The Reviews of State Content Standards in English Language Arts and Mathematics: A Summary and Review of Their Methods and Findings And Implications For Future Standards Development

Douglas A. Archbald
University of Delaware
July 1998

This paper is commissioned by the National Education Goals Panel (purchase order number ED-98-PO-2038). The opinions and recommendations expressed in this paper are those of the author and do not necessarily reflect those of the Goals Panel or its members.

Executive Summary

This report compares and contrasts the evaluation methods and findings of the AFT, CBE, and Fordham reviews of state English Language Arts and mathematics content standards. The reviews found that state content standards vary greatly in how they are organized, in the level of detail and specificity of their content prescriptions, and in the clarity of their expression. A number of the states' standards were found to be of exemplary quality and could be used as models for other states as they develop or refine their state content standards; however, an unacceptably large number of state content standards were judged to be of inadequate quality.

For English Language Arts standards, the average ratings on the 0 to 4 scale (with a midpoint of 2) by AFT, CBE, and Fordham were, respectively, 2.00, 2.29, and 2.34 – or "C to C+" in a letter grade scale. However, the Fordham sample did not include about 20 states with content standards that did not achieve at least a score of 2 in the AFT evaluation. Fordham's English Language Arts score otherwise (including all states) would have been much lower. For mathematics, the average ratings on the 0 to 4 scale by AFT, CBE, and Fordham were, respectively, 2.35, 2.88, and 1.43 – or in the high "D" to low "B" range.

There was considerable discrepancy in the reviewers' rankings of the states. That is, while there was more agreement than not – comparing one reviewer to another produced positive correlations – there was still considerable disagreement among the reviewers about which states had high quality and which states had low quality content standards. One reason for the discrepancies is that the reviewers used different criteria. While at a general level they were all concerned with the quality of standards, and they took into consideration such qualities as specificity, organization, clarity of language, and content coverage, at a more specific operational level, the reviewers defined these qualities and applied their evaluation criteria differently. Another reason for the discrepancies in the state ratings among the reviewers is that there is an unavoidable subjectivity in the evaluation process. State content standards are lengthy, complex documents, varying greatly in organization and content. There is inevitably a certain amount of inconsistency in the reviewers' application of their evaluation criteria. Both reasons help explain why in many cases the same state received different grades from different reviewers.

This report concludes by discussing ways research and policy initiatives could improve the quality of state content standards. National organizations should collaborate to promote greater consistency in the terminology used in content standards documents. Resources for such an initiative include existing research and theory on curriculum frameworks and content standards, the criteria and findings of the recently published reviews of state content standards (AFT, CBE, Fordham), this report, and the state content standards that received positive reviews from each of the reviewer groups. Research should be done to learn in more detail about how content standards are used by local educators and to determine the characteristics of content standards most effective in improving curriculum and instruction.

Introduction

The rise to prominence of state content standards for education over the last twenty years has coincided with the growing role of state education law and policy in local school and district governance. Twenty years ago most states did not have content standards; or they had them but the standards were in the form of optional supplementary materials local educators could order if they sought assistance for curriculum revision. In the 70s the state role in education greatly expanded with education finance equalization and the minimum competency testing movement; and in the 80s came the Nation At Risk report and a major national wave of state education reform. The role of state law and policy has grown steadily in public education funding, professional licensing, school accountability, student achievement testing, graduation standards, and curriculum. Now virtually all states have content standards, standards have grown more comprehensive in scope and detailed in coverage, and they are in most states no longer optional for schools to use.

In many states, content standards have been the linchpin of large scale reform programs aimed at upgrading curriculum and strengthening accountability. An enormous investment nationwide has gone into developing content standards. Often these are multi-year projects into which dozens and sometimes hundreds of professional people have invested weeks and months of time. Testing programs, textbook adoption, and teacher training in states have also been linked with content standards revision processes. Despite state content standards' prominent role in education reforms and the large investment in developing them, the quality of state content standards has received little critical scrutiny – until now.

A few years ago the American Federation of Teachers (AFT), the Council for Basic Education (CBE), and the Fordham foundation all decided to examine state content standards. The AFT, in 1995, was the first. This effort lead to an annual review, with it's latest report published in 1997 entitled Making Standards Matter. This report examines state standards, assessment programs, and incentives for students linked to achievement of state standards. A major part of AFT's report is its state-by-state review of the quality of state content standards. The Fordham foundation began its series of reviews in 1997 focusing on English Language Arts, and in 1998 published reviews of state standards in mathematics, science, history and geography. The CBE study, published in 1998, focused on English Language Arts and mathematics.[1]

These reports are the first systematic attempts to evaluate the overall quality of state content standards.[2] The present scrutiny of standards is needed and timely. For it is really only in the last ten years that there has been a concerted and sustained focus on developing, raising, and refining standards on both a national and state level.[3] And, now,

[1] See reference list for complete citations: Gandal (1997); Joftus & Berman (1998); Lerner (1998); Munroe & Smith (1998); Raimi & Braden (1998); Saxe (1998); Stotsky (1997).

[2] Many educators and state documents use the term "curriculum framework." While some analysts attempt to make distinctions, in common parlance these terms appear to have synonymous meanings. Years ago the term "curriculum guide" or "guidelines" was sometimes used.

[3] The year 1989 stands out because of President Bush's and the nation's governors' call at the Charlottesville Education Summit for a nationwide commitment to higher academic standards and the publication of three key national "standards" reports, Everybody Counts (National Research Council),

more so than in the past, content standards are occupying a central role in "systemic reform" as states are consciously linking other initiatives and policies to their state content standards to make reforms more coherent (Archbald, 1997; Curry & Temple, 1992; Floden et al., 1995; Fuhrmann, 1993). A lot is riding on state standards. It is therefore of critical importance to examine the content and quality of state standards to determine if they "will be strong enough to bear the considerable burden now being placed on them" (Finn, Petrilli, and Vanourek, 1998, p.1).

This report compares and contrasts the evaluation methods and findings of recent reviews of state content standards by the three organizations. I focus on English Language Arts (ELA) and mathematics because these two subjects were reviewed by each of the three organizations. The first section of this report discusses two critical issues affecting state content standards. The second section summarizes the reviews' results. The third section presents the criteria of the different reviews and explains the differing results. The fourth section discusses implications of the reviews and offers recommendations toward the goal of improved state content standards.

