

The Resource Costs of Standards, Assessments, and Accountability

Douglas N. Harris
Lori L. Taylor

with
Amy A. Levine
William K. Ingle
Leslie McDonald

March 10, 2008

A Final Report to the National Research Council

Acknowledgements: This research was commissioned and funded by the National Academy of Sciences, National Research Council. For their valuable comments, we thank Stuart Elliot, Margaret Goertz, Margaret Hilton, Lisa Towne and other participants in the NRC Workshop on Common Standards, January 17-18, 2008, Washington, DC. All errors are those of the authors.

Abstract

While most previous research on standards, assessments and accountability systems have focused on the potential benefits of common standards, we consider the real resource costs of those systems. Among the small number of previous studies on the subject, the largest cost estimate is 0.33 percent of total K-12 education expenditures. However, previous studies under-state current costs by focusing on costs before NCLB was put in place and by excluding important cost categories. We consider three questions related to the costs of SAA systems:

(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? We find that the real resource costs are in the range of \$125-174 per pupil, or \$6.1-8.5 billion total per year throughout the nation. This represents 1.7-1.9 percent of total public education expenditures, which is nearly six times more than previous estimates have suggested.

(2) What costs would the nation incur if the current state-based system of standards, assessments, and accountability required of each individual state by NCLB were implemented as common system? The potential savings from a common system come from the fact that some SAA system costs incurred in every state—the “fixed costs”—would be incurred only once in a common system. We find that fixed costs are a 3-8 percent of total costs. Combined with the above total national cost, this implies potential cost savings of \$160-\$680 million per year.

(3) What are the costs of some of the specific “add-ons” to SAA systems that are used in some states and districts, but not required by state or federal law? In answering this question, we consider the following specific add-ons: benchmarking systems, high-quality assessments, teacher merit pay and school performance rewards, funding support for low-performing schools and social promotion.

These findings inform both the general issue of standards, assessments and accountability and the specific design of the systems, including the degree to which systems should be centralized. The costs are of course not the only issue that needs to be considered in these decisions, but it is one of many that will be important as the nation considers changes in its policies to improve the nation’s schools.

I. Introduction

The increasing role of state and federal governments in public schooling in the United States represents one of the most fundamental changes that has ever occurred in public education. 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 NCLB obviously shifted some control over schooling from the local to the federal level but, by placing so much weight on state-determined standards and assessments, the new federal law has also shifted control to state governments. Not surprisingly, NCLB has led to a myriad of state standards and assessments—some very stringent, others virtually toothless. The array of standards has frustrated those who think either that there is one right way to approach standards-based reform, or that making comparisons across states is important in order to treat states and schools equally with regard to NCLB sanctions. On the other hand, the diversity of policy approaches has been celebrated by others who think that local control and experimentation are, in the long run, best for everyone. This tension between central and local control is a well-known feature of the U.S. federal system of government, and is at the heart of proposals for additional federal control over standards and assessments.

One way to approach this tension is to examine evidence about the costs and benefits of state-driven versus common systems of standards, assessments, and

¹ NCLB reauthorized the Elementary and Secondary Education Act of 1965 (ESEA). The previous 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).

accountability (SAA). The present study focuses on comparing the resource costs of state-driven SAA systems with those of a single common system. Specifically, 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 current state-based system of standards, assessments, and accountability required of each individual state by NCLB were implemented as common system?

(3) What are the costs of some of the specific “add-ons” to SAA systems that are used in some states and districts, but not required by state or federal law?

One of the most important factors affecting the answers to the first two questions is how ambitious the systems are in 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 “prototypical” system and then consider the costs of state implementation versus common implementation of that same prototype. 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 of the 50 states 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.

Our 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 about how common standards might differ from existing SAA systems. Also, it turns out that only a small percentage of total costs are fixed costs so that the costs of state-driven and common SAA systems are quite similar.

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, as noted above, such costs are clearly a function of how high the standards are set, and the purpose of this study is not to guess about the height of the bar. Second, even if we knew where the bar would be set, 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 not only on the height of the bar, but also on the research methodology used in the analysis and on 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, recognizing that the costs of adequacy are orders of magnitude greater than the costs of compliance.

While we do not directly address issues of adequacy, we do consider in the third question the costs of “add-ons,” or aspects of SAA systems that go beyond compliance. Some state governments go above and beyond federal requirements and some school districts go beyond what is required by their respective state governments. We consider, for example, more costly assessments that go beyond multiple choice questions to include open-ended responses, as well as quarterly assessment or “benchmark” systems that involve testing students many times throughout the year to gauge progress. The costs of these add-ons are not included in the answers to the first two questions.

We focus primarily on the “costs to the nation” by which we mean the “real resource” (opportunity) costs, as contrasted with “expenditures” (budgetary) costs. Some resource costs might not show up in a state or school district budget, because the budgets are not designed specifically to account for the costs of SAA systems. For example, a school principal might spend a significant fraction of her time training teachers on the state content standards, but all of her time would be attributed to “general administration” rather than an SAA system budget. On the other hand, some budgetary costs attributed to SAA systems might not be considered real resource costs because the funds are simply being shifted from one group to another without any real resources being used (sometimes called a “cash transfer”). While we focus on measuring real resource costs, we do utilize budgetary information as part of the calculations.

