CHAPTER 5

Rethinking Assessment

> Ask the teachers, ask the kids, ask the parents, and look at test scores and other measures of success. One single test score can't possibly tell the story of leadership.
>
> *Minnesota school principal, in response to a question about how to measure his leadership*

The results of the analyses in the previous chapters provide cause for concern but also for hope. Unfortunately, identification under No Child Left Behind appears to be related mostly to factors that principals and teachers cannot control. Identification also appears to produce the kinds of unfortunate bureaucratic responses that the theory predicts. NCLB is also making survival of large numbers of diverse public, alternative, and charter schools unlikely, particularly as we approach the later stages of its implementation. This large-scale failure remains highly likely even though the principals of many of these schools are doing the types of things that lead to higher-quality education. However, the underlying endeavor does not appear to be futile. Principals' leadership can have a positive effect on academic achievement, leaving open the possibility that a top-down accountability system that rewarded good leadership could lead to closing the achievement gap between minority and nonminority, advantaged

and disadvantaged students. This finding, combined with the fact that top-down accountability appears, for better or worse, to constrain and control principals' ability to lead and organize their schools, leaves open the possibility that No Child Left Behind can work if we get the assessments and the incentives right. In this chapter, I ask:

> If we wish to preserve the goals and achieve the promised benefits of top-down accountability, how can we do a better job?

Central to this prospect would be a measurement system that allowed policymakers to draw meaningful quality distinctions between schools that serve disadvantaged student populations without universally condemning them to identification and sanction. In this chapter, I consider three alternative approaches to identifying educational quality under a top-down system such as No Child Left Behind. Researchers and policymakers have already discussed the first two approaches but not the third. They represent increasingly significant modifications to NCLB's testing regimen. Briefly, they involve evaluating schools based on

> the yearly change in the percentages of students achieving state standards;
> the change in proficiency of a given student in a given year; and
> the conditions of production within the schools.

Though the specific provisions of evaluation systems may seem a bit esoteric and detailed, their effects are hardly inconsequential.

Growth Models of Assessment

The failings of the current approach of evaluating schools based only on the percentages of their students that meet a state set standard are now familiar. These systems do a very good job of singling out schools with high-minority, low-income student populations. These kinds of systems are sometimes called "status models," as they provide only a

snapshot of student performance at a given point in time and "cannot factor out year-to-year changes in student body composition or grade-to-grade changes in instructional design or teacher quality."[1] Perhaps, some observers have argued, status models are useful for descriptive purposes and for identifying schools with underperforming students so that more resources can be sent their way.[2] This, however, is not what No Child Left Behind was designed to do.

In contrast, several researchers and policymakers have proposed that we evaluate students based on changes in overall academic achievement from year to year. These are often called "gain" or "growth" models, because they reflect gains in proficiency rather than overall levels of proficiency.[3] In theory, their application is straightforward. One simply analyzes the year-by-year changes in the percentages of students—overall and, if desired, among racial, ethnic, or other subgroups of students—that meet the relevant proficiency targets.

The logic behind growth models makes intuitive sense. Schools that are coasting on the laurels of their already well-educated students will not automatically be identified as outstanding, though their high levels of proficiency would likely shield these schools from sanction in a regime that combined growth models with the status models currently in use; however, schools that took their student populations from very low levels of proficiency to significantly higher levels would be rewarded rather than punished for their performance, even if their schools still scored below a state's proficiency requirements.

Growth models present several issues, however. Perhaps the biggest challenge is the problem of regression to the mean, whereby outstanding gains in one year are likely to be followed by less spectacular results in the following year, not because of anything that is happening in the schools but because these numbers bounce around a lot and randomness and noise are inherent in any system: "A large portion of the change in test scores from one year to the next could be the result of sampling variation and other nonpersistent causes."[4] Regression to the mean is why investing in last year's hot mutual fund is usually a bad idea. In testing, like finance, past performance may not always be a good indicator of future performance.

Other problems arise as well. Any growth model must account for

the fact that given student mobility, especially in low-income areas, the group of students that takes the tests in one year will not be the same as the group tested in another year. The systemic consequences are also potentially significant. Schools that take in large numbers of lower-performing students in one year would be penalized for doing so, since the new population would be compared to a very different cohort of students. This is an issue of fairness, for sure, but it is also an issue of incentives. School administrators might be much less likely to open their doors to lower-achieving students under a public choice system if they felt that they might be penalized for doing so (although the status models currently in use also provide incentives not to reach out to lower-performing students). One researcher has commented on the political challenges likely to arise with the widespread use of gain-score models: "The public is unlikely to soon consider a school progressing from the 5th to the 10th percentile as more successful than a school declining from the 95th to the 90th percentile."[5] Finally, it is not at all clear that a lifetime of resource inequalities is less relevant to gains in academic achievement than to measured achievement in one year. Poverty's half-life may not be so short as this.