Section I.
Two Critical Issues:
No Single Definition of "Standards"
and
The Difficult Challenge of *Specific* Standards

No Single Definition of "Standards"

There are two points to be made which are helpful as background for this report. The first is about the *inchoate state* of content standards design – there is no standard language or model for content standards.[4] State content standards are a key component of the education policies of almost every state in the nation. Developing standards is generally a big project – time-consuming and involving many teachers and other experts. When content standards documents are finished they are unveiled with much fanfare. Their goals, contents, and intellectual ideals are widely applauded and create high expectations for curriculum reform and student learning. The extensive attention to content standards and their widespread acceptance belies the significant lack of consensus on how state standards should be organized, how specific they should be, and how they are supposed to transform instruction. There is no proven formula or model or best practice for guidance. This is evident in the startling variety in length, depth, terminology,

---

Science for All Americans (American Association for the Advancement of Science), and Curriculum and Evaluation Standards for School Mathematics (National Council of Teachers of Mathematics). Reports in other subjects followed shortly. Building a History Curriculum: Guidelines for Teaching History in Schools (Bradley Commission on History in Schools), Curriculum Guidelines (National Council for the Social Studies), and Charting a Course: Social Studies for the 21st Century (National Commission on Social Studies in the Schools).

[4] Joftus & Berman (1998, p.7) write, "The word standard means many things to many people. States use various terms to describe their expectations for student learning, including curriculum frameworks, goals, learning outcomes, proficiencies, and benchmarks." See also, Resnick & Nolan (1995).

organization, and specificity of state content standards. It is also evident in the differences in evaluation criteria used by the reviewers.

We need a "theory of design" for content standards linking purposes, content, and organization. This will require clear, hard-nosed thinking about the purposes and role of state content standards in improving education. What are realistic learning goals at each grade level? How should these be determined? How are standards intended to be used by local educators to improve student learning? How are standards supposed to increase schools' productivity? How do content standards link with other policies of the state/district education system? Too often, these questions receive insufficient attention in the rush to develop content standards. The reviews of the standards and the reaction papers they are generating will certainly be a catalyst towards more serious attention to these and other critical questions.

The Difficult Challenge of *Specific* Standards

The second point, related to the first, has to do with the question of optimal specificity in the design of content standards. The reviewers, quite reasonably from their perspective, expect state content standards to be specific, yet the specificity they expect has uncertain political and pedagogical implications. Generally, the more specific expectations for curriculum and learning become, the more transparent the gulf between the rhetoric of standards and the reality of learning in schools.

The reviewers expect state content standards to be specific enough to be unambiguous about the content to be covered and what students should know and be able to do at specified points in the K-12 sequence. For example: "Analyze spatial relationships using the Cartesian coordinate system in three dimensions" (Raimi & Braden, 1998, p. 12). "The student will differentiate between area and perimeter and identify whether the application of the concept of perimeter or area is appropriate for a given situation" (Gandall et al., 1997, p. 16). "Distinguish between fact and opinion, and main ideas and supporting details to draw meaning from discussions and oral presentations" (Joftus & Berman, 1998, p. 44). The reviewers argue for a level of specificity making it clear whether a standard has been achieved or not.

The expectation of specificity is warranted because a vaguely worded prescription for content or achievement cannot be a standard. A "standard" is "an acknowledged measure of comparison for quantitative or qualitative value" (American Heritage Dictionary); "something established for use as a rule or basis of comparison in measuring or judging capacity, quantity, content, extent, value, quality, etc." (Webster's New 20th Century Dictionary). For local educators to compare meaningfully their curriculum and their students' achievement to state standards, those standards must be specific. Standards like, "Compare and contrast communication in their writing and speaking," or "appreciate and respond to written, spoken, and audio-visual texts" (Stotsky, 1997, p.12) are way too general. Vague prescriptions like these can be met with widely differing forms of accomplishment or proficiency. Therefore, according to this conception, they cannot be considered a "standard."

The expectation of specificity is also warranted because of the kinds of claims made in state standards documents. In addition to the title of "standards" on most of the documents, their introductory statements typically describe goals such as "raising

standards," "creating a common core curriculum," "improving curriculum coherence," "promoting greater uniformity of curriculum," "promoting equal standards across schools and districts," and the like.[5] A necessary (though insufficient) condition for state standards to have these kinds of effects on local curriculum and instruction is that they are very specific in their prescriptions. If standards are not specific enough to prevent widely differing interpretations of their prescriptions, then different teachers or schools can follow their own preferred interpretations and generate their own different curriculum and expectations of achievement, undermining goals of uniformity of standards or improved coherence of curriculum.

It is important to understand that there are political and pedagogical barriers working against specificity. Content standards, particularly when linked to accountability, confront in most states strong traditions of local control over schools. Content standards committees – teachers, curriculum specialists, and academics – often do not see it as their task to be highly specific in their prescriptions about content and achievement. The writers of standards are prone to take a "middle ground" position, prescribing goals, topics, concepts, learning objectives, and skills at three or four points in the K-12 sequence, at a level specific enough to qualify as a guide or framework for curriculum, but not at a level specific enough to qualify as genuine "standards" in the strict interpretation of that term. Writers of more general standards would argue that they promote a vision of curriculum and learning, and give examples of that vision, but leave decisions about scope, sequence, and standards of achievement to users of the standards. The NCTM Standards, the ill-fated NCTE/IRA Standards, and many state standards apparently reflect this viewpoint.

Standards-writing teams are also influenced by the view that it is desirable for teachers, or groups of teachers in departments or grade levels, to have a reasonable amount of discretion over curriculum. This is typically justified on several grounds: the ideal of professional autonomy; the inevitability of some variation among teachers' interests and teaching strengths; the desirability of tailoring curriculum to local needs, opportunities, and resources; and the view that there really is no "one best" curriculum. The assumption is that students will gain more from enthusiastic teachers in charge of their own curriculum than from teachers following specific externally imposed topics and objectives. Thus, rather than write specific standards which threaten to supplant teacher discretion, it is believed more general prescriptions can be written which leave room for interpretation and respect teachers' professional autonomy.

Section II.
Findings of the Reviews

Can The Reviews Be Compared?

Is it reasonable to compare the reviews from the three organizations? The answer is "Yes," but with significant qualifications. First, the same state curriculum documents – the most recently available mathematics and ELA standards – were evaluated by each of

[5] Additional material on the purposes of content standards is in Joftus & Berman, p.6 and Gandal et al., p.2 & 15; see also Archbald (1997 and 1994).

the organizations, although Fordham in addition to mathematics and ELA evaluated, science, history and geography.

Second, serendipitously, each of the organizations used a 5 point scale. There is no indication this was planned, since the different reviewers make no reference to each others' methods, and the evaluations were designed autonomously by the respective organizations' reviewers. However, the scales of the different reviewers are defined differently (described in more detail later).

