to work harder after failure. But there is a real world that differs from this ideal world. In the real world, success on a test is often attributed to one's inherent intelligence, one's social class standing and its attendant advantages, or to just plain luck. Students who feel they succeed for these reasons feel little motivation to work hard to pass tests. These students have no reason to believe success is connected primarily to one's effort. Worse, however, is that those who fail at tests often attribute their failure to a lack of intelligence, a belief communicated to them also by schools that see those lower-achieving students as hurting the school's reputation because of their "low ability." A new form of discrimination is apparently creeping into our schools, and it is against the score suppressors, those children who keep the school from looking good on high-stakes tests. These students have little motivation to work harder at preparing for tests since they have learned from the messages they receive from other students and their teachers to accept that they are the "dumb" kids. On the other hand, some failing students attribute their test failure to bad luck or teacher discrimination. Students who attribute failure to low ability, to bad luck, or to teacher discrimination are not likely to make any attempt to study harder the next time the test is given. These students do not make increased efforts to pass the test after failure and, in fact, often choose to drop out of school rather than confront more indications that they lack the academic ability to succeed in a test-oriented environment. Dropping out of school is a perfectly sensible way to protect one's ego when it is under attack, as it always is for the least academically able in an environment that overly values high academic achievement as defined by standardized test performance.

4. *Students and teachers need high-stakes tests to know what is important to teach and to learn.* Response: For decades teachers and administrators have had scope-and-sequence charts at the state or district levels, clearly outlining what is to be taught and at what grade. In more recent years we have had the national professional associations, such as of math or science educators, informing us about what should be learned, and when, and even how these disciplines might best be taught. Textbook publishers responding to these ideas have produced texts that also help teachers and administrators to know what to teach and when. Even more recently,

states have developed their own curriculum standards that inform all teachers, parents, and citizens in the state what is to be taught throughout grades K–12. In each case the curriculum that is adopted by schools should influence the test-makers. The test-makers should not be the ones influencing the curriculum of the schools. Deciding on a curriculum is a difficult and time-consuming job, perhaps the most important in education. There really seems to be no reason at all to have schools if a community cannot answer the (deceptively simple) question: What is it we want the young to know and be able to do? From those answers come the reasons for schooling and the design of curricula. A test must be developed along with a curriculum or after it is decided upon. The test must never dictate what is to be taught and learned. Otherwise, it is a clear case of the tail wagging the dog. Sadly, in high-stakes testing environments, we often see the test overinfluencing teaching, resulting in a narrowing of the curriculum offered to students to just what is on the test.

5. *Teachers need to be held accountable through high-stakes tests to motivate them to teach better and to push the lazy ones to work harder.* Response: Here is the heart of NCLB. This is the theory of action behind the law. This law is designed to push lazy teachers and lazy students to work harder. It is based on the premise that children and teachers are not performing as well as they should, an easy belief to hold and an impossible one to verify empirically. Based on our hundreds of school visits, we have come to believe that the percentage of lazy teachers amid the 3.5 million teachers we have is considerably smaller than the percentage of lazy politicians who do not read the legislation they support. Our visits to schools, as well as some research reports, lead us to believe that NCLB and its reliance on high-stakes testing actually is motivating teachers of poor students to work much harder, though it seems to have little effect on teachers of more advantaged students. Although NCLB is influencing the work of some teachers, primarily those who teach the poor, are we concerned about whether they work smarter, better, and in the best interests of their children? Are they violating their own professional norms of behavior as they engage in preparing their students for high-stakes tests? Are they training their students much more than they are educating their students? Is the work involved in high-stakes testing environments so stressful that teachers

choose to leave the schools they are in for work at schools where the children are more advantaged? Are teachers leaving the profession altogether because of the stress they are under? Although NCLB seems to strongly influence the work of those who teach the lower social classes, it is not at all clear that the increased work is a source of benefit to the students or a source of satisfaction to their teachers.

