DOCUMENT RESUME

ED 375 159 TM 022 109

AUTHOR McDonnell, Lorraine M.
TITLE Policymakers' Views of Student Assessment.
INSTITUTION Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.; Rand Corp., Santa Monica, CA. Inst. for Education and Training.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO ISBN-0-8330-1542-7; MR-348-UCLA/OERI
PUB DATE 94
CONTRACT 0070-G-40250
NOTE 56p.
AVAILABLE FROM RAND, Distribution Services, 1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Accountability; *Attitudes; *Educational Assessment; Educational Change; Educational Policy; Elementary Secondary Education; Expectation; *Policy Formation; *Student Evaluation; *Test Use
IDENTIFIERS Experts; Reform Efforts

ABSTRACT

The gap between policymaker enthusiasm for the uses of student assessment and expert caution is analyzed by examining new forms of student assessment as an education policy strategy. The study is based on interviews with 34 national and state policymakers and focuses on their differing expectations of what assessment policy can accomplish and how they view the feasibility of assessment-based reforms. Some policymakers agree with testing experts that assessments should provide information about the overall status of the education system and aid in instructional decisions about individual students, but others want to use assessments for accountability or for certifying that individual students have attained specified levels of mastery. Testing experts, who have cautioned against assessment misuse in the past, are in danger of being ignored as they continue their dire warnings against using assessments in policy formation. (Contains 34 references.) (SLD)

RAND

Policymakers' Views of Student Assessment

Lorraine M. McDonnell

Institute on Education and Training

The research described in this report was sponsored by the National Center for Research, Evaluation, Standards, and Student Testing (CRESST) and funded by the Office of Educational Research and Improvement (OERI), U.S. Department of Education, Contract No. 0070-G-40250.

Library of Congress Cataloging in Publication Data

McDonnell, Lorraine, 1947–
Policymakers' views of student assessment / Lorraine M. McDonnell.
p. cm.
"Supported by the University of California, Los Angeles / Office of Educational Research and Improvement."
"MR-348-UCLA/OERI."
Includes bibliographical references.
ISBN 0-8330-1542-7
1. Educational tests and measurements—United States. 2. Educational evaluation—United States. 3. Educational accountability—United States. 4. Education—Standards—United States. 5. Educational surveys—United States. I. University of California, Los Angeles. Office of Educational Research and Improvement. II. Title.
LB3051.M26 1994
379.1'54—dc20 94-12141
CIP

Published 1994 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
To order RAND documents or to obtain additional information, contact Distribution Services: Telephone: (310) 451-7002; Fax: (310) 451-6915; Internet: order@rand.org.

RAND

Policymakers' Views of Student Assessment

Lorraine M. McDonnell

Prepared for the
UCLA Center for Research on Evaluation, Standards, and Student Testing
Office of Educational Research and Improvement, U.S. Department of Education

Institute on Education and Training

PREFACE

Much of the work of the National Center for Research, Evaluation, Standards, and Student Testing (CRESST) is aimed at improving the technical quality of new student assessments. However, the continuing prominence of assessment on the nation's education policy agenda means that technical issues cannot be considered independently from political ones. This monograph focuses on the political realm by examining the expectations that federal and state policymakers hold for new forms of assessment and their judgments about the feasibility of assessment-based reforms. This work was sponsored by CRESST and funded by the Office of Educational Research and Improvement (OERI), U.S. Department of Education.

The hope is that policymakers and testing experts alike will find this analysis useful in their continuing efforts to understand each other's interests and concerns.

CONTENTS

Preface iii

Summary vii

Chapter One
INTRODUCTION 1
Study Data 2

Chapter Two
WHAT POLICYMAKERS EXPECT OF ASSESSMENT POLICIES 5
Differing Views of Assessment 6
Competing Expectations Complicate Assessment Policy 12

Chapter Three
THE FEASIBILITY OF ASSESSMENT-BASED REFORMS . . 15
Technical Constraints 16
The Concerns of Testing Experts 17
Policymakers' Views About Technical Issues 20
Cost 23
Limited Information Available on Total Costs 24
Policymakers' Views About Cost Issues 27
Political Support and Opposition 30
Support for the Standards and Assessment Concept . . . 31
Consensus Assumed Likely on Content Standards 33
Delivery Standards as the Major Area of Political Disagreement 34
Policy Uses of Assessment Remain Unresolved 37

Chapter Four
CONCLUSIONS 41

References 45

SUMMARY

The sharpest disagreements between policymakers and members of the professional testing community have traditionally centered on the policy uses of student assessment. Of particular concern to experts is the use of test results to impose rewards and sanctions on schools and students. Over the past five years, testing experts and policymakers have altered the content and format of student assessments. The trend is away from a sole reliance on multiple choice tests toward tests that also require students to demonstrate the process by which they solve problems or to apply knowledge in new situations. Despite these changes, however, debate over the appropriate uses of student assessment persists.

In this monograph, the continuing gap between policymaker enthusiasm and expert caution is analyzed by examining new forms of student assessment as an education policy strategy. The study is based on interviews with 34 national and state policymakers, and focuses on their differing expectations of what assessment policy can accomplish and on how they view the feasibility of assessment-based reforms.

Policymakers have varying, and sometimes conflicting, expectations about what assessment can accomplish. Some agree with testing experts that assessments should primarily provide information about the overall status of the education system and aid in instructional decisions about individual students. Other policymakers want to use assessments as a tool for holding schools and educators accountable for student performance or for certifying that individual students have attained specified levels of mastery. In between these two views

are those who believe that assessments can be used to bring greater curricular coherence to schools, motivate students to perform better, and act as a lever to change instructional content and strategies.

Just as policymakers and testing experts view the purpose of assessments from differing perspectives, they also define feasibility issues differently. For testing experts, technical questions about generalizability and the ability to link diverse assessments to common content standards are sufficiently problematic to warrant caution in moving to widespread implementation of new strategies. For those policymakers who advocate new forms of assessment, however, the picture looks different. They see an extraordinary window of opportunity in which a broad spectrum of constituents has endorsed the idea of national standards and a system of linked assessments. But they recognize that this policy window could close just as quickly as it opened, as new issues in education and other policy areas crowd assessment off the national agenda. Therefore, they see their immediate tasks as twofold. First, they must devise strategies to deal with the cost constraints that alternative assessments pose. Second, they must resolve such political issues as the need to balance concerns about the unequal consequences that assessments might have on different students with concerns about preserving local autonomy. For these national and state policymakers, technical constraints are not irrelevant or even unimportant, they are simply seen as problems that can be solved over a longer time frame as new forms of assessment are implemented and modified with use.

The question arises, then, whether policymakers' enthusiasm for using student assessment as an instrument of education policy can be reconciled with experts' caution about its potential misuses. Will the move to alternative assessments and their policy applications repeat the recent experience with multiple choice tests, in which, over the last two decades, policymakers expanded the uses of multiple choice tests beyond their original, intended purposes, while testing experts documented the negative consequences on students and schools? The probable answer is yes. As long as policymakers see assessments as exerting a powerful leverage over school practice and, at the same time, are constrained by cost and other considerations, they will continue to use the same assessments for multiple purposes—some of which may have negative consequences for students, teachers, and schools. Consequently, as long as testing experts are

unable or unwilling to design assessments that can explicitly serve multiple purposes or be linked to other high-stakes policies, the impasse will continue.

Nevertheless, there may be lessons that policymakers and testing experts can learn from the history of assessment policy over the past 20 years Policymakers need to develop more realistic expectations about what assessments can accomplish. They should acknowledge that even the best assessments are imprecise measurement tools with real limits on their generalizability and appropriate use. Without such a recognition on the part of policymakers, some students will continue to be hurt by the use of assessment data.

The lesson for testing experts is about how many times one can cry "wolf" and still be believed. A number of testing experts now find themselves in an uncomfortable position. Policymakers have responded to their earlier criticisms of multiple choice tests that focused on low-level skills and downgraded the curriculum by agreeing to new forms of assessment and higher content standards. But now these same experts must tell policymakers that the solution carries its own set of problems. As a result, their warnings are in danger of being dismissed as those of "perennial naysayers." However, those testing experts whose professional judgment allows them to accept at least some forms of high-stakes testing can maintain a voice in deliberations about assessment policy by outlining the conditions under which the problems they have identified can be mitigated, and by estimating a reasonable timetable for implementing solutions. In proposing that timetable, they will need to keep in mind the resolve of many policymakers, particularly at the state level, to proceed with alternative assessments while simultaneously fine-tuning them.

Chapter One

INTRODUCTION

Political elites, business leaders, and the general public are looking once again to student assessment as a cornerstone of education reform because of its powerful leverage as a policy instrument. Although the effect of externally mandated assessments varies—depending on the type of test, the grade levels tested, and the students' socioeconomic status—a growing body of research indicates that school and classroom practices do change in response to these assessments (e.g., Corbett and Wilson, 1991; Madaus, 1988; Herman et al., 1990; Herman et al., nd). Scholarly debate about assessment has focused on questions of test content and format, but the sharpest disagreements between testing experts and the policy community have been over the policy uses of assessment. Experts warn that if assessments are used to advance policy objectives, particularly if they involve the imposition of rewards and sanctions based on test results, negative consequences are likely to result. These may include widening the gap in educational opportunities available to different kinds of students, a narrowing of the content and skills taught, a centralization of educational decisionmaking, and the deprofessionalization of teachers (Haertel, 1989; Airasian, 1987).