Finally, in one way or another, the reviewers were all concerned with dimensions of quality. The reviewers emphasized different dimensions and differed in the scope of their analyses, but all were concerned with at least two central qualities: the extent to which the standards were clear and specific enough to provide guidance to local users for curriculum planning and instruction, and the extent to which the standards embodied a core of academic content. Because the reviewers were each concerned about quality and because the reviewers did use some of the same criteria, it is reasonable to inquire into the extent to which the reviews produced similar findings.

Results of the Reviews in the Aggregate

Table 1 shows the average scores given by the different reviewers for the standards for each of the two subjects. A score of "2.0" is the midpoint of the 0-4 scale. Since some of the reports ultimately interpreted their scores in letter grades, 4.0 = A, 3.0 = B, 2.0 = C, 1.0 = D, and 0.0 = F. The first section below discusses the ELA results; the second section, the mathematics results.

Table 1 The Reviewers' Ratings of the Standards

	ELA		Mathematics	
	Mean	S.D.[6]	Mean	S.D.
AFT	2.00	1.01	2.35	.97
CBE	2.29	.73	2.88	.68
Fordham	2.34	.61	1.43	1.31

English Language Arts standards. For ELA state standards, the average ratings of Fordham, 2.34, and CBE, 2.29, are very close; AFT's is 2.00, exactly at the midpoint of the scale.

Two points are important concerning the Fordham-ELA rating. First, the Fordham ELA review excluded about 20 states – those not meeting the AFT's minimal "common core" criterion for standards in 1996. Thus, the Fordham ELA sample is skewed by the absence of the states with the most brief or vague content standards (in the eyes of the AFT reviewers). Fordham's 2.34 score would likely be significantly lower if this pre-selection had not been done.

The second point is that Fordham-ELA's state average of 2.34 is based on my computation of the Fordham scores. Finn et al. (1998, p.3) in their just-released summary of all of the Fordham-sponsored reviews reports an ELA state average of 1.21. Why the difference? The Fordham-ELA review was based on 34 separate criteria. So each state's

[6] S.D. = "standard deviation," a measure of dispersion of scores around the mean. The larger the standard deviation, the more "spread out" the scores; the smaller the standard deviation, the more "clustered" the scores around the mean.

content standards document received 34 scores, ranging from 0 - 4.[7] I calculated for each framework, its average score (average of the 34 scores), and then averaged the scores of entire sample of 28 standards reviewed. In the Finn et al. (1998) report, the derivation of the 1.21 score for ELA standards appears to result from a rescaling process so that the Fordham-ELA ratings could be reported as a letter grade. The algorithm used was to add the 34 scores for each framework, so each framework had a total score. Then – and this was not done in the original ELA evaluation (Stotsky, 1997) – cutoff points were set for grades of A through F, resulting in 12 states with Fs, 6 Ds, 4 Cs, 5 Bs, and 1 A, which produces an average of 1.21, or D+.

Mathematics standards. Turning to the mathematics evaluations, Fordham's average score for the states is 1.43, AFT's score is 2.35, and CBE's is 2.88. Note, that AFT simply gave each state one score on its 5-point scale of "unusable" to "exemplary," whereas Fordham's and CBE's state scores are average scores stemming from multiple criteria, each given a separate criterion score.

A need for improvement. Not surprisingly, the reviewers found a broad range in quality. Overall, the results of the reviews suggest there is much room for improvement in the state content standards. This conclusion is also drawn more or less strongly by each of the different groups of reviewers. The CBE report is the least critical in tone and sticks closely to its findings, merely stating that "fourteen states were found to have standards with low levels of rigor" (CBE, 1998, p.4); other CBE statements point out some specific strengths or weaknesses identified in the reviews. The AFT report concludes that, "Most states still need to improve some of their standards... and states need to be encouraged to revise and improve their academic standards" (AFT, 1997, p. v). The Fordham reports are considerably more critical and are replete with specific evaluative comments about the standards' weaknesses, and to a much lesser degree, strengths. For instance in the Fordham-mathematics executive summary are such statements as, "On the whole, the nation flunks... The failure of almost every state to delineate even that which is to be *desired* in mathematics education is a national disaster" (p. vii). Finn's forward in the Fordham ELA report points out that out of the 50 states, "just five emerge from this analysis with reasonably high marks" (p. i).

How Similar Are the Rankings?

"The way the American Federation of Teachers figures it, Michigan earns a C for the quality of its math and English standards. By the Council for Basic Education's reckoning, the grade rises to a B-plus. But on the Thomas B. Fordham Foundation's report card, Michigan plummets to an F. No wonder educators there are confused." So began a recent Education Week article entitled "An 'A' or a 'D': State Rankings Differ Widely" (April 15, 1998). According to the Education Week analysis, "more than half the

[7] Among the 34 criteria were 7 "negative criteria," also rated 0 - 4, but for which higher scores were worse. I reversed these (i.e., 4=0, 3=1, etc.) for the computation of the average. The Fordham review used a different system, adding the scores of the 27 "positive criteria" and then subtracting from this total, the scores on the "negative criteria" to produce each framework's final rating. In the Fordham scale, the state standards' final scores ranged from 3 to 94. This scoring system serves adequately to rank the states, but makes comparisons to the other reviewers' ratings impossible. For this, a simple average of the 34 scores is best.

states received marks in mathematics that varied by at least two letter grades across the three reports. In English/language arts, 19 states had such differences."

The concepts of inter-rater correlation and inter-rater agreement are useful in determining how similar or different the reviewers' ratings were to each other. These measures help answer the question – to what degree is one reviewer's ranking similar to another's ranking of state standards in ELA or in mathematics?

Inter-rater correlation refers to the statistical correlation between two different sets of scores, or more simply, the extent to which two sets of scores are ranked similarly. The possible range of a correlation is from –1.0 to +1.0, with a correlation of zero indicating absolutely no relationship between two sets of scores and a correlation of 1.0 indicating essentially identical scores given by the two raters.[8] For each subject, there are three sets of correlations to examine: CBE and AFT; CBE and Fordham; Fordham and AFT. The correlations are shown in Table 2. (Asterisks denote correlations that are statistically significant at the P=.05 level.)

