

-- PRELIMINARY AND INCOMPLETE. DO NOT QUOTE --

The Resource Costs of Standards, Assessments, and Accountability

Douglas N. Harris
Lori L. Taylor

with
Amy Albee
William K. Ingle
Leslie McDonald

January 10, 2008

A Preliminary Report to the National Research Council

Acknowledgements: This research was commissioned and funded by the National Academy of Sciences, National Research Council. We thank Stuart Elliot, Margaret Goertz, Margaret Hilton, and Lisa Towne for useful comments. All errors are those of the authors.

I. Introduction

The increasing role of state and federal governments in public schooling in the United States is one of the most fundamental education issues of the current era. The No Child Left Behind Act of 2001 (NCLB), dramatically increased the federal role by requiring annual testing of students in grades 3-8 and at the high school level, and by creating a cascade of sanctions associated with different levels of school performance.¹ The federal demand for increased accountability shifted some control over schooling from the local to the federal level. Because the federal requirements place so much weight on state-determined standards and assessments, and left states in charge of specific content and proficiency standards, NCLB has also shifted control to state governments. Not surprisingly, this has led to a myriad of state standards and assessments—some very stringent, others virtually toothless—that has frustrated those interested in making comparisons across states and fostered the suspicion that some states are artificially raising their passing rates by setting the bar low.

The tension between high and uniform standards and state flexibility has led to proposals for additional federal control over standards and assessments. One option could be voluntary standards and assessments, created by the federal government or a non-governmental body and available for adoption by any state that so chooses. Another would push the federal role further by requiring states to administer federal standardized tests, which would then become the basis for federal accountability.

¹ NCLB reauthorized the Elementary and Secondary Education Act of 1965 (ESEA). The 1994 reauthorization of ESEA, the Improving America's Schools Act of 1994, also required the states to develop standards and assessments, and imposed sanctions if students were performing poorly. All states were required to test students in reading/language arts and mathematics at the elementary, middle and high school levels. The states were also required to report disaggregate results for student subgroups. However, most states had received waivers or were otherwise not in compliance with the federal testing requirements. As of April 2002, only 19 states were fully compliant with the federal testing requirements under the Improving America's Schools Act of 1994 (Taylor 2002).

There are many issues surrounding the prospect of a stronger federal role in the system of standards, assessments, and accountability (SAA) and this report addresses one—the resource costs of SAA systems. In this report, we consider the following cost-related questions:

- (1) What costs are now being incurred by the nation to create, update, and minimally comply with standards, assessments, and accountability under current state and federal laws and rules?
- (2) What costs would the nation incur if the typical state system of standards, assessments, and accountability were replaced with a single common system?

One of the most important factors affecting the answers to these questions is how ambitious the systems are in terms of the goals they are trying to achieve. Given that part of the purpose of a common system is to prevent states from setting a low bar, it appears likely that the costs incurred in Question (2) would be greater than Question (1). However, we do not attempt in this report to predict what a common system might look like. Rather, we use existing information about what states are already doing to identify a “typical” system and then consider the costs of state implementation versus common implementation of that typical system.

Framing the analysis in this way, it is clear from the outset that the cost estimate arising from Question (2) will be lower than the cost estimate arising from Question (1). The reason is that some costs of SAA systems are the same whether the system serves one student or one million students. For example, the cost of developing standards and

designing tests for third grade math is the same regardless of the number of third graders. Because each state incurs these fixed costs of standards, the fixed cost of state implementation is roughly 50 times greater than the fixed cost of common implementation for any given type of standards. This approach to the analysis is not intended to stack the deck in favor of common SAA systems, but only to make the analysis more tractable and avoid guesswork.

By focusing on the “cost of creating, updating, and minimally complying with the system,” we deliberately exclude costs of achieving the objectives laid out in the SAA systems, sometimes referred to as the “adequacy” of the system. We do so for two reasons. First, such costs are clearly a function of the standards each state has set for itself. If the standards for proficiency are high, then the costs of achieving widespread proficiency by those standards are also high. If the standards are low, the costs are low. We have no way of knowing how high the common bar would be set, and therefore no basis for reliably estimating the costs of achieving common standards. Second, even if we knew where the nation would set the bar, we have few reliable guides to the costs of meeting those standards. A number of researchers have used a variety of techniques to estimate the costs of educational adequacy in individual states. (For a survey of the literature on educational adequacy, see Baker, Taylor and Vedlitz 2005). The estimates are all over the map and depend crucially on how high standards are set, the research methodology used in the analysis, and the extent to which state policies and regulations foster efficient behavior by school districts. Taylor (2005) finds that even after adjustments for inflation and state-to-state differences in the price level, estimates of the cost of an adequate education range from \$5,124 to \$15,655 per pupil (in 2004 dollars). A

reliable estimate of the costs of achieving a common proficiency standard in all 50 states would require a state-by-state analysis of institutional structures, student needs and the gap between existing state standards and the proposed common standards. Because of these difficulties in studying adequacy, we instead focus on the costs of compliance with SAA systems laws and rules.²

We focus primarily on the “costs to the nation” by which we mean the economic or “opportunity” costs of resources. These costs sometimes differ from the budgetary costs that individual governments incur. Some resource costs might not show up in a state or school district budget because the resources are donated (e.g., volunteers on a panel to create math standards) or because the resources are incurred by some other governmental or non-governmental body (e.g., a regional education agency). Also, some budgetary costs might not be considered economic costs because the funds are serving multiple purposes in addition to the SAA systems. While we focus on measuring costs based on the opportunity cost concept, we do utilize budgetary information as part of the calculations and provide some discussion of the differences between the budgetary and economic costs.

After providing a brief literature review of the costs of SAA systems (Section II), we describe the SAA systems in the three sample states chosen by the National Academy of Sciences: Florida, North Dakota and Texas (Section III). We then discuss the conceptual issues involved and our approach to data collection in the respective states (Section IV), including the interview protocol that served as the initial basis for data collection (shown in the Appendix). We answer the first research question, relating to the

² It is worth recognizing, however, that the recent vocal debate about the “under-funding” of NCLB crucially depends on whether one focuses on the costs of compliance or adequacy. The costs of adequacy are surely many times larger than the costs of compliance.

current costs of SAA systems, by first estimating the costs of the SAA systems in the three states (Section V), defining and measuring costs in the typical state under NCLB, and finally calculating the implied total national costs now being incurred (Section VI). To answer the second research question, we then compare these current national costs with those that would be incurred in a single common SAA system (Section VII). We discuss our results in further detail (Section VIII) and draw preliminary conclusions (Section IX). These estimates are, again, only one of the many issues that must be considered in any debate about common standards. The cost savings we find here must be balanced against the possible disadvantages of a common system.

II. Literature Review

There are many theoretical arguments that favor the use of educational standards, particularly national standards. As discussed in Betts and Costrell (2001), educational standards increase the information value of a diploma, while also providing students and teachers with additional incentives to improve learning. However, when all states apply their own standards, it is rational for each individual state to develop relatively weak standards so that more of its students will receive the coveted (and now more valuable) credential. As a result, the standards states set for themselves may be too low. Costrell's (1997) analysis demonstrates that as long as states retain the option to set higher standards, national educational standards always improve social well-being.

To reach the conclusion that standards are always good for society, however, the theorists presume, among other things, that standards are costless to develop and implement. In practice, standards and accountability are costly. Therefore, any

conclusion about the social desirability of a system of standards and accountability requires analyses of both the costs and benefits of such a system.

A number of researchers have examined the benefits of educational standards. For example, Betts and Costrell (2001) survey the empirical evidence and conclude that student achievement rises when standards rise. Harris and Herrington (2006) draw the same conclusion from an extensive review of the evidence and further conclude that these benefits are especially great for disadvantaged students. Other recent work in this area includes Bishop (2006), Carnoy and Loeb (2002), Hanushek and Raymond (2005), Harris (2007), and West (2007).

Whatever the effects of standards and assessments, very few researchers have attempted to quantify their costs. The General Accounting Office (GAO) surveyed state testing officials and school districts in 1991 and concluded that the costs of standardized testing were generally in the neighborhood of \$15 per student (GAO 1993). In 1991, \$15 per pupil was 0.3 percent of current operating expenditures.³ The GAO report found that “the personnel time devoted to test administration always comprised the majority of the costs, and these were, of course, costs only to the local school districts” (GAO 1993, p. 32). Phelps (2000) argues that the GAO should not have counted all of the teacher time spent administering exams as a cost of the exams because teachers can pursue productive activities while proctoring exams, and therefore that the \$15 estimate is an upper bound on the cost of administering standards.⁴ On the other hand, neither Phelps nor the GAO

³ According to St. John et al. (2007) average, per-pupil current operating expenditure in 1991 was \$4,902.

⁴ Similarly, if teacher aides rather than teachers proctored the exams, the opportunity cost of test administration would be lower because the opportunity cost of a teacher aide is lower than that of a regular teacher. However, the relevant comparison is between the resources used in the course of the usual non-assessment learning activities and we are unaware of any evidence that schools hire additional aides during assessment administration. Also, while some teachers (and aides) might pursue other productive activities while administering assessments, doing so arguably occurs at the expense of monitoring students therefore

study ascribes any costs to standard setting, so their estimates could be too low. In either case, their analysis predates NCLB and undoubtedly understates the current costs of SAA.

A decade after the GAO report, Hoxby (2002) approached the question of costs from two directions. First, she gathered data on the revenues of test development firms (most states outsource test development and scoring) for 2000, and found that the total revenues of test makers accounted for less than \$6 per American student, or considerably less than one-tenth of one percent of expenditures on K-12 education. The costs of assessment have obviously increased with NCLB, so Hoxby's first cut at the analysis is necessarily a lower bound estimate of the costs as of 2007.

Hoxby also collected data from 25 state education agencies on the budgetary outlays for school assessment and accountability. Her cost estimates following this approach included all state level costs of assessment and accountability, such as the costs of test development contracts, the costs of running a state government office of accountability, and the costs of publishing test results. Her estimates did not include any district level costs such as the opportunity costs of administration. She again calculated the costs per pupil of state accountability systems. Her findings were strikingly similar to those of the GAO, ten years earlier. Hoxby found only one state where the costs of accountability exceeded \$25 per pupil (Delaware), and nine states where the costs were less than \$10 per pupil. A contemporaneous survey of all 50 state departments of education by Achieve Inc. also found that the budgetary costs of developing, administering and correcting state tests (a narrower definition of assessment than Hoxby

reducing compliance with rules aimed at avoiding student cheating. For this reason, in the later analysis, we do count teacher time as part of the cost of administering assessments.

used) were \$9.04 per pupil, on average, and less than \$25 per pupil in all but two states (The Education Commission of the States 2001).⁵ On average, school districts spent \$7,485 per pupil in 2001 (St. John, et al. 2007), so both Hoxby's study and the Achieve survey indicate that assessment and accountability comprised no more than 0.33 percent of current operating expenditures.