Over the past five years, testing experts and members of the education policy community have been changing the content and format of student assessments. The trend is away from a sole reliance on multiple choice tests to ones that also require students to demonstrate the process by which they solve problems or to apply knowledge in new situations. New forms of assessment may require that students write short essays, engage in problem-solving activities, or prepare portfolios of their work. Regardless of the subject matter

being tested, the content of these new assessments tends to focus on critical thinking skills; the application of knowledge, often in "real world" contexts; and the integration of knowledge across discrete disciplines, such as science and social studies.[1] Despite these changes in content and format, however, the debate about the appropriate policy uses of student assessment data persists. Many policymakers still see testing as an appealing instrument of education reform, while a number of testing experts continue to express caution about the role of assessment in advancing policy objectives.[2]

This monograph examines the current push for new forms of assessment as an attempt to fashion policy tools that will influence the behavior of educators, students, and their parents. It explicitly analyzes assessment from a top-down, policy perspective and addresses two questions:

- What do policymakers expect assessment policies to accomplish?
- How do policymakers view the feasibility of assessment-based reforms in terms of technical soundness, cost, and likely political support or opposition?

STUDY DATA

This study is based on data that constitute the initial phase in a larger research project. Between August 1991 and February 1992, interviews were conducted with 34 national and state policymakers, including White House staff, congressional staff, state legislators,

[1]As of late 1993, 18 states were piloting or implementing new forms of student assessments. The subjects and grade levels tested vary across these states, but all are supplementing their traditional multiple choice tests with essays, open-ended items, and/or some form of portfolio assessment. In most cases, test content has also been made more rigorous.

[2]It is important to note that not all testing experts oppose the use of new forms of assessment for high-stakes purposes. Testing and measurement experts on contract to or working in universities and state departments of education in such states as Kentucky are clearly committed to the notion that new forms of assessment can be used for policy purposes, such as for holding schools and students accountable. However, at this point, most nationally visible testing experts either have withheld judgment until more assessments have been designed and piloted, or have expressed mild to strong cautions against their widespread application for policy purposes.

governors' education aides, and interest group representatives.[3] Relevant policy documents and recent research on the uses of assessment data were also reviewed. The purpose of the interviews was to ascertain what expectations policymakers hold for various national assessment initiatives, such as the work of the National Council on Education Standards and Testing (NCEST), *America 2000*, and the New Standards Project. Since these interviews were conducted, Congress has enacted the Clinton administration's "Goals 2000: Educate America Act." The legislation differs somewhat from the Bush reform agenda, most notably by the absence of any private school choice provisions. However, the Clinton bill is similar to that of his predecessor's in its emphasis on national standards and a voluntary certification system for state-run assessments. Consequently, the issues that shaped debate over national standards and new forms of assessment several years ago remain prominent today, and most of the respondents interviewed for this study continue to play a prominent role in that debate, whether from within or outside government.

Respondents were asked about their own support for, or opposition to, these assessment initiatives, what they and other policymakers expect such initiatives to accomplish, the relationship between as-

---

[3]Eighteen of these interviews were conducted face-to-face with national-level respondents. Respondents were selected because of their direct involvement in the national debate on standards and testing. Included in this group were three members of the Bush administration; seven congressional staff (two Republican, four Democratic, and one member of the nonpartisan congressional agency staff); and six members of interest groups representing organized teachers, the commercial testing industry, and state elected officials. In addition, a staff member of the National Council on Education Standards and Testing (NCEST) and a journalist knowledgeable about student assessment were also interviewed.

The remaining 16 interviews were conducted by telephone with respondents in California, Colorado, Delaware, Missouri, New Jersey, Ohio, South Carolina, and Texas. This group included seven state legislators, seven governors' education aides, and two chief state school officers; it was selected to represent a range of state political culture and degree of policy activism on assessment and education reform. Four states, California, Delaware, New Jersey, and Texas, are among the 18 that are piloting or implementing new forms of assessment. Policymakers in Missouri reported planning to add open-ended items to the state assessment, and content standards were being developed in South Carolina. But in neither state were the changes operational yet. Respondents in Colorado reported that they expected local districts to develop their own standards and assessment systems, with the state playing a review and quality control function. Ohio has no formal plans at this time to develop new state assessments. (All respondents were promised confidentiality.)

sessment and other reform initiatives, how much is known about the cost of these initiatives and the willingness of the federal and state governments to invest in them, how seriously technical and feasibility issues raised by experts are being considered in current debates, and what is likely to happen with regard to assessment policy over the next three to five years. State policymakers were asked about their own states' activities and the likelihood that they would move in directions consistent with the national discussion. The assumption was that while most assessment policy comes from state government, debate and discussion at the national level set a tone that influences state action. Subsequent phases of the project are focusing on the design, implementation, and effects of new forms of assessment in four states and a sample of schools in each of those states.

I report interview responses in terms of general patterns among respondents, rather than as counts or percentages. This approach seemed the most appropriate level of quantification, given the small sample size, the fact that respondents were selected strategically, and the open-ended nature of the interviews. Although all respondents were asked the same questions, their responses tended to be contextualized according to their own role position and political experience. Therefore, I have chosen to treat these data as elite interviews, using excerpts to illustrate patterns that can be characterized by adjectives such as "most" and "few." To provide a context for these excerpts, role position is always noted. The data and analysis presented in this monograph are not generalizable to all national and state policymakers who deal with education. Rather, they represent the views of a select group who either have been influential in shaping national discussions about new forms of assessment, or who play a key role in formulating new assessment policies in seven states, which are characterized by diverse approaches to education policy.

In Chapter Two of the monograph, the diverse expectations that policymakers hold for assessment are examined, and in Chapter Three, the major feasibility issues raised by new forms of assessment are discussed. The concluding chapter analyzes the extent to which the enduring tension between technical and political judgments can be reconciled in assessment policy.

Chapter Two

WHAT POLICYMAKERS EXPECT OF ASSESSMENT POLICIES

An analysis of policymaker responses to questions about what problems they see assessment policy addressing and what they expect such policy to accomplish elicits at least seven different types of purposes that policymakers expect assessments to serve. These purposes include

- providing information about the status of the education system
- aiding in instructional decisions about individual students
- bringing greater curricular coherence to the system
- motivating students to perform better and parents to demand higher performance
- acting as a lever to change instructional content and strategies
- holding schools and educators accountable for student performance
- certifying individual students as having attained specified levels of achievement or mastery.

The connection between assessment and these desired outcomes is complicated by several factors. First, as a result of the six national goals that emerged from the 1989 education summit between former President Bush and the nation's 50 governors, assessment has now become linked to the notion of national education standards defining what students should know in specific subject areas. Although the overwhelming majority of respondents in this study support the

notion of national standards and point to opinion polls as evidence of their constituents' support,[1] few had a clear conception of what the content of those standards should be beyond some broad generalities. Second, a strong consensus is lacking on which of the seven general purposes listed above are most important or on what combination of purposes is most appropriate. Opinions differ, depending on personal philosophy, political party affiliation, and the governmental level at which a respondent works.

DIFFERING VIEWS OF ASSESSMENT

Not surprisingly, given their preeminent role in the financing and governance of public schooling, state-level respondents see accountability as a major purpose of assessment. For example, three of the six chairs of legislative education committees spoke of assessment's role in this way:

> I think, in general, that people are willing to pay more if they know what the product is; standards are linked closely to financing. (Colorado)

> . . . the public really is demanding more accountability from public education. I am an educator as well as a politician, and I know how to respond. Substantive documentation is the proof of a successful

[1] For example, a number of respondents pointed to the results of the 1991 Gallup Poll of education, which found that 81 percent of a nationally representative sample of respondents support having their local schools "conform to national achievement standards and goals." In addition, 77 percent of the sample favor "requiring the public schools in their community to use standardized national tests to measure the academic achievement of students" (Elam et al., 1991, p. 46). This level of public support for the use of national tests has been quite consistent since the Gallup poll first investigated the issue in 1970. In addition, in the 1992 poll, a majority of respondents favored using standardized achievement tests to rank local public schools (65 percent), to determine if a student advances to the next grade level (60 percent), to identify areas where teachers need to improve their teaching skills (79 percent), and to identify areas in which students need extra help (85 percent). However, only slightly more than one-third of respondents supported the use of achievement testing to determine either how much teachers should be paid (38 percent) or the level of funding a local school should receive (36 percent) (Elam et al., 1992, p. 47). Although results of these polls over several years suggest that the public looks more favorably on using test results and evidence about school performance to reward rather than punish schools, 57 percent of those included in the 1991 poll favored not renewing the contracts of principals and teachers in those schools that do not show progress toward meeting the national goals (Elam et al., 1991, p. 44).

> educational system. Clearly, educators have dropped the ball and have been too busy promoting social activities while the kids have been sliding through the system. High school graduates lack a quality education today; a valid assessment system is the answer. (Ohio)

> Politicians have to be in tune with the demands of their constituents who want to see the results of what they have been paying for. However, with the present system one can't understand the results clearly. What happened to that student, was it college, job, or jail? The general public wants to see something on a regular basis that they can clearly understand. They are making judgments about professional educators based on student achievement whether it is fair or not. . . . (Texas)

But state-level policymakers were not alone in stressing the accountability function of assessment. A number of congressional and administration staff attributed the growing interest in assessment to pressure from the business community. For example:

> A dominant theme in education over the past twenty years has been accountability. That theme has been around a long time, but it is even more dominant now because of business involvement in education. Business is very bottom line-oriented. Assessment is on the agenda because it lies at the heart of accountability. A high school diploma is not an adequate measure of achievement, so the business community wants a new standard of achievement.

