Journal of Writing Assessment Vol. 1, No. 1, pp. 23-45

That Was Then, This Is Now: The Impact of Changing Assessment Policies on Teachers and the Teaching of Writing in California

SANDRA MURPHY
University of California, Davis

The objective of this study was to investigate the impact of changes in assessment policies on teachers of writing and the teaching of writing in California. Surveys of middle school and high school English teachers conducted in California in 1988 and 2001 when two different accountability systems were in place provided the data for the study. Results indicate that both of California's recent assessment systems have influenced what and how teachers teach, but in very different ways.

Americans have long had a love-hate relationship with educational testing, as Robert Linn (2001) reminded us. Debates about educational testing extend from the controversies that surrounded the first localized trial of standardized tests in the 1840s (D. Resnick, 1982) to more recent debates about the SAT and the use of high-stakes standardized tests in the efforts by states to jump start educational reform (Linn, 2001). In recent years, some educational reformers have

Sandra Murphy is a Professor in the School of Education at the University of California, Davis. She received her PhD in Language and Literacy from the University of California, Berkeley. She teaches graduate-level courses on research on reading and writing and has taught freshmen composition at the college level and high school English. She coauthored *Designing Writing Tasks for the Assessment of Writing* (with Leo Ruth), *Writing Portfolios: A Bridge from Teaching to Assessment* (with Mary Ann Smith), and *Portfolio Practices: Lessons from Schools, Districts and States* (with Terry Underwood). She has written several articles and book chapters on the teaching of reading and writing and their assessment.

Direct all correspondence to: Sandra Murphy, School of Education, One Shields Ave., University of California-Davis, Davis, CA 95616

looked to standardized tests as an efficient way to raise standards, and to monitor students and schools. However, others decry the negative impact of such tests, particularly when high stakes are attached. They condemn both the harmful direct effects of such tests on individuals and the unintended side effects they may have on teachers and schools (L. Jones, 2001). Assessment polices appear to mirror the ebb and flow of these controversies. During more progressive eras, tests tend to be backgrounded and learning foregrounded. During less progressive ones, tests are used to police education and are linked with policies such as rewards and sanctions for schools, or the elimination of social promotion, to enforce high standards. The purpose of this study was to examine the impact of changing assessment policies on teachers of writing and the teaching of writing in California. In no other state has the swing of the assessment policy pendulum between these two poles been quite as obvious and extreme as it has been in California since the 1980s.

Then

In the mid-1980s, California was touted as a progressive innovator in assessment development and in the professional development of teachers. The state supported six statewide teacher professional development projects—the California Subject Matter Projects—in writing, math, science, literature, history and art, exhibiting a commitment to the professional development of teachers unmatched by any other state. The California Assessment Program (CAP) had initiated what was arguably the most ambitious achievement test in writing in the nation, a direct assessment of writing that evaluated a variety of genres and forms. California's direct assessment was the product of a broader movement toward performance assessment that began in the 1970s and early 1980s when policymakers had become convinced that traditional fill-in-the-bubble exams were not tapping into the full range of students knowledge and abilities. In the field at large, an increasingly complex perspective on writing and on writing instruction had prompted a shift from the use of multiple-choice tests to the "direct" assessment of writing by way of the use of a writing sample and ultimately, to experiments with alternative forms of assessment (Camp, 1993a, 1993b; Lucas, 1988a, 1988b). California followed the trend, replacing its "bubble-in-the-answers," multiple-choice editing test with an assessment system that assessed multiple types of writing.

Many educators applauded the move toward direct assessment. They favored the direct collection of an actual writing sample rather than indirect methods because they considered the sample to be essential to the validity of the writing assessment. As Roberta Camp (1993b) explained, "Teachers and researchers in writing and writing instruction have argued that student writers should demonstrate their knowledge and skills not merely by recognizing correctness or error in text, as they do in multiple-choice tests of writing ability, but by engaging in the complex act of creating their own text" (p. 187). The earlier standardized multiple-choice test characterized knowledge about writing as discrete hierarchically arranged components (Camp, 1993b). At the time, the use of multiple-choice tests was widespread in the assessment field at large. But concern had been growing about the pernicious effects on education of indirect, multiple-choice assessments. Because multiple-

choice tests were limited in the range of student skills and knowledge that could be assessed, and because those skills were assessed as separable components, some theoreticians argued that multiple-choice tests led to a narrowed and fragmented curriculum (Haertel & Calfee, 1983; Madaus, 1988). In the writing assessment field in particular, there was a growing consensus that the better way to assess students' writing skills was through the direct assessment of writing (Greenberg, 1992).

By 1984, direct writing assessments were conducted in 22 states, and 5 more states were scheduled to begin conducting direct writing assessments by 1985 (Hadley, 1984). Moving far beyond the single sample writing assessment programs that were becoming more common in those days, CAP added a direct writing test to its battery of multiple-choice tests in reading, written expression, and mathematics. The test asked students to compose a piece of well-written prose in a variety of genres or forms. In all, when fully implemented, CAP planned to assess eight different types of writing at Grade 8: a proposal for solving a problem, a sketch of a person whom the writer knows (i.e., firsthand biography), a story, an evaluation (e.g., a judgment of a literary text, other book, movie, artwork, etc.), an autobiographical incident, observational writing, and an essay that speculates about causes or effects. At Grade 10, CAP planned to assess the last five of these writing types, along with an essay about a controversial issue, an interpretation of a literary work, and a reflective essay. At Grade 4, four broader, more developmentally appropriate types were introduced. These included expressive, informative, narrative, and persuasive writing. The first direct writing test was administered in 1987 in the eighth grade and focused on four types of writing: autobiographical incident, report of information, problem solution, and evaluation.

The state employed a matrix sampling system to gather information about how schools were performing with respect to the teaching of writing. Because CAP evaluated the performance of schools and school districts, as opposed to that of individual students, it was necessary for each student to write only one paper at the time of the test. However, across a classroom and a school, all of the types of writing were tested. The result was a composite picture of how well students performed in a spectrum of writing situations, in short, a report card for the state. CAP testing focused on assessment of school programs to provide data for evaluating student learning at the school, district, county, and state levels, but not for evaluating individual students or classes.