Other than the GAO report, there are, to our knowledge, no studies of the costs of SAA systems at the local level. There are, however, some previous studies that, while they do not discuss costs per se, do outline the resources that are involved. We use these in the process of estimating costs, as discussed later. As we show later in the analysis, cost estimates of SAA systems in previous studies are considerably lower than our own estimates, both because of NCLB and because we include a wider array of resources that go into SAA systems.

III. State Systems of Standards, Assessments, and Accountability

To provide a sense of the types of resources involved in SAA systems, and how they vary across states, we provide brief descriptions of the systems in Florida, North Dakota and Texas. These three states reflect a range of enrollments, expenditures and assessment histories.⁶ Texas is the largest of the three with more than 4.4 million students in 2005, and annual expenditures of \$7,246 per pupil. Florida is somewhat smaller, with 2.6 million students in 2005, but its expenditures per pupil (\$7,215) are very similar to Texas'. Both Texas and Florida also have a history of SAA that predates NCLB. In contrast, North Dakota is a small state (100,513 students in 2005) with

⁵ Those two states were Delaware and Alaska. Alaska was not part of Hoxby's sample.

⁶ Data on enrollments and expenditures in this paragraph come the NCES Common Core of Data.

relatively high annual expenditures per pupil (\$7,829 in 2005) and no history of criterion-referenced testing before the introduction of NCLB. (While many elements of these state systems have been influenced by the requirements of NCLB, we leave the discussion of NCLB until Section VII.)

Florida

The state with arguably the longest and most aggressive record of employing assessments and accountability has been Florida (Herrington & MacDonald, 2001). Florida was the first state in the nation to require annual testing of every student in select grades and subjects every year (1973), a high school exit exam for receipt of a high school diploma (1978), a college sophomore exit exam for advancement to upper division (1984). Beginning in 1990, the state legislature enacted *Blueprint 2000*, another accountability initiative resulting in the elaboration of seven (later eight) state education goals and ratcheting up the rigor and number of state assessments (Harris, Herrington, and Albee, 2007). A school improvement process was established, with incentives and sanctions based on progress on the state goals. The measure called for schools to assess their own yearly progress on a number of measures, including but not limited to performance on state achievement tests, and for district or state intervention if progress was not forthcoming after three years. In 1995, with this tighter protocol, the state identified 158 public schools in Florida as “critically low” performing (FDOE, 2004). This school grading system evolved over the next few years and served as the basis for

the A-F school grading system has been featured prominently since the adoption of the 1999 A+ *accountability program*.⁷

The cornerstone of the Florida accountability system is the Florida Comprehensive Assessment Test (FCAT). According to the DOE (2007), the Florida Commission on Education Reform and Accountability began conceptualizing the FCAT in 1995. The Commission recommended procedures for assessing student learning in Florida that would raise educational expectations for students and help them compete for jobs in the global marketplace. The State Board of Education adopted the Commission's recommendations in June 1995, including developing content standards and assessments in four broad areas: reading, writing, mathematics, and creative/critical thinking.

The state moved quickly to implement the policy, creating the Sunshine State Standards and requesting outside bids for the design, development, field testing, and implementation of (criterion-referenced) assessments. CTB/McGraw-Hill was chosen as the assessment contractor. Tests in reading at grades 4, 8, and 10 and in mathematics at grades 5, 8, and 10 were added to the ongoing assessment of writing at grades 4, 8, and 10. In 1996, the State Board of Education approved the Sunshine State Standards as Florida's new academic standards and distributed descriptions of the standards to school districts. In addition, the 1996 Florida Legislature passed laws recognizing the Sunshine State Standards as the academics standards for Florida students, and authorized the February 1997 field testing of the standards-based assessments. Four field-test forms were utilized at each grade level to obtain statistical information on a large pool of items. The purpose of the field test was to evaluate the quality of the items before they were

⁷ At one point, students in schools receiving a school grade of F for two or more consecutive years were eligible to receive a voucher to attend a private school. This voucher program was later judged unconstitutional under state law (Harris, Herrington, & Albee, 2007).

included on a test on which students received scores. Using these results, the first form of FCAT was developed. In 1998, nearly all students in grades 4, 5, 8, and 10 took the FCAT reading and mathematics tests for the first time. The tests, administered for baseline data, included some performance assessment items and measured students' skills in grades 4 (reading), 5 (math), 8 (reading and math), and 10 (reading and math). In May 1998, the districts, schools, and parents received the baseline FCAT results. Although the FCAT results were not used for accountability purposes in 1998, the preliminary results were released to schools.

In 1998, the Florida Legislature changed the language of Section 229.57, F.S., to permit the identification of students who would not be required to take and pass the High School Competency Test (HSCT). In the early fall of 1998, educators, citizens, and business leaders from across the state were involved in a process that led to the development of achievement levels on FCAT. The State Board of Education subsequently adopted the achievement level standards that enabled FCAT scores to be reported in a manner consistent with the legislative requirement. In addition, the Commissioner designated the score level for the exemption from HSCT. Students earning an FCAT total mathematics score of 315 or higher were not required to take the HSCT mathematics test; students earning an FCAT total reading score of 327 or higher were not required to take the HSCT communications test. Attaining a high score on the FCAT thus provided an exemption from the HSCT.

In February 1999, the second administration of the FCAT occurred for grades 4, 5, 8, and 10. In addition to releasing the results to districts, schools, and parents, the results were released to the general public and used as the basis for a new school grading

system. That same year, the legislature authorized an expansion of the state student assessment program to additional grade levels and to an additional norm-referenced test (NRT) component. Harcourt Educational Measurement received the test development contract for its SAT-9 (now SAT-10) norm-referenced assessment. National Computer Services (NCS), now NCS Pearson, received the contract for the scoring and reporting of the FCAT results. The first tests administered under the terms of the new NCS test support contract were those given statewide in February and March of 2000. (Because the state has subsequently administered both criterion- and norm-referenced tests, the term “FCAT” refers to both tests or, when referring to a specific test, we refer to the test based on the Sunshine State Standards as FCAT-SSS and the other as FCAT-NRT.)

Table 1: Recent History of Florida’s Assessment System

Academic Year	Reading and Language Arts	Mathematics	Science	Writing
1997-1998	4,8,10 CRT	5,8,10 CRT		
1998-1999	4,8,10 CRT	5,8,10 CRT		
1999-2000	3-10 CRT, NRT	3-10 CRT, NRT		4,8,10 CRT
2000-2001	3-10 CRT, NRT	3-10 CRT, NRT		4,8,10 CRT
2001-2002	3-10 CRT, NRT	3-10 CRT, NRT		4,8,10 CRT
2002-2003	3-10 CRT, NRT	3-10 CRT, NRT	5,8,10 CRT	4,8,10 CRT
2003-2004	3-10 CRT, NRT	3-10 CRT, NRT	5,8,10 CRT	4,8,10 CRT
2004-2005	3-10 CRT, NRT	3-10 CRT, NRT	5,8,11 CRT	4,8,10 CRT
2005-2006	3-10 CRT, NRT	3-10 CRT, NRT	5,8,11 CRT	4,8,10 CRT
2006-2007	3-10 CRT, NRT	3-10 CRT, NRT	5,8,11 CRT	4,8,10 CRT
2007-2008	3-10 CRT, NRT	3-10 CRT, NRT	5,8,11 CRT	4,8, 10 CRT

The third administration of the FCAT took place in February and March 2000 in grades all grades 3-10. Only students in grades 4, 5, 8, and 10 received FCAT score reports and used as the basis for assigning school grades. In February and March 2007,

the fourth administration of the FCAT included assessments in writing, reading and mathematics in grades 3-10. Science tests were added in 2002-2003 and, in 2008, the science scores will become part of a student's cumulative score and be assessed in school and district grades.

Florida continued to ratchet up the accountability attached to the increasingly frequent assessments. As part of Florida's accountability plan, the Florida legislature implemented the Special Teachers are Awarded (STAR) program in 2006. The STAR program, with a \$147.5 million appropriation by the 2006 Florida Legislature, recognizes and rewards educator and school personnel for outstanding performance and rewards teachers for improvements in student performance. Each school district is required: (1) to adopt a salary schedule that bases a portion of each instructional employee's salary on performance; and (2) to adopt a performance-pay policy for school administrators and instructional personnel. Both the salary schedule and the performance pay policy are based upon employee performance as demonstrated in the district's performance assessment system (something that each district had already required to have). A school district may submit a performance pay plan by either revising its existing performance pay plan, or creating a new one. Each plan must include an evaluation component focused on the improvement of student learning (at least half of the total evaluation).⁸ Plans are reviewed by the Florida Department of Education and must be approved by the State Board of Education. The state appropriation of funds was set so that districts can provide merit pay equal to five percent of district average salary to the top 25 percent of instructional personnel in the district.

⁸ The focus on improvement is noteworthy as it does not punish teachers whose student start at an initially low achievement level. Improvement is also partly the basis for the school grades discussed earlier.

In 2006, the Florida legislature passed the A++ Plan. The A++ Plan indicated that for each school in the district that earns a school grade of "C" or below, or is required to have a school improvement plan under federal law, the school improvement plan shall, at a minimum, also include: Professional development that supports enhanced and differentiated instructional strategies to improve teaching and learning; continuous use of disaggregated student achievement data to determine effectiveness of instructional strategies; ongoing informal and formal assessments to monitor individual student progress, including progress toward mastery of the Sunshine State Standards, and to redesign instruction if needed; and alternative instructional delivery methods to support remediation, acceleration, and enrichment strategies.

Schools that receive an A or schools that improve at least one performance grade category receive \$100 per student as a reward. The staff and school advisory council at each recognized school jointly decide how to use the financial award. As specified in statute, schools must use their awards for one or any combination of the following: nonrecurring faculty and staff bonuses, nonrecurring expenditures for educational equipment and materials, temporary personnel to assist in maintaining or improving student performance. The final school recognition list for 2007 included 1,613 schools and \$129,385,669 in awards.

