GREAT LAKES CENTER
FOR EDUCATION RESEARCH & PRACTICE

DOCUMENT REVIEWED:	"The State of State Standards 2006"
AUTHORS:	Chester E. Finn, Jr. and Michael J. Petrilli
PUBLISHER/THINK TANK:	Fordham Institute
DOCUMENT RELEASE DATE:	August 28, 2006
REVIEW DATE:	September 11, 2006
REVIEWER:	Ken Howe
PHONE NUMBER:	(303) 492-7229
E-MAIL ADDRESS:	ken.howe@colorado.edu

Summary of Review

The reviewed report assigns grades to the content standards of 49 states and the District of Columbia, on an A-F scale, and uses those grades as a basis for criticizing schools for lack of progress in improving standards. This review found no evidence supporting the validity of the grades and also found no evidence of a relationship to student academic performance, contrary to the report's conclusions. The report's claims in support of its grading practice were selectively data-mined and were seriously lacking in methodological rigor. Policymakers and educators would be ill-advised to base any decisions about policy or practice on the grades assigned by this report.

Review

I. INTRODUCTION

Educational standards have become a central issue in educational policy over the last several decades and have assumed particular importance with the passage and implementation of the No Child Left Behind Act of 2001 (NCLB). Standards are divided into two types: *content* and *performance*. Content standards specify the knowledge and skills to be learned in a given subject area; performance standards specify the level of learning deemed sufficient, typically labeled as "proficient."

Content and performance standards work in tandem in test-based accountability systems like NCLB. In theory, such systems "incentivize" educators to produce improved student learning by holding them accountable for improvement on performance standards, as measured by standardized tests. Performance standards must be "aligned" with content standards; content standards drive improvement in performance if they are sufficiently rigorous and provide guidance to educators by being clear, precise, and manageable in number.

The report under review, *The State of State Standards 2006*, rated each state's "subject" (i.e., "content") standards in U.S. History, English/language arts, mathematics, science, and world history, using an A-F scale. The report then compared those grades to earlier ratings from 2000. An accompanying document, *It Takes a Vision: How Three States Created Great Academic Standards*, provides a separate account of the development of the three state standards judged best: those of California, Massachusetts, and Indiana.

The report was produced by the Thomas B. Fordham Institute, which has "raising standards" and "strengthening accountability" at the forefront of its stated mission.[1] The authors of the report are Chester Finn, President of the Fordham Institute and Assistant Secretary of Education in the Reagan administration; Michael J. Petrilli, Vice President for National Programs & Policy of the Fordham Institute; and Liam Julian, Associate Writer and Editor, also of the Fordham Institute. Joanne Jacobs, author of the accompanying document ("It Takes a Vision"), is a freelance writer and blogger.

II. THE REPORT'S FINDINGS AND CONCLUSIONS

The main report reaches two primary conclusions, while the 'It Takes a Vision' document offers a third:

1. Between 2000 (pre-NCLB) and 2006 there has been no overall progress in raising the quality of state content standards. Whereas some have gotten better, this it offset by the finding that others have gotten worse. The average grade in 2000, C-, remained the average grade in 2006.

2. Students in states with better content standards do better on performance standards.

3. Effective leadership on the part of office holders, representatives of business, and academic content experts, against often significant resistance, is required in order for states to develop good content standards.

III. THE REPORT'S RATIONALES FOR ITS FINDINGS AND CONCLUSIONS

To support the first conclusion – that there has been no progress overall in raising the quality of content standards between 2000

and 2006 – the report compared the average letter grade Fordham Institute experts gave in 2000 with the average letter grade such experts gave in subsequent years, ending in 2006.

To support the second conclusion – that students in states with better content standards do better on performance standards – the report identified states that had made statistically significant gains in the percentage of students who attained proficiency in given subject areas of NAEP and then related this to the states' grades on the corresponding standards. Three examples are provided, one each from English/language arts, science, and mathematics.