In response to some of these concerns, other variants of growth models have been proposed. One would be to track the achievement of specific cohorts of students as they progress through the schools. This "cohort gain model,"[6] while arguably providing a more accurate reflection of the experiences of specific groups of students, still suffers from some of the same problems of any gain model of assessment: student mobility and the possibility that resource differences might affect growth in achievement just as much as absolute achievement.

In November 2005, U.S. Secretary of Education Margaret Spellings announced a national program designed to encourage and allow ten state pilot programs using growth models to measure adequate yearly progress (AYP): "A growth model is not a way around accountability standards. It's a way for states that are already raising achievement and following the bright-line principles of the law to strengthen accountability."[7] While acknowledging the move to be a step in the right direction, the president of the National Education

Association quickly criticized Spellings's plan as insufficient: "Unfortunately, the Department's move does not go far enough because students in a maximum of 10 states will be able to benefit from this more reasonable and valid growth model, leaving students in classrooms in the rest of the country with the very same model the Department has identified as flawed."[8]

Several states had already begun to experiment with growth models of measurement prior to NCLB's passage or during the early years of its implementation. Florida's A+ system, for example, scores schools on a point system that involves both percentages of students meeting state standards and gains in student achievement within schools. If Florida's experience is to be a guide, then state and national policymakers should prepare for some confusion when the pilot programs attempt to fuse static and growth models within the same accountability system. As a Florida Parent-Teacher Association president remarked, "'I think there is a real lack of understanding about how these school grades are put together.'"[9]

Given these potential benefits and limitations, it seems worthwhile to begin to explore what might happen if Minnesota adopted a growth model to see if those schools identified as performing or underperforming would differ from the patterns that emerge under the current status model of assessment. Table 5 presents summary data for the 5 percent of Minnesota's public and charter schools that made the greatest gains in the percentages of their students meeting proficiency targets for the third-grade reading tests between 2003 and 2004. Table 6 presents the same summary data for the top 5 percent of gainers on the third-grade mathematics tests.[10]

Minnesota's top gaining schools in grade three reading and mathematics scores made significant progress in the percentages of students meeting the state's proficiency targets. Though, as one would expect, the top-gaining schools started from much lower levels of success—less than half of the students in these schools met reading and math proficiency targets in 2003—they increased the proficiency rate by more than 50 percent in both subjects. These top-gaining schools, however, were only more likely to make AYP based on mathematics test scores. The AYP success rates were nearly the same for the top gainers in reading as they were for the rest of the state's ele-

TABLE 5. Minnesota's Top Gainers in Reading, 2003–4

	Top 5% by Grade Three Reading Gain (1)	All Other Schools (2)
Average 2003–4 change in percentage of students meeting standards	+24%	–1%
Average 2003 baseline of percentage of students meeting standards	47%	73%
Percentage failing to make AYP in 2004	14%	13%
Percentage of these schools that are charter schools	12%	3%
Percentage of students who are . . .		
Of minority race and/or ethnicity	42%	22%
Eligible for free or reduced-price lunch	47%	27%
In special education	14%	13%
LEP	14%	8%
Number of schools	42	792

Source: Author's analysis based on data from Minnesota Department of Education 2003c, 2003d, 2003f, 2004a, 2004d.

TABLE 6. Minnesota's Top Gainers in Math, 2003–4

	Top 5% by Grade Three Math Gain (1)	All Other Schools (2)
Average 2003–4 change in percentage of students meeting standards	+26%	–3%
Average 2003 baseline of percentage of students meeting standards	47%	72%
Percentage failing to make AYP in 2004	7%	14%
Percentage of these schools that are charter schools	14%	3%
Percentage of students who are . . .		
Of minority race and/or ethnicity	34%	22%
Eligible for free or reduced-price lunch	40%	27%
In special education	13%	8%
LEP	9%	9%
Number of schools	42	792

Source: Author's analysis based on data from Minnesota Department of Education 2003c, 2003d, 2003f, 2004a, 2004d.

mentary schools, probably because although these schools made safe harbor overall, they failed to do so for a particular subgroup category, which raises an important concern. Any growth model that focuses only on overall changes runs the risk of leaving subgroups behind if the model does not account for growth in proficiency among all subgroups.

Minnesota's charter schools do much better under this simulated growth model than they do under the state's current status model. While charter schools make up only 3 percent of schools in the state that offer grade three education, they constitute 12 percent (for reading) and 14 percent (for math) of the top gaining schools in these data. The biggest gainers in reading and math are also schools that serve higher-need student populations. Minnesota's best in these simulations have nearly twice the proportions of minority and low-income students.