North Dakota

North Dakota has a more modest history of statewide assessment and one much less well documented. During the 1990s, the North Dakota Department of Public Instruction (NDDPI) administered norm-referenced tests through a contract with

CTB/McGraw-Hill. In 2001-2002, NDDPI began shifting toward a criterion-referenced assessment that was aligned to state content and achievement standards (see Table 2). From 2001-2002 to 2003-2004, the NDDPI administered CTB/McGraw-Hill’s *Terra Nova*⁹ with a section of the “a dedicated State Supplement of uniquely aligned test items” (NDDPI, 2006, p. 9). Testing was initially administered to grades 4, 8, and 12 only. In 2004-2005, the NDDPI introduced a newly integrated criterion-referenced instrument—the North Dakota State Assessment (NDSA). The NDSA, also designed by CTB/McGraw-Hill, combines both selected-response items (e.g., multiple choice) and constructed-response items (e.g., essay) and is administered to students in grades 3 through 8 and 11 during a three week testing window in the fall of each year. Reading/language arts and mathematics tests have been administered in all grades since the inception of the NDSA, while science tests have been administered in grades 4, 8 and 11 since 2006-07. The norm-referenced *Terra Nova*⁷ is not currently used.

Table 2. History of North Dakota’s Assessment System

Academic Year	Reading/Language Arts	Mathematics	Science
2001-2002*	4,8,12	4,8,12	
2002-2003*	4,8,12	4,8,12	
2003-2004*	4,8,12	4,8,12	
2004-2005	3-8,11	3-8,11	
2005-2006	3-8,11	3-8,11	
2006-2007	3-8,11	3-8,11	4,8,11
2007-2008	3-8,11	3-8,11	4,8,11

*Academic years in which a hybrid assessment consisting of the *Terra Nova, Second Edition, Basic Multiple Assessment* and a North Dakota supplement (aligned with State Standards) was administered

⁹ Specifically, the *Terra Nova, Second Edition, Basic Multiple Assessment* was administered.

North Dakota state law requires that the state accountability system compile assessment data that can be used to compare performance of individual students, classrooms within a given school, schools within a district, and school districts within the State. Individual student reports, content standard performance reports, and summary reports are generated and delivered to North Dakota school districts by CTB/McGraw-Hill.

In compliance with the federal *No Child Left Behind Act*, North Dakota students have three testing options: (1) a standard statewide achievement testing program of the general education curriculum; (2) a standard statewide achievement testing program with valid and reliable testing accommodations; or (3) the Alternate Assessment if he or she is unable to participate in the state-wide testing program due to significant disabilities

The NDDPI's Standards and Achievement Unit is responsible for the administration of accountability efforts; the development of state content and achievement standards; administration of state and federal language acquisition programs; administration of the NDSA and the National Assessment of Educational Progress (NAEP) test programs; and the provision of statewide professional development opportunities under Title II and Innovative Programs under Title V. The unit has nine full time employees, consisting of a director, three assistant directors (state testing, bilingual/language acquisition programs, and alternative assessments), three program administrators (State/Federal Programs, Title II/Title V, NAEP), and two administrative assistants.

The North Dakota accountability system incorporates a number of sanctions. Schools and districts failing to make federal adequate yearly progress (AYP) under

NCLB for four consecutive years are required to take corrective measures, ranging from implementing a new curriculum, or extending the school day or school year, to operating under new management, replacing key staff, and/or complete restructuring.

If a district has multiple schools in program improvement and in corrective action, each school must choose the corrective action option that best addresses their unique needs. Furthermore, schools/districts must also write a program improvement plan, use 10% of the district's Title I allocation for professional development purposes (optional), receive technical assistance, offer school choice (if applicable), and offer supplemental services (if applicable). North Dakota does not provide rewards to high-performing schools or supports to low-performing schools beyond those outlined in NCLB.

Texas

As in Florida, school assessment in Texas has deep roots.¹⁰ In 1979, the Texas legislature directed the Texas Education Agency (TEA) to develop a criterion-referenced, basic skills assessment. Starting in 1980, the Texas Assessment of Basic Skills (TABS) was administered annually to all students in grades three, five and nine. School and district level results were made public, allowing parents and taxpayers to hold school districts accountable, but there were no official sanctions. Students who did not pass were not prevented from advancing to the next grade or from graduating.

The state increased the rigor of its assessment system in 1985, with the introduction of the Texas Educational Assessment of Minimum Skills (TEAMS). TEAMS was administered to students in odd numbered grades from first through eleventh, and in order to graduate, students had to pass the eleventh grade TEAMS in

¹⁰ The information in this and subsequent paragraphs is taken from TEA (2002).

mathematics, reading and writing. Summary reports continued to be published, but as with TABS, there were no official sanctions for schools or school districts.

Texas switched to the Texas Assessment of Academic Skills (TAAS) in the fall of 1990. TAAS was considered more rigorous than TEAMS, and more closely aligned with the state curriculum standards. TAAS was administered to students in odd numbered grades, starting with the third grade, but the eleventh grade test remained the only high stakes test. At the same time, the state rolled out its Academic Excellence Indicator System (AEIS), which made it much easier for parents and taxpayers to compare student performance across schools. At least initially, there were no official sanctions for schools and districts.

In 1993, the Texas legislature revamped the school finance formula to meet its court-mandated equity obligations, and simultaneously introduced the Texas public school accountability system. Under the accountability system, the TAAS test was shifted from the fall to the spring, and expanded to include all grades from third through eighth. The ninth grade test was dropped and the exit exam was moved from the eleventh to the tenth grade.

Starting in 1994, the accountability system assigned ratings to each school and district in Texas. Schools were rated as Exemplary, Recognized, Acceptable and Low Performing, based on TAAS passing rates, attendance rates, and dropout rates. Furthermore, the TAAS component of the accountability rating was based on the performance of the lowest-performing student subgroup (low-income, white, black, Hispanic and all students). In other words, if all of the other student groups were

performing at exemplary levels, but the low-income students were performing at only an acceptable level, then the school was rated only acceptable.

The accountability system incorporated a number of rewards and sanctions. The state provided each school with a “School Report Card” and required each school to provide a copy of the report card to every parent with children in the school. Schools had to be rated Exemplary or Recognized, or Acceptable with high performance gains, to be eligible for modest monetary awards through the Texas Successful Schools Awards program (TSSAS).¹¹ Exemplary school districts were exempt from a number of state rules and regulations, and low performing school districts were subject to sanctions ranging from the requirement to develop and action plan to complete state takeover or forced consolidation with another district.

Table 3: History of Texas’ Accountability System

Academic Year	Reading/ Language Arts	Mathematics	Social Studies	Science	Writing
1993-1994 to 2001-2002	3-8,10	3-8,10			
2002-2003*	3-11	3-11	8,10,11	5,10,11	4,7
2003-2004	3-11	3-11	8,10,11	5,10,11	4,7
2004-2005	3-11	3-11	8,10,11	5,10,11	4,7
2005-2006	3-11	3-11	8,10,11	5,10,11	4,7
2006-2007	3-11	3-11	8,10,11	5,10,11	4,7
2007-2008	3-11	3-11	8,10,11	5,10,11	4,7

*Although all tests were administered, no accountability ratings were assigned in 2002-2003, which was the first year TAKS tests were administered. **All tests are criterion referenced.

¹¹ Total state outlays for the TSSAS were only \$5 million per year, or less than 0.05 percent of operating expenditures.

In 1999, the legislature revisited both the state curriculum and the accountability system. As a result, Texas transitioned to the Texas Assessment of Knowledge and Skills (TAKS) test in the spring of 2003 (see Table 3). New subjects, including science and social studies, were added to the system, and the third, fifth and eighth grade tests became high stakes tests for students.¹² The state also introduced the Successful Schools Initiative (SSI) to provide additional support to students who are at risk of failing these high stakes tests. In 2007-08, the budget for SSI exceeded \$500 million.

TAKS measures student aptitude on five different tests. Students are tested in reading every year between third grade and ninth grade; in grades 10 and 11, they are tested in English language arts. In mathematics, students are tested annually from third grade until eleventh grade. Fourth graders and seventh graders take the writing assessment. The social studies exam is administered to eighth, tenth, and eleventh graders. Likewise, students in fifth, eighth, tenth, and eleventh grades take a grade-level science exam. Accountability ratings continue to be based on the performance of the lowest-performing student subgroup.

In addition, Texas added a parallel testing instrument for special education students. The State Developed Alternative Assessment II (SDAA II) is used to measure the performance of special education students whose instruction follows the state curriculum. Whereas other measurement tools are based upon the number of participants, this performance measure is based upon the percentage of SDAA II exams administered by a district or campus that meet passing standards. A student who qualified to take all three SDAA II exams (reading/English-Language Arts, writing, & math), for example,

¹² High stakes testing was introduced incrementally. In 2004, students had to pass the third grade test to be promoted. In 2006, they also had to pass the fifth grade test. In 2008, eighth graders will have to pass.

would be counted three times, rather than as a single student. A district or campus must administer at least 30 SDAA II exams to be evaluated under this performance measurement.

The 2007 accountability rating system also uses two methods to measure student retention. The percentage of students who either complete their high school education within four years or are continuing their high school studies four years later are reflected in Completion Rate I. This is measured as the ratio of students who complete their education to the number of students in their class. Although students who received a GED in previous years were considered to have completed their high school education, the 2006 assessment does not include students with GEDs as completers.

Student retention is also measured through a district or campus-level annual drop-out rate. This measurement is defined as the sum of seventh and eighth grade students from the previous academic year who are designated as official drop-outs divided by the sum of all seventh and eighth grade students who were in attendance at any given time during that academic year.

With the shift to TAKS, the consequences for a low performance rating became more severe. In addition, schools with a low performance rating are not eligible for the Texas Educator Excellence Grants program, which provides significant funds for teacher performance pay in schools that serve low income students.

Discussion

This historical background above provides some sense of how much has changed in SAA systems in recent decades and the potential for further evolution. More

importantly here, the basic descriptions of the policies in Florida, North Dakota and Texas provide a general sense of the types of activities involved with SAA systems, which is important background for identifying the specific resources necessary to carry out those activities.

The discussion here also highlights the very different histories of the three states; specifically, how Florida and Texas, in stark contrast to North Dakota, had adopted elaborate SAA systems even before NCLB had been adopted. We return to this importance difference again in the discussion of a possible shift of SAA systems to the federal level (Section VII). In the subsequent sections, we discuss other information we gathered to develop more concrete cost estimates.

IV. Conceptualizing and Measuring the SAA System Costs

Economists typically conceptualize costs as relating to a particular output such as automobiles in the private sector or student achievement in education. However, the present cost study is framed in terms of the costs of a particular system—specifically, a system of standards, assessments, and accountability. We therefore discuss below how we have made choices regarding the scope of the systems and therefore about the types of costs to include and exclude.