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). We answer the first research question, relating to the current costs of SAA systems, by estimating the costs of the SAA systems in the three states (Section V), defining and measuring costs in the prototypical state under NCLB, and calculating the implied total national costs now being incurred (Section VI). To answer the second research question, we then separate the portion of total costs that are fixed and compare these with current total national costs to estimate the costs that would be incurred in a single common SAA system (Section VII). Finally, we discuss the costs of “add-ons” (Section VIII) and draw conclusions (Section IX). These estimates are, again, only one of the many issues that must be considered in any debate about common standards. The costs we find here, as well as other potential benefits, must be balanced against the possible disadvantages of various types of SAA systems.

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 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 compared with the option of states’ controlling standards,

national educational standards always improve social well-being. A number of researchers have also gone beyond theory and tried to measure empirically the benefits of educational standards.²

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. Of course, 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.

Very few researchers have attempted to quantify SAA system 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 study ascribes any costs to

² 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).

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

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

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. Again, the costs of assessment have obviously increased since the year 2000 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 regarding their budgetary outlays for school assessment and accountability. Her cost estimates 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, but excluded district-level costs such as administering assessments and providing assessment-related professional development to teachers. She again calculated the costs per pupil of state accountability systems. 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.

At about the same time, a survey of all 50 state departments of education largely corroborated Hoxby's finding. The survey found that the budgetary costs of developing,

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

administering and correcting state tests (a narrower definition of assessment than Hoxby 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 Taken together, Hoxby's study and the survey described by the Education Commission of the States indicate that assessment and accountability comprised no more than 0.33 percent of current operating expenditures.⁶

We show later in the analysis cost estimates of SAA systems in these previous studies are considerably lower than our own estimates, both because of the recent adoption 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.⁷ As of the 2004-05 school year, Texas is the largest of the three with more than 4.4 million students, and annual expenditures of \$7,246 per pupil. Florida is one third smaller, with 2.6 million students, but its expenditures per pupil (\$7,215) are very similar to those in Texas. Both Texas and Florida also have a history of SAA that predates NCLB. In contrast, North Dakota is a very small state (100,513

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

⁶ This calculation is based on the fact that, on average, school districts spent a total of \$7,485 per pupil in 2001 (St. John, et al. 2007).

⁷ Data on enrollments and expenditures in this paragraph come from the NCES Common Core of Data for the 2004-2005 school year. More recent financial data are not available..

students) with relatively high annual expenditures per pupil (\$7,829) and no history of criterion-referenced testing before the introduction of NCLB.

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), and a college sophomore exit exam for advancement to upper division (1984). Beginning in 1990, the state legislature enacted *Blueprint 2000*, an 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. Schools were required 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 in schools where sufficient progress does not occur after three years. In 1995, with this tighter protocol, the state identified 158 public schools in Florida as “critically low” performing (Harris, 2001). This school grading system evolved over the next few years toward the A-F school grading system that has been a prominent part of school accountability since the adoption of the 1999 “A+” accountability program.⁸

⁸ 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 and was therefore eliminated (Harris, Herrington, & Albee, 2007).

The cornerstone of the Florida accountability system is the Florida Comprehensive Assessment Test (FCAT). According to the Florida Department of Education (FDOE, 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. 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 academic standards for Florida students, and authorized the February 1997 field testing of the standards-based assessments. In 1998, nearly all students in grades 4 (reading), 5 (math), 8 (reading and math), and 10 (reading and math) took the FCAT reading and mathematics tests for the first time and results were released to districts, schools, and parents.

With the content of the FCAT established, educators, citizens, and business leaders from across the state were involved in a process, also in 1998, to develop FCAT

performance standards or “cut scores,” which the State Board of Education subsequently adopted. In addition, the state education commissioner designated the score level on the FCAT that would allow students to be exempt from the High School Competency Test (HSCT).⁹

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.

In compliance with the federal NCLB, Florida students with learning disabilities 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 accommodations¹⁰; or (3) the alternate assessment for students who are unable to participate in the state-wide testing program due to significant

⁹ 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.

¹⁰ Based on information we collected in North Dakota, testing accommodations include: *setting and location accommodations* which are changes in the location in which a test is given or the conditions of the assessment setting; *presentation accommodations* which allow students to access information in ways that do not require them to visually read standard print; *response accommodations* which allow students to complete assessments in different ways or to solve or organize problems using some type of assistive device or organizer; and *timing and scheduling accommodations* which increase the allowable length of time to complete an assessment and perhaps changes the way the time is organized.

disabilities. However, based on information we collected from one Florida school district, no more than two percent of students take alternative assessments.¹¹ Though few students take tests with accommodations or alternative assessments, the cost per student in these few cases is no doubt considerably higher. Nevertheless, we were unable to collect information that would have allowed for valid cost estimates and we therefore exclude these costs from the remainder of the study.

In February and March 2007, the FCAT was expanded to include 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.

Table 1: Recent History of Florida’s Accountability System, by Grades and Subjects

Academic Year	Reading and Language Arts	Mathematics	Science	Writing
1997-1998 to 1998-1999	4,8,10	5,8,10		
1999-2000 to 2001-2002	3-10 *	3-10 *		4,8,10
2002-2003 to 2007-2008	3-10 *	3-10 *	5,8,10	4,8,10

Notes: Criterion-referenced tests were administered in all subjects and grades indicated. An “*” indicates that norm-referenced tests were also administered in the designated grades and subjects.

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

¹¹ We found similarly low percentages of students taking alternative assessments in North Dakota.

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 evaluation system (something the state previously required of each district). Importantly, student scores on the FCAT must comprise at least half of the total evaluation.¹² 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.