Largely because of the governance structure of American education, with its limited federal role, national-level respondents talked about a weaker form of accountability than state policymakers, whose scope of authority puts them in a better position to hold schools accountable for their performance. National-level respondents implied a concept of accountability that is much closer to the notion of disseminating information about the status of the education system than to a version of oversight with rewards and sanctions attached. Respondents talked about using an assessment to inform parents about the progress of their own child and to aid in comparing that child with other students or with some performance standard. Others said that politicians and the business community "had gone out on a limb in support of education, and they can't point to successes without an assessment system."

However, a few respondents argued that using assessment results for accountability purposes should go far beyond a purely informational function. The strongest advocate of such an approach was, surprisingly, a teacher union leader:

> I'm talking about a system where there are real consequences. Under this system, there will be school consequences: If they improve, there will be bonuses; if they don't, people can be fired. With consequences, there will be a change in attitudes because teachers' interests will be different. When there are no consequences, the emphasis is on not alienating your neighbors. Right now in New York City, if 75 percent of the teachers wanted to suspend a contract provision, they wouldn't impose it on the other 25 percent. . . .
>
> In other organizations, change occurs because there is something at stake. There's no question that people would behave differently with consequences.

Just as national policymakers agree with testing experts in their support of the use of assessment for informational purposes, several congressional staff indicated that the members for whom they work assign a high priority to the use of assessment in instructional decisionmaking—the other assessment function endorsed by the professional testing community:

> The issue for the subcommittee [on elementary, secondary, and vocational education] is how is this going to help Mrs. Ellison and her fourth grade.
>
> For Senator. . ., if tests provide useful information to teachers, then [the senator] has no problems with them. But if we were to spend millions or billions of dollars to develop tests to compare performance across countries, [the senator] would be opposed to it.

Few respondents questioned the idea of national standards, and most saw the purpose of these standards as bringing greater curricular coherence to the nation's education system. "Standards are a call to a common education agenda . . . the expectations are that with standards and assessment, we will get the coherence and focus the system has lacked," argued an NCEST staff member. Underlying the notion of national standards is a belief not just that there will be greater coherence and commonality in what is taught across the

country, but also that the overall level of standards will rise. In response to a question about what policymakers hope to accomplish with assessment, a staffer for a Republican member of Congress explained it this way:

> The hope is that with national standards, we'll get higher uniform standards. There's [a higher] level inherent in the drive for national standards. If you don't figure out what you're about, you may flounder around a lot. Also, the assumption is that if the standards are not national, they won't be set uniformly high, and there will be a lot of diversions. There's an assumed discipline to the standards, and assessment is part of the stick.

As we will see in the next chapter, although most respondents support or at least accept the concept of national standards, opinions differ about the ease with which consensus can be reached on a set of standards, the extent to which these standards should or can influence local curriculum, and the exact link between standards and assessment. Nevertheless, the belief that standards and assessment can achieve greater curricular coherence is a pervasive one among national and state policymakers.

Respondents also mentioned assessment as a motivating "wake-up call." This purpose was often attributed to Democratic Governor Roy Romer of Colorado and a former assistant secretary of education in the Reagan administration, Chester Finn. A newspaper reporter noted that "a lot of people look at assessment as a way of providing information that isn't there now and assume that the information will shake people up." A congressional staffer talked of state policymakers seeing assessment as a way to "embarrass people into change." A governor's education aide echoed the notion expressed by several respondents that even parents in affluent communities would be surprised to find that their children are not performing academically as well as they might have assumed:

> Every parent thinks that their school is doing okay, but with the implementation of a national standards and assessment system, there will be something in place to measure against and to compare results statewide to a national norm. This system would also replace the old bell curve, as well as clear up the fact that every child is doing well at his or her local school. Expectations will be elevated

> and, in general, this system will provide us with a wake-up call to action. . . .

Interestingly enough, most respondents talked of the motivational purposes of assessment in terms of parents, rather than students. Even when pressed about this issue, most respondents continued to talk about assessment results as a way to motivate parents to take action to improve the quality of local schools. A few respondents, however, expressed the hope that assessment results "will encourage kids who are currently not doing well to do better." But any connection between assessment and student motivation was not one to which respondents had given much thought.

Respondents who expect assessments to motivate students see it coming from some type of certification process—i.e., students who score above a certain level would be certified as having met some standard of achievement or mastery and would be judged qualified for future education or employment. But certification as a purpose of assessment is a decidedly minority view. Even a member of the White House staff seemed to have modified the position taken by the Bush administration in *America 2000* when she said:

> Our view is that it is not necessary to have consequences to drive instruction. We want incentives, not consequences. We wouldn't want to see children punished.
>
> It is important though that business send a message to kids. We envision it almost as a one-on-one mentor approach where, for example, a business says that it will pay an extra fifty cents an hour if students do well on assessments. This may sound corny. But we don't see this system as punitive, and tests should be one of a number of things employers consider.

As with the accountability purposes of assessment, national policymakers have offered certification as a possible use of assessment but have stressed that any nationally developed assessment should be voluntary and that its use should be decided by states and localities. Still, they offer the possibility of such use, as evidenced in the NCEST report (1992) and in materials from the New Standards Project:

> The Council concludes that the United States, with appropriate safeguards, should initiate the development of a voluntary system

> of assessments linked to high national standards. These assessments should be created as expeditiously as possible by a wide array of developers and be made available for adoption by states and localities. The Council finds that the assessments eventually could be used for such high-stakes purposes for students as high school graduation, college admission, continuing education, and certification for employment. . . . (National Council on Education Standards and Testing)

> Students passing final examinations in high school and completing all of their required tasks would be recognized with a certificate for their achievement, and students who possess that certificate will find it much easier to get a good job or get into college. So all students will see the connection between the effort they put into school and what they want for themselves when school is over. . . . (The New Standards Project)

Respondents saw the New Standards Project as the initiative that most clearly embodied a final purpose of assessment—with assessment acting as a lever to change instructional content and strategies. They noted that Lauren Resnick, the director of that project and a member of NCEST, has made clear that "this [the New Standards Project] is not about testing; it is about changing instruction and learning" (Licitra, 1991). Staff of the National Governors Association (NGA) noted that the staffs of several governors close to the national goals process believed that national standards could drive new forms of assessment, which, in turn, would reshape curriculum and instruction. However, other participants saw the outcomes of the goals process in less ambitious terms: As a result of the national goals initiative, measures of student achievement would be improved and, thus, provide a more accurate picture of the status of the U.S. education system. According to these respondents, Lauren Resnick and other avowed reformers such as Marc Tucker (President of the National Center on Education and the Economy) were able to convince key actors such as Governor Romer that new forms of assessment could do more than just produce more valid measures of student achievement; they could actually improve schools by changing instruction.

COMPETING EXPECTATIONS COMPLICATE ASSESSMENT POLICY

Multiple and diverse expectations for what assessment can accomplish translate into multiple policy targets, disparate notions of the process by which change occurs, and competing uses for the results of student assessments. In looking across the differing purposes of assessment that were articulated by respondents, we find that policymakers hope to change the behavior of students, teachers, administrators, and, in some cases, parents, employers, and even the general public. But they do not agree on how that change should occur. Some see it as a direct process with new assessment formats prompting specific changes in curriculum content and instruction that, in turn, will result in improved student achievement. Some proponents of this direct strategy assume that teachers will be motivated to alter their behavior simply because they accept the notion that such change will lead to more effective teaching and learning. Others assume, however, that rewards and sanctions are needed to reinforce the suasion of professional norms.

Still other policymakers conceive of the change process as more complex and varied. For example, those who stress the informational aspect of assessment assume that it can serve as a resource for those interested in improving schools, but that the route to changes in teaching and learning will be a circuitous one. They assume that not everyone will use testing results in the same way. While some states and localities might use student assessment as a lever for curricular reform, others might use results to reward and punish schools but not to make changes in curriculum. Still others might simply report test results and expect individual schools to make appropriate changes in practice with little or no central direction. In these latter cases, the causal processes linking assessment and improved student achievement are not only less direct, but they will vary significantly across states and local communities.

The question then becomes whether a single assessment system can serve these diverse purposes. Testing experts have warned about the difficulties inherent in relying on the same assessment to serve multiple purposes. For example, a number of experts have argued over the past decade that so-called high-stakes tests used for accountability purposes cannot also be validly used to provide information

about the status of the educational system or to shape a coherent curriculum. Not only are such assessments limited in the scope of what they test and are, thus, incapable of providing a comprehensive picture of student achievement, but high stakes in the form of rewards and punishments mean that teachers typically emphasize what is being tested, thus narrowing and fragmenting the curriculum (for recent arguments along these lines, see Koretz et al., 1992). Similarly, one can imagine that the kind of information and level of detail needed for making instructional decisions about individual students are not the same as those needed for reporting on the status of the educational system as a whole or for making state- or even district-level policy decisions. Nevertheless, not all the myriad purposes that policymakers expect assessment systems to serve are necessarily incompatible. Consequently, one criterion policymakers are likely to use in judging the feasibility of different assessment strategies is the extent to which multiple expectations can be met by the same system.

But questions about the valid uses of assessment represent just one dimension of policy feasibility. In addition, there are a variety of other technical issues as well as questions about the cost of new forms of assessment and who will bear those costs, and about political support and opposition to various assessment-based reforms. The next chapter examines these feasibility issues from the perspective of national and state policymakers. As we will see, the technical issues that most concern testing experts are often of considerably less importance to policymakers than questions about the additional cost of new assessments, how those assessments will affect different kinds of students, and how they might alter current distributions of educational authority across different levels of government.