6. *The high-stakes tests associated with NCLB are good measures of the curricula taught in school.* Response: A typical achievement test might have 50 or 70 multiple-choice items, perhaps with an open-ended item or two, but rarely are such tests much longer. On the other hand, the curriculum in, say, reading and language arts in the fourth grade, or algebra in high school, is quite enormous and varied. The test, therefore, is always a small sample of a much broader curriculum and can never be relied upon to be a completely trustworthy measure of what students know and are capable of doing. Furthermore, while standards-based instruction has narrowed the curriculum a great deal, there is still enormous variation in the implementation of any curriculum in a classroom. The intended curriculum, which the test covers, always differs substantially from the implemented curriculum, the curriculum actually covered in a class by teachers and students. The amount covered per year, the depth of the coverage attempted, the variety of the examples chosen for discussion in class, the use of homework for promoting transfer of learning, and so on: these vary considerably from teacher to teacher and school to school. The only way to control these variations in the implemented curriculum is through scripted lessons by teachers or online, highly controlled exposure to a single teacher and classroom format. In more traditional instructional settings, however, the variations in the implemented curriculum mean that any reasonably brief standardized assessment may not be a good measure of what was actually taught in schools. And this problem of a match between the test and the implemented curriculum is only magnified when it comes to special-education learners or those for whom English is a second language.

7. *The high-stakes tests provide a kind of level playing field, an equal opportunity for all students to demonstrate their knowledge and skill.* Response: The reason that lawsuits against high-stakes testing exist in California, Arizona, and elsewhere is simple: plaintiffs claim they have not had the opportunity

to learn what the tests assess. Imagine a low-income Mexican American student from a small town in an agricultural area of California or Arizona. When compared to a wealthier, white, suburban student, the chances are much greater that the rural Mexican American student had lower-quality teachers, newer teachers, more teachers teaching mathematics and science without mathematics and science majors or minors, less money spent per student per classroom, poor or no bilingual services, fewer opportunities to have advanced-placement courses, less counseling help, access to fewer libraries, and those libraries had fewer books in them. The high-stakes high school exit exams in California and Arizona are predicated on a belief that all who take them have had equal access to a high-quality education. That is patently not true in these cases, and that is why lawsuits challenging the existing exams are plentiful. Until white politicians of the middle class can look poor minority students in the eye and say, "You have had the same high-quality schooling my own children received," there is no level playing field. Legislatively requiring all students to achieve the same level of proficiency, in exactly the same amount of time and under vastly unequal conditions of schooling, appears to us to be more than a pipe dream—it is also immoral.

8. *Teachers use the results of high-stakes tests to help provide better instruction to students.* Response: If tests are formative, designed to provide information about who knows what in a class and whether a teacher needs to reteach a lesson or unit or not, then tests are being used appropriately. Most teachers know how to make adjustments to lessons and to accommodate the individual differences among students as a function of student performance on a test. This is as it should be. But the vast majority of high-stakes tests are given in spring, with results coming back in the summer, and so the testing schedule provides no time to adjust instruction for individual children. Typically, the test informs a teacher about last year's students and provides little guidance about teaching this year's students. Unfortunately, when stakes are high and scores are not at the level desired, the test is used to determine whether more drill or more test preparation is needed. Low scores could mean that more time is needed in the subjects not showing the expected results, thus leading to a decrease in the time teachers spend on nontested subjects. It is questionable, then,

whether receipt of high-stakes test results lead to "better" instruction. Test results may only lead to more efficient instruction that better prepares students for the test.

On the other hand, when test scores go up, it is likely that the inference will be that better instruction is taking place, supporting the argument made by those who approve of high-stakes testing. But this is not necessarily so. Better scores could be a result of more instruction, or more targeted instruction, not better instruction, and a score gain is equally likely to be the result of other forces affecting achievement. Typically, high-stakes testing takes place in an environment where other school reforms and systemic changes are taking place simultaneously. So, along with the imposition of high-stakes testing in every state, the states are also implementing curricular reforms; increasing teacher professional development; changing teacher-education programs and developing mentoring programs for the first few years of teaching; purchasing new standards-oriented textbooks; and increasing the time spent by students in summer school, after-school programs, tutoring programs, and so forth. So even when there are gains on state tests, it is not at all clear that a high-stakes testing program is the reason for those gains. One concrete example of this is that the number of substantive courses taken by high school students went up substantially in the 1990s. The number of years of English, mathematics, history, and science has gone up in every state. And advanced-placement courses have seen huge enrollment increases in the last decade. If scores on a state high-stakes test then go up, is it due to high-stakes testing or because a revised high school curriculum has been put into place? There is no way of knowing the answer to this question without enormously expensive research. Absent such research, we think that the curriculum changes are more likely to lead to higher test performance than the use of high-stakes testing. Actually, the best evidence currently available to answer the question whether high-stakes testing leads to better education comes not from a state's own test but from audit tests. These are tests of similar content, but not the tests for which a state has prepared its students. The National Assessment of Education Progress (NAEP) tests serve this audit function. Currently, there is no evidence that the introduction of high-stakes testing in the states has changed the

growth curves on the NAEP assessments.[22] Although the results are mixed, we find no convincing evidence to suggest that increases in students' learning on a state's own high-stakes test generalize to other indicators of achievement.[23]