In 2006, the Florida legislature added to the A+ plan with the A++ plan stipulating that each school earning a school grade of "C" or below, or that is required to have a school improvement plan under federal law, must have a school improvement plan that includes a variety of components: 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

¹² The focus on achievement gains 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.

alternative instructional delivery methods to support remediation, acceleration, and enrichment strategies.

At the other end of the A++ spectrum, schools that receive an “A” grade or improve at least one performance grade category in a given year 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 some 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 million in awards.

Assessments in Florida also involve high stakes for students. In 2002, the Florida legislature voted to require third-grade students to meet at least the Level 2 benchmark (the second-lowest of five levels) on the FCAT reading test in order to be promoted to the fourth grade. Prior to the legislative change, students were promoted to the next grade based on a “social promotion” policy, whereby students are promoted to the next grade regardless of academic preparation. The third-grade class of 2002–03 was the first to be affected by the new policy.

North Dakota

North Dakota has a more modest history of statewide assessment and one 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. From 2001-2002 to 2003-2004, the NDDPI administered CTB/McGraw-Hill’s *Terra Nova*¹³ with a section of the assessment “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* is no longer being used.

Table 2: History of North Dakota’s Accountability System, by Grades and Subjects

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

Notes: Criterion referenced tests were administered in all subjects and grades indicated. An “*” indicates that a hybrid assessment consisting of the *Terra Nova, Second Edition, Basic Multiple Assessment* and a North Dakota supplement (aligned with State Standards) was administered

¹³ 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.

The NDDPI's Standards and Achievement Unit is responsible for the administration of accountability efforts; the development of state content and achievement standards; the administration of state and federal language acquisition programs; the administration of standardized tests; 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 is limited almost entirely to the requirements of NCLB.

Texas

As in Florida, school assessment has deep roots in Texas.¹⁴ 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 3, 5 and 9. School and district level results were made public, allowing parents and taxpayers to hold school districts

¹⁴ The information in this and subsequent paragraphs is taken from TEA (2002).

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 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 or 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.

Table 3: History of Texas' Accountability System by Grade and Subject

Academic Year	Reading/ Language Arts	Mathematics	Science	Writing	Social Studies
1993-1994 to 2000-2001	3-8,10	3-8,10		4,8,10	
2001-2002	3-8,10	3-8,10		4,8,10	8
2002-2003	3-11*	3-11*	5,10,11*	4,7*	8,10,11*
2003-2004 to 2007-2008	3-11	3-11	5,10,11	4,7	8,10,11

Notes: Criterion referenced tests were administered in all subjects and grades indicated. *Although all tests were administered, no accountability ratings were assigned in 2002-2003, which was the first year TAKS tests were administered.

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). For example, if all of the other student groups in a school were performing at Exemplary levels, but the low-income students were performing at only an Acceptable level, then the school was rated only Acceptable. Each school was required to send school report cards to every parent with children in the school. Schools rated Exemplary, Recognized or Acceptable with high performance gains were eligible for modest monetary awards through the Texas Successful Schools Awards program (TSSAS).¹⁵ Exemplary school districts were exempt from a number of state rules and regulations, and low performing school districts were subject to sanctions ranging from

¹⁵ Total state outlays for the TSSAS were only \$5 million per year, or less than 0.05 percent of operating expenditures.

the requirement to develop an action plan to complete state takeover or forced consolidation with another district. The Public Education Grants (PEG) program allowed students in low performing schools to transfer to higher performing schools in other districts—and take their funding with them.

Starting in 1995, students in grades 3 through 6 who were not proficient in English could take the TAAS test in Spanish rather than English. Student performance on the Spanish TAAS was first incorporated into the accountability system during the 1998-99 school year.

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 accountability system, and the third, fifth and eighth grade tests became high-stakes tests for students.¹⁶ The state also introduced the Student Success 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 \$150 million.

In addition, Texas added a parallel testing instrument for special education students to the accountability system. The State Developed Alternative Assessment II (SDAA II) is used to measure the performance of special education students whose instruction follows the state curriculum. A district or campus must administer at least 30 SDAA II exams to be evaluated under this performance measurement.¹⁷

¹⁶ 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. Science and social studies tests have been administered in Texas since 1995, but were not included in the accountability system until 2003 and 2002, respectively.

¹⁷ A student who took SDAA II exams in reading/English-Language Arts, writing, & math would be counted as having taken three exams, so a school could meet this standard with as few as 10 students.

With the shift to TAKS, the consequences for a low performance rating became more severe. For example, 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

Table 4 compares the current accountability systems in Florida, North Dakota and Texas. As the table illustrates, all three states comply with the NCLB requirements with respect to reading, mathematics and science. Texas and Florida go beyond the NCLB requirements by testing in additional grades for mathematics and reading, by including a writing test, and by making some of their tests high-stakes tests from the student perspective. Texas also goes further by including social studies in its state accountability system, and by allowing elementary-grade students who are not proficient in English to take the TAKS tests in Spanish.

Table 4: Assessments for Accountability 2007-08

State	Reading/ Language Arts	Math	Science	Writing	Social Studies	Social Promotion for Students	High Stakes for Educators
Florida	3-10	3-10	5,8,11	4,8,10	-	3, 11	Merit pay, school bonuses
North Dakota	3-8,11	3-8, 11	4,8,11	-	-	-	---
Texas	3-11	3-11	5,10,11	4,7	8,10,11	3,5,8,11	Merit pay in some schools, PEG grants

Notes: Under NCLB, one could easily argue that all educators are in a “high stakes” environment under NCLB. We have therefore limited the table to those elements that go beyond NCLB.