Each student's writing was evaluated on "criteria emphasizing the critical thinking, problem-solving, and composing requirements unique to the particular type of writing it represented" (California Education Roundtable, 1988, p. 52). Student writing was also evaluated on features such as coherence, organization, and conventions. Results provided information about the strengths and weaknesses of current programs. For example, in 1988 then Superintendent Bill Honig reported that a majority of California's students appeared "well-grounded in the ability to write about learned or personal experiences" but " less skilled in writing arguments in support of their judgments or solutions to problems" (California Department of Education, 1988, p. 2).

The test was explicitly aimed at improving the teaching of writing in the state, but rather than aiming the effort at the least experienced and less capable teachers,

as some high-stakes test now appear to do, policymakers had decided to craft a test through which teacher leaders could improve instruction and enhance the professional accountability of all of the state's teachers (M. A. Smith, personal communication, December 21, 2002). At the time, resistance to bureaucratic systems was high. Bureaucratic systems were thought to tend to *control* teaching and learning in ways that can work against what we know about the principles of academic achievement motivation (Anderman, 1997; Covington, 1992). Scholars argued that such systems inhibit thoughtful, reflective learning and teaching and can lead to rote "delivery" of instruction, formulaic teaching, and the deskilling of teachers. As Darling-Hammond (1989) explained,

> In the bureaucratic conception of teaching, teachers do not need to be highly knowledgeable about learning theory and pedagogy, cognitive science and child development, curriculum and assessment; they do not need to be highly skilled, because they do not, presumably, make the major decisions about these matters. Curriculum planning is done by administrators and specialists. . . . Inspection of teachers work is conducted by hierarchical superiors, whose job is to make sure that the teacher is implementing the curriculum and procedures of the district. Teachers do not plan or evaluate their own work; they merely perform it. . . . The problem with the bureaucratic solution to the accountability dilemma in education is that effective teaching is not routine, students are not passive, and questions of practice are not simple, predictable, or standardized. (p. 64)

The alternatives to the bureaucratic model of accountability (and schooling) that Darling-Hammond suggested were "client-oriented" and "professional" accountability. "[C]lient-oriented accountability requires that teachers primarily teach *students* rather than teaching *courses*, that they attend more to learning than to covering a curriculum" (p. 73). Professional accountability means that teachers and their professional organizations accept responsibility for ensuring competence, standards, and appropriate practice. Professional accountability demands new roles for teachers, including their participation in the evaluative and decision-making functions of schools, their collective review of teaching practices and policies, and their collective investigation of problems (Darling-Hammond, 1989).

The CAP assessment system was clearly linked to a professional model of accountability. Teachers were involved in all stages of the test development and implementation and their work on the assessment was supported by the leadership of the California Writing Project (CWP). CWP also provided staff development linked to the types of writing assessed by the test throughout California. As Mary Ann Smith (1988), a co-director of the National Writing Project explained:

> At the center of the effort to prepare the test were California's classroom teachers. . . . The entire new assessment—writing tasks, scoring guides, classroom materials—was developed by teachers who also conducted field tests and led the subsequent scoring sessions. (p. 9)

Participating in the assessment development and/or the scoring provided a powerful professional development experience for teachers. Moreover, because the test was aimed at improving instruction, it was designed to be relatively transparent. A variety of support materials, including handbooks for middle and high school and "samplers" that contained full descriptions of the writing types and their characteristics and sample topics were widely distributed. More than one writing project teacher joked at the time that the Department of Education might just as well rent a plane and leaflet the state with test prompts. The scoring guides were also an important part of the support materials. As Martha Dudley (1997) explained, "There was never any attempt to keep them secret; they were part of the writing handbooks, they were distributed at workshops and in-service sessions" (p.17). The goal was to help teachers become fully aware of what was being assessed and the criteria used to score the papers.

Now

The education landscape changed radically when the assessment policy pendulum swung once again and California's Public School Accountability Act (PSAA) officially ushered in an era of high-stakes accountability and educational reform. In the years immediately following the passage of this legislation, most decisions about curriculum, instruction, assessment, and professional development were made at the state level, with the expectation that change would occur at the school level. State-approved academic content standards guided the development of curriculum frameworks and the adoption of curriculum materials. High-stakes testing, the cornerstone of the legislation, was introduced to monitor student promotion and retention, to measure school success, and to drive curriculum development and selection. In short, the state adopted the teacher's pedagogical authority and high stakes were seen as the lever for change.

Stakes, as scholars have explained, are the consequences associated with test results (Heubert & Hauser, 1999; Madaus, 1988). As Heubert and Hauser noted, tests, as policy instruments, can be used in two fundamentally different ways. Low-stakes testing policy lacks significant consequences based on results, with the assumption that information alone (test scores) is motivation enough to promote change. High-stakes testing policy, on the other hand, assumes that information alone is insufficient. Advocates of high-stakes testing assume that serious consequences for low scores such as loss of funding or retention in grade are required. Stakes are characterized as high when consequences that flow from tests are automatic and have direct impact on students or adults either in the form of rewards or sanctions (Madaus, 1988). They are high when results are used to make important decisions about students such as graduation or placement in programs, or about adults for purposes of evaluation or for allocation of resources (Madaus, 1988).

At the time this study was conducted, California's high-stakes Standardized Testing and Reporting (STAR) program system had three components, the California Standards Tests, The Spanish Assessment of Basic Education, 2nd edition (SABE/2), and the Stanford 9 (SAT9). The California Standards Tests were comprised of items developed specifically to assess students' performance on

California's Academic Content Standards, and included the collection of actual writing samples, although only at Grades 4 and 7. The SABE/2 was used to assess achievement in reading, language, spelling, and mathematics in Spanish. Limited-English proficient students who had been enrolled in California public schools less than 12 months were required to take this test. The third component, the SAT9, was a multiple-choice test used to test students in Grades 2 through 11 in "reading, language (written expression) and mathematics" (California Department of Education, 2002).

The Stanford 9 was designated as the STAR Program's achievement test and was first administered in 1998. High stakes were attached to this test. The students scores were used in the calculation of the Academic Performance Index (API). In turn, the API was used to track school performance and determine whether schools would be designated as "high-performing" or "under-performing." In other words, accountability was enforced. High rates of improvement meant monetary awards for schools and teachers. Low rates, or failure to improve, meant the threat of reconstitution. When schools were reconstituted, personnel were dispersed and the school was rebuilt, so to speak, from the ground up.