Table 2: Correlations Between The Different Reviewers

English Language Arts		
	CBE	AFT
CBE		.352*
Fordham	.328	.522*
Mathematics		
	CBE	AFT
CBE		.282
Fordham	.280	.495*

Table 2 shows moderate correlations between the AFT and Fordham rankings in both ELA and mathematics, and a weak, but statistically significant correlation between AFT and CBE rankings in ELA. The correlations between CBE and Fordham (both ELA and math) and between CBE-math and AFT-math are very small. Using conventional statistical criteria (a fairly strict standard), the magnitude of the correlations would be considered too small to confidently rule out the possibility that they are due to chance.[9]

Table 3 presents inter-rater agreement percentages, which is another way to look at the extent to which rankings between two reviewers are similar or dissimilar. Table 3 shows the percentage of the state standards getting the same score from two different reviewers. The highest percent of agreement between reviewers is in ELA, between Fordham and CBE, with about half the states being awarded the same letter grade; the lowest is between Fordham and CBE in mathematics, with about one-fifth getting the same grade. (It was necessary to round off the grades awarded by CBE and Fordham to the

Table 3: Agreement Between The Different Reviewers

English Language Arts		
	CBE	AFT
CBE	--	23.8%
Fordham	53.8%	35.7%
Mathematics		
	CBE	AFT
CBE	--	39.5%
Fordham	16.7%	19.1%

[8] A negative correlation, which did not occur here, would indicate an inverse relationship between two sets of scores – in other words, the higher reviewer A scores a state, the lower reviewer B scores the state (on average).

[9] This is at the P = .05 level. It should be noted, though, that it is not entirely appropriate to apply this concept from statistics since statistical significance is used to gauge the adequacy of making inferences about population parameters from samples. In this case, the sample *is* the population.

nearest whole-letter grade, since both CBE and Fordham used plusses and minuses. This was not necessary to do for the computation of correlations.)

Like the correlations, the inter-rater agreement percentages are fairly low. Thus, if the question is – were the different raters judging the same qualities of the standards and assigning the same values to those qualities? – the answer quite clearly is, "No." The above results show greatly varying opinions among the reviewers in their judgments about which state standards had the best qualities and which did not. This is not surprising, because as we shall see below, the different reviewers' definitions of "quality" are quite different.

Section III.
Examining The Evaluation Criteria
And Other Factors Affecting the Evaluation of Standards

The Evaluation Criteria

Below I present the evaluation criteria and evaluation processes used by the different reviewers. These are summaries. I have attempted to be as faithful to the original descriptions of criteria as possible, in some cases even excerpting where original language is needed. In the last part of Section III, I offer interpretive and comparative comments which may offer some insights into the different reviewers' findings.

The Council for Basic Education: Rigor

CBE chose to focus on the "rigor" of state standards. There are two key components of CBE's definition of rigor: (1) *essential concepts and skills* and (2) *sophisticated learning* – "[applying the] essential concepts and skills at a level of sophistication or complexity that is appropriate and challenging to students at a particular grade level" (p10). For a standard to be rigorous, it must be both essential and challenging. As an example, the report describes "solving equations" as an essential concept in mathematics, but lacking in sophistication unless the standard makes more explicit the kind of equation to be solved (e.g., simple arithmetic versus solving a quadratic equation with one term unknown). The report acknowledges that the quality of standards is determined by a number of factors, including clarity, organization, and specificity and that judgments about the rigor of a state's content standards will reflect these features (clarity, organization, and specificity). Rigor in the two subjects is defined as follows:

Mathematics. CBE's evaluation of mathematics standards compared state mathematics standards in grades 8 and 12 to a set of 81 CBE-developed *framework benchmarks* for those same grades. CBE's 81 benchmarks were developed by drawing on two sources: the National Assessment of Educational Progress (NAEP) in mathematics and the National Council of Teachers of Mathematics (NCTM).[10] CBE's report states (p.

[10] "The CBE [math] study therefore begins with a list of performance demands expressed in 51 clauses [benchmarks] for the 8th grade and 30 for the 12th grade. These clauses are largely drawn from the NCTM Standards, both in spirit and in wording, or from criteria used in the National Assessment of Educational

9), "[State] standards should require all students, at the appropriate grade level, to learn the essential concepts and skills of mathematics at the level of sophistication specified [in NAEP and NCTM.]"

Here are examples of CBE benchmarks: "Understand the meaning of percent and apply it in meaningful contexts." (8th grade) "Read, interpret and make predictions using tables and graphs; interpolate or extrapolate from data." (8th grade) "Use transformations (translations, rotations, reflections, dilations, and symmetry) synthetically and algebraically." (12th grade) "Use appropriate notation and terminology to describe functions and their properties, including domain and range." (12th grade) "Express mathematical ideas and generalizations orally and in writing, using appropriate vocabulary and notation." (12th grade) These benchmarks are organized into 9 mathematics topics, closely reflecting the categories of the NCTM Standards.

English Language Arts: CBE compared state English language arts standards in grades 4 and 12 to a set of 62 CBE-developed *framework benchmarks*. CBE's 62 benchmarks were developed by drawing on CBE's own *Standards for Excellence in Education* which "was written in consultation with subject experts who drew inspiration from exemplary state and national standards documents" (p. 9). The CBE report does not explain further about the 1995 *Standards for Excellence in Education* document, but explains its decision not to use NAEP standards in English (key areas not covered; not enough specificity) and not to use the standards produced by the National Council of Teachers of English/International Reading Association (overly broad).

Here are examples of CBE benchmarks: "Distinguish between fact and opinion and main ideas and supporting details to draw meaning from various texts and media presentations." (4th grade) "Recognize and employ the distinguishing features of different types of writing, such as instructions, narratives, journals, stories, poetry, drama, letters, news articles, and speeches." (4th grade) "Present a report that is a culmination of a research process." (4th grade) "Analyze texts from a variety of literary genres with regard to author's craft (e.g., literary techniques, command of language); support opinions with evidence from a variety of sources besides the text (e.g., experience, news, other texts)." (12th grade) "Support an original thesis by conducting sustained research on a topic that synthesizes knowledge from more than one discipline and uses a variety of sources, such as technical journals and government publications." (12th grade) These benchmarks are organized into categories covering reading, writing, and speaking.

CBE evaluated the state mathematics and ELA standards using a 5-point, 0 - 4 scale. The five points of the CBE scale reflect the extent to which a given state standard reflects the evaluation criteria CBE established. Each point of the CBE scale has a definition specifying a set of conditions which must be met to achieve that score of the scale. Thus, each state's mathematics and ELA standards received, respectively, a total of 81 and 62 scores – a score for each *frameworks benchmark*. The final "rigor" score for a state's standards was its average score.

Progress. The NAEP Framework provided the "concept strands," while the NCTM Standards provide the "skill standards" (Raimi letter, 1998).