This historical background above provides some sense of how much has changed in SAA systems in recent decades, how much variation remains across states, 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 in these states.

IV. Conceptualizing and Measuring the SAA System Costs

Economists typically conceptualize costs in terms of the resources necessary to produce 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 (SAA). We therefore discuss below how we have made choices regarding the scope of the SAA 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, 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. Benchmark systems are clearly a type of standardized assessment and therefore intuitively part of the assessment portion of SAA systems. On the other hand, one might argue that benchmark systems are a supplemental tool that some school districts use to meet state and federal pressures for academic performance. As such, the costs of benchmark systems would fall into the excluded “adequacy” category. 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 treat the cost of quarterly assessments as an add-on to SAA systems. We discuss the cost of specific add-ons—including quarterly assessments—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, compliance with governmental requirements. We therefore include SAA resources that are required by centralized governments *and* serve the function of SAA, but we exclude resources that do not meet both of these criteria.

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 (personnel, facilities, equipment, etc.). Resources within these categories are then measured in raw units (e.g., the number of hours spent on a particular task), which are then translated into cost figures by identifying the opportunity cost per unit. For example, the opportunity cost of a teacher’s time can be measured by her hourly rate of compensation (which includes employer contributions to health and pensions). This is preferred to the expenditure/budgetary approach because the real resource approach ensures that all types of costs to society are accounted for.

To illustrate the distinction between expenditures and real resources, consider that many states—including Florida—incorporate financial incentives for educators directly into their accountability systems.¹⁸ Taking a budgetary cost perspective, these costs clearly should be included in any estimate of the cost of SAA systems, but the situation is less clear when taking a real resource perspective. 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 and therefore an important cost. On the other hand, depending on how they are implemented, 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. Whether one focuses on real resources versus budgetary costs can have a dramatic impact on cost estimates, as we will see in section VIII.

¹⁸ Texas has a number of incentive pay programs that tie teacher awards and school eligibility to the accountability system. However, these awards are not considered part of SAA in the Texas budget.

Because real resources are sometimes hard to measure, and budgets are often readily available, it is common in cost analysis to rely on budgetary information even when trying to measure real resources. In our analysis, we are able to measure real resources directly 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, in these cases, it is generally impossible to disentangle the costs of specific activities. We are careful to discuss possible discrepancies between expenditure and real resource costs where such differences arise.

New Resources versus Re-allocation

One of the central issues in the debate about standards-based reform is the concern that assessments cannot 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 when estimating the costs of the current SAA system.

If we were to take a “strict budgetary” approach to the problem, re-allocation of existing 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 as additional real resources. 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 at all. In most of the analysis below, we avoid the strict budgetary approach and calculate costs assuming that all SAA resources are new resources.

The Real Resource Cost of Educator Time

The primary cost associated with SAA systems is the opportunity cost of educator time. Therefore, one of the keys to our analysis is an accurate estimate of educator wages. We rely on data from the Bureau of Labor Statistics (BLS) National Compensation Survey (NCS) for 2005. The NCS is a quarterly survey of civilian employers in 154 metropolitan and non-metropolitan areas. The BLS uses the NCS to provide national estimates of salaries and benefits for over 800 occupations. We base our cost estimates on the 2005-06 school year. Therefore, our estimates of labor cost are the average of hourly compensation in the third and fourth quarters of 2005, and the first and second quarters of 2006.

The NCS indicates that primary, secondary and special education school teachers in the public sector earned \$48.07 per hour, on average, during the 2005-06 school year.¹⁹ Of that total, \$35.08 was wages while \$12.99 was benefits (mostly paid leave, health

¹⁹ On average, all elementary, secondary and special education teachers (public and private) earned \$45.51 per hour in total compensation in 2005-06. However, private schools are not subject to the forms of government accountability described here, so we limit costs to teachers in public schools.

insurance and retirement benefits). As we will see below, some of the people involved in the SAA systems are school administrators, state government officials and private-sector business people. In order to value their time, we use NCS estimates of the average hourly compensation for management, business and financial occupations, which was \$48.50. Nationally, benefits comprised 30.5 percent of the total compensation of managers, while they comprised 27 percent of the compensation of public school teachers.²⁰

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 & 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 NCS

²⁰ Considering that over time the share of compensation going to fringe benefits has been rising, the NCS estimates for benefits are reasonably consistent with those from other sources. According to the U.S. Census (2005), fringe benefits for school instructional staff are 23.9-27.3 percent above the salary level. Podgursky (2003) reports that fringe benefits for teachers are 20.2 percent of total compensation.

indicates that the national average hourly wage for management occupations is \$48.50, we presume that such a person would earn \$44.91 in Florida, \$38.89 in North Dakota and \$47.99 in Texas. Similarly, the average hourly wage for primary, secondary and special education school teachers in the U.S. was \$48.07 in 2006, implying that average hourly teacher wages in Florida, North Dakota and Texas were \$44.51, \$38.54 and \$47.55, respectively.

Generalizing from the Sample to the Nation: The “Prototypical 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 NRC 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. We then use the information from the prototypical 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 with officials in state education agencies (SEAs) and local education agencies (LEAs); 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, we were referred to another knowledgeable person within the agency. Four or more officials were interviewed in each state, two at the SEA level and two at the LEA level. All interviews were conducted by phone.

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 LEAs, the district names will also remain anonymous.