9. *Administrators use the results of tests to improve student learning and to design professional development.* Response: This rationale is flawed for many of the same reasons we have already discussed. Test results are imperfect measures of what students know and of how well teachers are teaching. Therefore, they do not provide enough information regarding what needs to be changed or how it needs to be changed in order to produce test-score gains. Administrators simply cannot make sensible decisions regarding what the strengths and weaknesses of their teaching staff are on the basis of test scores alone. Administrators require a range of information to make decisions about the personnel they supervise and the nature of instruction at their school, including teacher observations, teacher reports, student reports, and meetings with parents (to name a few). Test scores provide only a small piece of the information administrators need to make the complex decisions they must make. Even when scores on tests go up or down, there is no reason to believe that "good" or "bad" instruction is taking place. What has to be taken into consideration is the demographic makeup of the new class and the old. Two special-education children, two English-language learners, two divorces by the parents of students in the classroom—all make a teacher's classroom from one year to the next highly variable. Even the reliability of scores for a class size of 25 suggests that swings in achievement-test performance from one year to the next could be substantial and unrelated to a teacher's efforts. Moreover, given the known correlation of poverty to school achievement,[24] even when scores do not move up there is no reason to think that teaching is poorly done. Test scores cannot stand apart from these other factors and play only a limited part in judging teachers' performance.

10. *Parents understand high-stakes test scores and can use them to interpret how well their child is doing in school.* Response: Citizens clearly do understand what it means to pass and fail a test, something that is celebrated and admonished over and over again in the press and the subject of conversations in the home. But neither newspapers nor parents understand

the complexity and completely arbitrary nature of determining a cut-off score for a test, the score that is used to determine who will pass and who will fail. Determination of a cut-off score is always difficult to justify, even when done carefully and rationally. No one, not the expert psychometrician, wise business leader, or concerned housewife, has any way to justify that some point on a scale really separates out those who understand something from those who do not. Even among scholars there is no consensus on how to establish a fair way to judge what it means to be proficient and what it means to be less than proficient. There is no way to defend that a test score of 47 indicates proficiency in a subject while a test score of 46 does not. Since experts cannot agree on these things, there is no reason to believe that parents and other citizens will have the knowledge or access to resources to understand the results of high-stakes testing. For example, someone who believes parents are helped by the clarity assumed to be a part of high-stakes testing programs might want to explain to a Pennsylvania parent why their student is a failure (that is, not proficient) when in eighth grade they can do the following:

> An eighth-grade student performing at the Basic Level solves simple or routine problems by applying skills and procedures in the five Pennsylvania Mathematics Reporting Categories. The student performs simple computations on fractions, integers and decimals, including powers; uses the order of operations to simplify basic numeric expressions. The student converts basic customary and metric units of length, capacity and time to one unit above or below (e.g., seconds to minutes); selects and uses correct formulas to calculate basic measures of simple two- and three-dimensional geometric objects. The student matches simple prisms with nets; recognizes properties of angles formed by intersecting lines; identifies or locates points on a coordinate plane. The student extends basic numeric or algebraic patterns; solves simple equations and uses substitution to check the accuracy of the solution; matches a linear graph to a table. The student identifies correct graphical representations for sets of data; calculates simple probability for mutually exclusive events; identifies basic correlations in scatter plots.[25]

A student who fits the above description is a failure, possibly to be retained a grade, because she is not "proficient." But we think this is a

description of a quite competent student, and were we a parent, we would wonder why our child and our school was declared to be failing. In fact, there is even empirical evidence that neither subject-matter specialists nor parents can easily interpret the descriptors that accompany some high-stakes tests.[26] We see more of an illusion of clarity than we see genuine understanding by parents and other citizens of what a test score means.

