Introduction and Overview

Williamson M. Evers and Herbert J. Walberg

Schooling is one of the top domestic policy issues of the day, and testing and the effectiveness of teaching, broadly considered, are among the top issues in education. Nearly all states have developed standards and have begun state testing programs in the last several years. The 2002 federal No Child Left Behind act makes testing and accountability policies even more crucial because poorly performing schools may be closed; already many failing schools must allow and pay for their students to attend successful schools.

More than ever, parents want to know how their children are achieving and how their children's school ranks compared with others or with standards. Testing and evaluating districts', schools', and staff members' teaching results are enduring concerns. Today they are particularly timely and are the reasons for this book.

Purposes of Testing and Public Policy

Of course, tests can serve a variety of purposes. For example, educators can use them to pinpoint students' strengths and weaknesses to plan curricula and adopt teaching practices tailored to their needs, both as individuals and groups. State legislators increasingly want to know how schools rank, and local school

boards should be studying the results of their programs, curriculum offerings, and staff efforts. Parents can contribute more to their children's learning if they understand their progress, and, increasingly, they can choose their children's school partly on the basis of publicly available school report cards, which reveal, to a greater or lesser degree, the quality or effectiveness of teaching.

Achievement test scores certainly do not reveal all the important outcomes of schooling nor can they form a comprehensive index of the quality of teaching. Nonetheless, well-designed commercial and state-developed tests usually provide reliable indications of the academic knowledge and skills children acquire largely in school, and no one has shown that these reasonable goals require sacrificing other objectives, such as character development. Even though some education experts and even some testing experts may disagree, Congress, state legislators, and citizens are increasingly insistent on such objective testing and accountability for measuring the results of teaching.

Such priorities are matters of public policy to be decided by citizens and their representatives rather than only by educators and testing experts on professional and technical grounds—since their interests may not be identical. The authors of the chapters in this book unabashedly and critically examine controversial professional, technical, and public policy issues that may divide educators and experts from citizens and their representatives.

Setting the Stage

Herbert Walberg begins by showing why citizens and legislators have become increasingly concerned about American student achievement and why they increasingly maintain that tests and standards are necessary. Though the achievement of American students is comparable with that of students in other countries when American students begin school, they fall increasingly behind as they progress through the grades. By the end of high school, their achievement is near the bottom of advanced countries, despite American schools' being close to the top in per-student

spending among economically advanced countries. Walberg attributes this productivity problem to a lack of school board and staff accountability—which, in turn, requires systematic testing and standards. Among the other problems he identifies are defective tests and standards and the proclivity of educators to promote and graduate students even when they have not met proficiency standards. Ending on a positive note, he identifies ways that tests can be used to help solve America's achievement crisis.

Oddly, in this period of national crisis, some prominent testing experts have objected to testing's having an enlarged role when it comes to making "high-stakes" decisions about student promotion and graduation and to evaluate the teaching provided by districts, schools, and individual staff members. Richard Phelps describes eight common objections and shows why they are false. Among the myths he debunks are that learning is narrowed because teachers concentrate solely on what is tested; that standardized tests measure only facts; that standardized test are biased against minorities; and that standardized tests are too expensive.

Constructive Uses of Tests

This section presents several constructive uses for tests including (1) diagnosis of children's learning difficulties and evidence-based procedures for solving them, (2) the study of curriculum impacts on specific aspects of achievement, and (3) assessment of teachers' strengths and weaknesses. First, Barbara Foorman, Jack Fletcher, and David Francis cite research revealing that a weak start in reading usually prevents children from catching up with their peers, then they show how tests can help even in the child's earliest years of learning to read. They are careful to point out, however, that their own research shows assessment of early reading skills is useful only to the extent that teachers understand and act upon the results. Such assessment can help teachers, parents, administrators, and policy makers in judging the effectiveness of programs as children move forward in school.

Though agreeing that testing of students is critical for educational reform, Stan Metzenberg takes issue with the validity of the NAEP Science and Mathematics tests, finding it suspicious that student performance should correlate strongly with Reading test scores. He suggests that the learning of mathematics and science may depend upon foundational reading skills that are not supported by hands-on activities and that mathematics and science tests that are largely based on constructed-response questions may fail to serve their intended purposes. Metzenberg also finds that previous interpreters of TIMSS data are mistaken in their suggesting that too many topics are covered in the U.S. curricula. Overall, Metzenberg calls for exercising great caution in interpreting the results of NAEP Science and Mathematics tests and in drawing causal conclusions from sloppy educational research.

Alan Siegel, on the other hand, finds much value in the TIMSS filming of teaching practices. As part of the TIMSS achievement survey, researchers filmed eighth-grade mathematics lessons in Germany, Japan, and the United States. Siegel's detailed analyses of the films suggest the reasons for the outstanding performance of Japanese teachers. Like Metzenberg, Siegel finds previous assertions about TIMSS mistaken. He concludes, for example, that Japanese teachers actively teach students rather than letting them discover mathematical ideas on their own. In fact, Japanese teachers engage in more lecturing and demonstration than even the most traditional American teachers. In short, Siegel's analyses refute the claim of many education theorists that student discovery rather than expert teaching primarily determines outstanding performance of Japanese students on the mathematics tests.