High stakes were also attached to the other key test in the system, the California High School Exit Exam (CAHSEE). The plan was for students to receive a "certificate of completion" but not a diploma, if they finished high school but failed to pass the CAHSEE. A writing sample was collected as one component of the CAHSEE, but at the time this study was conducted, the sanctions associated with the CAHSEE were only on the horizon, and not yet in place. As a result, the primary means for assessing writing in high school were the "written expression" components of the multiple-choice SAT9/STAR system.

The contrasts between the two systems that operated in California within such a relatively short span of years offered an interesting opportunity to investigate and compare the impact of the two assessments on curriculum and instruction. An extensive literature on the subject suggests that tests do more than yield achievement scores. They define what achievement is. "Through the test," wrote Madaus and Kellaghan (1993), "the teacher, and later the policymaker defined what was expected of students" (p. 6). Frederickson and Collins (1989) coined the term *systemic validity* to refer to this phenomenon: Systemic validity takes into account instructional changes brought about by the use of a test and asks whether such changes are good or bad.

According to Madaus and Kellaghan (1993), standardized tests had little influence on state or federal policymakers from the 1920s to the 1960s. By the 1970s, however, policymakers began to use standardized test data to make high-stakes decisions, and the design and content of these tests began to influence what and how teachers taught.

Highly regarded scholars and professional organizations now agree that state standardized testing programs of the 1970s and 1980s had negative effects on students, teachers, and learning:

> Since about 1970, when standardized tests began to be used for a wider variety of accountability purposes, basic skills test scores have been increasing

> slightly, while assessments of higher order thinking skills have declined in virtually all subject areas. Officials of the National Assessment of Educational Progress, the National Research Council, and the National Council of Teachers of English and Mathematics, among others, have all attributed this decline in higher order thinking . . . to schools' emphasis on tests of basic skills. They argue that . . . the uses of the tests have corrupted teaching practices. (Darling-Hammond, 1994, p. 16)

Recent trends in the literature suggest that to some extent, history may be repeating itself. Numerous arguments have appeared in the last decade or so about the negative effects of important tests on teachers, students, and schools. Some scholars argue that tests have unequal consequences for different students (Jacob, 2001; Kohn, 2000; Madaus & Clarke, 2001; McNeil & Valenzuela, 2000). Others are concerned about conflicts between research based views of the construct to be assessed and the construct underlying the assessment (Haney, 1984; Pearson & Valencia, 1987). Still other scholars have focused on cause-and-effect relationships between teaching to the test and a narrowing of the curriculum (Clotfelter & Ladd, 1996; Corbett & Wilson, 1991; Firestone & Mayrowetz, 2000; Gitomer, 1993; Haertel, 1989; Koretz, 1998; McNeil, 2000a, 2000b; McNeil & Valenzuela, 2000; Mehrens, 1998; Mehrens & Kaminiski, 1989; L. Resnick & Klopfer, 1989; M.L. Smith & Fey, 2000). Evidence suggests that teachers will base instruction on the content and form of tests, especially when high stakes are attached (Madaus, 1988; M.L. Smith, 1991). Districts, as well as teachers, alter their curriculum to reflect the form and content of tests (Corbett & Wilson, 1991; Dorr-Bremme & Herman, 1986; Haney, 1991). In some cases, substantial amounts of time are spent preparing students for a particular test, time that is then unavailable for other curricular objectives (Koretz, Linn, Dunbar, & Shepard, 1991).

Scholars are also concerned about the impact of standardized assessments on teachers. Some suggest that mandated tests adversely affect the quality of teachers' lives at work (Cochran-Smith & Lytle, 1992; DiPardo, 1999). Evidence also suggests that standardized tests constrain the professional development of teachers, weaken the authority of their professional judgment, and work against helping teachers learn how to teach effectively (Corbett & Wilson, 1991; Hillocks, 2002; Pearson & Valencia, 1987; Shepard, 1991; M.L. Smith, 1991). Along with workbooks, canned lessons, drills, and other "teacher-proof" instructional packages, standardized tests tend to devalue the professional competence of teachers. M.L. Smith (1991), after analyzing teacher interviews and classroom observations in order to understand the ways in which testing affected instructional and curriculum decisions, reported that teachers experienced a range of negative emotions about having test scores published in newspapers each year by the state Department of Education. Teachers who experienced anger or embarrassment at having low test scores made public indicated that they would "do what is necessary to avoid such feelings in the future" (M.L. Smith, 1991, p. 9). This often meant foregoing curriculum choices that deviated from test preparation. Other studies reinforce the idea that loss of teacher autonomy occurs when classroom instruction becomes synonymous with test preparation (Jones, Jones, Hardin, & Chapman,

1999; Rapp, 2001). Rapp reported that 88% of National Board Certified Teachers surveyed in Ohio believe that high-stakes tests have lessened teacher autonomy in the classroom.

This study investigated the impact of California's different assessment systems on curriculum and instruction and on teachers. In particular, the study explored the amount of time teachers spent on particular language arts activities, what they emphasized when they were responding to student writing, what kinds of writing they assigned, and the kinds of professional programmatic activities they engaged in at school. The study also investigated teachers' attitudes and opinions about the different assessment systems and their influence on instruction.

Methods

Data for the study came from three surveys of California high school English teachers collected at two different points in time—1988 and 2001. The first two surveys were conducted in 1988 by Charles Cooper and Sandra Murphy as part of a research project funded by the national Office of Educational Research and Improvement (OERI) on the impact of assessment on curriculum and instruction at Center for the Study of Writing (CSW) in Berkeley. One survey was directed at middle school teachers of English and was designed to investigate the impact of the test on teachers of writing, what they knew about the test, and what they thought about it (Cooper & Murphy, 1989). A second survey was directed at high school teachers of English. It also investigated the impact of the test, but probed as well for detailed information about teachers' practices in the teaching of writing and the kinds of professional development activities they engaged in at school. In 2001, a third survey, supported by the National Council of Teachers of English (NCTE) and the University of California's Education Research Center (UCERC), was conducted in California as part of a larger project investigating the impact of assessment on the teaching of writing and factors that influence the preparation of students for college-level writing in three states (Huot, Murphy, & O'Neill, 2002).

The same procedures were followed for all three surveys to obtain a random sample of teachers. The surveys were mailed to principals, who then identified a teacher following a procedure that had been devised by a statistician at UC Berkeley. The principal was asked to write all the schools' teachers' names alphabetically on a numbered list and then to identify the teacher whose number on the list corresponded to a random number provided by the statistician. Principals sent the identified teacher's name to the project directors so that follow-up reminders could be sent directly to the teacher.