We have now presented many of the arguments supporting the use of high-stakes testing as a lever for improving public schools. They appear to be rational and seemingly straightforward explanations for why high-stakes testing will successfully improve American schools. Closer analysis of all these arguments, however, uncovers many flaws. The arguments often consist of half-truths—that is, truths for some students and teachers but not for most or all students and teachers. Some of the arguments are clearly debatable and likely will always be unconvincing to someone who already has their mind made up. But some of the arguments are simply unfounded. What follows in this book are more reasons to question the use of high-stakes testing as a lever for school change. Although we are sure that high-stakes testing is fatally flawed in theory and in action, these tests have become a part of contemporary American life. Next, we examine why that is so.

WHY HAS HIGH-STAKES TESTING SO EASILY SLIPPED INTO CONTEMPORARY AMERICAN LIFE?

Why has high-stakes testing so easily, so dramatically, and so recently become a part of contemporary American life? We offer five reasons for that change.

First, and the most popular explanation, is one that notes the co-evolution of the prominence of business and accountability in our daily lives. Accountability in education is modeled on corporate efforts to increase productivity. This reflects a larger trend toward seeing society as modeled on the corporation rather than the family. Tax policy, government spending, health care, employment training, and educational policy have all been strongly influenced by the corporate model. Policymakers have

Box 1.1 What Have We Done to Our Students?

From a *Dallas Morning News* editorial, July 2006

When I was teaching sophomore English in 1998, one of my students, a stocky 16-year-old football player, came up to me one day after class to say he wanted to transfer out. His last English teacher, he said, spent much more time preparing her class for the state's standardized assessment test, mostly by having students bubble in sample tests. He had decided my class, where we analyzed poetry and wrote essays constantly, wasn't going to help him pass the test. "If I fail, Miss, it's going to be all your fault."

Macarena Hernandez, "Test Pressure Is Getting to Our Schools: It's Inspiring Cheaters and Stifling Real Learning," editorial, *Dallas Morning News*, July 28, 2006.

applied basic Business 101 models to our schools—namely, to find ways to monitor productivity, then increase it, and to do so without spending any more money. Tests were chosen as the mechanism to measure productivity. Like those in the business community, legislators believed that productivity could be increased without more money needing to be spent simply by holding schools and educators accountable through the practice of high-stakes testing. Lazy teachers and students would be detected and made to work harder. The models of accountability used in business could be applied to the inefficient school systems of America and—voilà!—the schools would improve. For many Americans, the analogy to business seems sensible and worth pursuing, so it was easy to buy into the high-stakes accountability movement.

The analogy doesn't really fit, however. The production of widgets is easier to count than the knowledge and skill possessed by students. Essentially, a widget is a widget but a well-educated student is a good citizen and a caring person, has aesthetic sensibilities, good habits of health, and so forth. These are outcomes our citizens demand that we produce through our schools, but they are never assessed by tests. So productivity for our teachers and our schools means something vastly different than

Box 1.2 What Have We Done to Our Teachers?

From Ann, a first-year teacher

Last year, when I was a college student, I had great ideas for using hands-on activities and cooperative learning in my classroom as a way to get students to be internally motivated for learning. With the testing programs we have in this school, there isn't much leeway for me to be creative or innovative and create excellent lessons. The test is the total goal. We spend every day doing rote exercises. Forget ever doing hands-on activities, science or math games, or creative writing experiences. We do one hour of sit and drill in each of the subjects of math, reading, and writing. We use a basal reader, math workbook pages, and rote writing prompts. It is all step by step; the same thing every week. I have to teach the letters by saying "A, what is A?" I repeat this over and over in a scripted lesson that my principal makes me use. You can't improvise, add, or take away. You read exactly what it says. This is how testing has impacted my school and my teaching. As a first-year teacher I feel like I don't have a choice to deviate from this awful test preparation.

M. Gail Jones, Brett Jones, and Tracy Hargrove, *The Unintended Consequences of High-Stakes Testing* (Lanham, MD: Rowman & Littlefield, 2003).

productivity in a manufacturing plant or productivity in the delivery of routine services. Furthermore, when inputs cannot be controlled, it is difficult to assess outputs. Measuring the production of widgets assumes control over the quality of the inputs needed for the production of the widgets. But in education we have little control over the input side. For example, students from poor and middle-class families, from immigrant and nonimmigrant families, from two-parent and single-parent families, from medically insured and noninsured families, those with and without disabilities, and those with and without English as a native language may all be in the same classroom or school. This represents heterogeneity in school inputs that would drive quality-control personnel from manufacturing crazy! Ordinary ways of measuring productivity appear to be sensible, but they do not work as easily in educational settings. The high-stakes tests, with their threats and incentives, are not well matched to the