Constructive Tests for Accountability

Essay examinations, live performances, and portfolios of students' work can provide insights for classroom teachers about what their students have learned. But should they be used, as many contend,

for purposes of large-scale accountability? The chapters in this section refute this common contention.

Brian Stecher reviews California, Kentucky, Pittsburgh, Pennsylvania, and Vermont "portfolio assessments," that is, ratings of collections of students' work products. He finds the poor reliability of the ratings severely limits their validity in measuring student progress and the teaching effectiveness of districts, schools, and teachers. Portfolio assessments, moreover, are expensive to score compared with multiple-choice tests, and they require large amounts of teacher and student time without adding to the validity of accountability programs. Portfolios are, however, effective tools for changing instructional practices, and their greatest potential may lie in their use as classroom assessment tools rather than large-scale accountability measures.

William Mehrens similarly concludes that "performance assessments" are problematic in providing useful information for holding educators accountable. Performance assessments require students to "construct" answers rather than to choose the best answers, as on multiple-choice tests. Performance assessments usually do not meet technical standards of reliability, validity, and objectivity and are subject to legal challenge when used for purposes of accountability. They are more expensive and more subject to bias and breaches of security than multiple-choice tests, which have a long record of measuring student knowledge and skills effectively, efficiently, and objectively.

State Testing Policies

State legislatures have largely initiated testing initiatives to hold school boards and educators accountable for student achievement. Even the new federal No Child Left Behind act gives states considerable latitude in determining the nature and content of accountability tests. For these reasons, the last section of this book presents case studies of successful and unsuccessful state testing initiatives and concludes with recommendations for improved policies.

In a case study of Kentucky's state testing and accountability program, George Cunningham describes the lessons that can be drawn from a state's failure to carefully plan and execute education policies. In his view, these lessons include (1) avoiding the claim that all students can learn at the same high level, (2) delineating the content to be covered by the curriculum and tests, (3) employing normative rather than absolute standards, and (4) preferring multiple-choice tests. Cunningham provides a number of other insights that should be considered by policy makers.

Darvin Winick and Sandy Kress offer a counterexample. State testing policy, they show, has worked well in Texas. They identify four factors as most important: leadership, accountability, decentralization, and external pressure for achievement results. Their extensive consideration of accountability is highly pertinent to this volume. In Texas, the state's successful accountability system included clear curriculum standards, objectives for each grade and campus, widely distributed reports on student achievement, and substantial consequences for accomplishments and failures. Several of the elements of the Texas accountability system form the basis of the recent federal No Child Left Behind act, which is intended, among other things, to influence state-level testing and teaching accountability in other states.

Conclusion

Well-educated young people tend to prosper and contribute much to our economic, cultural, and civic life. Yet American educators and students are not living up to their potential, even though taxpayers have generously supported their efforts. The chapters in this book show the ways forward: Tests results can show educators' and students' strengths and weaknesses as a basis for planning. Test results can inform educators and students of their progress or lack thereof and thus serve to reward and sanction their actions. Test results can reveal the degree to which educational products, programs, and practices are working and thus inform state and local school boards about choices they face. In

these and other ways described in the subsequent chapters, tests can play a vital role in improving American schools.

This book stems from an October 1998 Hoover Institution symposium entitled "Testing America's Schoolchildren." The talks by George Cunningham, Barbara Foorman, Sandy Kress, Stan Metzenberg, Brian Stecher, and Herb Walberg at that conference were based on papers written for this volume. These papers have been revised in the intervening time to bring them up to date. In addition, the editors selected three important articles written in the past decade by Richard Phelps, Alan Siegel, and William Mehrens to round out the book's coverage of testing and teaching practices.[1]

The editors wish to thank all those who contributed their papers or wrote original chapters for this volume. In addition, we thank Marion Joseph and Jerry Hume for inspiration and suggestions on participants for the original symposium. Hoover Institution director John Raisian supported this endeavor as part of the institution's Initiative on American Public Education. Senior associate director Richard Sousa aided in the preparation for the symposium and the publishing of the book. Executive editor Patricia Baker oversaw the production of the book, and Ann Wood and Joan D. Saunders were responsible for the copyediting. Kate Feinstein and Elizabeth Maples provided the editors with research assistance for the preparation of this volume.

Notes

1. Original places of publication: Richard P. Phelps, "Why Testing Experts Hate Testing," Fordham Report (Thomas B. Fordham Foundation) 3, no. 1 (January 1999); Alan R. Siegel, "Effective Teaching and the TIMSS Observational Study," http://www.cs.nyu.edu/cs/faculty/siegel/; William A. Mehrens, "Using Performance Assessment for Accountability Purposes," *Educational Measurement: Issues and Practice* 11, no. 1 (spring 1992), 3–9, 20.