The research questions of this study involve the costs of “creating, updating and minimally complying with” SAA systems and deliberately exclude the costs of reaching the goals (adequacy). These general principles are helpful in making decisions about what costs to include and exclude, but there is also considerable middle ground that these principles do not clearly address. For example, Burch and Hayes (2007) measure the cost

of quarterly assessment or “benchmark” systems that use standardized testing of student throughout a given school year as a means to identify and address student needs in advance of high-stakes state assessments. On the one hand, one might argue that benchmark systems respond to local demands for academic performance, and the costs of benchmark systems would therefore be incurred whether or not there were state or common standards—a clear case for exclusion from this cost analysis. On the other hand, benchmark systems are clearly a type of standardized assessment. While we acknowledge the ambiguity that this example highlights, we maintain a strict “compliance” definition throughout the main cost estimates provided in this study and therefore exclude the costs of quarterly assessments from our analysis. In order to highlight the potential magnitude of policies such as quarterly assessments, we provide some brief references to cost estimates from other studies in Section VII.

The example of the quarterly assessments systems highlights a more general distinction between costs associated with the *function* of resources versus the *reasons* those resources are being used. Quarterly assessment systems clearly serve the function of assessment, but we have also framed our research questions in terms of the reasons—that is, state and common requirements. We therefore include SAA resources that are required by state and federal governments *and* serve the function of SAA, but we exclude resources that do not meet both of these criteria.

Many states, including Florida and Texas, link their accountability systems to financial incentives for educators (e.g., Florida’s merit pay system, and Texas’ Educator Excellence Grants). It is not obvious whether such costs should be included or not. On the one hand, bonuses paid for increased student performance could be compensating

teachers for increased effort. Increased effort is a real resource cost of accountability. On the other hand, performance bonuses could simply represent a windfall to teachers. If so, then they represent merely “cash transfers” from one group (citizens) to another (teachers) and therefore impose no net cost on society. (This is a good example of the distinction between budgetary and economic costs.) In either case, we are principally interested in the difference in costs between a common SAA system and a system of state-level SAAs, and the “costs” of these financial incentives would be unchanged between these two alternative SAA approaches. Even if these cash transfers are excluded, the costs of administering these programs are fixed and therefore clearly relevant to the present analysis and the second research question.

Expenditures versus Real Resources

The ideal approach to measuring costs is the “ingredients method” described by Levin and McEwan (2001) by which resources are identified under each of several categories (human resources, capital, equipment, etc.). Resources within these categories are then measured in raw units (e.g., the number of hours spent on a particular task). These raw units are then translated into cost figures by identifying the opportunity cost per unit. For example, the hourly rate of a teacher’s time can be measured by her hourly wage rate (plus employer contributes to health and pensions).

In our analysis, we are able to use the ingredients method for some resources, especially those incurred at the school district level, but we rely on budgetary information in many cases where the budgets appear to coincide with the parts of the SAA systems that are of interest. Many budget figures incorporate a variety of activities and these

cases it is generally impossible to disentangle the costs of specific activities. We are careful to discuss possible discrepancies between budgetary and economic costs.

New Resources versus Re-allocation

One of the central issues in the debate about standards-based reform is the concern that assessments cannot possibly measure all of the important outcomes of schools, and that by focusing only on formal assessments, SAA systems lead to a narrowing of the curriculum and a reallocation of resources away from unmeasured, but potentially quite desirable, activities. This issue relates closely to the difference of budgetary versus economic costs and raises a particularly important issue in the measurement of the cost of SAA systems.

If we were to take a “strict budgetary” approach to the problem, re-allocation of resources (e.g., from music and arts to math and reading) would not count as costs because total expenditures would be unaffected. However, this strict budgetary approach is clearly problematic because it implicitly assigns zero value to the activities that are losing resources. At the other logical extreme, we could assume that all resources apparently being directed to SAA systems are new resources and we can value these based on the concept of opportunity cost. For example, if an hour of instructional time is shifted from music to math, and the value of the teacher’s time is \$30 per hour, then we are essentially saying that the loss in music time has a social value of \$30. While this estimate of the social value of music is far from perfect, the alternative (the strict budgetary approach) is even more problematic because it assumes that music instruction

has no social value. Therefore, in most of the analysis below, we avoid the strict budgetary approach and assume that all resources are new resources.

State and National Unit Cost Measures

When using data on expenditures, the figures reflect the specific costs of doing business in the respective states and school districts. Using the state budget figures is reasonable when making estimates for each specific state. However, estimating the costs to the nation under any type of system requires combining data across states. There are significant differences in labor cost across states that need to be addressed before aggregating state-level estimates.

The National Center for Education Statistics' Comparable Wage Index (CWI) indicates that all three of the states under analysis have below average labor costs (Taylor and Fowler 2006). In 2005 the CWI indicates that labor costs in Florida were 7 percent below the national average, labor costs in North Dakota were 20 percent below the national average and labor costs in Texas were 1 percent below the national average.

Generalizing from the cost profiles of these three states requires adjustments for regional cost differences. We use the CWI for those adjustments. Thus, where the Bureau of Labor Statistics indicates that the national average hourly wage for management occupations is \$40.07¹³, we presume that such a person would earn \$37.10 in Florida, \$32.13 in North Dakota and \$39.64 in Texas. Similarly, the average hourly wage for primary, secondary and special education school teachers in the U.S. was

¹³ The national wage estimates are from the Bureau of Labor Statistics National Compensation Survey for 2006. These wage estimates do not include fringe benefits.

\$33.06 in 2006, implying that average hourly teacher wages in Florida, North Dakota and Texas were \$30.61, \$26.51 and \$32.71, respectively.

Generalizing from the Sample to the Nation: The “Typical State”

This study is part of a larger project by the National Research Council (NRC) to understand state standards and the potential of common standards. In order to integrate our study with this larger project, we are studying the same three states that are the focus in the larger project. Florida, North Dakota, and Texas vary in their SAA systems, but they are not necessarily representative of the nation, so even the average costs in these states might not reflect the costs of the average state. Therefore, in the analysis, we use additional research to define a prototypical SAA system and then use the cost information from the three specific states to estimate the costs of this prototype. In our future work on this subject, we will then use the information from the typical state to better understand the cost savings of moving toward national standards.

Data Collection

Much of the information about the basic outlines of the SAA systems came from online document searches of the web sites of the state education agencies. Two other sources of information are being used to create more concrete cost estimates: (a) interviews officials in the state education agencies (SEA) and local education agencies (LEA); and (b) additional document collection, especially budgetary information.

The first initial contact in each state education agency was to the director(s) of assessments, chief academic officer, chief financial officer, and/or assistant

superintendent for policy. In some cases, the researcher was referred to another knowledgeable person within the agency. At least two officials have now been interviewed in each SEA and at least one official has been interviewed in one LEA in each state. Officials were interviewed by phone using the interview protocols shown in the appendix.

One of the main purposes of the initial interviews was to identify knowledgeable individuals and additional documentation about specific elements of the SAA systems—what is sometimes called a “snowball” sample. These follow-up interviews were unstructured because the positions of the individuals varied. The names of all individuals interviewed will remain anonymous in order to protect their identities. In cases where we report information from officials from local school districts, the district names will also remain anonymous.

In some cases, especially the costs at the local level, it was necessary to rely on previous research that, while not focused on costs per se, does provide a sense of how districts have re-allocated resources in response to SAA systems.

V. Analysis: SAA Costs in Florida, North Dakota, and Texas

The first step in the analysis is to measure the SAA costs for the three specific states in the sample. Below, we discuss the costs of each of the three separate elements of the SAA systems—standards, assessment, accountability—each of which is broken down by whether the costs are incurred primarily at the state versus local levels. The focus here is on the first research and on the costs of compliance, which leads directly into specific cost estimates.

We divide the discussion into “state” and “local” cost categories, though it is important emphasize again that we are not primarily interested in which governmental units are bearing the burden of these costs. These two categories therefore primarily reflect the source of the data (state official versus local officials). We also divide costs into sections on standards, assessment, and accountability, though these lines, too, are somewhat blurry. However, dividing the discussion in this way helps us to distinguish between fixed and variable costs, and therefore to isolate the potential cost savings from national standards.

Standards

All three of the states under analysis went to great lengths to include stakeholders in the development of their educational standards. The following brief discussion sketches the process in each state.

Florida. Florida’s Sunshine State Standards were first approved by the State Board of Education in 1996 and have evolved considerably since then. The original standards were written in seven subject areas and divided into four separate grade clusters (PreK-2, 3-5, 6-8, and 9-12). As Florida moved toward greater accountability for student achievement at each grade level, the Sunshine State Standards were further defined with specific K-8 “Grade Level Expectations” added in 1999.

When standards were first developed in Florida in 1996, the state plan was to revise them on a ten-year cycle. In keeping with that rule, the State Board and the Florida Department of Education (FDOE) began the process of revising all of the academic

standards 2006. Each subject as well as students with disabilities and English language learners have standards to meet and instead of dividing by grade level, the standards were distinguished by subject. (The details on the required standards can be found at www.flstandards.org.) According to the DOE (2007), this move went far beyond increasing the rigor of the standards, and included the alignment of the new standards with assessments, instructional materials, professional development, and teacher licensure exams. In 2006, the state also adopted a shorter six-year cycle or subsequent reviews of the state standards.

The process for revising the Sunshine State Standards entails a variety of activities including multiple opportunities for stakeholder input. These activities include meetings with content supervisors, teachers, content specialists, professional organizations, and other stakeholders. According to DOE, “continued stakeholder input is encouraged through both hard copy and a Web-based input system that ensures stakeholder ease in providing meaningful feedback so that the Sunshine State Standards will be world class” (flstandards.org, 2007). The language arts and mathematics standards were revised and approved by the State Board of Education in 2007. As of December 2007, the science standards are ready for review and open to public comment on www.flstandards.org. The social studies standards have just started the review process and framers will meet in 2008.

North Dakota. The NDDPI laid out a development protocol for all standards (tested or otherwise) consisting of three phases: drafting/dissemination, approval/implementation, and professional development/feedback NDDPI (2002). The first phase consists of the

project proposal; the appointment of the director; the selection, contracting, convening, and training of the design team. Once the initial drafts of the standards are written, reviewed, and revised, the standards are then distributed to stakeholder groups for review and comment. The final drafts for each subject area and grade level are made available at the NDDPI website.

In the approval/implementation phase, the State Superintendent approves the final standards document and the standards documents are disseminated electronically to school districts, libraries, universities, and other relevant organizations. The NDDPI provides technical assistance to school districts regarding the use of the standards. The school districts, in turn, can go beyond state standards, either by including additional detail or by adding standards related to content not covered by state standards.

In the final phase, teachers and administrators use the standards document as the basis for standards-based professional development, incorporate the standards into improvement planning; and submit recommendations to the NDDPI as to how future standards can be improved upon. According to state documents, a stated goal is that “continual feedback on the effectiveness of the standards document will be ongoing through regional and statewide education conferences, graduate programs, and long-term local curriculum mapping and development initiatives” (NDDPI, 2002, p. 8).

NDDPI initially developed reading/language arts and mathematics state content and achievement standards. Content and standards have also been developed for the various other subjects, but for purposes of accountability, only reading/language arts, math, and science are currently tested.