The American Federation of Teachers' "Common Core" Criterion

The AFT rated state content standards in four core subjects (mathematics, ELA, science, social studies) on their adequacy in providing a framework for a "common core curriculum." Content standards from 49 states received a "common core curriculum" rating for each of the four subjects. In making their judgments about a state's standards, the AFT reviewers used these five criteria:

#1 References to grade levels or clusters of grades. The AFT rating system does not prescribe a particular clustering approach, but states, "Strong standards … tend to use smaller grade clusters (e.g., K-2, 3-5, 6-8, 9-10, 11-12)." (p. 3) In general it appears states with grade-by-grade standards received higher sores: 8 of 14 state standards rated as "exemplary" have grade-by-grade standards.

#2 Detailed and comprehensive. "[S]trong standards should provide the basis for 60 to 80 percent of the academic curriculum [and] reflect the breadth and depth of each subject area." (p. 3).

#3 Firmly rooted in the content of the subject area. The standards must specify content – items of knowledge. For mathematics, examples given include the periodic table, the Pythagorean theorem, and the area of a circle. For ELA, specific examples of content are not given. The report states, "It is inadequate for an English standard to state that students should 'read a variety of genres' without specifying which genres and giving some examples of works, authors, or literary traditions" (p. 4).

#4 Clear and explicit. The AFT report does not directly define this criterion other than by explaining what is not: "It is not enough to provide selected detail of the content students should learn or the level of performance they should achieve and then claim these are only 'models' or 'examples' because this implies that other ideas of content or performance are just as acceptable" (p. 4). This criterion appears to require clear and explicit standards in all the topic areas prescribed in the content standards.

#5 Course-based standards must specify which courses are required of all students. This criterion applies primarily to secondary school standards, which are often organized by course rather than by discipline. The issue, according to the AFT report, is that not only must standards be prescribed for courses in the core subjects, a common core of courses required for all students must also be prescribed.

The AFT evaluation judged each of the state content standards as either "not meeting" (score of 0 or 1) or "meeting the AFT common core criterion" (score of 2, 3, or 4). State content standards which do not link standards in any way to grades or grade clusters received a 0 (only 2 states), and those that link standards in some way to grades or clusters of grades, but fall short on one of the other four criteria received a score of 1. Content standards meeting the common core criterion were judged to range in quality, scoring 2, 3, or 4. A "2" was "borderline ... meet[ing] our criterion, but only be a narrow margin" (p. 5). Content standards receiving a "3" "are in our view, strong enough to provide the basis for a common core of learning across the state. They embody the qualities of clarity, content, and precision described earlier" (p. 5). Standards receiving a "4" are described as "exemplary."

In the AFT review process, the content standards were not given separate scores on each of the individual criteria; rather, the criteria were used as guidelines in producing

the single rating. This is "holistic scoring" to use terminology from the field of performance assessment (Archbald & Newmann, 1988; Herman et al., 1992).

Fordham-Mathematics

Fordham's evaluation of state mathematics standards was based on four criteria:

#1 Clarity of the standards' statements. This criterion has three subcriteria: (a) clarity of prose; (b) use of mathematically definite language (unambiguous mathematical prescriptions); and, (c) prescriptions about knowledge and skill that are potentially subject to being tested.

#2 Adequacy of content refers to the right amount of content prescribed at appropriate grade levels -- "whether a state is asking its children to learn the right things at the right times, and enough such things, and not an unreasonable amount either" (p. 13). The state standards were rated at each of the three levels: primary, middle, and secondary.

#3 Mathematical reasoning should be pervasive among the state's content standards. Higher ratings were awarded when mathematical reasoning was an integral component of standards and reflected throughout content prescriptions; lower scores were awarded when mathematical reasoning was isolated from other content prescriptions and treated as a separate skill.

#4 Absence of the negative qualities "false doctrine" and "inflation." Regarding false doctrine, the authors write, "A standard must not offer advice which, if followed, will subvert instruction in the material otherwise demanded" (p. 15). This criterion results in lower ratings for standards containing prescriptions that reflect misconceptions or errors about the nature of mathematical learning and knowledge. The authors cite such examples as standards which indicate students should "discover" mathematical conventions or facts or which over-prescribe use of calculators or manipulatives on faulty assumptions about learning. Inflation refers to excessive verbosity and jargon in standards or standards which "suggest a profundity not possible for the level in question, especially if the indications are that the author doesn't understand the words being used" (p. 16).

The Fordham-Mathematics evaluators used a 5-point, 0 - 4 scale. The mathematics scale is defined "in more or less the usual way; '4' representing the best available, and '0' representing the least useful... these figures must simply be seen as a ranking of value" (p. 18).

Each state standards document received a total of 9 scores: 3 for the "clarity" subcriteria, 3 for "content" (primary, middle, secondary), 1 for "reasoning," 1 for "false doctrine" and 1 for "inflation." The two negative criteria were scored inversely, that is, more points being added for the absence of the negative quality. Each of the four major criteria (above in italics) received <u>equal</u> weight in the calculation of a content standard's final score. That is, the subcriteria scores were first averaged to create a single criterion score; then the 4 main criteria scores were averaged to create the standard's single total score.

Fordham-ELA

This evaluation did not attempt to examine ELA standards from all of the states. The main intent was to evaluate state ELA standards meeting AFT's "common core criterion" in its 1996 review. However, standards from several others states were included in the evaluation to insure review of standards from as many of the large states (by population) as possible and to include in the review a small selection of state standards not meeting the "common core criterion" for comparison purposes. The resulting sample included 28 state content standards, 21 out of the 22 meeting the AFT "common core criterion."

The ELA evaluation was based on five main evaluation criteria, each of which contained a set of more specific subcriteria. Here, summarized, are the main criteria (with the number of subcriteria given in parentheses):

#1 Purpose, audience, expectations, and assumptions. This criterion encompasses a variety of more specific criteria, with the following emphases: Only Standard English will be used, taught, and expected of students' writing and speaking in ELA classes; the standards acknowledge a corpus of American Literature, however diverse its origins and portrayals of American society; the standards expect decoding instruction in the primary grades, use of meaningful reading materials, and regular independent reading, with prescribed K-12 benchmarks of quality and quantity; the standards are clear enough for statewide testing. (8 subcriteria used.)

#2 Organization of the standards. Standards are presented at a minimum at 3 levels (e.g., primary, middle, secondary), are grouped into coherent categories of scholarship and research in ELA, and distinguish higher from lower order skills. (3 subcriteria used.)

#3 Disciplinary coverage. The standards cover listening, speaking, reading, literature, writing, history of the English language, and research processes in ELA. (7 subcriteria used.)