All of the surveys had high rates of return, likely reflecting educators' interest in and concern about the topic. The first CSW survey was mailed to 600 (37.5%) of the public junior high and middle schools in California, and 387 were returned, resulting in a return rate of 65%. The second CSW survey was mailed to 856 (100%) of the public high schools and 635 surveys were returned, resulting in a rather astounding return rate of 74%. In the present study, surveys were mailed to 770 public and private high schools, and 419 surveys were returned, resulting in a

return rate of 54%. Of the returned surveys, 337 were from public schools, representing 35% of the public high schools in the state. Data from the 1988 and 2001 public school surveys were used in the chi-square tests conducted for this study.

Survey Questions

Several of the questions on the two public high school surveys were identical, and those particular questions provided the data for statistical analyses of the impact of changing assessment policies on curriculum, instruction, and teachers in this study. Other data from the surveys helped to provide a more complete picture of the changes in the educational landscape between 1988 and 2001.

The questions that overlapped on the two high school surveys were about the frequency of different writing types that teachers assigned, typical assignment length, the amount of time allowed for assignments, the typical number of drafts required; the kinds of things teachers emphasized in responding to student writing, the amounts of time teachers typically spend on various kinds of language arts activities, the kinds of professional programmatic work teachers engage in at school, and the degree that statewide tests and other factors influence the curriculum.

A subset of seven questions from the larger set of overlapping questions on the high school surveys were included in the analysis. It was expected that the form of each test would influence what teachers did on several fronts, but in different ways, depending on whether the test was multiple-choice in form and emphasized the basics, or whether the test called for actual writing in a variety of genres. Differences were expected in what teachers reported about the amounts of time they typically spent on particular language arts activities, the kinds of things they emphasized when they responded to student writing, the kinds of writing they assigned most frequently, and the kinds of professional programmatic work that they engaged in at school. Seven chi-square tests of significance were employed to determine whether the responses of teachers in 1988 were statistically different from the responses of teachers in 2001. To avoid probability pyramiding, the overall alpha for the study was set at .05 and distributed over the seven tests. The patterns of responses across these questions suggested that the two different assessment systems influenced curriculum in very different ways, but in ways that were consistent with the form of each test.

Results

Questions about teaching practices asked teachers to focus on a particular class (second period, or the next class in which they taught English) as the basis for their responses to several questions. One question that was asked in both 1988 and 2001 on the high school surveys was: "During this Fall Semester *in this English class,* how much time are you likely to spend on each of the following activities in your teaching? PLEASE CHECK ONE ANSWER FOR EACH."

Time Spent on Language Arts Activities

Teachers' responses to the questions about the amount of time they spent on particular language arts activities in the designated class suggested that there were somewhat different patterns of emphasis in the curriculum in 1988 as opposed to 2001. As Table 1 shows, a majority of the teachers in both 1988 and 2001 indicated they spent a lot of time on teaching writing. However, more teachers in 1988 (76%) reported they spent a lot of time teaching writing than did teachers in 2001 (67%), a difference of approximately 9 percentage points, χ^2 (3, N = 961) = 11.94, $p < .01$.

Table 1: Time Spent Teaching Writing

	CAP		SAT9/STAR	
Time Spent	**No.**	**%**	**No.**	**%**
No time on this	2	.32	1	.30
Very little time on this	21	3.36	7	2.08
Some time on this	129	20.64	102	30.36
A lot of time on this	473	75.68	226	67.26
Total	625	100	336	100

The pattern was nearly reversed when teachers responses to the question about the amount of time they spent teaching grammar were considered. Relatively small percentages of teachers in both groups indicated that they spent a lot of time teaching grammar (11% in 1988 and 16% in 2001). Nevertheless, as Table 2 indicates, more teachers in 2001 reported spending some or a lot of time teaching grammar and usage than did teachers in 1988 χ^2 (3, N = 959) = 40.74, $p < .005$. In 1988, 47% of the teachers indicated they spent some or a lot of time teaching grammar, compared to 68% in 2001. Responses to questions about the amount of time teachers spent on teaching writing and grammar support the interpretation that the form of the test influences what teachers teach.

Table 2: Time Spent Teaching Grammar

	CAP		SAT9/STAR	
Time Spent	**No.**	**%**	**No.**	**%**
No time on this	66	10.59	12	3.57
Very little time on this	262	42.05	96	28.57
Some time on this	226	36.28	173	51.49
A lot of time on this	69	11.08	55	16.37
Total	623	100	336	100

Features Emphasized in Response to Student Writing

Survey questions in both 1988 and 2001 asked teachers how much emphasis they gave to a variety of features of writing when they were responding to student writing, including coherence, development of ideas, voice, paragraph structure, sentence fluency, sentence variety, word choice, correct usage (grammar), introductions and conclusions, and characteristics of the particular type of writing assigned (genre characteristics). Responses to questions about emphasis on usage (grammar) and genre characteristics support the interpretation that the form of the test influences how teachers teach. Recall that in 1988, the state assessment system sampled multiple types of writing. When teachers were asked how much they emphasized genre characteristics when they were responding to students' writing, more teachers in 1988 (88%) indicated that they put a lot of emphasis on genre characteristics than did teachers in 2001 (56%), a difference of 32 percentage points, χ^2 (3, N = 953) = 140.80, $p < .005$ (see Table 3).

Table 3: Emphasis on Genre Characteristics in Response to Student Writing

	CAP		SAT9/STAR	
Amount of Emphasis	**No.**	**%**	**No.**	**%**
No emphasis at all	5	.81	1	.30
Very little emphasis	7	1.13	38	11.45
Some emphasis	62	9.98	108	32.53
A lot of emphasis	547	88.08	185	55.72
Total	621	100	332	100

When teachers were asked how much they emphasized usage (grammar) when they were responding to students' writing, a majority of both groups of teachers indicated some or a lot (see Table 4). However, more teachers indicated they emphasized usage a lot in 2001 (50%) than in 1988 (36%), $\chi^2(3, N = 954) = 20.96$, $p < .005$.

Table 4: Emphasis on Usage (Grammar) in Response to Student Writing

	CAP		SAT9/STAR	
Amount of Emphasis	**No.**	**%**	**No.**	**%**
No emphasis at all	12	1.93	2	.60
Very little emphasis	87	14.01	34	10.21
Some emphasis	300	48.31	129	38.74
A lot of emphasis	222	35.75	168	50.45
Total	621	100	333	100

Responses to questions about what teachers emphasized when they were responding to student writing support the interpretation that the form of the test influences how teachers teach.