To support the third conclusion – that effective leadership, against often significant resistance, is required in order for states to develop good content standards – case studies were provided documenting the development of content standards in the three states that Fordham judges to have the best standards: California, Massachusetts, and Indiana.

IV. REVIEW OF THE REPORT'S USE OF RESEARCH LITERATURE

The text of the report refers to work by the American Federation of Teachers (AFT), as well as Fordham's own previous work, but the report includes no citations. The report makes fleeting reference to the controversy surrounding testing and accountability regimens and again provides no citations. The report does not have a reference list.

The Fordham Institute exhibits a strong prior commitment to the centrality of education standards both in its mission statement and in the report under consideration (e.g., subject matter standards "are the foundation of standards-based reform, the dominant education policy strategy in America today"…and…"exert enormous influence over what actually happens inside the classroom," p. 6). By not including a meaningful discussion of the research literature, the report is able to simply assume that the "dominant education policy" is unproblematic. Research-based arguments on both sides question whether standards-based accountability regimes like NCLB improve student performance.[2] A rating or grading system like the one used in the report is based necessarily on a belief in a strong connection between the policy and an outcome goal that is accepted as beneficial. The issues of outcomes and of failing to address the research base are also important here because researchers must struggle with the fact that content standards are only one among many factors that might influence student (and teacher) performance.

V. REVIEW OF THE REPORT'S METHODOLOGY

The accompanying document by Joanne Jacobs employs, in a broad sense of the term, a case-study methodology—an approach that, generally speaking, is both useful and defensible. However, the methodology that the report itself uses to support its conclusions is highly problematic.

States' grades were determined by raters who were deemed experts by the report's authors. How many raters were used and what their qualifications might be are not addressed in the text of the report. A number of individuals are mentioned in the acknowledgements, many of whom would appear to be employees of the Fordham Institute.

State standards are judged on the general basis of whether they are "clear, rigorous, and right-headed about content" (p. 6). A few slightly more specific subject area examples are provided from English, science,

and history. The report does not specify the criteria that are employed; nor does the report tell the reader what the various grades mean in terms of such criteria. The reader is instead referred to the individual evaluations for each state for more specificity. But the individual state reports offer little that is more informative than what is provided in the text of the report. These state reports give the reader short descriptions of Fordham's summary judgments, but the reader is not given specific criteria or even the state standards to which Fordham's judgments are applied.

A much more detailed description of the grading criteria – which presumably remained the same – can be found in *The State of State Standards 2000*.[3] Still, this earlier document provides little description or defense of the procedure by which the criteria were developed and validated. In the case of English grading criteria, for example, the reader is referred to a 1997 one-page document by Sandra Stotsky, contained within "State English Standards," which vaguely describes a procedure that makes unspecified use of outside reviewers (who, how many, what they attended to, etc.) and heavily depends on her individual judgments. The resulting standards[4] include criteria such as English-only instruction in English/language arts (A2) and also "anti-criteria," i.e., things to be avoided, such as relating lived experiences to literature (F2) and addressing contemporary social issues (F3). This means that Fordham's standards are quite at odds with those of authoritative groups such as the National Council of Teachers of English.[5] Although the value judgments embedded in standards like Stotsky's may accurately reflect the beliefs of Fordham's leadership, readers are ill-served when important information about the character of its standards is hidden from view, as is the case with the report under review here.

In general, the report provides no evidence for the reliability of the grades—either that grades assigned by the same expert are consistent over time (test-retest reliability) or that different raters agree on grades assigned (inter-rater reliability). The grading process also apparently had no control for rater bias by insuring that raters were blind to information that might distort their judgments, e.g., that a given state had done well or poorly on NAEP or that "progressives" or "postmodernists" were influential in negotiating the standards. In sum, no evidence is offered that the grades are valid measures of the quality of state content standards. Readers are asked simply to rely on the overall conclusions reached by Fordham and its graders, supplemented by a few cursory statements in the state documents regarding strengths and/or weaknesses.