Growth models also appear to identify public school principals who allocate more of their time to developing the school's mission and guiding development of the curriculum and less time supervising faculty and focusing on facilities and security (table 7). This, of

TABLE 7. Leadership Patterns of Minnesota's Top Gainers in Reading and Math between 2003 and 2004

	Grade 3 Reading (1)	Grade 3 Mathematics (2)
Difference between top 5% of schools and all others in percentage of total time spent on . . .		
Facilitating achievement of the school's mission	+2%	+3%
Supervising faculty	–10%	–4%
Guiding development of the curriculum	+26%	+15%
Building relationships with the parent community	–11%	+4%
Maintaining the physical security of students and staff	–18%	–4%
Managing facilities	–19%	–10%
Completing administrative tasks	+12%	–2%

Source: Minnesota Schools Survey 2003. Test score data from Minnesota Department of Education 2003f, 2004d.

Note: The survey question was worded as follows: "During the past month, about how much of *your time* was spent on the following activities?" Each variable is coded as 1 "None or almost none," 2 "Slightly less time than on other activities," 3 "About as much time as other activities," 4 "Slightly more time than other activities," 5 "A great deal of time."

course, does not necessarily mean that these environments are less clean and safe but only that the principals probably spend less time putting out physical or social fires in spite of the fact that they tend to serve higher-need student populations.

Incorporating any variant of a growth model into No Child Left Behind's system of assessment is of course much more complicated than simply saying it should be done. Doing so involves detailed questions about actual targets and theoretical questions about building in the proper system of incentives. A particular challenge is the use of growth models when analyzing the achievement of specific subgroups. Problems of noise and variation are only compounded when one is looking at test score changes in groups of as few as twenty students, where "a dog barking in the playground on the day of the test, a severe flu season, one particularly disruptive student in a class, or favorable chemistry between a group of students and their teacher"[11] can have a significant impact on measured gains. I will take up these issues in more detail in the final chapter. For now, however, it is useful to note that, based on the small analysis here, growth models have the potential to identify schools that are doing well, especially in resource-poor communities. Any system that incorporated these models would likely not come down as hard on urban and rural public schools or charter schools as the current status models do.

Value-Added Assessment

In contrast to growth models of quality measurement, value-added assessment (VAA) systems seek to analyze educational quality in the way that the term describes—by extracting what value a school or teacher is adding to total achievement of the student in a way that accounts for all of the correlates to success or failure that I have been discussing.[12] In practice, VAA does not differ substantially from my analyses in this book, although of course I have used aggregate rather than student-level data. The similarity lies in the fact that value-added analysis uses regression techniques on a database of individual student test scores over time to account for the myriad of factors that predict academic achievement.

VAA models have gained a considerable amount of support, espe-

cially from those researchers who have focused on the methodological side of implementing No Child Left Behind's mandates. Unlike status models, which capture schools that serve disproportionately high numbers of disadvantaged students, VAA measures "can be designed to eliminate the effects of mobility, affluence, and other extrinsic factors"[13] from evaluations of school quality.

Like growth models, value-added models of quality assessment have been and continue to be part of the debate about making NCLB work in several states. Researchers employing a "massive, longitudinally merged database"[14] of Tennessee student test scores and demographic characteristics along with teacher, school, and district characteristics argued that their analyses extracted a truer measure of teacher quality than could be attained under simple status models of quality assessment. The best schools with lower aggregate test scores actually did better on these measures than schools with higher levels of total proficiency, a finding that confirmed a 2002 study of the effectiveness of value-added measures in the Milwaukee public schools.[15] Neither the Tennessee study nor the Milwaukee study, however, examined the relationship between the implementation of value-added analysis and the responses and perceptions of principals and teachers. Interestingly, the author of the Milwaukee study found that the utility of value-added analysis declined in the higher grades, a result that makes perfect sense in light of the fact that a given student's ability to make achievement gains becomes progressively more dependent on all prior educational experiences, reinforcing the predictions about the cumulative challenges of experience raised in chapter 2.

In contrast to the Tennessee and Milwaukee studies, however, other studies of states that had implemented value-added systems prior to the passage of No Child Left Behind have found that wealthier school communities fare better than students in poorer communities, raising concerns that value-added systems will lead to the same kinds of disincentives for schools and teachers to serve traditionally underperforming students.[16]

In use, value-added models have raised some new issues and requirements—in particular, some nontrivial methodological challenges.[17] To fully implement a VAA system, state agencies require a

considerable amount of demographic data for each student, and this data must be measured with as little error as possible.[18] Doing so requires at least yearly testing of all students, tracking and recording the scores of individual students as they progress through a school or move to a new one, and test compatibility across grade levels.[19] These agencies must decide which data are important to measure.[20] Is, for example, a student's race important because race by itself determines achievement, or is race a proxy for resource inequalities? In addition, many details would have to be worked out and contingencies anticipated, such as what to do if the fifth-grade test is so much harder than the fourth-grade test that every school looks like it is detracting from rather than adding to children's education during that year.