Variety of Genres and/or Forms Most Frequently Taught

Teachers in both 1988 and 2001 were asked "How often are you likely to assign each of the following types of multi-paragraph writing in this class during the current fall semester?" A subsequent question asked "Which one type of writing will you assign most often?" (see Table 5). Although it is clear that a large number of teachers in both groups indicated they taught response to literature (called "interpretation of a literary text" in the 2001 survey) most frequently, the data indicate that teachers in 1988 taught a somewhat wider range of writing types than teachers in 2001, as one would expect given the assessment of multiple types of writing in the CAP system. The contrast is more dramatic if the categories

Table 5: Writing Most Frequently Taught

	CAP		SAT9/STAR	
Type of Writing Assignment	**No.**	**%**	**No.**	**%**
Short story	11	2.17	3	.94
Summary	41	8.07	24	7.52
Argument	28	5.51	11	3.45
Response to literature	269	52.95	231	72.41
Proposal for solving a problem	2	.39	—	—
Sketch of person	3	.59	—	—
Autobiographical narrative writing	63	12.40	12	3.76
Reflective essay	37	7.28	17	5.33
Report	15	2.96	3	.94
Other	39	7.68	18	5.64
Total	508	100	319	99.99

are collapsed into response to literature versus "other" (see Table 6). Fifty-three percent of the teachers in 1988 indicated that they taught response to literature most frequently, whereas 72% indicated they did so in 2001, a difference of 19 percentage points, χ^2 (1, N = 827) = 31.04, p <.005. Responses to the question about what writing type of writing teachers taught most frequently support the interpretation that the test influences the content of the curriculum.

Table 6: Writing Most Frequently Taught

	CAP		SAT9/STAR	
Type of Writing Assignment	**No.**	**%**	**No.**	**%**
Response to literature	269	52.95	231	72.41
Other	239	47.05	88	27.59
Total	508	100	319	100

Program Development Activities Undertaken During the Past Year

Teachers in both years were asked about the kinds of program development activities on which a subcommittee of teachers in their English department (or the entire department) may have worked during the previous year. Teachers in 1988 and 2001 responded similarly when they were asked whether they had worked on developing a systematic writing program (see Table 7). The chi-square test was not significant $\chi^2(1, N = 957) = 1.58, p > .005$. However, when asked whether or not they had worked on devising ways to strengthen the teaching of usage (grammar) during the past year, teachers in 1988 responded very differently from teachers in 2001 (see Table 8). Substantially more teachers in 2001 (59%) said they had worked on ways to strengthen the teaching of grammar than in 1988

Table 7: Worked On Developing a Systematic Writing Program

	CAP		SAT9/STAR	
Response	**No.**	**%**	**No.**	**%**
Yes	348	55.59	198	59.82
No	278	44.41	133	50.18
Total	626	100	331	100

Table 8: Worked on Devising Ways to Strengthen the Teaching of Usage (Grammar)

	CAP		SAT9/STAR	
Response	**No.**	**%**	**No.**	**%**
Yes	144	22.9	196	59.21
No	484	77.1	135	40.79
Total	628	100	331	100

(23%), a difference of 36 percentage points, $\chi^2(1, N = 959) = 124.70$, $p < .005$. Responses to the question that asked teachers whether they had worked on ways to strengthen the teaching of usage (grammar) support the interpretation that the test influences the content of the curriculum.

Changes Made to the Curriculum to Prepare Students for SAT9/STAR

Responses to questions that were included only in the 2001 high school survey also support the interpretation that the form of the test influences what teachers teach. In the 2001 survey, teachers were asked whether they had made any changes in their curriculum to prepare students for the SAT9/STAR. Of the 327 public school teachers who responded to this question, 232 indicated they had made changes, 85 indicated they had not, and 10 of the public high school teachers did not respond. Of the teachers who indicated they had made changes, 192 described those changes in an open-ended response. The open-ended responses were coded into several categories. After the initial coding, some categories were collapsed into a broader category called "other impact" because some types of responses were very rare (see Table 9). This "other impact" category included comments about teachers' efforts to boost student effort, teacher-training, changes in class size for remedial students, and rewards.

Table 9: Reported Changes Made to the Curriculum to Prepare Students for SAT/STAR

Type of Change	No.	%
Test preparation	76	39.58
Reading	29	15.10
Basics	52	27.08
Align	27	14.06
Other impact	8	4.17
Total	192	99.99

Of the comments, 4% were in the "other impact" category, and included comments such as the following:

- Schoolwide Awareness Week—consists of a schoolwide Jeopardy/SAT9 competition.
- As a result of very high scores on the SAT9 test, our school and individual teachers received financial bonuses.
- We cut the class size for our more remedial students in math and English to 15 to 20 students.

Comments about other kinds of changes were much more frequent. For example, nearly 40% of the comments were about test preparation. This category included comments such as the following:

- We discuss test-taking strategies. Practice test-taking strategies.
- Test-taking strategies Kaplan (about 1 week).
- I have taken out pieces of literature so that students can work on/practice type of test; more objective tests.

Fifteen percent of the changes were in the area of reading. This category included comments such as the following:

- Actively teaching and understanding reading comp [comprehesion] strategies.
- Formal reading instruction.
- Added reading comprehension exercises; beginning to add more nonfiction.

Twenty-seven percent of the comments were about additions or revisions to the curriculum in basics (grammar and vocabulary). This category included comments such as the following:

- Personally, some of the work I do with vocab [vocabulary] has changed. I ask students questions more like the way they are on the STAR test.
- Spent more time on mechanics and grammar.
- I have been preparing more materials for grammar study to add to what I already do in that area.

Fourteen percent of the comments were about general efforts to align the curriculum with the test or the state standards, such as the following:

- I've included more lessons specifically tied in to STAR skills. I also refer to the test more often.
- With the help of a consultant, the curriculum now addresses the state standards. Each unit is designed to teach to specific standards that are on the SAT9 and HSEE.

In general, the teachers' comments support the interpretation that the statewide test has influenced curriculum in the schools.