Chapter Three

THE FEASIBILITY OF ASSESSMENT-BASED REFORMS

In examining the various expectations that policymakers hold for assessment, the previous chapter outlined what policymakers consider to be desirable goals—i.e., what they hope to accomplish by using student assessment as an instrument of education policy. Certainly, "the tendency to equate the desirable with the feasible is always strong, especially in politics" (Majone, 1989, p. 69). But there are always constraints that set limits on the implementation of the desirable. Some of these constraints are technical, defined by the state of the technology available to address a particular policy problem; others are organizational—i.e., whether or not there is an appropriate institution with sufficient capacity to put a policy in place. A related constraint deals with resources—what are the total costs associated with a particular policy and who is willing to bear those costs? Finally, there are political constraints. These stem from three facts: Policymakers need to be responsive to constituents if they are to be reelected; coalitions must be built and bargains struck if a policy is to be enacted; and the design of new policy is shaped by political ideologies, by relatively stable divisions of power and authority among governmental levels and societal groups, and by past policies.

Many constraints are not fixed and can be relaxed. Sufficient time and an increased research and development budget might modify technical constraints; new organizations can be created or the capacity of existing ones improved; more funding can be appropriated; and political support can be mobilized in the face of opposition. In fact, skilled politicians have long understood that under the right conditions, they can convert constraints into opportunities for policy action. Nevertheless, not all constraints can be relaxed, and even

those that can be will usually require the skilled application of time or other resources. Therefore, when we examine assessment as a policy instrument, it is important to evaluate the feasibility of this strategy and understand the constraints it poses and how policymakers view those constraints.

TECHNICAL CONSTRAINTS

The technical dimensions of feasibility focus on the technology of testing and ask whether alternatives to traditional multiple choice tests will work in the ways intended. At one level, we are talking about questions of reliability and validity. For example: will test scores be consistent across different types of performance tasks designed to measure similar skills; can these assessments be scored reliably; will student performance on test items be generalizable across the full range of a subject-matter domain; is the assessment really measuring the skills or content knowledge it purports to measure? These are some of the questions that must be asked of any assessment whatever its format or intended uses (for an extended discussion of these issues, see Messick, 1989; United States Congress, Office of Technology Assessment, 1992).

But there are two other levels at which one needs to consider technical feasibility constraints. To use an assessment for policy purposes, one must determine whether the information generated can be utilized in the ways that policymakers expect. This dimension focuses on the valid uses of a particular assessment and relates to the questions raised in the previous chapter. For example, can an assessment, such as the National Assessment of Educational Progress (NAEP), which is designed to supply information about the status of the education system as a whole, also be validly used to provide individual-level scores? Could such a test also be used to guide instruction or to certify levels of individual achievement? These validity-in-use questions lie at the heart of the debate over assessment as an instrument of public policy and remain major sources of disagreement between testing experts and policymakers.

A final level that needs consideration is whether a particular assessment strategy can be implemented. Is sufficient expertise available to design, administer, score, and use the assessment? What kind of capacity-building will be necessary to ensure that educators can use

the assessment appropriately? Because assessments are typically linked to other policies dealing with school organization, curriculum, and instruction, the same questions also need to be asked of those other policies. For example, if an assessment is intended to encourage the use of new curriculum frameworks, a major feasibility question is whether teachers are willing and able to teach according to those frameworks.

The Concerns of Testing Experts

Largely in response to the growing state use of minimum competency tests for high school graduation, beginning in the 1970s, a number of testing and measurement experts analyzed the problems associated with the policy uses of assessment (e.g., Airasian and Madaus, 1983; Frederiksen, 1984; Haertel, 1989). Their critiques of the policy uses of traditional multiple choice tests are now well known. They questioned the disjuncture between the actual curriculum in schools and what was being tested, the assessment's emphasis on basic skills, the lack of opportunity for some students to gain even the basic knowledge and skills needed to score well on these tests, and the corruption of the tests as valid and reliable measuring devices because of the strong sanctions keyed to their results. Experts argued that tests whose results determined whether students graduated from high school or whether schools received extra resources would change school behavior and, in the process, the tests themselves would be altered as valid measures of student achievement.

To a large extent, new forms of assessment were developed in response to the identified shortcomings of multiple choice tests. In fact, if one looks at some of the recommendations coming from critics of traditional tests, one finds an almost direct correspondence between these recommendations and the goals of various forms of alternative assessment. For example, Haertel (1989) recommends that other kinds of learning outcomes be recognized, including "not only better tests of critical thinking and higher order skills, but also ways to recognize students' exceptional individual accomplishments, from written works or science fair projects to artistic creations" (p. 31). Those who espouse the use of such assessment devices as student portfolios see themselves as having responded to the

identified shortcomings of more traditional forms of student assessment (e.g., see NCEST, 1992, p. 28).

But, at least to this stage of their development, new forms of assessment are not without their problems, and in light of those problems, experts have cautioned against moving too quickly toward their widespread implementation. In an analysis of current knowledge about performance assessments, Linn (1993) focuses particularly on the generalizability problems associated with new forms of assessments. He notes that one of the major stumbling blocks to the implementation of performance-based assessment systems is "the limited degree of generalizability of performance from one task to another" (p. 9). Citing a variety of evidence from performance assessments in history, mathematics, and science, and even licensure examinations in law and medicine, Linn concludes that because performance on one task has only a weak to modest relationship to performance on another, a large number of tasks (or increased testing time for more complex tasks) will be necessary to ensure that assessments produce comparable information and that results are fair to the individuals being tested.[1] Others have warned of the same potential problem, noting that "some students who fail on the basis of one overly limited or non-representative sample of tasks [might] have passed if given an equally defensible alternative set" (Koretz et al., 1992). If a test has consequences, but valid generalizations cannot be drawn from the exercises included on it, then test-takers will be treated capriciously. However, the remedy of increasing the number of tasks creates its own feasibility problems in terms of cost and time burden.

Up to this point, the technical constraints that have been described apply to alternative assessments regardless of the organizational ar-

[1]As one example of the increased testing needed to achieve acceptable levels of generalizability on performance assessments, Linn cites the case of the Advanced Placement (AP) exams. If these exams, which are currently a combination of multiple choice and performance items, were converted to be solely performance based, the amount of testing time required to achieve a generalizability coefficient of .90 or higher would increase dramatically, particularly for some subjects. The estimated amounts of required testing time would range from a low of an hour and 15 minutes for Physics C: Electricity and Magnetism to 13 hours for European history. Six or more hours would be required for 8 of the 21 subjects tested in the AP series (Linn, 1993, p. 12).

rangements under which they are administered. However, another set of technical issues arises because of the particular way the notion of a national assessment system is being conceived. For a variety of political reasons (discussed below), a consensus has developed that if the United States moves to a system of national standards and assessments, there should be voluntary clusters of assessments, all geared to the national standards. Under this arrangement, states, groups of states, or local districts would design their own assessments, with this voluntary "system" operating in lieu of a single national examination. In its report, NCEST (1992) argued that these assessments should be aligned "with high national standards and [have] the capacity to produce useful, comparable results" (p. 4). During the discussions preceding the NCEST report, proponents of this strategy talked of "calibrating" locally developed assessments to the national standards. However, there was never any clear agreement about what the term calibration meant, either practically or psychometrically.

Since the issuance of the NCEST report, a number of testing experts have considered the question of whether different assessments can be linked and how it might be done. In the most detailed analysis to date, Mislevy (1992) concludes the following:

> No single score can give a full picture of the range of competencies of different students in different instructional programs. Accordingly, multiple sources of evidence—different question types, formats, and contexts—must be considered. Some of these will be broadly meaningful and useful; others will be more idiosyncratic at the level of the state, the school, the classroom, or even the individual student.
>
> . . . it *isn't* possible to construct once-and-for-all correspondence tables to 'calibrate' whatever assessments might be built in different clusters of schools, districts, or states to provide different kinds of information about students. (p. 73)

Mislevy then suggests several strategies that will meet "less ambitious, but more realistic goals" than those assumed in discussions of calibration. These include comparing students' skills across localities on a selected sample of valued tasks administered under standardized conditions, supplementing those assessments with others that measure a broader range of skills and are tailored to specific

states and localities, and relying on linking studies in which students are administered portions of assessments from other localities to identify the extent of commonalities and differences in results and the sources of those patterns.

In sum, while acknowledging that newer forms of assessment may address some of the shortcomings of multiple choice formats, testing experts see significant technical constraints limiting these alternative strategies. They caution that if performance assessments are implemented on a widespread basis before technical problems are substantially remedied, invalid information may be generated and, at worst, students could be harmed. The issue, then, is do policymakers view these constraints in the same way as the professional testing community.

Policymakers' Views About Technical Issues

Both national and state-level respondents in this study were asked the following question: *A number of testing experts have raised technical and feasiblity questions (e.g., problems with validity, generalizability) about a national assessment system. How seriously are those issues being considered in discussions about various proposals?* At a general level, respondents were aware that testing experts had raised cautions about the various assessment systems being considered. For example, most national-level respondents either had read a paper on the subject prepared for the National Goals Panel by Linn (1991) or had been briefed about its content by their staffs. Many of the state-level respondents had heard similar cautions expressed by experts in their own state departments of education.

Respondents did not dispute that new forms of assessment presented these problems, but they rejected them as a reason either to abandon performance assessments or to slow their implementation. In fact, a wide variety of respondents expressed the sense that experts are cautious by nature, but that policymakers need to move ahead. A chief state school officer articulated that sentiment in this way:

> Honestly speaking, I classify these remarks [about the limitations of new assessments] as made by some people who make a career out of having reservations about other people's work. Of course, these

> are real concerns and there will be some real problems, but that is no reason not to proceed and move ahead. For example, in 1961, we put a man on the moon without all the technology needed for such a venture. We need to continue to work together in developing new systems.