Minnesota has jumped on the value-added bandwagon, and although VAA has not yet become part of Minnesota's compliance with No Child Left Behind, it will soon be incorporated.[21] In March 2004, the Minnesota Office of the Legislative Auditor recommended that the state's Department of Education devise a "plan that outlines how value-added measures of student achievement could be incorporated into the annual AYP determination process."[22] Later that year, Governor Tim Pawlenty encouraged state education officials to take up the recommendations: "As a state, it's time to move to the next level of accountability in our schools. Based on new technologies, we now have the ability to measure individual students compared to where they were the year before. I have instructed the Department of Education to move forward in developing and implementing this 'value-added' system."[23] Minnesota lawmakers followed suit in 2004, requiring state education officials to "include, by the 2006–2007 school year, a value-added component to measure student achievement growth over time."[24]

Minnesota state officials have based their value-added hopes on the promising results of Tennessee's efforts in implementing value-added measures in comparing schools based on the quality of the instruction that they provide. I lack value-added data from Minnesota that can be matched to my survey results because Minnesota does not yet employ VAA in its No Child Left Behind compliance. Future researchers might, however, want to incorporate the theoretical perspective employed here in looking at the future VAA systems that are

sure to emerge under No Child Left Behind. A few of the school principals whom I interviewed offered, without prompting, the idea of a growth or value-added model measure of educational achievement as a possible fix for the law:

> We need to embrace longitudinal factors and readiness if we are to measure true success of schools. Someone must have derived a model formula of how to do so somewhere.

> I would expect as the AYP focuses more on student achievement, there will be a "value added" element that eventually teachers will need to respond to should students not improve from the prior year. Not to punish them, but to make them aware of individual student performance on these tests, on attendance, and on graduation rates. The district employees are in this together to help children succeed.

Missing from the recent wave of enthusiasm for VAA systems is a sufficiently detailed discussion of how the results of these regressions will be used. Theoretically, the idea makes a lot of sense; however, implementation makes the idea much trickier. State administrators initially offered VAA as a means to reward schools. With this goal in mind, it is a much more straightforward approach, although it may have inequitable consequences given recent findings that wealthier districts fare better under VAA systems. Much less thought has been given to the implementation of VAA in a system that relies solely on punishment.

A frequent argument in favor of value-added analysis is that it aims to get inside the "black box" of schools, but this is not quite the case. Rather, value-added analysis uses statistical techniques to shake the black box until the educational quality falls out. Where VAA attempts to use these techniques to measure educational quality, I have had much more modest objectives: making a case for the argument that quality does in fact lurk within these test scores. The question is whether "quality" can be measured more directly rather than shaken out of the black box. This is the third, as yet undiscussed, possibility for quality assessment under No Child Left Behind.

A Production Model of Quality Assessment

The production model of evaluation that I propose for inclusion in NCLB is nothing more than an attempt to incorporate principals' and teachers' actions into an accountability system. Rather than jumping through statistical hoops, which bring in a lot of noise and their own problems, to extract a measure of school quality from student test scores, this model is based on the premise that the quality of the principalship should be measured as directly as possible.

For guidance in devising a production model of educational assessment, I turn to several sources. The first is the effective schools literature discussed in chapter 2. Ronald Edmonds and his successors found that effective schools were characterized by safety and order, clear expectations, and strong leadership. To this, it is useful to add the recognition that schools constitute learning communities and that, as in all communities, high levels of trust and interconnectedness are crucial. Cultures, communities, and organizations that are characterized by these high levels of trust and reciprocity are said to possess high levels of "social capital."[25]

In education, the idea of focusing on the conditions of production has its roots in arguments in favor of developing "process indicators" of school performance, an approach that recognizes the fact that "schools provide educational opportunity; they do not directly produce learning."[26] The process approach in education identifies the importance of teacher and curricular quality and the critical role that nonschool factors play in producing academic achievement; however, it has placed less emphasis on school principals' role in maintaining effective school communities and has not been proposed in a system as consequential as No Child Left Behind.