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. We divide the discussion into “state” and “local” cost categories, 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 primarily reflect the source of the data (state officials versus local officials). We also divide state costs into sections on standards, assessment, and accountability, though these lines, too, are somewhat blurry because standards are the basis for assessments and so on. Using these

categories helps us to structure the discussion and make useful generalizations about the sources of fixed and variable costs and therefore to isolate the potential cost savings from national standards. We begin by discussing all the state costs, followed by local costs.

The Standard Setting Process

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. Here, we distinguish “content standards,” referring to the academic content that students are part of the curriculum, from the “performance standards” or “cut points”, referring to how students are placed into categories based on their numeric scores.

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 content 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. (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 for 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. The language arts and mathematics standards were revised and approved by the State Board of Education in 2007 and the science standards are currently undergoing public input. The social studies standards have just started the review process and framers will meet in 2008.

North Dakota. NDDPI initially developed reading/language arts and mathematics state content and achievement standards. Content 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. For each set of standards, the NDDPI has 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 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.

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 labor intensive. For example, the social studies TEKS were developed by a 35-member writing team composed teachers, campus administrators, college/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. “More than 1,500 response forms were received, compiled and summarized, and given to the writing team for consideration and action.” (See <http://www.tea.state.tx.us/ssc/teks> and [taas/teks/teksqa.htm](http://www.tea.state.tx.us/taas/teks/teksqa.htm).) 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.²¹ 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

²¹ 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 (See http://www.tea.state.tx.us/ssc/teks_and_taas/teks/teksqa.htm)

recognized experts.²² 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.

State-Level Cost of Standards

The costs of developing and updating standards depend principally on three factors: the number and specificity of the standards, the opportunity cost of those participating at each stage, and the frequency with which the standards are updated. 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

²² For names and biographies, visit <http://www.tea.state.tx.us/student.assessment/taks/standards/index.html>.

drafts for comments before submitting their final draft to the State Board of Education for approval. Assuming that the writing team met twice, with a day of preparation and two days of meetings each time, then writing the curriculum standards required 174 days teacher labor and 36 days labor from other professionals. The State Board of Education designated a 15-member review committee. Assuming that their review took two days, this adds another 30 person-days of professional labor. Finally, the state hired four national experts to provide comments. We assume that those individuals add another eight days to the total. Thus, developing the TEKS for social studies cost approximately 174 days of teacher labor, 74 days of professional labor, and travel expenses.²³ We presume that the opportunity cost of a teacher day is eight hours at the Texas prevailing wage for teachers ($\$47.55 \times 8 = \380.43) while the opportunity cost of a professional day is eight hours at the Texas prevailing wage for managerial occupations ($\$47.99 \times 8 = \383.90). We use the 2007 IRS per diem rate for Austin, Texas ($\$139$ per day) as our best estimate of food and lodging costs for the standard-setting groups, and assume that travel costs averaged $\$200$ per person.²⁴ At current wage rates, therefore, developing the TEKS curriculum standards for social studies cost $\$135,234$, plus the costs to the agency.²⁵ Lacking a better estimate of the costs to the agency, we assume that developing the TEKS for social studies required one full-time-equivalent TEA

²³ 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.

²⁴ 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.

²⁵ Total cost equals the opportunity cost of 174 teacher days (174×380.43) plus the opportunity cost of 74 days for other professionals (74×383.90) plus the travel cost for the 35 members of the writing team ($35 \times 2 \times 200$) plus the lodging cost for the writing team ($35 \times 4 \times 139$) plus the travel and lodging cost for the 15 member review panel ($15 \times 2 \times 139 + 15 \times 200$).

employee, or \$54,000.²⁶ TEKS social studies standards were set for grades K-8 and for ten high school courses. Assuming that each of the nineteen standards was equally costly, that works out to roughly \$10,000 per subject-grade standard.

Another set of stakeholder panels developed performance standards for the TAKS test. The National Technical Advisory Committee's 13 members met four times in Austin and twice by telephone (TEA, 2004). Assuming that each Austin meeting took two days, and each telephone meeting took half a day, we estimate that the opportunity cost of time and travel for the 13 members was just under \$70,000.²⁷ 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.²⁸ At current wage rates, 1,228 days is equivalent to \$469,365.²⁹ Travel expenses

²⁶ The average employee in the curriculum division of TEA was expected to earn \$56,492 in 2007-08 (TEA 2007b). We use the consumer price index to adjust for inflation between 2005-06 and 2007-08.

²⁷ Total cost=the opportunity cost of time (13*8*\$384 for the 4 Austin meetings + 13*\$384 for the two half day phone meetings) plus the cost of travel for the Austin meetings (13*4*\$400) + the cost of lodging for the Austin meetings (13*8*\$139). We assume that the cost of travel to Austin from outside the state is double the cost of travel inside the state.

²⁸ Total time spent setting performance standards equals the total days spent by the advisory panel (19*2) plus the total days spent by the regular panels (15*3*20+ 6*2*20) plus the supplemental day spent by the 50 regular panelists (50*1).

would add another \$258,492. Thus, the cost of setting the performance standards for the 36 TAKS tests was approximately \$22,000 per subject-grade standard.³⁰ Because the cost of setting content standards was approximately \$10,000, we estimate that the cost of setting content and performance standards in Texas was \$32,000 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 subjects 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

²⁹ Lacking more precise data on their composition, we assume that the regular panels were comprised of half teachers and half other professionals. Also, note that costs of setting performance standards appear to be higher than the costs of setting content standards. This may be because, while content standards are more complex and performance standards are inherently arbitrary, performance standards do have significant consequences for students and educators.