Several other respondents noted that this interest in technical problems was a peculiarly American concern:

> The technical questions that have been raised are absolutely unimportant and nonsensical. No other country raises these questions. (Member of NCEST)

> It's perhaps desirable in theory to have assessment systems stand up to reliability and validity criteria. But while we have endless attention to psychometric issues, there will be delays and delays. There is little use of psychometric and social science research in the European tests such as the French national exam. The trend is that the experts are being grouped with the naysayers—the cup is half-empty crowd. It is not helpful. (Senior administrator, National Endowment for the Humanities)

A congressional staffer argued that given the intense interest in standards and assessment,

> The debate over the uses of alternative assessments won't turn on research findings because those findings and conclusions aren't compelling enough to stop policy movement. The debate will turn on whether people buy into the expectations that the loudest advocates have for alternative assessment. That is, whether people believe that students will learn more if they are tested in new ways.

A representative of the commercial testing industry argued that policymakers' disregard of the cautions raised by testing experts related to how they viewed the experts' role in the current system, which elected officials were now trying to change:

> Politicians latched on to the idea of alternative assessments, and they are not necessarily talking to state testing directors. With [Governor] Romer, it's like you could amend Shakespeare and say he believes, "First kill all the psychometricians and then the lawyers." The belief is that those with the expertise are part of the problem. The [test] publishing industry has been totally excluded

> from the process. It is as if you ordered solar power for cars and excluded the auto and gas companies from the development process because they had a vested interest.

Most respondents did not see calibration of state and local tests to national standards as a problem because they view calibration as a political, not a technical, issue. A member of NCEST argued that "calibration is a concession to get people aboard." In a similar vein, a congressional staffer noted that he was not certain whether any kind of calibration or linking of tests was possible but viewed it as a trade-off: "It's probably a good idea not to have a single national test but [as a result], the possibility of comparisons in a psychometric way has declined." In this case, needing support from those who fear that a single national assessment could lead to a centralized curriculum, policymakers have advocated calibration, which may be a technically infeasible idea. But they seem to understand the trade-off they are making.

One respondent summed up the tension between experts and policymakers by stating very simply: "There's a difference between the policy people who want it now, and the technical experts who say it can't be done." The reason for this disjuncture is not that policymakers do not understand the nature of these constraints or disbelieve that they exist. Rather, they view the confluence of factors that put the spotlight on national standards and assessment as an opportunity with a limited time frame, and as one Bush Education Department official argued, "We can't wait 20 years for all the research." Policymakers will move ahead regardless of expert cautions. Comments by both a senior Democratic congressional staffer and a member of the Bush White House staff reflect the extent of consensus on this point:

> Policymakers are listening and saying that over time the technical problems will be taken care of. Nothing that is being said is stopping the process. (congressional staffer)

> Policymakers are taking the technical people seriously, but they are not accepting their cautions as excuses not to continue. Campbell's [Governor of South Carolina, co-chair of NCEST] attitude is that "we'll do the best we can right now; this is an evolutionary process which will improve over time." (Bush White House staff)

It is important to note that, with a few notable exceptions, most policymakers do not dismiss technical constraints as unimportant. Instead, they see them as problems that must be remedied, but in an iterative fashion that occurs simultaneously with the implementation of new assessment strategies. Respondents talked of "mid-course corrections," but they did not see technical problems as a barrier to broad-based implementation of new forms of assessment.

COST

One of the traditional appeals of student assessment has been its low cost as compared with other education reforms, such as the development and implementation of new curriculum or the reorganization of individual schools. In discussing the policy uses of minimum competency tests, Airasian and Madaus (1983, p. 108) note,

> There was little that policy makers could do to reform instruction directly; they could not mandate and enforce a better technology of instruction, even assuming one were available. So instead of making policy about instruction, they made policy about testing, an available, well-developed, relatively cheap and administratively simple technology.

Although new forms of assessment will cost more than older, multiple choice tests, policymakers view them as among the least expensive strategies for reforming schools. A congressional staffer expressed this sentiment, "People settle on assessment as a cheap way to fix problems. One of the most prominent governors sees assessment as an important lever to change American education. . . . It's a lever for change without having to spend a lot of money."

Nevertheless, those involved in the move to implement new forms of assessment recognize that alternative approaches will cost more than traditional multiple choice tests, if only because scoring them is more complicated. But no one agrees, and few hard data exist, on exactly how much the additional costs will be. Nor is there agreement about what items should be included in calculating additional costs. For example, most agree that the costs of research and development, administering and scoring the tests, and training educators in their use ought to be included in any cost calculations. However, if one purpose of new forms of assessment is to change curriculum

and instruction, should the cost of training teachers in new pedagogies also be included? Research by Cohen and Peterson (1990) on the implementation of the new mathematics frameworks in California suggests that the training requirements necessary for the profound changes envisioned in California's assessment-curriculum link are extensive and long term.

Limited Information Available on Total Costs

A full enumeration of the costs of assessment needs to include not just financial costs, but also costs calculated in other currency such as time.[2] One of the implications of expert recommendations that a large number of performance tasks are necessary to generate valid scores is that considerable opportunity costs will be incurred—i.e., additional time needed for assessment will be lost to direct instruction. On the other hand, those who believe that there should be a strong link between assessment and curriculum argue that some assessments will be so "embedded" in the classroom instructional process that concerns about excessive time burdens are unfounded. The performance tasks that students will have to perform singly or in groups will actually reinforce their learning of new skills as well as test their mastery of prior knowledge. Cost, then, is a major feasibility issue not just because testing expenditures are likely to increase with new forms of assessment, but also because there is little consensus on how to calculate the full cost.

To date, Monk (1993) has prepared the most comprehensive analysis of the likely costs associated with large-scale efforts to introduce student performance assessment. His analysis is based on the model embodied in the New Standards Project, and it estimates costs for a large state (Texas), a mid-size state (Virginia), and a small one

[2]As Monk (1993, p. 7) notes, there is an important distinction between *costs* that measure what must be forgone to realize some benefit, and *expenditures* that are measures of resource flows regardless of their consequences. In education policy, when we talk about the cost of a particular service or program, we are typically referring to the level of resources expended and often, in even narrower terms, of how much policymakers have appropriated rather than the full range of expenditures, some of which are less visible and borne by a variety of sources. As a result, we often lack a complete estimate of total expenditures, and we rarely have a full accounting of costs, as compared with benefits.

(Vermont). In Monk's estimates, performance assessments would be administered in mathematics and language arts to students in grades 4, 8, and 10. His middle-case estimates of the operational costs in the sixth year of the program range for the large state between $67.94 and $34.69 per pupil tested,[3] between $69.59 and $43.20 in the mid-size state, and between $69.11 and $46.02 per pupil tested in the small state.[4] Monk assumes that development costs will be spread over all students in the 17 states participating in the New Standards Project, so that the per student cost is only about twenty cents for each of four years.

Other estimates have been extrapolated from existing assessments and vary widely. At the high end are estimates based on the cost of the Advanced Placement (AP) examinations. The AP exams are deemed relevant because they test higher order, analytical skills and combine multiple choice and performance items. Using the current $65 cost per AP exam, Koretz and his colleagues (1992) estimated the total cost of national testing to be $3 billion a year for the five subjects and three grade levels recommended in the NCEST report. They suggested that these costs could be higher if students in more grade levels were tested as part of an assessment-based effort to guide instruction. At the low end, the U.S. General Accounting Office (GAO) (1993) estimated that the assessment system proposed by NCEST would require about $100 million in development costs and then cost about $330 million a year to operate, with costs likely to de-

[3]Monk reports his estimates by amortizing the costs over all the students in a state, rather than just the students being tested. So, for example, his estimates for the large state range from $7.47 to $11.93; for the mid-size state, from $15.19 to $9.43; and for the small state, from $14.59 to $9.72. I have chosen to report his estimates in terms of the *cost per pupil tested* to permit comparisons with other cost estimates, and because it seems a more appropriate metric for considering the unit costs of alternative assessments.

[4]Monk generates different scenarios based on the extent of diminishing marginal productivities—e.g., the amount of variation in the ability of teachers to benefit from inservice and the variation in students' ability to benefit from the feedback provided by performance assessment. A low-cost scenario includes a teachers' and students' learning curve, which causes the assessments to become less costly over time, and by scale economies because assessments developed in one locale can be transferred to another. A high-cost scenario makes the opposite assumptions.

Monk's estimates include the operations costs of supplemental lead teacher training, scorer training, continuing scorer training, outside auditing, administration of tasks, scoring, utilization of results, and administration and overhead.

crease over time. The GAO's estimate was based on survey data collected on the 1990–91 school year from the testing directors of the 48 states that administered statewide tests that year and from a survey of 368 local districts. The GAO collected data about the costs of performance assessments from local districts in the six states that used performance items in more than one subject. Because all of the state tests except one combined multiple choice and performance items, the GAO notes that its calculations could underestimate the cost of pure performance-based tests.