Texas. Texas has followed a two part process for setting standards. The first part of the process determined the content of the required curriculum. The second part, which was conducted some five years after the first, determined the performance standards by which that content would be assessed.

State law obligates the Texas Education Agency to develop a required curriculum in foundation and enrichment subject areas. The foundation subject areas are English reading and language arts, mathematics, science, social studies, Spanish language arts, and English as a second language. The enrichment subject areas are languages other than English, fine arts, health, physical education, technology applications, and career and technology education. The required curriculum is known as the Texas Essential Knowledge and Skills (TEKS). School districts are required to provide instruction in both the foundation and the enrichment TEKS, but only the foundation TEKS are part of the accountability system.

The process for developing the TEKS was very labor intensive. For example, consider the social studies TEKS. The social studies TEKS were developed by a 35-member writing team composed teachers, campus administrators, college and university professors, members of business and industry, and parents. Twenty-nine members of the writing team were educators with significant expertise in the subject area.

The team members developed two drafts of the social studies TEKS that were circulated for public comment. The State Board of Education designated a 15-member committee to provide additional comments. Four national experts were also asked to

review the near-final draft.¹⁴ After revisions in response to the array of comments, the social studies TAKS was formally adopted by the State Board of Education in 1997.

Similar processes were followed for the other TEKS.

The Texas Assessment of Knowledge and Skills (TAKS) testing system was developed to measure student comprehension of the TEKS. To set the passing standards on 36 separate TAKS tests (26 in English and 10 in Spanish), the State Board of Education and the Texas Education Agency designated three types of advisory panels. The first was a National Technical Advisory Committee comprised of thirteen nationally recognized experts.¹⁵ The second was a standard setting advisory panel comprised of 19 Texas educators and representatives from prominent Texas advocacy groups (e.g. the National Association for the Advancement of Colored People, The Texas Association of School Boards, The Texas Association for Bilingual Education, The Texas State Teachers Association, and The Texas Business and Education Coalition). This panel provided initial guidance and oversight and review for the third type of advisory panels—the “regular” standard setting panels. There were 21 regular standard setting panels, each comprised of 15 to 22 Texas educators and other stakeholders. By design, most of the participants in the regular standard setting panels were teachers. The regular panels met for two or three days to develop performance standards for the designated tests. Each regular panel was responsible for setting standards for one or two of the 36 TAKS tests. All told, nearly 300 individuals participated in the regular panels.

¹⁴ The four experts were T. R. Fehrenbach, author of *Lone Star, A History of Texas and the Texans*; John Fonte, a Fellow at the Alexis de Tocqueville Institution; John J. Patrick, Professor of Education at Indiana University; and Diane Ravitch, Senior Research Scholar at New York University.

¹⁵ For names and biographies, visit <http://www.tea.state.tx.us/student.assessment/taks/standards/index.html>.

State Level Cost of Standards. The costs of developing and updating standards depend principally on three main factors: the number and specificity of the standards, the opportunity cost of those participating at each stage, and the frequency of updating. Therefore, in order to estimate the state cost of standards, we use the above information on the standard-setting and updating in each of the three sample states to calculate the cost of developing a single educational standard for a grade and subject (e.g., third grade math).

In Texas, a 35-member team (29 teachers and six other professionals) developed the social studies curriculum standards for grades K-12. The writing team circulated two drafts for comments before submitting their final draft to the State Board of Education for approval. Assuming that the writing team met twice, for two days each time, then writing the curriculum standards required 116 days teacher labor and 24 days labor from other professionals. The State Board of Education designated a 15-member review committee. Assuming that their review took two days, adds another 30 days of professional labor. Finally, the state hired four national experts to provide comments. Those individuals add another eight days to the total. Thus, developing the TEKS for social studies cost approximately 116 days of teacher labor, 62 days of professional labor, and travel expenses.¹⁶ In addition, the Texas Education Agency incurred meeting costs associated with coordinating the various standard-setting groups. We presume that the opportunity cost of teacher labor is the prevailing wage for teachers (\$32.71) while the opportunity cost of professional labor is the prevailing wage for managerial occupations (\$39.64). We use the 2007 IRS per diem rate for Austin, Texas (\$139 per day) as our best estimate

¹⁶ We do not attempt to estimate the opportunity cost of individuals who provided public comments on the TEKS. The national experts did not travel to Austin to provide feedback, and therefore did not incur travel expenses.

of food and lodging costs for the standard-setting groups, and assume that travel costs averaged \$200 per person.¹⁷ At current wage rates, therefore, developing the TEKS curriculum standards for social studies cost \$90,650, plus the costs to the agency. TEKS social studies standards were set for grades K-8 and for ten high school courses. Assuming that the cost of setting each of those nineteen standards was equivalent, that works out to roughly \$4,771 per subject-grade standard.

Another set of stakeholder panels developed performance standards for the TAKS test. The National Technical Advisory Committee's 13 members were paid a total of \$75,000 for their contribution to setting the 36 TAKS standards, or \$2,000 per standard (preliminary figures). The Standard Setting Advisory Panel met twice for one day each time, first to provide guidance to the regular panels and then to review the panel recommendations. Although this group was not compensated for their time, the opportunity cost of time for the 19 panel members is clearly a cost of standard setting. The final cost of standard setting was the opportunity cost of time for the 21 regular panels. Fifteen panels met for three days, while six panels met for two days. On average, there were 20 members in each panel. Therefore, the regular panel meetings cost 1,140 days labor. In addition, 50 members of the regular panels served an additional day to review all of the standards for their respective content areas (math, English language arts, science and social studies). All told, the regular and advisory panels devoted 1,228 days to setting performance standards for the 36 TAKS tests. At current wage rates, 1,228 days is equivalent to \$356,452. Travel expenses would add another \$258,492. Thus, the cost of setting the performance standards for the 36 TAKS tests was approximately

¹⁷ At the IRS allowable rate of \$0.485 per mile, this assumption presumes the average panel member drove 206 miles. The average Texas teacher lives 183 miles from Austin, and nearly 75 percent of Texas teachers live within 206 miles of Austin.

\$17,000 per subject-grade standard. We estimate that the cost of setting curriculum and performance standards in Texas was just under \$22,000 per subject-grade standard.

In Florida, there are standards in 12 subjects: mathematics, reading, writing, science, special education, limited English proficiency, social studies, physical education, health, workforce education, voluntary pre-kindergarten, and fine arts all have standards. (As indicated earlier, not all of these standards are associated with assessments.) The initial standards for all subjects and grades were developed by 30 “framers” who were experts in their fields and 33 “writers” (mostly senior teachers from around the state) who translated these general guidelines into concrete descriptions. The standards are updated every six years. To develop the Florida math and science standards, separate groups of writers were created to compose the standards for K-8 and the specific bodies of knowledge covered in grades 9-12. These writers met before finalizing the standards to ensure alignment between the K-8 and 9-12 standards. In order to develop standards for special education and Limited English Proficient (LEP) students, the writers and framers included LEP and special education experts in their meetings; however, the primary role in development was that of the math and science experts. There were no stipends to the framers and writers, so we use the opportunity wage for teachers to estimate the costs. As with our analysis of Texas, we use the IRS per diem for Tallahassee (\$133), and assume that travel costs average \$200 per person per trip.

We obtained the most detailed documentation regarding the creation [adoption?] of the state’s science standards. This included nine days of meetings for framers and 18 days of meetings for the writers. We use the average hourly wage of people with master’s degrees in Florida for the framers (\$32.13) and the average hourly wage for

teachers for the writers (\$30.61). Nine days of meetings (72 hours) of meetings for each of 30 framers yields time costs of \$69,400. Eighteen days of meetings (144 hours) for each of 33 writers yields additional time costs of \$145,459, for a total time cost of \$214,859. Travel costs for 27 meeting days for each of 63 participants yields \$226,233. There were apparently three trips for the framers and five for the writers, yielding a total of travel costs of \$51,000. The total time and travel costs for the framers add up to \$492,092. These costs pertain to a single subject and 13 grades, thus the cost per standard is \$37,853.

Other activities included “identification of research, experts, framers, and writers,” “meeting with education foundation members,” “preparation of draft document review,” “online public review,” “meeting with state science supervisors,” “expert review,” “public hearings” (four), and “revision of draft.” There is little information to cost out these items.

The costs of setting standards in Florida appear to be considerably lower than those of updating. The reason is that the initial standard-setting was done somewhat hasty and relied heavily on national standards, whereas the updating process has been more deliberative. Therefore, while the costs of creating standards would generally seem to be more expensive than the updates, the reverse appears to be true in this particular case.

North Dakota used a somewhat similar approach, including the English/language arts standards team (26 teachers), math standards team (28 teachers), science standards team (33 teachers), two project consultants, one program evaluator, and one NDDPI program coordinator. Each team participated in 5-8 sessions for a total of 2-3 days of

meeting. Each teacher was paid a \$200 stipend per day, plus lodging and travel (assumed to be \$300 per teacher).¹⁸ We use the stipend as the estimate of the opportunity costs in this case. A total meeting time of 2.5 days translates into \$800 per person (combining opportunity and travel costs) for each of 29 teachers (the average of the three subjects), yields \$23,200. In addition, apparently because of the state's small size, they also contracted with one of the federal regional education laboratories, McREL (\$125,000 per year) to provide technical assistance in the process of drafting and updating the standards. There is no concrete information available about the time of the two consultants, program evaluator, and program coordinator. We assume 0.5 annual FTE per person across all the standards, yielding 2.0 FTE total for these four people. Using the North Dakota salary for managerial occupations (\$32.13 per hour) this yields \$128,500.

In math and reading, tests are administered in grades 3-8 and 11 in North Dakota, for a total of seven grades. We assume that standards were created only for these grades. Thus, the cost of teacher time for each of these individual subjects (\$23,200) comes to \$3,314 per grade-subject standard. (We assume this same rate for science per standard for science, though this is somewhat unclear because of the lower number of grades that are being tested.)

The costs of the senior personnel and McCREL contract are divided evenly over the standards. Assuming that standards are being created in all subjects (not just the tested grades and subjects 4, 8, 11), this yields per standard costs of \$12,071 ($\$128,500 + \$125,000 = \$253,500$ divided by the assumed 21 grade-subject standards). Combining this with the teacher costs (\$3,314) yields \$15,385 per subject-grade standard.

¹⁸ The value of the teacher time can be valued either by the stipend or an estimated hourly wage. The stipend itself is only a budgetary cost. The cost of substitute teacher's time would also count as a budgetary cost, but not as an economic cost.