#4 Quality of the standards. Standards are clear, specific, measurable, comprehensive, and demanding. They prescribe increasing levels of difficulty at higher educational levels, indicating expectations with specific reading levels or titles, examples of writing, and sample assignments. The standards prescribe a specific common core of high academic expectations. (9 subcriteria used.)

#5 Absence of anti-literary or anti-academic qualities including claims that literary or popular culture is monolithic, requirements that students relate what they read to their life experiences, statements that "all literary and nonliterary texts are susceptible to an infinite number of [equally valid] interpretations" (p. 3), standards containing political ideology or dogma, and insistence on one instructional approach only for all teachers to follow. (7 subcriteria used.)

The ELA evaluation used a 5-point, 0 - 4 scale. The ELA scale used for each criterion is defined as follows: 0 = "no," 1 = "to some extent," 2 = "unclear," 3 = "for the most part," and 4 = "yes," with the score awarded based upon the reviewer's judgment as to how well the state standard met the criterion.

The state standards were scored on each of the subcriteria contained in the above 5 main criteria. This produced for each state content standards document a total of 34 subcriteria scores in 5 categories (the main criteria).

The Fordham-ELA review graded and ranked the states based on their total scores in the following way. For each state, the subcriteria scores were totaled to produce subtotal score for each of the 5 main criteria. Then, the subtotals for criteria #1 through #4 were added, and from this sum was subtracted the subtotal for criterion #5 (since this is a negative criterion). This produced for each state a final total score. The state scores ranged from a low of 3 to a high of 94. These sums were the basis of the state grades and rankings presented in the Fordham report (Stotsky, 1997).

How The Evaluations Were Done

Each of the groups doing the mathematics and ELA standards evaluations used a somewhat different approach. Understanding these differences is helpful in understanding the different outcomes of the evaluations.

The CBE reviews were done by a nine-person team, including the report's two authors. There were also advisory panels for each subject – an eight person advisory panel for ELA and a nine person advisory panel for mathematics. The advisory panels included "subject area experts, a parent representative, teacher representatives, and a business community representative... responsible for overseeing CBE's work and making recommendations for improvement" (p. 27). The panelists reviewed and made several recommendations for CBE-drafted definitions of rigor, frameworks, and rubrics.

The CBE reviewers spent "several days of training" refining the evaluation rubric and developing closer agreement on how to interpret and apply it. During this process the reviewers worked together as a group or in pairs scoring a small sample of the content standards (five), while reviewing each others' scores and discussing decisions for scores. This is called "calibrating" and it help create more consistency in scoring. The scoring rubric was modified during this process, with feedback from the subject matter advisory panels. Then, after additional discussion and calibration, the remainder of the ELA and mathematics state standards were scored independently, though with conferencing as necessary when scoring questions arose. The CBE report contains a 4-page section that describes, with examples, excerpts from state standards scoring 1, 2, 3, and 4 in both ELA and mathematics.

AFT's evaluation was conducted by a four person evaluation team -- AFT research staff lead by the report's principal author. The report does not discuss the question of consistency (reliability) in the rating process, nor whether the content standards were reviewed as a group or divided up among the four person group and reviewed independently. The report does include a chart (p. 16) show contrasting pairs of standards, one "strong" and one "weak," in each of the four subjects.

Fordham's reviews were commissioned from subject-matter experts in their respective disciplines. Fordham's five reviews (ELA, history, geography, math, science) were done by one or two people apiece. Fordham's evaluation of mathematics standards was conducted by a university mathematician and a school mathematics teacher with assistance by two advisory panelists. Fordham's evaluation of ELA standards was conducted by a university language arts scholar. The Fordham reviewers do not discuss the question of consistency in the application of their criteria during the review process, but each report cites repeatedly and extensively from the state content standards

illustrating high and low quality standards. Numerous excerpts are provided in the reports' overview sections and in the state-by-state sections.

Comparing the Different Evaluation Criteria and Methods

Why The Different State Rankings?

There are two probably equally important reasons that the different reviewers ranked the states differently. The first is the most obvious: they used different criteria.

Reason #1: Different criteria. AFT focused more on readily observable features of form than on the specifics of content. That is, they did not make judgments about disciplinary content or whether the standards were covering the "right" material. Rather, they asked if the statements of goals, topics, and skills were presented in a well-organized fashion, covered the K-12 sequence, and were reasonably specific in stating expectations at 3 or 4 points (as a minimum) in the K-12 sequence.

CBE, in contrast, developed its own set of content standards – the *framework benchmarks.* It is important to emphasize that these *benchmarks* are at just two selected grade levels (one primary and one secondary). Technically speaking, CBE's is a review of the extent to which state content standards matched the CBE content standards at two grade levels. It is an inference that the unreviewed portion of each state's content standard has the same qualities as the reviewed portion.

The Fordham reviews had the most comprehensive criteria in that qualities of coverage, specificity, rigor, and adequacy of content, as well as the presence of qualities deemed inappropriate, were all considered. The Fordham reviews were also the most explicit about their own epistemological perspectives, which is evident in their choice of negative qualities. (All five of the Fordham reviews use a "negative qualities" criterion.) Also, the Fordham reviewers simply seemed to have "graded harder" – they were less tolerant of weaknesses of organization or ambiguity in the standards' statements.

Reason #2: Inconsistency in scoring. The second reason for the differences of grades and rankings among the different reviewers has to do with the issue of reliability of scoring. Measurements must be reliable to be trusted. If a rating system gives inconsistent results, there is a problem of reliability. An example of this would two different people use the same scoring rubric on the same sample of items to be evaluated and producing different results.

Take for example the case of the early 90s statewide assessment in Vermont based on student portfolios. The intent was to evaluate students and schools by evaluating the quality of actual academic work rather than simply performance on multiple choice tests, and to do this using uniform criteria and scales to allow comparisons across schools. Vermont students developed portfolios of work based on a set of specific criteria, which were uniform statewide. The student portfolios were evaluated by teachers using the criteria and scoring scales. The results were intended to provide richer, but still measurable and quantifiable, assessments of student academic achievement and school performance. School performance measures were derived by aggregating the scores on the individual student portfolios. The problem was that the portfolios could not be rated reliably. Different raters using the officially prescribed and agreed-upon scoring rubrics would come up with differing scores. Inter-rater correlations for the most part were in the

.3 to .5 range (Koretz et al., 1993). The result was that the scores could not be trusted enough to use for school accountability purposes, and were used primarily as supplementary information for student evaluation purposes.

The reviews of the state content standards are a similar kind of performance assessment. The reviewers used scoring rubrics and evaluated written work – the content standards produced by teams of educators in each state. While each of the reviewers attempted to apply their own criteria consistently, the actual reliability of the reviewers' judgments is not known. We do not know, for instance, if a second rater using the Fordham-ELA evaluation criteria would come up with a similar set of scores for the 28 states; or, if the one of the review teams evaluated the same set of state content standards a second time (e.g. several months later) whether the second review process would produce scores similar to the first.