Finally, I add studies of U.S. criminal justice policy by James Q. Wilson and John J. DiIulio Jr. In thinking about the problem of measuring the quality of services provided by those involved with the criminal justice system—individuals with the same high levels of autonomy and discretion that seem to confound top-down educational policymakers—DiIulio, a political scientist, proposed a "new paradigm"[27] for evaluating quality in the criminal justice system that was based on incorporating those closest to the production process

into the system of assessment. In addition to high levels of discretion, the criminal justice system shares with the educational system a need to recognize culture and climate in producing quality services. The "broken windows" model of police reform recognizes that police officers in the field are part of local cultures and societies, rejects attempts to dictate behavior from the top down, and realizes that successful police work can often be better observed in good relationships and orderly communities than with arrest data alone.[28]

Measuring production differs fundamentally from measuring outcomes.[29] It involves attention to the types of details that only a student of incentives and bureaucratic minutiae could appreciate. DiIulio's approach and that of his colleague, Charles H. Logan, was to translate an agreed-upon set of goals for these kinds of organizations and to use detailed surveys of those who deliver and receive services to measure attainment of these objectives.[30] The key to DiIulio's and Logan's suggestions was grounding assessment in the conditions of production: "Realistic measures account for the daily activities of justice agencies and for the constraints under which they operate."[31] Performance measures are no panacea, but they can—if accompanied by carefully designed sanctions and rewards—have meaningful consequences for the behaviors of the targeted agents. The trick is to align the assessment and the consequences with the ultimate goals. As DiIulio cautions, "Be careful of what you measure, for you may (or may not) get it."[32]

These ideas are consistent with the theoretical underpinnings of experiential organizations but also point out a few issues that any production model of assessment will inevitably encounter when applied to institutions such as schools, police departments, and prisons whose outputs and outcomes are measured imperfectly if at all. Organizations subject to customer choice do not face the same challenges and ambiguities as those subject to bureaucratic evaluations. They do not have the time. Customers choose. Firms respond. This is the logic behind charter schools and many other forms of public school choice. In the absence of bottom-up control, however, those aspects of production measured at the "micro-level," are likely to be among the few valid measures of police performance."[33] This requirement extends beyond the narrow scope of criminal justice agencies, and

Wilson explicitly connects the challenges and promises of the use of performance measures in criminal justice policy to the illogic of using test scores—with their multiple causative factors—and the promise of the apparently neglected effective schools literature.

I have, in a sense, come full circle, back to the conditions of effective schools; now, however, the conditions of schools are themselves quality measures rather than precursors to quality assessed by other means. A production model of assessing educational quality would, therefore, incorporate detailed surveys of students, teachers, and parents. These surveys would be used to measure the quality of leadership; the commitment to high standards and a focused curriculum; the maintenance of a safe, orderly, and inquisitive environment; and the involvement of parents in school processes.

When asked about how their leadership should be measured, the small group of Minnesota principals interviewed pointed nearly unanimously toward some sort of system based on observing and measuring the quality of what they do. Even the principal of one of Minnesota's most successful schools had a broader sense of measuring the actual conditions of production in his school:

> I believe there are numerous ways to measure my leadership. Do I have a good rapport with my staff? Do my students respond to my leadership?

His sentiments were echoed by the principal of an elementary school with large percentages of students of minority ethnicity and eligible for free or reduced-price lunch. Though his school has avoided AYP identification so far, the fact that only two-thirds of his students scored at Minnesota's level of proficiency on the last round of tests makes future sanction a near certainty without a dramatic and sustained rise in test scores.

> Academic testing needs to be one of the tools. I have no qualms with [Minnesota's] testing. I do believe there must be a better way than testing in March and returning the scores in June. One of my favorite quotes came from a workshop about ten years ago. I don't remember the speaker but he said, "Remember, cows don't gain weight while on the scales."

> Assessment should be constant and should be directly tied to lesson or instructional design. Putting kids in special testing settings, interrupting instruction for four days and then returning the results four months later in order to get [a] meaningful test score is pretty unreliable. Here are some other assessment tools to measure leadership:
>
> 1. Administrative evaluation by the superintendent. After all, he/she is the boss.
> 2. Staff surveys. Teachers and non-certified staff.
> 3. Parent surveys. Ask them how well they think they (parents) are informed.
> 4. Student surveys. Depending on age of child.
> 5. Administrative achievement checklists. Let administrators work towards achievement goals just like students.
> 6. Student achievement checklists like Work Sampling.

Others agreed that the assessment of educational quality can only occur in the actual settings where that quality is produced:

> [Spend] real time in an elementary building, follow me around for a day or two, talk with parents, look at our School of Excellence improvement plan, talk to teachers, see where we were 3 years ago and see where we are today.