Teachers' Opinions About the Tests and Their Influence on Curriculum

Data from the different surveys indicates that teachers held very different opinions about the state tests that were in place in 1988 and 2001. Teachers were very positive in their opinions about CAP, but mostly negative about SAT9/STAR. When middle school teachers were asked how they thought the CAP

writing assessment would affect their school's English curriculum, 92% indicated they thought it would improve or strengthen the curriculum. Forty-nine percent agreed strongly. When asked whether they thought that students in California would write more as a result of CAP, 96% agreed, 49.9% strongly. When asked whether they thought CAP was an improvement over multiple-choice tests, 97.6% agreed, 76.3% strongly. When asked whether the test would increase teacher expectations for students' writing achievement at their school, 93.5% agreed, 42.2% strongly. Ninety-seven percent agreed that students' chances for learning and achievement would be improved if English teachers started assigning and teaching seriously the types of writing that CAP assessed and 70.1% agreed that it would improve their chances "considerably." When asked if teacher responsibility and professionalism would erode as a result of CAP, 92% disagreed, 62.4% strongly. Ninety-three percent agreed that teachers themselves would learn more about the nature of written discourse as a result of CAP, 54.3% strongly. Teachers clearly endorsed the CAP assessment.

In the 2001 survey, teachers' opinions about the impact of the SAT9/STAR test were not solicited directly. But teachers gave their opinions anyway. At the end of the survey, teachers were asked: "Has the SAT9/STAR assessment had any impact on your school that we haven't already asked about." One hundred and forty –two teachers responded to this item with comments. Eighteen of the comments were positive in tone (13%), but as the examples illustrate, they tended to focus on successful school performance or rewards received instead of the quality of the test or its impact on curriculum and students.

- We improved enough to get the bonuses!
- We were one of six schools in LAUSD to meet our API goal—this has brought positive attention.
- SAT 9/Star assessment is a big focus at XXXX. We are the number 1 most improved high school in LA county. This has really helped to motivate both teachers and students.
- We're proud of our scores.

Thirty-three (23%) of the comments were about changes made or actions taken and were more or less neutral in tone:

- Scheduling changes.
- More math classes and class size reduction.
- We arranged for more tutoring.

However, 91 (64%) of the comments were decidedly negative in tone and content. The negative comments were about the test, the test policy, the time it took from instruction, and the negative impact it was having on morale, the curriculum, the school in general, and students in particular. The following quotes are representative:

- We are losing 3 days of instructional time. Students are very stressed. Teachers feel sometimes that they are being judged on superficial and sometimes irrelevant matters, things students haven't been taught and shouldn't be taught.

- Our school raised its API (Academic Performance Index) by 14 points last year. Focus has been on maintaining this mundane surge of our students ability to work with the letters A-E. Instead of promoting higher learning, critical thinking, and preparation for the real world, students are focused to be good test-takers.

- It has put my English colleagues under such pressure to "perform" on standardized tests that measure things that we do not believe should be the measure of an accomplished student. These tests do not measure true composition skills nor creative thinking. SHAME ON SACRAMENTO!

- Huge time demand. Shortens our classes for 6 days.

- We do not buy breakdown scores from the testing service. To be quite honest, our district/teachers assessed several tests 4 years ago and ranked the STAR test dead last. . . . We don't feel that it measures the skills laid out in the State Frameworks.

- Oh yes, it's caused panic. Soon we'll be teaching directly to the test I'm afraid. The ESL students I teach sit for two hour periods and stare at their papers: a form of child abuse.

- Yes, our teachers are demoralized because we know our students will not score well due to the lack of consistent, quality education in K-8 from the elementary/ middle school districts.

- Fear.

Survey responses to this open-ended question suggest that few of the teachers surveyed endorsed the SAT9/STAR program and many held negative opinions about it.

The contrasting opinions held by teachers in 1988 and 2001 are especially interesting in the light of other data that indicates that teachers in both time periods acknowledged the influence that the two statewide tests had on curriculum. As Table 10 indicates, in 1988, approximately 87% of the high school teachers said the CAP test had some or a lot of influence on their teaching. In 2001, approximately 77% of the high school teachers said that the SAT9/STAR test had some or a lot of influence. However, although a majority of teachers in both time periods acknowledged the influence of their respective statewide tests on curriculum, it is also clear that teachers in 1988 responded very differently from teachers in 2001. In 2001, less

Table 10: Teachers' Opinions About the Influence of the Tests on Curriculum

	CAP		SAT9/STAR	
Amount of Influence	**No.**	**%**	**No.**	**%**
No influence	21	3.34	17	5.13
Very little influence	58	9.24	58	17.52
Some influence	237	37.74	159	48.04
A lot of influence	312	49.68	97	29.31
Total	628	100	331	100

than one third of the teachers indicated that the SAT9/STAR test had a lot of influence on their teaching, whereas almost half of the teachers who responded to the 1988 survey indicated that CAP had a lot of influence on their teaching.

Because these data refer to different tests as well as to different populations of teachers, tests of statistical significance were inappropriate. Nevertheless, the data suggest that teachers in 1988 were more influenced by the accountability test used at the time than were teachers in 2001. In combination with the data on teachers' attitudes toward these assessments, the data support the interpretation that "carrot and stick" approaches in assessment policy, such as the policy currently operating in California, may not be as effective in promoting change as other, more professionally oriented approaches to accountability.

Conclusion

Although achievement tests may, on the whole, be only weakly associated with what might actually be taught in the classroom, they appear to have a pervasive effect on the curriculum. This study supports the findings of others in the literature that teachers shape instruction to match the content and form of tests. Although this might not be seen as a problem if the tests are truly worth teaching to, the literature on assessment suggests that narrowing and fragmentation of the curriculum occurs when teachers are overly influenced by multiple-choice tests. In an early article on the subject of narrowing and fragmentation of the curriculum, Norman Frederiksen (1984) called the influence of multiple-choice tests on instruction—not discrimination against minorities or women—"the real test bias." Frederiksen noted the potential of such tests to focus too much attention on the basic skills:

> Improvement in basic skills is of course much to be desired, and the use of tests to achieve that outcome is not to be condemned. My concern, however, is that reliance on objective tests to provide evidence of improvement may have contributed to a bias in education that decreases effort to teach other important abilities that are difficult to measure with multiple-choice tests. (p. 195)

The study reported here appears to confirm Frederiksen's fears. The findings also suggest that Lauren and Daniel Resnick (1992) were right when they elaborated on the thinking behind the idea that teachers will teach to tests. The Resnicks proposed three caveats that are useful in thinking about assessment policy:

> 1) *You get what you assess*. Educators will teach to tests if the tests matter in their own or their students' lives.
> 2) *You do not get what you do not assess*. What does not appear on tests tends to disappear from classrooms in time. (p. 59)

In the present study, there was evidence that more teachers in 2001 were emphasizing isolated skills in usage and vocabulary than teachers in 1988. Whereas instruction in grammar is clearly a valuable activity, we should be concerned if exercises in grammar and vocabulary are supplanting instruction in writing.