More precise cost data should become available as more states move to performance-based assessments. At this point, even in states quite far along in the use of performance assessments, data are incomplete. For example, with its five million students, California expects to spend about $35 million to $50 million a year to test three grade levels in five subjects once its new assessment system is fully operational. But this cost range does not include the cost of either the testing-related workshops that local districts and county offices of education are sponsoring or the various subject matter projects for teachers run by the state universities and various education agencies. Also, it is unclear whether these cost estimates will hold if the number of performance tasks on a given test has to be expanded to provide valid individual-level scores.[5] Similarly, Kentucky has a fully operational assessment system, which includes portfolio assessments in two subject areas and performance events in five subjects for three grade levels, and expects to spend $25–30 million over six years, or about $38 per student tested. But the Kentucky costs represent a scaled-back effort from the state commissioner of education's original plan and do not include the costs of considerable state and local school district staff development, or the initial portfolio scoring,

[5]In 1993, California tested all its fourth, eighth, and tenth graders in two subjects, using a combined multiple choice and performance item format, on a budget of approximately $12 million or about $10 per student tested. Although this budget included some pilot testing in another subject, this initial round of testing will not produce individual-level scores, and the budget did not include most related teacher inservice costs. If California were to spend $35 million a year for its testing program, the cost per student tested would be about $30.

which is done locally with costs borne by individual schools and districts.[6]

Clearly, at this point, there are no reliable cost estimates for alternative assessments. With estimates varying by a factor of ten from a high of about $325 per student tested in AP-like exams to the GAO's estimate of an average $33 per student for state tests that combine multiple choice and performance items, only limited conclusions can be made about alternative assessment costs. First, we do know that they will be higher than from the $2 to $20 per student cost for commercially developed multiple choice tests. Second, the nature of alternative assessments is such that their full cost may be very difficult to calculate. Depending on the intended purpose of these tests, costs could be quite straightforward and well defined (development, administration, scoring, etc.), or they could be embedded in a wide variety of teacher training and curriculum development activities and, therefore, be difficult to estimate reliably.

Policymakers' Views About Cost Issues

Given the lack of solid information about costs, how do policymakers view this feasibility issue? Three conclusions emerge from interviews with national and state policymakers. First, the majority of respondents noted that limited cost data are available, and that cost considerations have not yet played much of a role in policy deliberations about new forms of assessment. Some argued that it was too soon in the process: The precise contours of the national assessment system have not yet been decided and no state system is well enough developed yet to determine actual costs. A few respondents felt that serious considerations of cost early in the process might have put a

[6]Advanced Systems, the contractor for the Kentucky assessment, also sells to individual districts and schools "scrimmage" tests, which can be administered to students outside the three grades tested as part of the state assessment program. Districts and schools are currently charged $7.25 per student tested for tests that parallel the state assessment, but that do not include the portfolio assessments. These tests supplement the mandated assessment of three grades. As such, they represent no additional development costs, and the scrimmage tests take advantage of scoring arrangements already in place. In essence, the price reflects only those marginal costs over and above ones already borne by the main assessment system.

damper on the move to new forms of assessment. For example, one congressional staffer noted,

> The willingness-to-invest issue is one that no one has been willing to look at from the beginning. Also test burden is a hidden cost. It would have been helpful along the way if there had been different cost models, then something could have been done. But politically speaking, it is unwise to put a cost figure on anything.

Second, despite the lack of reliable information about cost, a majority of respondents indicated that they understood alternative assessments would cost more than current multiple choice tests, and that there were costs in addition to the direct expenses of testing. A number of national-level respondents mentioned Governor Romer's reference to the "meat in the sandwich"—i.e., the funds needed for teacher training and adequate materials to accompany new forms of testing. Respondents at both the national and state levels echoed that testing is only one ingredient in the "sandwich" of education reform—and not necessarily the most important. For example,

> . . . You can't teach world class math and science without more resources—labs, manipulatives, etc. (National Goals Panel staff)

> We will have to redirect money for the assessment system [because] massive funding will be necessary to pay for new staff development and retraining programs for our teachers, counselors, and principals. (Chief state school officer)

Third, some policymakers have already begun to think about how the additional costs of alternative assessments might be handled. When cost is considered as a feasibility issue, more than just total costs need to be taken into account. Two other major questions are who will bear those costs and what opportunity costs will be incurred because other goods and services cannot be purchased. It is very clear that once a consensus was reached not to have a single national exam, policymakers at all levels assumed that the federal government's role in bearing the costs of a new assessment system would be limited, like its role in funding education expenditures generally. Respondents saw the federal government bearing a significant portion of the research and development costs associated with standards-setting and test design. Members of NCEST, the Bush and

Clinton administrations, and a few state-level respondents also assumed that a national body would be established to develop criteria for certifying state assessments and, on a voluntary basis, to award "Good Housekeeping seals of approval." A few congressional staff noted that the federal government could further coax state assessments in particular directions by the types of testing requirements they impose in the upcoming Chapter 1 reauthorization. The federal government might also expand the number of students taking NAEP and the scope of that exam, but all of these federally funded activities were expected to represent a fraction of nationwide assessment activities.

In five of the seven states in which policymakers were either implementing or actively considering new forms of assessment, they indicated a commitment to funding the additional costs even though it might be difficult. For example,

> Yes, definitely, this state will make the investment. However, it will be necessary to reallocate money from some current activities to create a system and this will be a hard sell. This new system will have to be perceived as quite compelling and necessary in the long run to take money from another educational effort. (Governor's education aide)

One of the ways that states are planning to deal with the additional costs of new tests is to work in collaboration with other states. For example, the governor's education aide in New Jersey explained that the governor and commissioner of education were already discussing cooperating on a regional basis with New York to share the costs of developing a new battery of fourth grade performance tests.

In the other two states, respondents were taking a "wait and see" attitude because of competing priorities or the condition of their state economies:

> I am doubtful about investing in a costly enterprise such as this [alternative assessments] right now because we have many more pressing needs for our money. (State legislator)

> It will depend on what is offered and how much it costs, but I suspect that the new assessment system will cost extra money and not be in lieu of some current activities. We have a tight budget now

> and work hard to take care of the priorities we have on our plate at this time. (State legislator)

Cost, then, is a feasibility concern for policymakers, and although it has not played a dominant role in discussions to this point, it is likely to be a more critical issue for policymakers than technical constraints. All respondents understand that new forms of assessment will cost more and some have made a commitment to bear at least part of those costs. But others are waiting for better information about full costs before they consider the inevitable trade-offs between improved tests and other funding responsibilities.

POLITICAL SUPPORT AND OPPOSITION

Even if technical problems were solved and the costs of new forms of assessment turned out to be reasonable, questions of political support and opposition would still remain. By virtue of their status as elected officials, policymakers must also judge new forms of assessment on an additional set of criteria. They need to ask, for example: What consequences will alternative assessments have for different kinds of students in different types of communities? Does a sufficient consensus exist about the content of new tests and how those tests will be used? Will new assessments result in an altered distribution of power and authority across governmental levels?

It is important to understand the context in which discussions about the political dimensions of assessment are occurring. The most salient feature is the policy's origins in the executive branch of both the federal and state governments. One can trace much of the impetus for the national assessment movement back to the six national goals promulgated at the 1989 Charlottesville education summit by President Bush and the nation's governors. President Clinton, then governor of Arkansas, played a key role in that meeting, and, like Bush, has used the goals as a framework for his education reform proposals. Although two senators, two representatives, and two state legislators served as members of NCEST, its co-chairs and the two most visible participants, Carroll Campbell and Roy Romer, are governors.

Congressional staff respondents from both political parties noted that Congress was placed in a reactive position after the Char-

lottesville meeting. One congressional staffer described Congress' role as trying to put "a harness on a raging bull." Or, as another staffer noted, "In most areas, Congress is a conservative institution that mulls over new ideas and sees what people can live with. Right now, Congress is in the mulling-over stage." Partly as a result of their role as scrutinizers of presidential proposals, those members of Congress interested in student assessment issues have been more willing to listen to the cautions of testing experts than their counterparts in the executive branch.

Participants in the assessment policy process tend to believe that Congress has been systematically excluded from executive branch deliberations. Republican members of Congress first found themselves in the position of having to acknowledge that they had not been consulted on the Bush administration's assessment proposals and then asking the administration to make modifications in exchange for their support. Two years later, the Democratic members found themselves in exactly the same position with Clinton's "Goals 2000" legislation. Not only has Congress been in a reactive position, but only a few members of both houses and their staffs are interested in standards and assessment, well informed about the issues, and actively engaged in the policy process. Similarly, those states that have chosen to move ahead on alternative assessment have largely done so at the initiative of their executive branch, either through the governor or the chief state school officer. In most cases, state legislatures have entered the process much later, in their role as funds appropriator.

Support for the Standards and Assessment Concept

Respondent interviews and an examination of the assessment debate over the past four years indicate that within this political context, there is broad-based support for the concept of national standards and new forms of assessment. A wide range of groups, from teacher unions to business organizations, has expressed support for this strategy as an alternative to the lack of clear educational expectations in this country and the shortcomings of multiple choice tests. Despite this consensus around the concept of new assessment strategies, key individuals and groups in the process have very different reasons for supporting such a change. For example, as discussed

in the previous chapter, a number of respondents argued that some see assessment as an inexpensive reform strategy. An NCEST staff member, who saw new forms of assessment as a way to curb the commercial testing industry, expressed another reason, which other respondents echoed:

> The education system has totally abdicated responsibility for assessment to the commercial test developers whose obligation is to their stockholders. If the test developers could sell the same test for 10 years, they would. The focus on national assessment is seen as seizing back that responsibility from the commercial test developers. If that can be done, then we won't have de facto, backdoor standards any more.

A teacher union leader who strongly supports the concept of national standards and testing with consequences sees such a strategy as an alternative to proposals for private school choice. He argued that choice advocates are raising the right question in asking what the incentives are to improve the system. But, in his view, the choice alternative would only provide incentives to attract students, not necessarily to educate them. Tests with consequences would provide a greater incentive for improving learning in the longer term.