For some, the culture of the school itself provides a measure of educational quality. The principal of a K–6 school in a relatively low-income neighborhood who had twenty-three years of experience as a teacher believed that test scores were not capable of measuring the kind of leadership that had brought her school, in spite of its challenges, to a five-star ranking:

> One aspect that is provided in our school that is not able to be tested is the social/emotional component. We spend a good deal of time working with students to build self-esteem and socially acceptable behaviors. This must be done to bring children to the teachable moment. This cannot be measured in a test.

> If you are looking at the administrative leadership of a school you cannot see that in a score. There are things to consider such as a successful working environment where there is a positive atmosphere and staff members' efforts are met with mutual respect. You also must consider communication skills with the staff, students, parents, and community. Finally, there is the matter of professionalism. There are policies to be followed, performance under physical and mental stress, decision making, and commitment to improving the quality of educational programs. Few of these can be assessed by the test scores of students.

Another five-star principal concurred:

> I am not sure but measuring confidence, self esteem, commitment to others, world perspective, service to society, respect for the environment would be a better test of leadership. I have a five-star school in reading and math and I pray these other areas are also 5 stars.

A critical worry that a few principals expressed is that the failure to measure educational quality in its entirety—excluding the social conditions within the schools—will damage the school culture itself. These worries echo the predictions of the theory of experiential organizations laid out in chapter 2: that experience matters to quality but that a system of measurement has the potential to change the school's experience.

> I spend much less time on non-mandated endeavors that do not show real time benefits, but are important to the overall culture and climate of our school. This includes consensus building, system reform, vision and team building, character education and at-risk programs. As the principal of a K–12 school with an enrollment of about 400 I end up taking a little away for everything. I feel the overall impact has been to move us backward, not forward. We are losing the intangibles that make this school special.

> Our test scores were quite good this year (given the socio-economics (reality) of our community) and for that we are very pleased. Two areas were exceptional—reading and writing; and one area, math, was okay compared to the state, but not as good as reading and writing. We will celebrate our successes and continue to work on improving our weaker areas and helping those students that need more help. But, we are realistic enough to know that while proficiency is an admirable goal for every student, it is not worth changing the culture of a school from one that is filled with the "joy of learning" to one that emphasizes results based on test scores alone.

The kind of approach that I propose—if used in combination with traditional outcome measures and if accompanied by meaningful consequences—offers several advantages over the use of outcome measures alone. First, by broadening the range of school quality indicators, it might mitigate the worst of the narrowing effects of test-based accountability in that principals and teachers would no longer have incentives to focus only on a narrow part of the curriculum, itself a narrow part of the education that students receive. This approach would broaden the incentives focus to include much more of what the schooling experience is actually about. In addition, "even though we do not fully understand how schools produce the results that they want, context information may provide clues to policy makers about why we get the outcomes we do."[34] In this way, production indicators potentially constitute a powerful tool for making more sense of the test-based data that we collect. Finally, it may be possible to draw quality distinctions among schools in low-income communities, a process that is not possible when the only indicator—success or failure under AYP—has already or will very soon condemn all inner-city schools to the same failure category.

The Conditions of Educational Production in Minneapolis

To explore the possibility of incorporating production measures of school quality into No Child Left Behind, I turn to a slightly different but related set of data from the Minneapolis Public Schools. As part

of a yearly self-evaluation, each spring Minneapolis administrators conduct surveys of students, teachers, and staff about school conditions.[35] The data for this analysis combine the results of the 2003–4 Minneapolis School Information Report with the same AYP, demographic, and test score data used earlier. This analysis thus uses more direct data than I used previously. I relied on indirect measures up to this point because national policymakers have not yet decided that measuring the processes of education as the outcomes is important, thereby resulting in a lack of sufficient data.[36]

The Minneapolis public schools, like their counterparts in other American cities, are failing to make AYP at an alarming rate. Every one of Minneapolis's regular middle and high schools in the district survey failed to make AYP in 2004, and the results for the city's elementary and combined elementary/middle schools were not much better (table 8). Minneapolis, like other large urban centers, also has disproportionately large high-need student populations, not only a challenge under NCLB but also a population of students that desperately needs the law to work. As in the rest of the country, many of the city's schools are progressing quickly down the AYP path, with more than 20 percent of its schools already in at least the second year of sanction as of 2004. Presuming that quality differences exist in the educations students receive in Minneapolis's various middle and high schools, then NCLB is not doing a very good job at distinguishing between the best and the worst of the city's public schools.

TABLE 8. Adequate Yearly Progress in Minneapolis, 2004

Percentage of Minneapolis public schools that failed to make AYP in 2004	
Elementary	51%
Combined elementary/middle	80%
Middle	100%
High	100%
Percentage of students in Minneapolis public schools who were . . .	
Of minority race and/or ethnicity	73%
Eligible for free or reduced-price lunch	61%
In special education	13%
LEP	24%

Source: Author's analysis based on data from Minnesota Department of Education 2003c, 2003d, 2003f, 2004d.