³⁰ $(469,365 + \$258,492 + \$70,000) / 36 = \$22,162$. We round these and may other numbers in order to avoid a false sense of precision in our results.

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 initial development of the state's science standards. This included nine days of meetings for framers and 18 days of meetings for the writers. We use average hourly compensation for managers in Florida for the framers (\$44.91) and the average hourly compensation for teachers for the writers (\$44.51). Nine days of meetings (72 hours) of meetings for each of 30 framers yields time costs of \$97,000. Eighteen days of meetings (144 hours) for each of 33 writers yields additional time costs of \$211,512, for a total time cost of \$308,512. 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 \$585,745. These costs pertain to a single subject and 13 grades, thus the cost per standard is roughly \$45,000.

Other standard-setting 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 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). Given teacher compensation of \$38.54 per hour, a total meeting time of 20 hours (2.5 days) translates into \$1,071 per person (\$771 plus \$300 in travel and lodging) for each of 29 teachers (the average of the three subjects), or a total of \$31,053. In addition, apparently because of the state's small size, North Dakota 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 annual compensation for managerial occupations (\$38.89 per hour times 50 weeks per year times 40 hours per week) this yields \$155,560.

In math and reading, tests are administered in grades 3-8 and 11 in North Dakota, for a total of seven grades. Thus, the first portion of the costs, from teacher time, for each of these individual subjects (\$31,053) comes to \$4,436 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 McREL 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 (\$155,560 +\$125,000 = \$280,560 divided by the 21 grade-subject standards). Combining this with the teacher cost per standard (\$4,436), yields nearly \$18,000 per subject-grade standard.

Across the three states, the costs of developing standards therefore range from \$18,000 per standard in North Dakota to \$32,000 in Texas and \$45,000 in Florida.

State-Level Costs of Assessments

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 (TEA 2007b). This is equivalent to \$20.46 per pupil. In Florida, these costs have been estimated at \$15.10 per pupil in 2006 (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 \$34.02 per pupil.³¹

³¹ 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

Contract outlays are not the only resource cost for test development, however. Much like the process of updating standards, states 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). At \$380 per day, the opportunity cost of educator time for those reviews is approximately \$842,000. 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 approximately \$61,000 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.³²

The state agencies maintain permanent staff who are responsible for 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 was \$26,547,890 in 2007 (FDOE, 2007), or \$10.21 per student. Excluding testing company contracts, the Assessment and Performance Reporting Divisions of the Texas Education Agency have a combined

standardized assessments and each teacher creates assessment systems individually. This approach is almost certainly more costly than even the most costly state assessment. Consider that there are more than two million teachers in the country, so that completely decentralizing the system would involve creating two million separate tests per subject. However, there is little evidence that standardized tests have reduced the prevalence of teacher-created tests and, in this case, the standardized tests represent additional costs and the above calculations are correct.

³² One eight-hour day at \$311 per day plus \$300 per day in travel and lodging for 100 people.

budget of \$12,731,389 or \$2.77 per student (TEA 2007b). Excluding external contracts, the annual budget for the NDDPI Division of Standards and Achievement in 2007 was \$975,000, or \$10.20 per student.

State-Level 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 and teachers. 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 may be just cash-transfers and, regardless of their intentions, almost certainly 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. We leave further discussion of financial awards and support funds for section VIII and the discussion of SAA “add-ons.”

Excluding the costs of adequacy, the costs of accountability per se are modest and are comprised principally of data management and reporting results. North Dakota’s

Department of Standards and Assessment has a contract of \$125,000 per year with an outside vendor (\$1.33 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 2008 budget for Accountability and Data Quality was \$1.21 per pupil. In Florida, this figure was \$19,738,000, or \$7.60 per pupil, perhaps reflecting the states large number of standardized tests and extensive data management and analysis system.

Local Costs of SAA

There are three main costs of SAA at the local level: administering the assessments, providing professional development to teachers and administrators regarding the content of standards and assessments, and managing and distributing assessment information. We consider these below in turn.

Test administration. The most obvious local costs of SAA systems involve the time that school and district staff spent in following the administrative rules associated with testing (e.g., they must ensure that students and teachers cannot see the exams in advance and that a minimum percentage of students show up and take the tests) 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. We assume 3.5 days of testing, 4 hours per day and \$38.54 per teacher-hour. According to the Common Core of Data, the pupil-teacher ratio in North Dakota is 12.5, so we initially estimate that the

opportunity cost of time for test administration is \$44 per student. Because only 7 of the 13 grades (K-12) are tested, we reduce this figure accordingly to \$24 per student.

Texas administered 8.8 million standardized tests in 2007 (Texas Education Agency 2007a). Assuming that each exam took half a day, the exams required 4.4 million student days. According to the Common Core of Data, the pupil-teacher ratio in Texas is 15 students per teacher, so 4.4 million student days is equivalent to 293,000 teacher days. Given our estimate of teacher wages, that also works out to \$24 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 that each test took half a day, and given a pupil-teacher of 16.2, this yields 262,000 teacher days and a cost per student of \$35.³³

Professional development. District officials in at least one state (Florida) indicated that they employ two full-time staff devoted to providing professional development to teachers to help them understand the standards. Using the state wage for management occupations to estimate the direct costs of these staff (\$89,820 as outlined in

³³ Cost per student equals 262,000 teacher days times \$356 per day (\$44.51*8) divided by 2.7 million students.