Still other respondents noted that the move to new forms of assessment was so strong that politicians put themselves at political risk if they either opposed the idea or were even too critical of it. As one congressional staffer noted, "If you criticize the movement, it seems as if you are against information and accountability." Or, as another respondent phrased it, "There's a certain amount of political correctness here."

The consensus around new forms of assessment extends beyond support for the concept to agreement about several key operational issues. The first operational issue has already been discussed: Early in the process, policymakers agreed that there should not be a single national exam. In the United States, local control over education is so well ingrained that even those who initially favored one exam recognized that there would be insufficient political support for such a strategy. Hence the notion of a voluntary system of linked exam clusters emerged as the preferred alternative. Nevertheless, a few respondents recognized that a voluntary, decentralized strategy does

pose some trade-offs for the quality of the assessment. One respondent, who is now an official in the Clinton administration, put it this way:

> In some ways, the standards and assessment initiatives are caught between politics and science. This tension manifests itself in the question of whether you can get generalizable data if the system is voluntary. In this case, science is trying to respond to what it perceives to be demands from the political arena. But the political arena may find that what it wants—assessment information collected on a voluntary basis—is of no use.

Consensus Assumed Likely on Content Standards

Another issue that might have been a source of political dispute is the establishment of national standards in curriculum content areas. Some observers assumed that while a consensus on what students should learn was possible in a subject such as mathematics, agreement might be much more difficult to reach in subjects such as science, English literature, and social studies—some of whose content reflects geographic, ethnic, and ideological divisions in society. A few observers have also raised the question of who has the right in a democracy to set educational standards. Sizer (1992), for example, questioned whether subject-matter experts, who neither have been elected nor represent the interests of all parents and citizens and who operate at a distance from most local communities, should be the ones to decide what students should learn.

However, most study respondents felt that a broad-based consensus could be reached on curriculum standards. They argued that it might be considerably more difficult in subjects other than mathematics and that there would likely be a zone of variation across communities in some subjects, but those differences would be small as compared with a much larger core on which experts, parents, and citizens could agree. For example, students in Colorado might have a unit on the history of the West, while students in South Carolina would have a different kind of local supplement. Yet in both cases, the bulk of what students learned in U.S. history would be similar. Most respondents agreed with the NCEST member who said that "if we had a referendum on what topics should be included in stan-

dards, there would be agreement on content. People tend to overstate the disagreement." Others talked about the "de facto" national curriculum that already existed and argued that the standards would only serve to raise the level of that curriculum. Two national-level respondents supported this view by noting that, when they traveled around the country, people said, "Look California did it with their frameworks; if that fractious state can do it, the rest of us can as well." Despite the expected agreement, however, a number of respondents conceded that the standards-setting process could be "bumpy."

Delivery Standards as the Major Area of Political Disagreement

In contrast to curriculum content standards, delivery standards for schools have emerged as the most visible focus of political debate. In an appendix to the NCEST report, the Standards Task Force established by NCEST argued that: "*School delivery standards* should provide a metric for determining whether a school 'delivers' to students the 'opportunity to learn' well the material in the *content standards*" (NCEST, 1992, p. E-5). This argument assumes that benchmarks could be established to measure whether a school's curriculum covers the material included in the content standards, and whether the school has well-trained teachers, suitable instructional materials, and adequate facilities to provide students with the opportunity to meet the expected performance standards.

Although not initially included in the concept of national standards and assessment, school delivery, or opportunity-to-learn, standards emerged as a way to address some policymakers' equity and access concerns. As one congressional staffer asked rhetorically, "What are you going to do once you find out which students aren't doing well?" School delivery standards were first proposed by several members of NCEST. They later represented the key demand that House Democrats made in asking the Clinton administration to revise its Goals 2000 proposal (Zuckman, 1993).

Supporters argue that without such delivery standards, students may be penalized for not meeting performance goals just because they attend schools that lack the capacity to teach adequately. According

to this view, delivery standards need to be in place before new forms of assessment are used for any high-stakes purposes. Otherwise, students in poor schools will be treated unfairly. Representative Dale E. Kildee (D-Mich.), the chair of the House Subcommittee on Elementary, Secondary, and Vocational Education, framed the argument in this way: "Without delivery standards, you don't know if the school is failing, or if students are failing" (Rothman, 1993, p. 21).

Groups such as the NAACP, the National Education Association, the Council of Great City Schools, and the Children's Defense Fund have argued that poor students will be unjustly penalized unless schools are held accountable for the resources they provide for all students.

Republicans in Congress have opposed the establishment of national delivery standards because they see them as limiting state and local flexibility and potentially leading to costly mandates. Congressional Republicans have been joined in their opposition by some governors from both parties (Rothman, 1993). Representatives of the Business Roundtable have argued that school delivery standards are likely to shift the national debate away from the key issue of what children are actually learning. Albert Shanker, president of the American Federation of Teachers and a member of NCEST, told a National Governors' Association hearing that "it would be totally wrong for the development of content standards to be held hostage" while school delivery standards are written (*Report on Education Research*, 1993). Opponents have also argued that the existence of national delivery standards might serve as the basis on which students could sue states to spend more on schooling inputs.

The lack of agreement among experts on the technical feasibility of such standards complicates the political debate over delivery standards. Some researchers have argued that it is possible to gauge whether students have had the opportunity to learn the more challenging subject matter that will be specified in national content standards. These researchers have recommended that such standards concentrate on teachers' ability to teach to high curricular standards and on the availability of an appropriate pedagogy and curriculum (O'Day and Smith, 1993). However, Porter (1993) argues that since research has not yet been able to specify with sufficient precision exactly what factors need to be present for schools to produce desired student outcomes, schools should be held accountable for the out-

comes they produce, not for the inputs or processes they use to produce those outcomes. In Porter's view, holding schools accountable for student performance keeps attention where it belongs and leaves schools "free to be inventive and experimental in the approaches they take" (Porter, 1993, p. 27). Furthermore, by considering student performance at the school level, as well as at the individual level, policymakers can still determine whether it is schools or students that are failing.

Porter and others have also noted that school delivery standards present difficult measurement problems. For example, everyone who advocates delivery standards believes that data are needed about the extent to which instruction in a given school is aligned to content standards. The least expensive way to collect such data is to survey school administrators and teachers, asking them what courses they offer, the topics covered, and the instructional strategies used. However, such data have typically been collected at such a high level of generality that they could not be used to make valid determinations about the alignment of individual schools with the content standards (Porter, 1993; McDonnell et al., 1990). Even data that are more expensive to collect, such as course syllabi, instructional materials, and assignments, are only proxies for the curriculum as it is actually implemented in individual classrooms. Yet due process standards would require that valid and reliable measures of a school's curriculum be established before it could be held accountable for its instructional activities. Such measurement problems are not insurmountable, but solving them will require a period of intense research and experimentation.

The Goals 2000 legislation that President Clinton signed into law on March 31, 1994, represented a compromise on the question of opportunity-to-learn standards. As a condition of receiving federal funds for education reform activities, states have to set cont t and performance standards. However, the legislation allows states to establish either opportunity standards or strategies, and it specifically asserts that participation in Goals 2000 is voluntary and that the law should not be construed to mandate school-finance equalization or school-building standards.

Some states may decide to forgo participation in Goals 2000 and not develop standards or strategies for ensuring that students have an opportunity to meet state content and performance standards. The more likely scenario, however, is that most states will develop such standards, but these standards will serve a largely hortatory purpose in encouraging, rather than requiring, school districts to improve their services to students. Nevertheless, forces other than congressional action may push states in the direction of stronger school delivery or opportunity-to-learn standards. For example, school finance lawsuits, such as the recent *Rodriguez* case in Los Angeles, are focusing increasingly on the distribution of such resources as experienced teachers and course offerings across schools within the same district. If such trends continue, school delivery standards may be established de facto through a series of judicial decisions.

Policy Uses of Assessment Remain Unresolved

The fact that school delivery standards have become a focal point for mobilizing political support and opposition is strong evidence that policymakers do expect new forms of assessment to be used for high-stakes purposes. If assessment results were only to provide data about the overall status of the education system, to act as a goad to students and parents, or to inform individual instructional decisions, school delivery standards would be largely irrelevant. Delivery standards would be more salient if the purposes of assessment were to bring greater curricular coherence to the educational system or to change instructional content and strategies. However, they are most relevant if assessment results are to serve as a basis for rewarding and punishing schools, individual students, or teachers or to certify levels of student mastery. The greater the consequences attached to the use of new assessments, the more germane other contextual information will be.

Once again it is the policy uses of assessment that engender the greatest disagreement among policymakers, between policymakers and the research community, and even among researchers themselves. A congressional staffer summarized those differences in this way:

> If you ask what problems different advocates [of standards and assessment] are trying to address, it's like the five blind men touching different parts of the elephant. For some it is to motivate students; for others, to achieve world class standards and compete internationally, to tell Americans which states and local districts are doing well; and for still others, testing is a way of supporting learning. These different goals and definitions of the problem will affect what tests will be designed to do. Will they be for accountability purposes or to improve instruction? If they are designed for one purpose and used for another, then we're back to the old problem.