Although the results of the study reported here tend to prove the truth in the Resnicks' words, they also suggest that there is more than one way to make things matter to teachers. One way is to use a "carrot and stick" approach like the policy that was in place in California when this study was conducted. Another way is to test things that teachers find relevant and worth teaching and to involve them in the assessment process and treat them as professionals. Recall that almost 50% of the high school teachers who responded to the 1988 survey indicated that CAP had a lot of influence on their teaching, but only 26% of the teachers who responded to the 2001 survey said that the SAT9/STAR had that amount of influence. Recall that most (96%) of the teachers who responded to the 1987 middle school survey about CAP thought that students in California would write more, and 94 % thought that the English curriculum in their school would be strengthened as a result of the test. Clearly, the data suggest that testing things that are relevant and worth teaching and involving teachers as professionals in the assessment process may be the more effective way to use assessment in educational reform.

The Resnicks' third caveat is also important:

> 3) *Build assessments toward which you want educators to teach*. Assessments must be designed so that when teachers do the natural thing—that is, prepare their students to perform well—they will exercise the kinds of abilities and develop the kinds of skills and knowledge that are the real goals of educational reform. (L. Resnick & Resnick, 1992, p. 59)

Designing tests that encourage teachers to "do the natural thing" is not a simple matter. At the least, it calls for congruence between research supported views of the construct to be assessed and the construct underlying the assessment (Pearson & Valencia, 1987). It also calls for test formats that mirror good instruction. But designing a model test is not enough. George Hillocks' (2002) stated the issue this way: "At the center of the K-12 testing fury is the myth that testing alone is able to raise standards and the rates of learning. Certainly, testing assures that what is tested is taught, but tests cannot assure that things can be taught well" (p. 204). His research on state assessments of writing reveals how assessments may work against

helping teachers learn to teach effectively. It is more than likely that the structure of the state assessment in California worked against teachers learning to teach composition during the years when the state test did not include a writing sample. And it may do so even now that writing samples have been added at the high school level in the CAHSEE. Security procedures that prevent teachers from examining test items in conjunction with student responses, and assessment procedures in which teachers do not do the scoring or develop the rubrics and scoring guides cut teachers off from opportunities to learn from the assessment process (Myers, 2002). Such policies undermine the professional development of K-12 teachers. Effective professional development is one way to guard against the promotion of formulaic writing, a symptom that plagues many state systems, according to Hillocks (2002).

There is an alternative and it was demonstrated by the CAP program. During scoring sessions for CAP, when teachers immersed themselves in student responses to writing in a variety of different genres, and as they learned to evaluate those responses in precise and specific ways, using rubrics that made it clear that writing entailed more than generic features such as sentence structure, usage, and punctuation, they learned that texts take different forms with variation in social purpose. Dudley (1997) described other benefits for teachers who participated in the scoring sessions: "they learned what students do when they read and write, where they succeed and where they struggle. It followed very naturally that many teaching ideas, both philosophical and practical, were generated at those sessions" (p.17). Teachers' understanding deepened as a result of their direct participation in the assessment.

So where does this leave us? Findings from research on teacher policy have emphasized the importance of recognizing the teacher as a decision maker and agent of informed change. Yet many education reform policies treat teachers as technicians and as the recipients of the reform agenda (McNeil, 2000a, 2000b; McNeil & Valenzuela, 2000; Sloan, 2000; M.L. Smith, 1991). Bureaucracies can be effective when the work is routine, mechanistic, and highly predictable, but teaching and learning are not routine kinds of work. The work of teachers and students as literate human beings in a democracy is complex, situationally and culturally bound, often unpredictable. Systems that attempt to change teachers from the outside and discount this fact sometimes have the unintended consequence of marginalizing teachers and making them less enabled to respond to student needs. Assessment-development-as-staff-development can lead teachers to make significant and positive changes in their beliefs and classroom practices (Sheingold, Heller, & Paulukonis, 1995). Perhaps it is time for the pendulum to swing once again. Perhaps policymakers should revisit the idea that assessment systems can be designed to promote the kinds of abilities we want our students to have and the professional development of teachers. Perhaps it is time for policymakers to make a better investment in our children's futures.

REFERENCES

Anderman, E. (1997). Motivation and school reform. In M. Maehr & P. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 10, pp. 303-337). Greenwich, CT: JAI Press.

California Department of Education (1988). *Honig releases results of first statewide writing test* (news release). Sacramento, CA: Author.

California Department of Education (2002). Standardized Testing and Reporting Program: Help about STAR 2001. Retrieved 11-18-02 from http://star.cde.ca.gov.

California Education Roundtable. (1988). *Systemwide and statewide assessment in California: A report of the Subcommittee on Student Assessment to the California Education Roundtable*. Sacramento, CA: Author.

Camp, R. (1993a). Changing the model for the direct assessment of writing. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 45-79). Cresskill, NJ: Hampton Press.

Camp, R. (1993b). The place of portfolios in our changing views of writing Assessment. In R.E. Bennet & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 183-212). Hillsdale, NJ: Erlbaum.

Clotfelter, C. T., & Ladd, H. F. (1996). Recognizing and rewarding success in public schools. In H. F. Ladd (Ed.), *Holding schools accountable: Performance based reform in education* (pp. 23-63). Washington, DC: Brookings Institute.

Cochran-Smith, M., & Lytle, S. L. (1992). Communities for teacher research: Fringe or forefront? *American Journal of Education, 100*(3), 298-324.

Cooper, C., & Murphy, S. (1989). *What we know about teachers' perceptions of CAP and it's impact on instruction*. Paper presented at the California Assessment Program "Beyond the Bubble" Conference, Anaheim, CA.

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

Covington, M. (1992). *Making the grade: A self-worth perspective on motivation and school reform*. New York: Cambridge University Press.

Darling-Hammond, L. (1989). *Accountability for professional practice*. Teachers College Record, 91(1), 60-80.