Local Costs of Standards. The costs of SAA systems at the local level are somewhat more difficult to define. District officials in at least one state (Florida) indicated that they employ staff two full-time devoted to providing professional development to teachers to help them understand the standards. Use the state wage for management occupations to estimate the direct costs of these staff, as \$74,200 (as outlined in Section III). Two full-time staff yields \$148,400. Given the approximate number of students in the school district, this yields a cost per pupil of roughly \$1.¹⁹ The cost of teacher time participating in this professional development is also relevant. We assume that each teacher participates in one day of standards-related professional development during each year in which standards are created or changed. Based on the average hourly wage of \$30.61, this yields roughly \$15.31 per student per year.

Assessments

State Level Costs. All three states under analysis rely on outside vendors for test development and scoring. Over the last five years, Florida has contracted with CTB/McGraw-Hill, Harcourt Educational Measurement, and NCS Pearson. In contrast, North Dakota and Texas have contracted with the same testing firms for decades. North Dakota has relied on CTB/McGraw-Hill since approximately 1990, while Texas has contracted with NCS Pearson, Inc. since 1981.

Because test development and scoring are outsourced, it is relatively straightforward to trace state outlays for this activity. For example, the state of Texas awarded assessment contracts totaling \$94 million in 2007. This is equivalent to \$20.46

¹⁹ This figure is approximated in order to maintain the district's anonymity.

per pupil. In Florida, these costs have been estimated at \$15.10 per pupil in YEAR (FDOE, 2007). The state of North Dakota's contract with CTB/McGraw-Hill, according to one official in the North Dakota SEA, is \$3.2 million per year—approximately \$33.47 per pupil.

Contract outlays are not the only resource cost for test development, however. The states also rely on a variety of commissions and advisory panels to refine their testing instruments. For example, during 2005-06, the Assessment Division of the Texas Education Agency convened more than 20 educator committee meetings attended by a total of 2,216 educators to “review all newly developed test items and all new field test data” for the states various standardized tests (TEA 2006). The opportunity cost of educator time for those reviews is approximately \$580,000. Similarly, North Dakota conducts an annual review of the potential biases in the state exams (e.g., racial bias). This involves meetings of 75-100 teachers, administrators, and community members (7-8 hours total over 5 days). As with the state's standard-setting process, participants were paid \$200 dollar stipend plus room and board. We estimate that these meetings cost \$67,000 per year. We also note that these costs, while recurring, are not dependent on the number of students in the state, and are therefore for our purposes, fixed costs.

The state agencies maintain permanent staff whose primary responsibility is the state assessment system. Budgetary outlays to support those personnel are also costs of assessment. In Florida, the cost for administration, reporting, and scoring of the FCAT, beyond the cost of contracts with testing companies, in 2007 was \$26,547,890 (FDOE, 2007), or \$10.21 per student. In Texas, the Assessment Division of the Texas Education Agency has a budget of \$2.38 per student. Excluding external contracts, the annual

budget for the NDDPI Division of Standards and Achievement in 2007 was \$975,000, or \$1.02 per student.

Local Costs. School districts are not directly involved in creating and updating SAA systems, but they are involved with compliance. The most obvious costs in this regard include the time that school and district staff in following the administrative rules associated with testing (e.g., they must ensure that students and teachers cannot see the exams in advance) and the time teachers spend administering the exams.

In North Dakota, students spend 3-4 days taking assessments. While the cost of student time is not considered relevant, the cost of teacher time is. Using 3.5 days as the relevant estimate, this yields 8 hours per teacher; at \$26.51 per teacher, this yields \$13.26 per student.

Texas administered 8.8 million standardized tests in 2007 (Texas Education Agency 2007). Assuming that each exam took half a day, then the exams cost 4.4 million student days. The average class size is 20 students, so 4.4 million student days is equivalent to 220,000 teacher days. At current wages, that works out to \$12.60 per student.

Florida administers both CRT and NRT tests and in a larger number of grades. With 2.6 million students and testing in grades 3-10 in reading and math, we estimate that 1.6 million tests were administered in reading and math (respectively) and 0.6 million in science and writing (respectively) with each of the two tests (NRT and CRT). This yields a total of 8.8 million tests. (The fact that this number is identical to Texas is a coincidence. The smaller number of students in Florida is offset by the larger number of

tests.) Assuming the same time per test as in Texas, this yields the same number of teacher days (220,000) but, because of this is spread over many fewer students, the cost per student ends up higher in Florida (\$20.72).

It is important to emphasize that the apparent costs of these systems may be replacing other forms of locally created assessments. At the extreme, we can imagine a school system in which there are no standardized assessments and each teacher creates assessment systems individually. This approach is almost certainly more costly than benchmark systems, at least in terms of opportunity cost. To get a rough estimate, consider that there are roughly two million teachers in the country and having each of them make their own assessments costs would clearly require more resources than, for example, each of the approximately 25,000 school district or 50 states.

Costs of Accountability

Below, we discuss the costs of accountability, which come principally from the supports provided to low-performing school and the rewards provided for high-performing schools. To clearly understand these costs, it is important to re-emphasize the important distinction between expenditures and real resources and between compliance and adequacy costs. Helping low-performing schools would appear to be in the adequacy category. However, some of these costs are “built in” to the accountability system, e.g., when schools deemed low-performing in the accountability system are given extra resources to improve. In these cases, where the supports are woven into the accountability system, it is reasonable to think of these as compliance costs.

By this reasoning, financial awards to high-performing teachers and schools are also built into accountability and are considered costs of accountability, but only in a budgetary sense. Recall that such incentive payments are largely just cash-transfers and, regardless of their intentions, almost certainly and dramatically over-state the value of new teacher and school resources (e.g., greater effort). Thus, we treat these as expenditures, but not than real resources.

State. North Dakota's Department of Standards and Assessment has a contract of \$125,000 per year with Data Driven Enterprises contract (\$1.31 per pupil) for data management and analysis (e.g., determinations of federal adequate yearly progress (AYP)). The state does not have performance awards for successful schools and teachers. The Texas Education Agency's annual budget for Accountability and Data Quality 2008 was \$1.60 per pupil.

Local. There are few costs of compliance with accountability at the local level, beyond those discussed above with respect to standards and assessments. School districts do of course incur costs in trying to meet the objectives of SAA systems, but these fall into the category of adequacy, which we have excluded.

Summary of Costs

Table 4 provides a summary of the above cost analysis. Most of the costs are reported as costs per student to provide some sense of the magnitude of the costs in relation to the total education expenditures. We continue to report the cost of developing a standard as a cost per standard because these costs are mostly fixed, converting them to

a per-student estimate could be misleading. Also, the cost of developing and/or updating a standard occurs only sporadically, whereas the other costs in the table arise annually.

Excluding the costs of standards, the reported preliminary estimates range from \$35 in Texas to \$61 in Florida, with North Dakota falling in between at \$48 per student. Using the 2005 figures reported in Section II, this yields cost as a percentage of total spending of 0.49, 0.61, and 0.85 in Texas, North Dakota, and Florida, respectively. These costs are somewhat less than one percent of total expenditures in all three states and are all larger than those reported in the GAO and Hoxby reports discussed in Section II. There are two main reasons: (a) the scope of costs considered here is broader; and (b) this is the first study to use data post-NCLB when these costs undoubtedly increased.

*Table 4: Summary of Preliminary Cost Estimates
(real resource costs, lower bound estimates)*

	Florida	North Dakota	Texas
Standards			
State	\$37,853/standard	\$15,385/standard	\$21,853/standard
Local	\$15.31/student	n.a.	n.a.
Assessments			
State – contracts	\$15.10/student	\$33.47/student	\$20.46/student
State – admin.	\$10.21/student	\$ 1.02/student	\$ 2.38/student
Local	\$20.72/student	\$13.26/student	\$12.60/student
Accountability			
State	n.a.	\$ 1.31/student	\$ 1.60/student
Local	\$ 0	\$ 0	\$ 0

It is important to re-emphasize some of the important caveats and clarifications discussed earlier. First, these estimates are preliminary and, because we have yet to estimate some costs, represent a lower bound on the actual total costs of the systems. Note that no cost estimates are available in four of the cells in Table 4 (indicated by

“n.a.”). Also, even in some cells where estimates are available, these do not reflect all of the cost sub-components that belong there. One specific excluded cost category relates to state and local data systems. Second, while we rely on expenditure information in many cases, we use these data to estimate “real resources.” Third, the distinctions between state and local costs, and between standards, assessment, and accountability, are less a focus here than the total costs of the whole systems. We have used these distinctions above only to organize the discussion.

Finally, these estimates alone do not of course answer either of the research questions of interest. They do, however, provide a basis for doing so, as we show in the subsequent sections.

VI. Analysis: Defining the Typical SAA System and Current System Costs

The first research in this study is, “What costs are now being incurred by the nation to create, update, and minimally comply with standards, assessments, and accountability under current state and federal laws and rules?” The sections above move toward an answer by providing cost estimates for the three sample states. In this section, we identify the SAA system of the typical state and use this to estimate the current costs to the nation as a whole.

NCLB, Standards, and Assessments

State standards and assessments have become increasingly similar as NCLB requires states to have standards and administer standards-based assessments in most

grades and basic academic subjects.²⁰ The law requires annual assessments in reading and math for grades 3-8 and in at least one high school grade (grade 10-12). Beginning in 2007-2008, science assessments will be required in at least three grades (one each within grades 3-5, 6-9, and 10-12). There are no requirements for tests in social studies and writing, though many states do have them.

NCLB distinguishes between “academic content standards,” which refers to what students are expected to know in each subject, by grade and “academic standards,” which refer to the level of achievement on the assessment that constitutes proficiency. States are given considerable latitude over both types of standards. They must create content standards in reading, math, and science, but the specific content is left to their discretion. Likewise, they must define proficiency and other levels of student performance, but federal guidelines are vague. The law stipulates only that academic standards should be “challenging,” “specify what children are expected to know and be able to do,” “contain coherent and rigorous content,” and be “aligned with the State’s academic content standards” (Section 1111 Part D).

For limited English proficiency (LEP) students, the law requires reading assessments using tests written in English for any student who has attended school in the US for three or more consecutive years. (School districts may, in addition, use tests in the student’s native language for up to two additional years.) In addition, states must annually assess English proficiency for all LEP students.

For special education students, states are allowed to: (a) administer regular assessments to students with accommodations such as extra time; (b) develop “alternate

²⁰ Unless otherwise specified, the information on NCLB in this section comes from the Elementary and Secondary Education Act of 2001, 115 STAT. 1445.

academic achievement standards” and alter state assessments (intended for student with the most severe cognitive disabilities; no more than one percent of students in each district and state can take the alternative assessment); or (c) develop and administer “modified” assessments that are based on regular standards, but which set a lower cut score in the definition of proficiency.