It is the last point that is perhaps the most important. Policymakers and their constituents are unlikely to agree on a single purpose, or even several related purposes, that new forms of assessment should serve. Consequently, the result will be that, over time and across different states and local communities, alternative assessments will come to serve a broad range of purposes—just as earlier multiple choice tests have done. In the view of many policymakers, the potentially strong leverage over local practice combined with feasibility concerns about cost and burden make multiple uses of student assessments inevitable. This sentiment was voiced by a number of study respondents who expressed their frustration with expert claims that the same test cannot be used for different purposes, and who maintained that they could see no reason why new forms of assessment could not be used for both accountability and other purposes.

When we examine the feasibility of new forms of student assessment, we see that experts and policymakers define the issues quite differently. For many testing experts, technical questions about generalizability and the ability to link diverse assessments to common content standards are sufficiently problematic to warrant caution in moving to widespread implementation of these new strategies. For those policymakers who advocate new forms of assessment, however, the picture looks different. They see an extraordinary window of opportunity in which a broad spectrum of constituents have endorsed the idea of national standards and a system of linked assessments. But they recognize that this policy window could close just as quickly as it opened, as new issues in education and other policy areas crowd assessment off the national agenda. Therefore, they see their immediate tasks as twofold. First, they must devise strategies to deal with the cost constraints that alternative assessments pose. Second,

they must resolve such political issues as the need to balance concerns about the unequal consequences that assessments might have on different students with concerns about preserving local autonomy. For these national and state policymakers, technical constraints are not irrelevant or even unimportant, they are simply seen as problems that can be solved over a longer time frame as new forms of assessment are implemented and modified with use.

Chapter Four

CONCLUSIONS

Most analyses of new forms of student assessment have focused on technical issues related to their reliability and validity. This monograph has taken a different approach. It has examined assessment as a policy strategy and has tried to understand policymakers' expectations for what student standards and new forms of assessment should accomplish. In doing so, it has highlighted the gap between the enthusiasm of policymakers who advocate alternative assessments and the cautions of testing experts who warn about the potential pitfalls.

It is not unusual for policymakers to be people of action, who want to move quickly before competing claims on their attention push a preferred solution off the policy agenda, or before electoral constraints restrict the range of politically acceptable alternatives. It is also not unusual for experts to be the skeptics who identify problems likely to be encountered in implementing a given policy, and who urge a careful weighing of the probable costs and benefits of any new policy. Each group acts according to well-established norms and incentives for their respective professional roles. But in the case of assessment policy, this traditional gap between the policy and research communities has been particularly notable.

The question arises, then, whether policymakers' enthusiasm for using student assessment as an instrument of education policy can be reconciled with experts' caution about its potential misuses. Will the move to alternative assessments and their policy applications repeat recent experience with multiple choice tests, in which policymakers expanded the uses of multiple choice tests beyond their

original, intended purposes, while testing experts documented the negative consequences on students and schools? The probable answer is yes. As long as policymakers see assessments as exerting a powerful leverage over school practice and, at the same time, are constrained by cost and other considerations, they will continue to use the same assessments for multiple purposes—some of which may have negative consequences for students, teachers, and schools. Consequently, as long as testing experts are unable or unwilling to design assessments that can explicitly serve multiple purposes or be linked to other high-stakes policies, the impasse will continue.

Nevertheless, there may be lessons that policymakers and testing experts can learn from the history of assessment policy over the past 20 years. Policymakers need to learn that assessment has not lived up to its expected potential as a vehicle for major education reform. It did alter classroom practice, but often not in the ways that policymakers intended. And while new forms of assessment may be a vast improvement over older basic skills and multiple choice tests, their ability to alter classroom practice in desired ways will still depend on factors over which policymakers have little control or will require investments in capacity-building far in excess of the costs of new tests. Furthermore, new forms of assessment may measure some domains of student learning more reliably and validly than older approaches, but their role as providers of truly objective information about overall student performance will remain largely a myth. Like most myths in public policy, the belief that assessments can produce impartial and comprehensive data about student achievement is a powerful one (De Neufville and Barton, 1987). But policymakers need to acknowledge that even the best assessments are imprecise measurement tools with real limits on their generalizability and appropriate use. Without such a recognition on the part of policymakers, some students will continue to be hurt by the use of assessment data.

While the lesson for policymakers centers on developing more realistic expectations for what assessments can accomplish, the lesson for testing experts is about how many times one can cry "wolf" and still be believed. A number of testing experts now find themselves in an uncomfortable position. Policymakers have agreed to new forms of assessment and higher content standards in response to testing experts' criticism that multiple choice tests focused on low-level

skills and downgraded the curriculum. Now these same experts must tell policymakers that the solution carries its own set of problems. As a result, their warnings are in danger of being dismissed as those of "perennial naysayers."

To maintain their voice in continuing deliberations about assessment policy, testing experts need to do two things. First, they should be explicit about how much of their criticism stems from principled opposition to high-stakes tests and how much is specific to particular types of assessments and their policy uses. For those testing experts whose professional judgment allows them to accept at least some forms of high-stakes testing, the second step is to outline the conditions under which the problems they have identified with new forms of assessment can be remedied, and to estimate a reasonable timetable for implementing solutions. In proposing that timetable, they will need to keep in mind the resolve of many policymakers, particularly at the state level, to proceed with alternative assessments while simultaneously fine-tuning them.

Policymakers and testing experts will continue to disagree about the appropriate policy uses of student assessments. But the experience of the past 20 years should have taught both groups that in failing to accommodate each other's values and interests, the interests of students are ill served.

REFERENCES

Airasian, P. W. (1987). State mandated testing and educational reform: Context and consequences. *American Journal of Education, 95*(3), 393–412.

Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*(2), 103–118.

Cohen, D. K., & Peterson, P. L. (1990). Special Issue of *Educational Evaluation and Policy Analysis, 12*(3), 233–353.

Corbett, H. D., & Wilson, B. L. (1991). *Testing, reform, and rebellion.* Norwood, NJ: Ablex Publishing.

De Neufville, J. I., & Barton, S. E. (1987). Myths and the definition of policy problems. *Policy Sciences, 20,* 181–206.

Elam, S. M., Rose, L. C., & Gallup, A. M. (1991). The 23rd annual Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 73,* 41–56.

Elam, S. M., Rose, L. C., & Gallup, A. M. (1992). The 24th annual Gallup/Phi Delta Kappa Poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 74*(1), 41–53.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*(3), 193–202.

Haertel, E. (1989). Student achievement tests as tools of educational policy: Practices and consequences. In B. R. Gifford (ed.), *Test policy and test performance: Education, language and culture,* pp. 25–50. Boston: Kluwer Academic Publishers.

Herman, J. L., Dreyfus, J., & Golan, S. (1990). *The effects of testing on teaching and learning.* CSE Technical Report 327. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, UCLA.

Herman, J. L., & Golan, S. (no date). *Effects of standardized testing on teachers and learning—another look.* CSE Technical Report 334. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, UCLA.

Koretz, D. M., Madaus, G. F., Haertel, E., & Beaton, A. E. (1992, February 19). *National educational standards and testing: A response to the recommendations of the National Council on Education Standards and Testing.* Testimony before the Subcommittee on Elementary, Secondary, and Vocational Education, Committee on Education and Labor, U.S. House of Representatives.

Learning Research and Development Center of the University of Pittsburgh, and the National Center on Education and the Economy. (1991). *The new standards project: An overview.* Pittsburgh, PA & Washington, DC.

Licitra, A. (1991, June 4). Value of national standards, tests will become clear, Resnick says. *Education Daily, p. 5.*

Linn, R. L. (1991). *Technical considerations in the proposed nationwide assessment system for the National Education Goals Panel.* Boulder, CO: Center for Research on Evaluation, Standards, and Student Testing, University of Colorado.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(1), 1–16.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (ed.), *Critical issues in curriculum, Eighty-seventh yearbook of the National Society for the Study of Education,* pp. 83–121. Chicago, IL: The University of Chicago Press.

Majone, G. (1989). *Evidence, argument, & persuasion in the policy process.* New Haven, CT: Yale University Press.

McDonnell, L. M. (forthcoming). Assessment policy as persuasion and regulation. *American Journal of Education.*

McDonnell, L. M., Burstein, L., Ormseth, T. H., Catterall, J. S., & Moody, D. (1990). *Discovering what schools really teach: Designing improved coursework indicators.* Santa Monica, CA: RAND.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement.* 3rd ed., pp. 13–103. New York, NY: American Council on Education.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service.

Monk, D. H. (1993). The costs of systemic education reform: Conceptual issues and preliminary estimates. Paper prepared for the New Standards Project.

The National Council on Education Standards and Testing. (1992). *Raising standards for American education.* A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American People. Washington, DC: U.S. Government Printing Office.

O'Day, J. A., & Smith, M. S. (1993). Systemic reform and educational opportunity. In S. H. Fuhrman (ed.), *Designing coherent education policy,* pp. 250–312. San Francisco: Jossey-Bass.

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

Report on Education Research. (1993, May 26). Education advocates split on school delivery standards. Alexandria, VA: Capitol Publications Inc., pp. 5–6.

Rothman, R. (1993, April 7). 'Delivery' standards for schools at heart of new policy debate. *Education Week, pp. 1, 21–22.*

Sizer, T. R. (1992, January 30). A test of democracy. *The New York Times,* p. 21.

United States Congress, Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions.* OTA-SET-519. Washington, DC: U.S. Government Printing Office.

United States General Accounting Office. (1993) *Student testing: Current extent and expenditures, with cost estimates for a national examination.* GAO/PEMD-93-8. Washington, DC.

The White House. (1991). *America 2000: The president's education strategy.* Washington, DC: Office of the Press Secretary.

Zuckman, J. (1993). Clinton's school reform plan has high hopes, low funds. *Congressional Quarterly, 51*(17), 1027–1028.