Darling-Hammond, L. (1994). Setting standards for students: The case for authentic assessment. *The Education Forum, 59*, 14-21.

DiPardo, A. (1999). *Teaching in common: Challenges to joint work in classrooms and schools*. New York: Teachers College Press.

Dorr-Bremme, D., & Herman, J. (1986). *Assessing student achievement: A profile of classroom practices*. Los Angeles, CA: Center for the Study of Evaluation.

Dudley, M. (1997, January). The rise and fall of a statewide assessment system. *English Journal*, pp. 15-20.

Firestone, W. A., & Mayrowetz, D. (2000). Rethinking "high stakes": Lessons from the United States and England and Wales. *Teachers College Record, 102*(4), 724-749.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*(1), 193-202.

Frederickson, J., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Gitomer, D. (1993). Performance assessment and educational measurement. In R. E. Bennet & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 241-264). Hillsdale, NJ: Erlbaum.

Greenberg, K. L. (1992). Validity and reliability issues in the direct assessment of writing. *Writing Program Administration, 16*(1-2), 7-22.

Hadley, C. (1984, September). Direct writing assessments in the states. *Issuegram*, *Education Commission of the States, 54*, 1-7

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.

Haertel, E. H., & Calfee, R. C. (1983). School achievement: Thinking about what to test. *Journal of Educational Measurement, 20*, 119-32.

Haney, W. (1984). Testing reasoning and reasoning about testing. *Review of Educational Research, 54*(4), 597-654.

Haney, W. (1991). We must take care: Fitting assessments to functions. In V. Perrone (Ed.), *Expanding student assessment* (pp. 142-163). Alexandria, VA: Association for Supervision and Curriculum Development.

Heubert, J. P., & Hauser, R. M. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Hillocks, G. (2002). *The testing trap: How state writing assessments control learning*. New York: Teachers College Press.

Huot, B., Murphy, S., & O'Neill, P. (November, 2002). The impact of state-mandated writing assessment on high school English curriculum, teachers and students' preparation for first-year college writing. In S. Kaplan (Chair), *Does it get any better? The impact of state-mandated writing assessment on high school English curriculum, teachers and students' preparation for first-year college writing*. Symposium conducted at the annual convention of the National Council of Teachers of English, Atlanta, GA.

Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis, 23*(2), 99-121.

Jones, L.V. (2001). Assessing achievement versus high-stakes testing: A crucial contrast. *Educational Assessment, 7*(1), 21-28.

Jones, M.G., Jones, B.D., Hardin, B., Chapman, L. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan, 81*(3), 199-203.

Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools*. Portsmouth, NH: Heineman/Boynton Cook.

Koretz, D. M. (1988). Arriving in Lake Wobegone: Are standardized tests exaggerating achievement and distorting instruction? *American Educator*, *12*(2), 8-15, 46-52.

Koretz, D. M., Linn, R. L., Dunbar, S.B., & Shepard, L.A. (1991, May). *The effects of high stakes testing on achievement: Preliminary findings about generalizations across tests*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Linn, R.L. (2001). A century of standardized testing: Controversies and pendulum swings. *Educational Assessment, 7*(1), 29-39.

Lucas, C. (1988a). Toward ecological evaluation: Part one. *The Quarterly of the National Writing Project and the Center for the Study of Writing, 10*(1), 1-3, 12-16.

Lucas, C. (1988b). Toward ecological evaluation: Part two. *The Quarterly of the National Writing Project and the Center for the Study of Writing, 10*(2), 4-10.

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

Madaus, G., & Clarke, M. (2001). The adverse impact of high-stakes testing on minority students: Evidence from one hundred years of test data. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 85-106). New York: The Century Foundation Press.

Madaus, G., & Kellaghan, T. (1993). Testing as a mechanism of public policy: A brief history and description. *Measurement and Evaluation in Counseling and Development, 26*, 6-10.

McNeil, L. (2000a). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge.

McNeil, L. M. (2000b). Creating new inequalities: Contradictions of reform. *Phi Delta Kappan, 81*, 729-734.

McNeil, L., & Valenzuela, A. (2000). *The harmful impact of the TAAS System of testing in Texas: Beneath the accountability rhetoric*. Cambridge, MA: The Civil Rights Project: Harvard University. Retrieved, August 31, 2001, from http://www.law.harvard.edu/civil-rights/…testing98/drafts/mcneil_valenzuela.html.

Mehrens, W. A. (1998). Consequences of Assessment: What is the evidence? *Education Policy Analysis Archives, 6*(13), 30.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice, 8*(1), 14-22.

Myers, M. (2002). Foreword. In G. Hillocks, *The Testing Trap: How state assessments control learning* (p. vii-ix). New York: Teachers College Press.

Pearson, P. D., & Valencia, S. (1987). Assessment, accountability, and professional prerogative. In J.E. Readance & S. Baldwin (Eds.), *Research in literacy: Merging perspectives*. Thirty-sixth Yearbook of the National Reading Conference (pp. 3-16). Rochester, NY: National Reading Conference.

Rapp, D. (2001). Ohio teachers give tests an "F." *Rethinking Schools: An Urban Education Resource, 15*(4). WWW.rethinkingschools.org/archive/15-04/Ohio 154.shtml.

Resnick, D. (1982). History of educational testing. In A.K. Wigdor & W.R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies: Part II. Documentation section* (pp. 173-194). Washington, DC: National Academy Press.

Resnick, L., & Resnick, D. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M.C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction*. Boston: Kluwer Academic.

Resnick, L. B., & Klopfer, K. (1989). *Toward the thinking curriculum: Current cognitive research. 1989 Yearbook of the Association for Supervision and Curriculum Development*. Alexandria, VA: The Association for Supervision and Curriculum Development.

Sheingold, K., Heller, J. I., & Paulukonis, S. T. (1994). *Actively seeking evidence: Teacher change through assessment development* (Tech. Rep. No. 94-04). Princeton, NJ: Educational Testing Service, Center for Performance Assessment.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 73*, 232-238.

Sloan, K. (2000). *Teacher agency and the TAAS: maintaining the ability to "act otherwise."* Paper presented at the annual meeting of the American Education Research Association, New Orleans, LA.

Smith, M. A. (1988). Teachers play lead role in developing state writing test. *The Educator, 2*(1), 9.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8-11.

Smith, M. L., & Fey, P. (2000). Validity and accountability in high-stakes testing. *Journal of Teacher Education, 51*(5), 334-344.