NCLB and Accountability

In addition to the requirements for state standards and assessments, NCLB requires states, districts and schools to intervene in schools that are “in need of improvement.” On the one hand, the costs of these interventions is irrelevant because their application depends on whether students have achieved specified levels of academic achievement—thus, it focuses on the adequacy of resources and programs. On the other hand, there is good evidence that nearly all schools and districts will eventually be “in need of improvement” unless the laws are changed. This is due to the law’s ambitious requirement that all students be proficient by the year 2014.

A plausible starting point for estimating the costs of these interventions is the Title I budget that each district receives from the federal government based on the characteristics of its student population. These funds, however, must be used for instruction rather than the administrative costs that are of interest here and these funds are not directly tied to the number of schools in need of improvement. Moreover, school districts could supplement Title I funds with funds from their general fund. Therefore, Title I budgets, either at the federal or district levels, are poor indicators of the costs of accountability.

Another important part of NCLB is the provision for Supplemental Education Services (SES) in which district Title I funds are given to private providers of tutoring services. Because this represents a shift in resources from districts to private entities, we these are changes in government expenditures rather than changes in real resources. This is a good example of the distinction between expenditures versus real resources, and between new resources and re-allocation, as discussed in Section IV.

Defining the Typical State

We define the typical state by using NCLB as a “minimum bar” and through other sources that describe the SAA systems being used around the country that exceed this bar. Two of the sample states—Florida and Texas—are well known for having gone well beyond NCLB requirements, which aids in identifying the costs of these systems.

The primary ways in which states use additional resources in their accountability systems is by providing funding to support improvement in low-performing schools and by providing financial rewards to teachers, schools, and districts that perform well. Some examples of this were described earlier in the discussion of the sample states. Teacher merit pay is extremely uncommon and there are reasons to expect that even those systems now in place will share the fate of those from previous decades, being dropped due to a combination of constrained state budgets, changes in political leadership, and lack of evidence of success (Murnane and Cohen, 1986). We therefore do not include teacher merit pay as part of the typical state SAA system.

Support and incentive funds for schools and districts are much more common. Goertz and Duffy (2001) mention six states (Kentucky, Maryland, New Mexico, New

York, North Carolina, Ohio, Pennsylvania) that had financial awards to successful schools as of 2000. In addition, it is clear from the discussion of the samples states that Florida has added such systems since then, while Texas had such a system both before and after 2000. Given the NCLB focus on low-performing schools, we suspect that rewards for successful schools will become even more widespread as states seek to highlight and reward success stories, rather than only punishing failures. The use of funds for supporting low-performing schools is now required by NCLB because states are required to intervene in schools that are labeled as “in need of improvement.”

Based on this discussion it is reasonable to define the typical state simply in terms of the requirements of NCLB. An increasing number of states are reaching its requirements and few states go substantially beyond its requirements. Moreover, to the degree that some states, such as Florida and Texas, go beyond the law’s requirements, they do so in ways that are reflected only in expenditure costs, not the real resources that are of greatest interest here.

Using the Sample States to Measure the Costs of the Typical State SAA System

Using the above definition, North Dakota is arguably the most typical state of the three considered in the above analysis. Unfortunately, North Dakota is also distinctive in its small size and probably large fixed costs. Perhaps the best evidence of this comes from comparing the discussion of the state SAA systems in Section III with the cost estimates of Table 4. Of the three states North Dakota has the least aggressive SAA system, but yet has relatively high costs and—by far—the smallest number of students. For this reason, as well as the preliminary nature of the above analysis, we cannot simply

In subsequent versions of this report, we will attempt to isolate the fixed costs.

Separating the fixed and variable costs of assessment systems is difficult because most of these costs are incurred through testing companies and most of the fixed and variable costs are combined in the state contract figures. To get some sense of the fixed costs, consider that the main fixed costs are those of developing and maintaining tests that are aligned with the standards in a particular state, including the development of test items, creation of multiple formats, validity and reliability testing, and test scaling. The main variable costs, in contrast, are the costs of printing and disseminating test booklets and scoring and reporting tests for each student.

Evidence from Hoxby (2002) helps to shed light on the specific proportion of these costs that are fixed versus variable. If costs were mainly fixed, then we would expect a relatively clear pattern of similar costs across states, whereas if they are mainly variable we would expect the costs to vary by the size of the student population. Unfortunately, in conducting this analysis, it would be necessary to control for the number of tests per year that each state administers and other factors, which is not presently available. In subsequent analysis, we will conduct additional analysis to more clearly separate the fixed and variable costs of assessment.

Again, the cost savings from moving to a common SAA system would be 1/50 of the total fixed costs.

VIII. Discussion: Just Beyond Compliance

Because the costs in the previous sections have been limited to the costs of compliance, we have excluded some costs that are directly related to SAA systems and

that many states and school district are choosing to incur. Below we discuss two general categories of costs that are excluded from the compliance definition: (a) costs that are arguably necessary, but not required; and (b) costs, such as maintaining alignment between standards and other education policies, that are arguably not new.²¹

Necessary, but non-required, costs. Some responses to SAA systems are arguably necessary, even if they are not required. Quarterly assessment or “benchmark” systems represent a useful example. These systems involve periodic testing during the school year, using assessments that are similar in content and structure to the high-stakes state assessments. In many cases, they are provided by the same testing companies that produce the statewide assessments discussed above. These quarterly assessment systems are not required by state or federal governments and therefore do not fit the compliance definition that is the focus above, but these are forms of assessments and therefore are part of the three-pronged standards-assessments-accountability systems that are the topic of this study. Further, there is good reason to think they are being adopted in direct response to state and federal SAA requirements.

Burch and Hayes (2007) study the percentage of large districts using these systems, the reasons districts adopt these systems, the timing of adoption, and, most importantly here, the costs of the systems. Of the initial sample of the 30 largest school districts in the country, they found that 23 used these systems. At least 70 percent of the districts had purchased the systems since the passage of NCLB, suggesting a strong influence from federal accountability. It also seems likely, though there is no data to test, that the states that had adopted the systems in advance of NCLB had done so in response

²¹ We thank Margaret Goertz for pointing out this category.

to state accountability systems. Many of the largest districts in the country are in Florida and Texas which, as indicated above, have a long history of high-stakes testing.

Of those districts that had adopted the systems, the average district spent \$5-10 per student on them.²² Because their sample relates only to large districts, it is unclear how common the usage of these systems is around the country or whether the per-student costs would differ in smaller school districts. Again, we do not include the costs of quarterly assessment systems as part of our overall cost calculations, but this example does serve to highlight some of the costs that we are excluding, which are at least indirectly attributable to state and federal SAA systems.

Apparently new resources: The case of policy alignment. State governments invest considerable resources in teacher preparation and certification. However, it is not clear that the costs of aligning these with SAA systems should be considered a cost of SAA. All states have accreditation processes for teacher education programs and all states have some form of certification, all of which are based on some set of standards, implicit or otherwise. In this sense, the SAA system only provides a clearer basis for the other policies. The same is true with regard to aligning the curriculum. States incur costs in recommending and approving textbooks and school districts, in turn, incur costs in keeping their textbooks inline with SAA systems. In a completely decentralized systems, in which each teacher could choose the content to be taught, there would still be a curriculum and textbooks are still a necessary part of the instructional process. One could even argue that the costs of textbooks are lower because the range of the curriculum is narrowed and therefore a smaller number of textbooks can supply the

²² Specifically, 30.4 percent of the districts report spending \$5-10 per students, 17.4 report spending \$0-5 per student, and 13.0 percent report spending \$10-15 per student (the remaining districts either reported inconsistent answers or did not provide the information).

market. Having to develop, update, and print 100 different textbooks is more expensive than 20 different textbooks.

In another sense, however, some of these costs of alignment might be considered important costs of SAA systems. SAA systems are subject to political winds, including changes in governors, state superintendents, and state legislatures. The current movement towards standards-based reform has certainly created regular changes in SAA systems, such as those described in Section II with regard to the SAA system histories in Florida, North Dakota and Texas. To the degree that SAA systems change more often than teachers' own views change, the costs of alignment are very real.

IX. Conclusion

The costs of standards, assessments, and accountability systems discussed in this draft are preliminary. Additional data collection and analysis will be conducted to quantify costs that are presently only discussed and to more precisely measure the costs for which preliminary estimates are already available.

While we do not report a total cost figure in this preliminary draft, we do argue that the costs estimated in what appear to be the only previous studies on the subject (GAO 1993, Education Commission of the States 2001, Hoxby 2002), under-state the true costs by most reasonable definitions. The Hoxby study excludes costs incurred at the local level and under-states some of the state costs. The GAO study describes a period when SAA systems were modest or non-existent in most states. Nevertheless, it appears likely that the total costs of SAA systems remain a relatively small percentage of total educational resources.

This preliminary draft outlines an approach to answering the second research question pertaining to the costs of switching from state to national SAA systems. Once the current costs of the SAA systems have been established, the authors will isolate the portion attributable to fixed costs, which represent the potential savings in shifting to a federal system.

Again, the fact that such cost savings exist does not mean that common standards represent a wise policy direction. Indeed, the cost savings are likely to be quite small relative to the overall spending on education and the main issue remains how these costs affect the nature and quality of the education system.

References

- Baker, B.D., Taylor, L.L. and Vedlitz, A. (2005). Measuring educational adequacy in public schools,” Bush School Working Paper #580.
- Betts, J. R., and Costrell, R. M. (2001). Incentives and equity under standards-based reform, in *Brookings Papers on Education Policy*, Diane Ravitch, ed.(Washington DC, Brookings Institution.
- Bishop, John, (2006). Drinking from the Fountain of Knowledge: Student Incentive to Study and Learn – Externalities, Information Problems and Peer Pressure. *Handbook of the Economics of Education*, Volume 2, 2006, Pages 909-944.
- Burch, P. and Hayes, T. (2007). “Accountability for Sale: The K-12 Testing Industry, District Contracting, and NCLB,” manuscript.
- Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4), 305-331.
- Costrell, R. M. (1997). Can centralized educational standards raise welfare? *Journal of Public Economics*, 65(3), 271-293.
- Costrell, R. M. (1994). A simple model of educational standards. *American Economic Review*, 84(4), 956-971.
- CTB/McGraw-Hill. (2004, May 6). CTB/McGraw-Hill expands assessment relationship with North Dakota; Wins contract to provide reading, language arts, mathematics, and science assessments in 2004-2007. PR Newswire.
- Education Commission of the States. (2001). A closer look: State policy trends in three key areas of the Bush education plan--testing, accountability and school choice. Special Report No. GP-01-02), Denver, CO: author.
- Florida Department of Education (2007). Briefing Book, DOE Division of Accountability, research, and Measurement. Retrieved December 20, 2007 from <http://fcats.fldoe.org/pdf/BriefingBook07web.pdf>.
- Goertz, M. E., & Duffy, M. C. (2001). Assessment and accountability systems in 50 states: 1999-2000. CPRE Research Report No. RR-046. Philadelphia, PA: Consortium for Policy Research in Education.
- Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2), 297-327.