Section III), two full-time staff yields \$179,640. Given the approximate number of students in the school district, this yields a cost per pupil of roughly \$1.³⁴

Much more costly is the time of teachers participating in this professional development. In Florida, our sample district indicated that the average teacher spends 20 hours per year in professional development related to assessments and accountability. With roughly 150,000 teachers in the state, this implies costs of \$135 million per year, or \$52 per pupil in Florida. Assuming the same of time is spent in Texas and North Dakota, the costs in both of those states are approximately \$64 and \$63 per pupil, respectively. Given hundreds of thousands of students, it becomes clear that the greatest cost of changing standards is the cost of training teachers about those new standards. It is also worth noting here that it is virtually impossible to separate professional development for standards from that of assessments and accountability. Thus, while we include these costs in the standards category, we recognize that these relate to the whole system of SAA.

Managing data. One moderately sized Texas school had an office of eight full-time employees devoted to assessment and accountability, led by a district accountability coordinator. Another, somewhat smaller Texas district has two full-time-equivalent employees who are responsible for managing the district's obligations under the state assessment system. Given enrollments in those districts (both well below 60,000) and managerial wages in Texas, these local districts have administrative costs of SAA systems that are between \$6 and \$15 per pupil. We average these figures and therefore

³⁴ This figure is approximated in order to maintain the district's anonymity. Also, we suspect that our salary estimate here overstates what the district professional development specialists earn, but we have no other information on this and this particular cost is an extremely small portion of the costs of SAA professional development, and an even smaller percentage of total local costs.

assume \$11 per student in Texas for these administrative costs. We do not have data on these SAA administrative costs for North Dakota and Florida. Because Florida administers more tests, we use the \$15 from Texas as the estimate for that state and the smaller value of \$6 for North Dakota.

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 5 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 very small costs of standards, the reported preliminary estimates range from \$123 per student in Texas to \$139 per student in North Dakota. Using the 2005 figures reported in Section II, this yields costs 1.7-1.9 percent of total K-12 education expenditures, an estimate that is considerably higher than that reported in previous studies. Again, these higher costs partly reflect the wider range of cost elements we have considered, so that our estimates are more complete. In addition, and especially in the case of North Dakota, the higher costs also reflect the fact that the state is now

using considerably more resources for SAA than it was before NCLB. (See discussion in section II regarding North Dakota’s history with SAA.)

*Table 5: Summary of Compliance Cost Estimates
(real resource cost approach)*

	Florida	North Dakota	Texas
Standards	\$45,000/standard Or \$0.06/student	\$18,000/standard Or \$0.52/student	\$32,000/standard Or \$0.04/student
Assessments			
Annual reviews	n/a	\$61,000/year Or \$0.62/student	\$842,000/year Or \$0.19/student
State – test contracts	\$15/student	\$34/student	\$20/student
State – admin.	\$10/student	\$10/student	\$ 3/student
Accountability	\$ 8/student	\$ 1/student	\$ 1/student
Local			
Test administration	\$35/student	\$24/student	\$24/student
Prof. development	\$53/student	\$64/student	\$65/student
Data management	\$15/student	\$ 6/student	\$11/student
Total cost per student	\$136/student	\$140/student	\$124/student
Cost as % of total annual school expenditures	1.9%	1.8%	\$1.7%

Note: A single standard is for one subject and grade (e.g., third grade reading). The zero figures for local accountability costs reflect only the way in which we have structured the review. There are of course compliance costs associated with accountability (e.g., test administration and accountability-related professional development), but because these cannot be disentangled from the local cost associated with standards and assessments, we have chosen to put them in the other local cost categories. We assume that standard are revised every six years (as is the case in Florida), so the per-pupil annualized cost is calculated by multiplying the cost per standard by the number of state assessments listed for each state in Table 4 (and recalling that Texas administers 10 tests in Spanish as well as English), and then dividing by six times enrollment.

The similarity in the results across states is not surprising for two reasons: first, we have focused on estimating cost within specific categories that are common across

states; and, second, we have, for some SAA elements, used information from one state to estimate costs for those same elements in other states. The fact that the costs per student are somewhat higher in North Dakota, and smallest in Texas, partly reflects the state's small population and consequently lower economies of scale. As we will see in section VIII, Florida and Texas have considerable more "add-ons" than North Dakota and adding these to the mix would clearly make the total costs in Florida and Texas greater than those in North Dakota.

These estimates alone do not of course answer any 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 Prototypical 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 prototypical state and combine this with the cost information in Table 5 to estimate the current costs to the nation as a whole. It is important to emphasize at the outset that we are *not* assuming that the SAA systems and system costs in Florida, North Dakota and Texas are representative of the nation. Instead, we make the much more plausible assumption that, for those SAA elements that are common across states, and after adjusting each of the three state's cost figures based

on the comparable wage index (CWI), the national cost per student is within the *range* of estimates from these three states.

NCLB, Standards, and Assessments

The choice of cost elements considered in Table 5 was based, implicitly, on the idea that there is an SAA prototype—really, a minimum bar—that states have to reach in the design of their SAA systems in order to comply with NCLB. Federal law now requires states to have standards and administer standards-based assessments in most grades and basic academic subjects.³⁵ 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 also allows considerable latitude over both content and performance 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 about how to do this. 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” (20 U.S.C.S. § 6311).

³⁵ Unless otherwise specified, the information on NCLB in this section comes from the Elementary and Secondary Education Act of 2001, codified at 20 U.S.C.S. §6301 et. seq.

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.” These interventions range 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. Further, 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).