Thus, scoring unreliability in the sense described above must also be seen as a contributing factor to the discrepancies among the reviewers in the state rankings. A primary source of inconsistency is the multi-trait nature of the content standards. Many are lengthy, complicated documents. Content standards are commonly 40, 50, 60 pages or more of text, containing many hundreds of variously termed items of content ("goals," "objectives," "themes," "strands," "standards," "concepts," "processes," "vignettes," "activities," and more). Any given trait, such as "specificity" will vary throughout the document. Even if the evaluation criterion is defined precisely and clearly, and each content prescription can be accurately and precisely scored, the document as a whole will contain dozens or hundreds of such content prescriptions, so the reviewer is still faced with a subjective estimation of the proportion of such content prescriptions that meet the criterion. This is a challenging assessment task for one trait, let alone the many on which the content standards have been evaluated. While distinguishing the highest from the lowest quality content standards can be done with a high degree of reliability, discriminations among the "middle of the pack" content standards is considerably more difficult and this is where inconsistency becomes a factor in accounting for differing evaluations of quality.

Inevitably, reviewing content standards is a combination of objective measurement and literary critique. The evaluation of any given state's content standards reflects a set of choices about "the right criteria" on which to evaluate standards and a reviewer's judgment about the extent to which a given state's standards meet the criteria that have been established. Even if in theory "the right criteria" could be determined, there is still no way around a certain amount of measurement error in the assessment process due to the reliability issues described above. It is an imperfect science.

Key Points of Contrast Among The Reviews

I am concerned here primarily with the utility of the reviews for state education planners and policy makers. Below I discuss three qualities of the reviews -- comprehensiveness, distinctness of evaluation criteria, and level of detail.

Comprehensiveness. This refers to the breadth or scope of the reviews which is determined largely by the number of different criteria on which the content standards were evaluated. The Fordham reviews use the greatest number of criteria. They assess the content standards' purposes, the quality of their content, the clarity of their organization,

the specificity of their content prescriptions, and the standards' propensity for excessive "vogue" jargon or unproven pedagogical prescriptions that (at least in the eyes of some). The content standards are scored on each of these criteria, and so, in addition to a total quality score, readers have scores for each of these criteria of overall quality. This allows a more detailed analysis regarding each criterion.

The CBE and AFT reviews, by contrast are less multi-dimensional. The CBE benchmarks reflect the qualities of "essentialness" and "sophistication," but these are not scored separately and the benchmarks are applied at only two grade levels, reducing the comprehensiveness of the review. AFT's "common core criterion" also comprises more specific criteria, but they were not separately scored and also are more narrow in focus than Fordham's reviews.

Distinctness of criteria. It is useful for analytical purposes not just to use multiple criteria, but it is important that the criteria be distinct from each other. "Distinct" means that the different criteria reflect clearly different features of the work being evaluated. In other words, the separate criteria must have clearly distinguishable definitions; they must not overlap in their meaning. On these grounds the Fordham and CBE criteria are adequately distinct and for the most part clearly defined. However, the multiple criteria constituting AFT's "common core" definition are not as distinct as they should be. For instance, "detailed" and "comprehensive" (both under criterion #2) are very different qualities. A curriculum framework could be detailed in a given area, without being comprehensive (i.e., broad, complete) in its coverage. Secondly, "clear and explicit" (criterion #4) seems potentially similar to "detailed." If a standard is detailed in a given area, chances are it can also be described as "clear and explicit." Thirdly, the definition of the "firmly rooted" criterion (#3) seems also to be similar to the criterion "detailed." While "firmly rooted" could and perhaps should mean "central" or "core" content as distinguished from "peripheral" or "less essential" content, the AFT definition does not stress this principle; rather, "firmly rooted" is explained as requiring the inclusion of specific items of content in standards – which, once again, seems similar to criteria like "detailed" or "explicit." Finally, part of the elaboration of the "firmly rooted" criterion for ELA states that standards should "give guidance regarding the complexity and level of sophistication of literature students should be reading at a given grade level" (p.5). This requirement should probably be under #4, "clear and explicit" standards.

These problems of precision and clarity in the AFT's explanation of its evaluation criteria do not necessarily mean that the evaluators were unclear or inconsistent in their application of their rating criteria. It is possible that their own interpretation of their criteria were applied consistently, however, the report does not describe how the different individual criteria were applied in the evaluation process.[11]

[11] The AFT report does give a modicum of information about documents not meeting the "common core" criterion. "These documents either don't provide enough detail, are too light on content, provide only 'models' but no explicit standards, or they don't establish a common core in high school." Documents meeting the "common core" criterion at the middle range of the scale are "strong enough to provide the basis for a common core of learning across the state. They embody the qualities of clarity, content, and precision described earlier….The best standards are those that combine rich content and skills n a grade-by-grade format with precision, efficiency, and coherence" (p.5).

Level of detail. The Fordham reviews are by far the most detailed. The Fordham ELA report is book length, 155 pages of small print. The Fordham mathematics report is also lengthy -- 55 very full pages. Each Fordham report (all five subjects) has a lengthy review of each state's content standards complete with numerous excerpts and examples.

AFT's report -- the section focusing on state standards -- consists of a 5 page introduction and about 2 to 4 short paragraphs on each state. However, the AFT report in its entirety is much larger, since it also reviewed each state's policies on testing and "making standards count" (i.e. incentives for students).

The CBE report does not provide state-by-state descriptions (just scores). The CBE report does have summary information drawn from its overall findings. It reports on which of its benchmarks the state's as a whole score well (above 3.0) and poorly (below 2.00). For instance, CBE reports that most state ELA standards address basic skills and the processes of writing, but few specify desired amounts of reading (at 4th or 12th grade). CBE also reports aggregate level summary findings from its mathematics reviews.

Section IV.
Implications and Recommendations
For "Next Generation" Content Standards

Implications for State Content Standards Policy

Periodically state committees are assembled to write or revise standards. It is important that this process be efficient and effective, given the substantial resources and effort involved and the prominent and visible role of content standards in the framework of state education policy. The typical standards development process includes appointing committees for each subject area for the actual work of writing the standards. The committees typically review other standards documents and invite input from people throughout the state as they develop their own draft standards. Based on the findings of the review reports, it appears there is room for improvement in the preparation of the standards development committees and the attention and resources devoted to the content standards development process. Here, then, are some recommendations.