In this analysis, the Minneapolis survey responses serve as the basis for school production indicators in four categories: safety and discipline, curriculum and high standards, parental involvement, and social capital. The safety and discipline indicator is the average (for each school) of the percentage of "students who say that they are [not] often kept from doing their work by students who misbehave," the percentage of "classroom teachers who are [not] often kept from teaching because of student misbehavior," and the percentage of "students who feel safe in their building."[37] The curriculum and standards indicator is the average of the percentage of "classroom teachers who agree there is a high rate of consistency in how their school's curriculum is implemented across classrooms, and across grade levels," the percentage of "classroom teachers who agree that homework policies/practices among teachers are consistent at their school," and the percentage of classroom teachers who agree that they are "able to incorporate the state high standards into the . . . curricula and in their instruction." The parental involvement indicator is the percentage of "classroom teachers who report sixty percent or more of the students' families attend teacher conferences." Finally, the social capital indicator is the average of the percentage of "students who believe that students in this school show respect for the teachers," the percentage of "students who believe that the teachers in this school treat them with respect," the percentage of "students who agree that their teachers have high expectations for them," and the percentage of "students who trust the adults in this school to keep them safe."

I will use only elementary and combined elementary-middle schools for the rest of the analysis because all of the middle and high schools surveyed failed to make AYP in 2004, thereby providing little information regarding the relationship between school performance indicators and success or failure under No Child Left Behind. I am not concerned with the overall level of satisfaction on the part of students and teachers but with whether these measures of school conditions offer any help in predicting academic achievement and, therefore, with their viability as objective measures of school quality.

Figure 15 presents the results of the same kinds of simulations as I presented earlier, focusing this time on the predicted probability

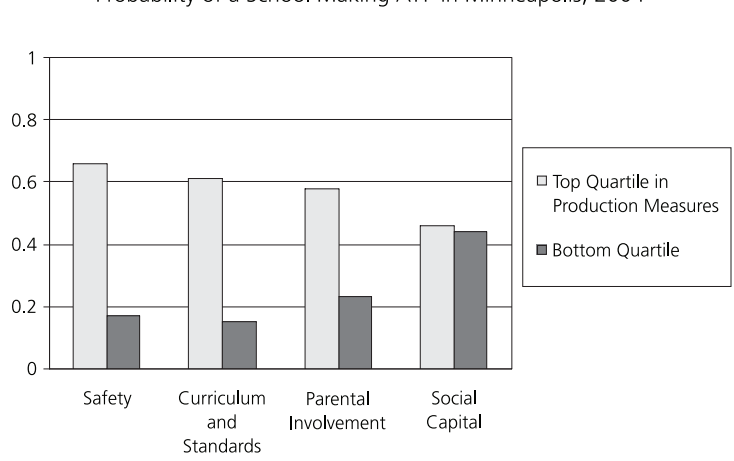

Fig. 15. AYP failure and the conditions of educational production in Minneapolis. (Data from Minneapolis Public Schools 2004; demographic and AYP data from Minnesota Department of Education 2003c, 2003d, 2004a.)

that a Minneapolis elementary or combined elementary/middle school will make AYP depending on how well the school fared on the student and teacher evaluations of the conditions of production within the school. In other words, the simulation results in figure 15 allow us to consider the effects of better conditions of production on the probablity of AYP success, controlling for the many influences on test scores and outcomes that principals and teachers cannot control. The data lack direct evaluations by parents, though these would be a necessary addition to a collection of school production indicators. As before, regression results are presented in the appendix.

The simulations and the underlying regressions present striking results. The quarter of Minneapolis's regular elementary and combined elementary/middle schools whose principals achieved the highest levels of parental involvement, focus on standards, and safety and discipline were between two and a half and three times as likely to make AYP than the schools in the lowest quartile. All three of these relationships are statistically significant, which is remarkable given the small sample size in these analyses. The social capital variable is not significant, though it would seem almost essential to have a mea-

sure of social capital that included data on parental–public school trust, which these data lack.

The diversity of Minneapolis's urban school system has survived suburbanization better than the schools in many other cities. Even within such a system, the concern arises that the results in figure 15 capture only the fact that a number of Minneapolis's public schools are wealthier and less diverse, score higher on many of the production indicators, and made AYP only because of their community characteristics. In other words, perhaps figure 15 illustrates only different patterns of principalship in different communities, not the effects of that leadership on a school's success or failure. If that were the case, then these results would offer little of value to educational policymakers.