- Harris, D. (2007). "Educational Outcomes of Disadvantaged Students: From Desegregation to Accountability," In *AEFA Handbook of Research in Education Finance and Policy*, (eds.) Helen Ladd and Edward Fiske. Hillsdale, NJ: Laurence Erlbaum.
- Harris, D.N. & Herrington, C.D. (2006). "Accountability, standards, and the growing achievement gap: Lessons from the past half-century," *American Journal of Education*, 112(2), 209-238.
- Harris, D., Herrington, C., & Albee, A. (2007). "The future of vouchers: Lessons from the adoption, design, and court challenges of Florida's three voucher programs," *Educational Policy*, 21(1), 215-244.
- Herrington, C.D. & MacDonald, V.M. (2001). Accountability as a school reform strategy: A 30-year perspective on Florida. In C.D. Herrington & K. Kasten, *Florida 2001: Educational Policy Alternatives* (pp. 7-34). Jacksonville: Florida Institute of Education, University of North Florida.
- Hoxby, C. (2002) The Cost of Accountability. In *School Accountability*, Williamson M. Evers and Herbert J. Walberg, editors. Stanford, CA: Hoover Institution Press.
- Levin, H. & P. McEwan (2001). *Cost-Effectiveness Analysis*, 2nd Edition. London: Sage Publications.
- Murnane, R.J., & Cohen, D. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review*, 56, 1-17.
- NDCI. (n.d.) The North Dakota Curriculum Initiative: About us. Retrieved December 17, 2007 from http://ndcurriculuminitiative.org/index.php/about_us
- NDDPI. (2002). North Dakota standards and assessment development protocols. Retrieved December 17, 2007 from <http://www.dpi.state.nd.us/standard/protocols.pdf>
- NDDPI. (2006). Understanding student achievement within the North Dakota State Assessment: A primer. Retrieved December 7, 2007 from <http://www.dpi.state.nd.us/testing/assess/understand0406.pdf>
- NDDPI (2007). *Title I school/district program improvement: Guidance and required documentation on corrective action and alternative governance options for schools and districts as outlined by the No Child Left Behind Act*. Retrieved December 20, 2007 from <http://www.dpi.state.nd.us/title1/progress/CorrAction.pdf>
- Phelps, R. P. Estimating the cost of standardized student testing in the United States. *Journal of Education Finance* v. 25 no. 3 (Winter 2000) p. 343-80.

- St. John, E., Hill, J., and Johnson, F. (2007). An Historical Overview of Revenues and Expenditures for Public Elementary and Secondary Education, by State: Fiscal Years 1990-2002 (NCES 2007-317). U.S. Department of Education: Washington, DC: National Center for Education Statistics.
- Taylor, L.L. (2005). Measuring educational adequacy in Kansas, *Kansas Policy Review*, 27(2), 31-5.
- Taylor, L.L. (2002). A Dose of Market Discipline: The New Education Initiatives, *Southwest Economy*, Issue 3, p. 1-12.
- Taylor, L. L. & William J. Fowler Jr. 2006. "A comparable wage approach to geographic cost adjustment." Washington, DC: U.S. Department of Education, National Center for Education Statistics.
- Texas Education Agency. 2007a. 2007 Comprehensive Annual Report on Texas Public Schools, Austin TX: author.
- Texas Education Agency. 2007b. Legislative Appropriations Request for Fiscal years 2008 and 2009, Austin TX: author.
- Texas Education Agency. 2006. 2006 Comprehensive Annual Report on Texas Public Schools, Austin TX: author.
- Texas Education Agency. 2005. 2005 Comprehensive Annual Report on Texas Public Schools, Austin TX: author.
- Texas Education Agency. 2004. 2004 Comprehensive Annual Report on Texas Public Schools, Austin TX: author.
- Texas Education Agency. 2002. 2002 Comprehensive Annual Report on Texas Public Schools, Austin TX: author.
- Texas Education Agency. 2001. 2001 Comprehensive Annual Report on Texas Public Schools, Austin TX: author.
- Texas Education Agency. 2000. 2000 Comprehensive Biennial Report on Texas Public Schools, Austin TX: author.
- Texas Education Agency. 1999. 1999 Interim Report on Texas Public Schools, Austin TX: author.
- Texas Education Agency. 1998. 1998 Comprehensive Biennial Report on Texas Public Schools, Austin TX: author.

Texas Education Agency. 1997. 1997 Interim Report on Texas Public Schools, Austin TX: author.

Texas Education Agency. 1996. 1996 Comprehensive Biennial Report on Texas Public Schools, Austin TX: author.

Texas Education Agency, A Standard-setting Plan for the State Board of Education. January 2002.

U.S. General Accounting Office. 1993. Student Testing: Current Extent and Expenditures, With Cost Estimates for a National Examination. PEMD 93-8, Washington, D.C.: author.

West, M. 2007. Testing, learning, and teaching: the effects of test-based accountability on student achievement and instructional time in core academic subjects. In *Beyond the Basics: Achieving a Liberal Education for All Children*, Finn, C.E., Jr, and Ravitch, D. editors. Thomas B. Fordham Institute.

Appendix A

State Education Agency Interview Protocol

Thank you for agreeing to be interviewed. As I indicated by email, this is part of a larger project by the National Academy of Sciences and National Research Council about the advantages and disadvantages of states systems of standards, assessments, and accountability. The purpose of this interview is to gather information about the costs and resources that the [STATE] incurs with its systems.

I want to emphasize that this project is independent of the federal government and unrelated to compliance with any state or federal mandate. Your identities will remain confidential and all documents pertaining to this interview will refer only to “officials of [RELEVANT AGENCY]” in [STATE].

Background

(1a) Are you now, or have you recently been, actively involved state policy design and implementation related to academic standards, assessments, and accountability?

(1b) [If yes] What is/was your role and in what years did you serve in this role?

Standards

(2) In what subject areas and grades does [STATE] have academic standards? Are there separate standards for students in special education or limited English proficiency programs?

(3) Have any of these standards been introduced or changed since 2001? Do you think these changes were made as a result of NCLB?

(4) Could you briefly describe the process by which standards were originally set? When did this occur?

(5a) Did you participate in the process of creating the standards? [If yes] Could you describe your memory of the number of people participating in meetings and who specifically was involved? [With this last question, and similar ones below, try to get names and contact information here for follow-up.]

(5b) Do you know of any documentation indicating how many people participated in the process, how many meetings were involved, whether there were public hearings, and so on?

(5c) How long did it take to develop academic standards when they were first implemented?

(6) Can you briefly describe the process by which standards are reviewed and updated? How often do these reviews and updates occur?

(7a) Did you participate in the process of reviewing and updating the standards? [If yes] Could you describe your memory of the number of people participating in meetings and who specifically was involved?

(7b) Do you know of any documentation indicating how many people participated in the updating process, how many meetings were involved, whether there were public hearings, and so on?

(8) Are there plans for other academic subjects to be added in the near future?

Assessments

(9) In what subject areas and grades does [STATE] have at least one standardized test? Are there separate assessments for students in special education or limited English proficiency programs?

(10) How many of these grade and subject tests have been added since 2001? [Unlike standards, NCLB has clear requirements regarding testing, so I don't see any need to ask questions about that.]

(11a) Are the tests norm-referenced or criterion-referenced? [In Florida, and perhaps other states, there are multiple tests and test types per subject.]

(11b) [If there is at least one criterion-referenced test] Could you briefly describe the process by which standards for the criterion-referenced tests were originally set? Are the tests updated each time the standards are updated?

(12) Did your state contract with a testing company to develop these tests? [If yes] Which companies?

(13a) Did you participate in the process of developing the assessments? If so, could you describe your memory of the number of people participating in meetings, etc.?

(13b) Do you know of any documentation indicating how many people participated in the development of the criterion-referenced tests, e.g., how many meetings were involved, whether there were public hearings, and so on?

(14) Are there plans for expanding the grades or subjects in which standardized testing will be administered in the near future?

Accountability

(15) Could you briefly describe what you see as the key elements of the state's school accountability system?

(16) I noticed on the web site that there is a program called [X], which you did not mention. Would you consider this to be part of [STATE]'s accountability system?

(17) Which accountability programs would you say require substantial financial or other resources from the state?

Other

(18) Is there someone within the [STATE] Department of Education who is especially knowledgeable about how state efforts related to standards, assessment, and accountability are funded, e.g., how much is spent and from which budget categories?

(19) Are there other people within the [STATE] Department of Education with whom you would recommend I talk to better understand the resources going into [STATE]'s standards, assessment, and accountability?

(20) Could you suggest anyone outside the Department of Education, e.g., former department of education officials?

(21) Is there anything else you can think of that would help us to better understand the costs of the states standards, assessments, and accountability?

School District Interview Protocol

Background

(1) In what ways have you been involved with implementing state standards, assessments, and accountability in your district?

(2) How long have you served in these roles?

(3) Was anyone in your school district involved in setting or updating the state standard, assessments or accountability systems? [If yes] Who?

Standards and Assessments

(4a) Does your district provide professional development to teachers or administrators to educate them about the content of state standards and state assessments?

(4b) [If yes] Who provides this professional development? [We'll need to follow-up with these people. Some of them might not be district employees.]

(4c) [If no] Do school personnel seek out and/or pay for professional development on their own? [If yes] Could you describe the sources of the professional development that teachers use?

(5a) How many hours or days per year do students spend taking state assessments?

(5b) How many hours or days per month do students spend learning test taking skills?

(6) Do teachers typically administer tests themselves? If not, who else is involved and how common is it for others to administer the tests?

(7) Does your district administer additional standardized tests that are not required by the state accountability system (such as the Stanford 10 or DIBLES)?

(8) Did your district administer standardized tests before they were required by the state accountability system? [If yes] How can I find out how much those tests cost?

Accountability

[This section is harder. I think we need to distinguish between those costs incurred to minimally comply with NCLB and state accountability versus costs that are induced to meet the accountability objectives.]

(9) Do you hold public meetings to communicate with parents, teachers, and/or the public about the state's accountability system and your school's rating? If so, how and how often?

(10) How much time do district and school personnel spend communicating about the state's accountability system to school district personnel?

Other

(11) Are there other costs incurred by the district with regard to standards, assessments, and accountability that we haven't yet discussed?

(12) Are you aware of any documentation (e.g., budgets) that describe the districts activities related to standards, assessment, and accountability?

Thank you again for participating. I may have some follow-up questions and I hope you won't mind my contacting you again in the near future.