312.
- Peterson, P., Hess, F., 2006. Keeping an eye on state standards. *Educ. Next* 6 (3), 28–29.
- Reback, R., 2008. Teaching to the rating: School accountability and the distribution of student achievement. *J. Public Econ.* 99 (5–6), 1394–1415.
- Richards, C., Sheu, T., 1992. The South Carolina School Incentive Reward Program: A policy analysis. *Econ. Educ. Rev.* 11 (1), 71–86.
- Rockoff, J., Turner, L., 2008. Short run impacts of accountability on school quality. National Bureau of Economic Research, NBER Working Paper No. 14564.

- Romberg, T., Zarinia, E., Williams, S., 1989. The Influence of Mandated Testing on Mathematics Instruction: Grade 8 Teachers' Perceptions. National Center for Research in Mathematical Science Education, University of Wisconsin-Madison.
- Rosenshine, B., 2003. High stakes testing: Another analysis. *Educ. Policy Anal. Arch.* 11 (24).
- Rothstein, R., Jacobson, R., Wilder, T., 2008. *Grading Education: Getting Accountability Right*. Teachers College Press.
- Rouse, C., Hannaway, J., Goldhaber, D., Figlio, D., 2007. Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. Princeton University and University of Florida, Working paper.
- Shepard, L.A., Dougherty, K.C., 1991. Effects of high-stakes testing on instruction, Paper presented at the annual meeting of the American Educational Research Association, April.
- Sims, D., 2009. Going down with the ship? The effect of school accountability on the distribution of teacher experience in California. The Urban Institute, Working paper.
- Smith, S.S., Mickelson, R.A., 2000. All that glitters is not gold: School reform in Charlotte-Mecklenburg. *Educ. Eval. Policy Anal.* 22 (2).
- Stecher, B., Barron, S., Chun, T., Ross, K., 2000. The Effects of the Washington State Education Reform on Schools and Classrooms. Center for Research on Evaluation, Standards and Student Testing.
- Stecher, B.M., Barron, S.I., Kaganoff, T., Goodwin, J., 1998. The effects of standards-based assessment on classroom practices: Results of the 1996-97 RAND survey of Kentucky teachers of mathematics and writing. Center for Research on Evaluation, Standards and Student Testing, CSE Technical Report 482.
- Stiefel, L.A., Schwartz, E., Rubinstein, R., Zabel, J., 2005. Measuring School Performance and Efficiency: Implications for Practice and Research. *Eye on Education*.
- Stullich, S., Eisner, L., McCrary, J., Roney, C., 2006. National Assessment of Title I: Interim Report. U.S. Department of Education.
- Vegas, E., Petrow, J., 2008. *Raising Student Learning in Latin America: The Challenge for the 21st Century*. World Bank.
- West, A., Pennell, H., 2000. Publishing school examination results in England: Incentives and consequences. *Educ. Stud.* 26 (4), 423-436.
- Wiggins, A., Tymms, P., 2002. Dysfunctional effects of league tables: A comparison between English and Scottish primary schools. *Public Money Manage* 22, 43-48.
- Wilson, M.B., Bertenthal, M.W., 2006. *Systems for State Science Assessment*. National Academies Press.
- Woessmann, L., 2003. Central exit exams and student achievement: International evidence. In: Peterson, P., West, M. (Eds.), *No Child Left Behind? The Politics and Practice of School Accountability*. Brookings Institution Press, pp. 292-324.
- Woessmann, L., 2005. The effect of heterogeneity of central exams: Evidence from TIMSS, TIMSS-Repeat and PISA. *Educ. Econ.* 13 (2), 143-169.
- Woessmann, L., 2007. International evidence on school competition, autonomy, and accountability: A review. *Peabody J. Educ.* 82 (2-3), 473-497.
- Wong, M., Cook, T.D., Steiner, P.M., 2009. *No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series*. Northwestern University Institute for Policy Research, Northwestern University, Working Paper Series WP-09-11.
- Yin, L., 2009. Are school accountability systems contributing to adolescent obesity? University of Florida, Working paper.
- Zimmer, R., Bettinger, E., 2008. Beyond the rhetoric: Surveying the evidence of vouchers and tax credits. In: Ladd, H.F., Fiske, E.B. (Eds.), *Handbook of Research in Education Finance and Policy*. Taylor and Francis, Inc., pp. 447-466.