VI. REVIEW OF THE VALIDITY OF THE FINDINGS AND CONCLUSIONS

Because of its methodological shortcoming regarding reliability and rater bias, the report's first conclusion, about how states' content standards have changed or not, is supported poorly. Worse, there is no indication that the grades were in any way validated.

The report's second conclusion, that students in states with better content standards do better on performance standards, is even more poorly supported than its first conclusion. In the section entitled "Do Good Standards Improve Student Achievement?" the report begins its analysis by acknowledging that "there is no simple relationship" between Fordham's grades and student performance. Indeed, the figure presented in the report, plotting fourth grade NAEP proficiency percentages against Fordham grades (p. 13), suggests not only that there is no "simple relationship," it suggests no relationship exists at all.

Yet the authors of the report find the evidence inconclusive and decide to keep looking for a relationship. They argue, in effect, that the straightforward comparison is wanting because it is based on only one moment in time. They thus offer an alternative: "what matters is whether any reform, including adoption of rigorous standards, leads to progress over time" (p. 13). Using this approach, the report presents three analyses of the data, based on states whose NAEP scores have improved over time. These alternative analyses aim to establish a positive relationship between Fordham's grades and state-level student performance. Below is a brief description and critique of each of these analyses.

Analysis 1, English/language arts

Of 10 states that made statistically significant progress for at least one group in fourth-grade reading on NAEP between 1998 and 2005, nine received at least a C from Fordham for their English/language arts standards.

Inspection of the data for the analysis (provided in the table on p. 14) raises a number of questions. Of the 40 states (39 plus the District of Columbia; Iowa is not included) that failed to produce statistically significant gains, 33 (82.5%) had a grade of C or above, and their average grade was 2.25. This compares with 90 percent of the 10 states that did make significant gains that had an average grade of 2.4. Any conclusion about a positive relationship between Fordham's grades and student performance drawn on the basis of these relatively small quantitative differences is exceedingly dubious, particularly because the small number of states (only 10) that produced significant gains decreases the stability of the summary statistics associated with them.

Other observations further weaken Fordham's analysis. For example, California and Massachusetts, the two states with A grades among in 10 that produced significant gains in fourth-grade reading on NAEP, did so for only one of four relevant groups, the same number produced by Wyoming, with a grade of F. Of the four states that produced gains for the most student groups, three had C grades and one had a B grade.

Another – and better – way to compare Fordham's grades and student performance is in terms of changes in each, i.e., how improvements in Fordham grades between 2000 and 2006 are related to improvements in performance.[6] Indeed, this is more faithful to their own view, quoted above, that "what matters is whether any reform, including adoption of rigorous standards, leads to progress over time" (p. 13). If content standards drive performance, then improvements in performance should reflect improvements in content standards. Although still burdened by all the validity questions concerning the assignment of state grades, this approach more directly addresses the key question and also uses the whole dataset rather than the subset limited to those states that produced significant gains in NAEP. The results of this form of analysis further diminish Fordham's case. Gains in Fordham grades simply fail to be reflected by gains in student performance. The 10 states that produced significant gains in fourth-grade reading on NAEP had a mean improvement in their Fordham English grade of 0.2; three of these states (30%) improved their grade, two (20%) got worse, and five (50%) stayed the same. The remaining 39 (Idaho and Iowa are excluded) that failed to produce significant gains on NAEP had a mean improvement in their Fordham English grade of 0.62 (over three times the improvement of the states that produced significant gains on

CSAP); 20 states (51%) improved, eight (21%) got worse, and 10 (26%) stayed the same. States with lower grades showed greater improvement.

Analysis 2, science

Of the five states that made statistically significant gains on the science NAEP between 2000 and 2005 in both fourth and eighth grade, three had A grades.

Less impressive is the fact that the remaining two states that made significant gains received a D and an F. Fordham's rating system seems to have 'worked' for three out of five states – just over the 50 percent mark one would expect from random guessing – which does not seem noteworthy.

Analysis 3, mathematics

Four of six states that received "honors" grades from Fordham produced statistically significant gains in the percent proficient on the eighth-grade NAEP mathematics test between 2000 and 2005.