A considerable amount of variation exists in Minneapolis's schools, both in the percentages of minority students that they enroll and in the scores that these schools received on student and teacher evaluations of educational production within the schools (table A14). The two most significant differences between the four groups of Minneapolis public schools are the percentages of minority students that they enroll and the levels of parental participation that they observe. These findings are consistent with what we know about resource inequalities and participation in politics and within schools.[38]

Figure 16 presents the simulated results of the same underlying regression models as in figure 15, but this time the quarter of Minneapolis elementary and combined elementary and middle schools that had the lowest percentages of minority students have been deleted from the data set. Although part of the Minneapolis school system, these schools are located in urban enclaves with higher home prices and long lines at the school information fairs that the school district holds every year. Their neighborhoods are characterized not by liquor stores and check-cashing facilities but by boutiques, coffee shops, and organic food.

Two things are remarkable about the results in figure 16. First, given a sample size of only forty-two schools, any of the underlying differences between the best- and worst-scoring schools on my production indicators are significant, which they are for all but the social capital measure. Second, excellent leadership appears to matter

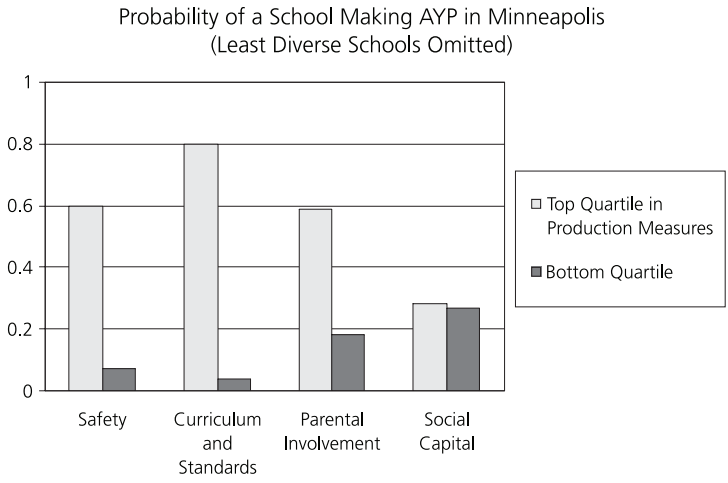

Fig. 16. Producing quality in Minneapolis's neediest schools. (Data from Minneapolis Public Schools 2004; demographic and AYP data from Minnesota Department of Education 2003c, 2003d, 2004a.)

more for lower-income schools than for those schools gifted with students who know more about privilege than about deprivation.

Principals of Minneapolis's public schools with the strongest focus on curriculum and standards—as determined by the evaluations of students and teachers—are twenty times more likely to make AYP than those schools with the lowest levels of safety and discipline. Those schools with the highest levels of safety and discipline were more than eight times as likely to make AYP, while those that did the best job of involving their parents were three times as likely to make AYP.[39] Those schools that failed in these three areas were almost certain to fail to make AYP. These findings confirm what Edmonds and those who followed him predicted: effective schooling has the greatest benefit in the schools that need it the most.

Conclusions

Test scores at best measure only indirectly what we really care about in education. No Child Left Behind seeks to create a group of princi-

pals and teachers who focus on what really matters in education: maintaining a safe and orderly environment, reaching out to and involving the parents, and guiding the curriculum by setting an overall tone of high expectations and by having excellent teachers and letting them do their jobs. No matter what kinds of machinations we undertake to extract these realities from test score results, we will always be doing so in a way that is less than perfect. Perhaps most troubling about No Child Left Behind, the law exhibits a lack of faith in the ability of those closest to educational production—teachers, principals, parents, and students—to tell us about the quality of education in their schools, opting instead to place this power in the hands of those very far from the experience of educational quality.

As I have shown, however, another possibility exists, and it has deep roots in thinking about schools and can avoid many of the unintended consequences of No Child Left Behind's indiscriminate nets. Critical readers may challenge the assertion that we should incorporate production measures into NCLB's assessment regimen on the grounds that because leadership can matter for AYP success and failure, perhaps we should stick with what we have. However, the models in these analyses control for community characteristics and compare schools within similar communities. NCLB currently does neither of these things. Rather, I conclude with some confidence that if leadership matters—even within a rigid framework such as No Child Left Behind—then we can use more direct measures of leadership to make quality assessments between schools in similar communities without condemning the public schools of an entire city to sanction and closure. The system of assessment and identification proposed here—to return to the analogy with which I began my empirical investigations in chapter 3—is dolphin friendly. It has the potential to keep the best public, charter, and alternative schools out of our nets and, if we are bold enough, to reward their principals and teachers for performance under pressure. We can make No Child Left Behind work, but doing so will require no small amount of humility, dispassionate analysis, and creativity. We need to get back to the basics of educational quality.