State content standards development committees now have the additional resource of the standards reviews. Thus, in addition to the usual processes, the standards development committees should read the AFT, CBE, and Fordham reviews, published responses to those reviews, and selected high-rated and low-rated state content standards. The objective is for committee members to

- understand what constitutes "quality" in state content standards,
- understand the justification for the criteria of quality, and
- agree on the goals, tasks, and issues involved in writing content standards.

The process of writing content standards will require that committee members deliberate over a number of questions. Key questions include:

- How should the standards be used by local educators?
- How should the standards be organized?
- What terminology should be used?
- At what grade levels or clusters should standards apply?

- How specific should standards be?
- How comprehensive should standards be in scope of coverage?
- At what achievement level should standards be set?

None of these questions have clear, definitive answers, but informed, thoughtful deliberation should be expected. As noted later, research and theory on curriculum frameworks and content standards can improve the quality of deliberation and provide working solutions.

Finally, it is also important for committee members to learn and be informed about other state education policies, particularly testing, accountability, professional development, professional certification, and textbook adoption policies. These policies affect teachers' and schools' capacity to achieve state standards. Writers of state content standards should offer recommendations for how other state education policies can support schools to achieve standards. While it is unreasonable to expect writers of content standards to become thoroughly informed about all other state education policies, it is important that standards are developed with an understanding of the state education policy context. The goal is to have standards that are reinforced by a coherent system of state education policy.

A Guide for Content Standards Development

A priority is to develop greater consistency among state content standards in terminology. Browsing through state standards documents one finds literally dozens of virtually synonymous terms. Rarely are definitions offered. This babel of terms stands in the way of the development of a more standard, precise terminology. A standard terminology would allow people from different states, organizations, and roles to communicate with and understand each other in working toward improved content standards.

It would be useful for a credible national group to promulgate a set of "essential features for content standards" that people in states and districts could use a guide (with the knowledge that others are using the same guide). The essential features would reflect the kinds of criteria used in the reviews discussed in this report -- a set of "shoulds" and "shouldn'ts," concerning the content and design of standards. The "essential features" would recommend terminology and provide answers to the kinds of questions posed earlier in this section. A good start would be define the term "standard. "

Three Research Priorities

#1 Can evaluation criteria be applied reliably? A credibility issue has arisen from the discrepant results of the different reviews. Since it is unlikely that the CBE, AFT, and Fordham reviews are the last to be done, it is worth investigating what might be done to minimize the discrepancy problem in future reviews. It would be informative to select certain criteria (e.g., testability, appropriateness of content, clarity of language) and then, following appropriate training, have different reviewers evaluate the same set of standards documents. It would be informative to learn whether peoples' professional backgrounds (e.g., mathematics teachers versus mathematicians) make a difference in how they rate standards; whether some criteria (e.g. testability) yield more reliable ratings than others (e.g., appropriateness of content); and whether standards in one subject (e.g.,

mathematics) can be evaluated more reliably than standards in another subject (e.g. ELA). Knowledge generated from this research would improve the quality and credibility of future reviews of state standards or other curriculum documents.

#2 Content standards documents should be designed to be as comprehensible, readable, and user-friendly as possible. Research involving teachers and district curriculum specialists, such as described above, can serve this goal. However, there is already a good deal of usable knowledge in the field of document design. Studies could explore how existing knowledge in the field of document design can help improve the effectiveness of content standards for their intended purposes.

#3 Earlier, a set of "key questions" was posed that would be useful for standards committee members to deliberate over. Research can provide some answers, or guidelines at least, to these questions. The reviews of the standards documents are a first step in providing descriptive information on the design and contents of state standards. Research now must focus on learning about how content standards are actually used by teachers and school districts and about relationships between variables of content standards design and variables of instructional practice and student achievement (e.g., are more specific standards more likely to influence instruction?). Research on these questions should have great value for content standards policy.

References

Archbald, D. (1994). Reflections on the design and purposes of state curriculum guides. Center for Policy Research in Education, Rutgers University, New Brunswick, NJ.

Archbald, D. & Newmann, F. (1988). Beyond standardized testing: Assessing authentic academic achievement in the secondary school. Reston VA: National Association of Secondary School Principals.

Archbald, D. (1997). Curriculum control policies and curriculum standardization: Teachers' reports of policy effects. International Journal of Educational Reform, **6**(2), April, 155-173.

Armstrong, J., Davis, A., Odden, A., & Gallagher, H. (1989). Designing state curriculum frameworks and assessment programs to improve instruction. Denver CO: Education Commission of the States.

Curry, B. & Temple, T. (1992). Using curriculum frameworks for systemic reform. Reston VA: Association for Supervision and Curriculum Development.

Finn, C., Petrilli, M., & Vanourek, G. (1998). The state of state standards. Washington D.C.: Fordham Foundation.

Floden, R., Goertz, M., & O'Day, J. (1995). Capacity building in systemic reform. Phi Delta Kappan. **77** (1), 19 - 22.

Fuhrmann, S. 1993. Designing Coherent Education Policy. San Fransisco: Jossey-Bass Publishers.

Gandal, M. (1997). Making standards matter: An annual fifty-state report on efforts to raise academic standards. Washington D.C.: American Federation of Teachers.

Herman, J., Aschbaker, P., & Winters, L. (1992). A Practical Guide to Alternative Assessment. Reston, VA: Association for Supervision and Curriculum Development

Joftus, S. & Berman, I. (1998). Great expectations? Defining and assessing rigor in state standards for mathematics and English language arts. Washington D.C.: Council for Basic Education.

Koretz, D., Stecher, B., Klein, S., McCaffrey, D., and Deibert, E., (1993). Can portfolios assess student performance and influence instruction? Santa Monica, CA: Rand

Lerner, L. (1998). State science standards: An appraisal of science standards in 36 states. Washington D.C.: Fordham Foundation.

Munroe, S. & Smith, T. (1998). State geography standards: An appraisal of geography standards in 38 states. Washington D.C.: Fordham Foundation.

Raimi, R. & Braden, L. (1998). State mathematics standards: An appraisal of math standards in 46 states, the District of Columbia, and Japan. Washington D.C.: Fordham Foundation.

Resnick, L., & Nolan, K. (1995). Standards for education. In Ravitch, D. (Ed.) Debating the future of American education: De we need national standards and assessments? Washington D.C.: Brookings Institution.

Saxe, D. (1998). State history standards: An appraisal of history standards in 37 states. Washington D.C.: Fordham Foundation.

Stotsky, S. (1997). State English language arts standards: An appraisal of English language arts/reading standards in 28 states. Washington D.C.: Fordham Foundation.