Why the reversal in the analysis strategy here? Why not an analysis parallel to examples 1 and 2 in which the starting point for the analysis is the identification of states that produced a statistically significant improvement on NAEP proficiency percentages? The apparent answer is that 23 states produced significant gains yet did not have content standards that were praised by Fordham. Using a consistent approach would yield answers inconsistent with the report's conclusions.

In summary, these three analyses were selectively mined from data gathered by Fordham – data which themselves are flawed and for which there is no evidence of validity. No rationale for Fordham's unorthodox and ad hoc analyses is provided, and those analyses are sorely lacking in methodological rigor. Indeed, the post-hoc massaging of the data reaches the point of absurdity, as the authors search for some approach to the data that might lend support to Fordham's conclusion that content standards of the kind it rates highly do result, in fact, in improved student performance.

The case studies of California, Massachusetts, and Indiana provided by Jacobs are less problematic. There is good reason to be cautious, however. Readers should not demand that Jacobs live up to an unattainable ideal of perfect objectivity; on the other hand, it is possible to go too far in the direction of subjectivity, which Jacobs does. The account she produces exhibits a marked bias in favor of Fordham's position on standards. It reads like a morality tale, in which sensible, hard-headed, altruistic "reformers," who support rigorous, precise, and clear standards, are "hand-to-hand combatants" engaging in the "good fight" to outmaneuver and defeat the muddled, soft-headed, self-serving "progressives."

VII. THE REPORT'S USEFULNESS FOR GUIDANCE OF POLICY AND PRACTICE.

Because there is no evidence supporting the validity of Fordham's grades, it would be unwise to base any decisions about policy or practice on them. It may very well be true that higher-quality content standards help improve results on performance assessments. But the Fordham report fails to offer a valid and reliable grading system to judge high-quality content standards. It also fails to establish that the grades it does present are associated with improved student performance.

Jacob's collection of case studies has some potential usefulness with respect to under-

standing the various dimensions of standard setting. But the collection pales in comparison to the kind of understanding provided by more rigorous and scholarly treatments of the subject.[7]

NOTES & REFERENCES

[1] Retrieved August 31, 2006 from http://www.edexcellence.net/foundation/global/page.cfm?id=6.

[2] See, for example,

Amrein, A.L. & Berliner, D.C. (2002, March 28). High-stakes testing, uncertainty, and student learning. Education Policy Analysis Archives, 10(18). Retrieved August 30, 2006, from http://epaa.asu.edu/epaa/v10n18/;

Carnoy, M. & Loeb, S (2002). Does external accountability affect student outcomes: A cross-state analysis. Educational Evaluation and Policy Analysis, 24(4), 305-331;

Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends.* Cambridge, MA: The Civil Rights Project at Harvard University;

Center on Education Policy (2006). *From the capital to the classroom: Year 4 of the No Child Left Behind Act*. Retrieved September 2, 2006, from http://www.cep-dc.org/nclb/Year4/Press/

[3] *The State of State Standards 2000.* Retrieved September 1, 2006, from http://www.edexcellence.net/foundation/publication/publication.cfm?id=24

[4] *The State of State Standards 2000*, p. 129-130.

[5] See NCTE Standards (retrieved September 1, 2006, from http://www.ncte.org/about/over/standards/110846.htm).

Number 10 states: "Students whose first language is not English make use of their first language to develop competency in the English language arts and to develop understanding of content across the curriculum." Number 3 states: "Students apply a wide range of strategies to comprehend, interpret, evaluate, and appreciate texts. They draw on their prior experience…."

[6] The 2000 English grades upon which this analysis depends are reported in *The State of State Standards*, p. 136 (retrieved September, 1 2006, from http://www.edexcellence.net/foundation/publication/publication.cfm?id=24).

[7] See, for example,

McDonnell, L. (2004). *Politics, persuasion, and educational testing* Cambridge, MA: Harvard University Press.

The Think Tank Review Project is made possible by funding from the Great Lakes Center for Education Research and Practice.