We do not count the costs of these interventions for several reasons, some of which we alluded to in the discussion in previous section regarding the program in Texas that, like NCLB, provides support funds to low-performing schools. First, because these funds are predicated on low student achievement, they arguably fall into the adequacy category. Second, the costs of these interventions are even harder to measure in the case of NCLB. 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. However, the amount of Title I funds given to a school is only very loosely related to the costs of the interventions required to comply with NCLB for schools in need of improvement—indeed, the fact that the law allows for a menu of somewhat vague options means that the costs could vary dramatically depending on choices made by the schools. Also, even if all the funds went into providing such supports, school districts could supplement Title I funds with funds from

their own funds. Therefore, Title I budgets, either at the federal or district levels, are poor indicators of the costs of accountability and we could not find sufficient information from other sources to make reasonable estimates. We discuss the issues further in section VIII, but do not include these costs in our analyses of the first two research questions.

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

The prototype state then is one that follows federal law and therefore has standards in core subjects and assessments in grades 3-8 and where, at the local level, some basic support is provided to teachers to inform and update them about those standards and assessments. All of the costs categories listed in Table 5 arguably are required to have a system that is consistent with this prototype, though the table also makes clear that some states spend more on certain elements than other states.

Another factor affecting cost estimates for each SAA element is the size of the state and the possibility of economies of scale. Since our goal here is to establish a reasonable range of estimates, having one of the largest states in the country (Texas) and one of the smallest (North Dakota) is an advantage of the sample of states under study.

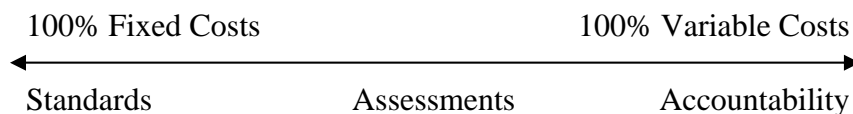
Given this prototype definition, the range of estimates in Table 5—\$124-140 per student—provides a useful starting point for establishing the range of total costs incurred in the nation as a whole. After applying the CWI to adjust for regional variations in cost to this range, we estimate that current national SAA systems cost between \$125 and \$174 per pupil. With 49 million students in the U.S.³⁶, this implies total national costs of \$6.1-8.5 billion per year.

³⁶ Data on enrollments and expenditures in this paragraph come from the NCES Common Core of Data for the 2004-2005 school year.

VII. Cost Savings from a Common SAA System: Fixed versus Variable Costs

The second research question in this study is, “What costs would the nation incur if the prototypical state system of standards, assessments, and accountability were replaced by a single common system?” As discussed earlier, fixed costs in education are those that are unrelated to the number of students and variable costs are those that depend principally on the number of students. At the one extreme, the costs of accountability are almost entirely variable, as the cost of any specific system depends mainly on the number schools that receive financial support and/or rewards. At the other extreme, the costs of setting standards are almost entirely fixed, although the costs of providing professional development to educators about standards are largely variable. Assessment costs fall somewhere in between as the costs of data systems and test design are largely fixed, while the costs of scoring and reporting test scores and administering tests at the local level are variable. We represent these generalizations about the cost categories in Figure 1 below.

Figure 1: Proportion of SAA System Costs that are Fixed versus Variable



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. In theory, we could compare the costs in each state to the measured cost per pupil and estimate the fixed costs directly. However, because the systems vary so much from state to state, this would require either completely specifying the costly elements of the assessment systems and estimating the fixed cost directly or gathering evidence directly from the testing companies. We have little confidence that the first option is possible and we tried, unsuccessfully, to gather the information from testing companies. Therefore, we have instead assumed, we think reasonably, that the testing contractor costs are essentially all variable. We have made similar qualitative judgments for other cost categories as well.

Table 6 shows the total costs incurred in each state by category. The cost categories are the same as those in Table 5, though we have re-organized them by whether they are primarily fixed or variable costs. To compare the portion in the fixed versus variable categories, we also now calculate the annualized total cost figures, multiplying the per student cost by the number of students in each state. From this simple calculation, fixed costs are 3-8 percent of total costs. This is likely an underestimate because, as indicated above, there are almost certainly some fixed costs in each of the cost categories labeled as variable costs.

*Table 6: Fixed versus Variable Costs
(total costs per state; in millions)*

	Florida	North Dakota	Texas
<i>Fixed costs</i>			
Standards – State	\$ 0.2 million	\$ 0.1 million	\$ 0.2 million
Assess – Annual reviews	0.0	\$ 0.1	\$ 0.8
Assess – State admin.	\$ 26.4	\$ 1.0	\$ 13.2
<i>Total</i>	<i>\$ 26.6 million</i>	<i>\$ 1.1 million</i>	<i>\$ 14.2 million</i>
<i>Variable costs</i>			
Assess – State test contracts	\$ 40	\$ 3.4	\$ 88
Acct – State	\$ 21	\$ 0.1	\$ 4
Local – Test admin.	\$ 92	\$ 2.4	\$106
Local – PD	\$140	\$ 6.4	\$284
Local – Data manage.	\$ 40	\$ 0.6	\$ 48
<i>Total</i>	<i>\$335 million</i>	<i>\$ 12.9 million</i>	<i>\$531 million</i>
Fixed Costs as % of Total	7.4 %	7.9 %	2.6 %

Note that the fixed costs are borne largely at the state level and the variable costs occur mainly at the local level. This is because the local costs reflect services being provided directly to students (and teachers) and variable costs are, by definition, those that vary based on the number of students. One exception is that the costs of contracts for state assessments are paid for by the state but are listed as being almost entirely variable costs. We did try to contact one of the testing companies to find out more about how they developed their bids for state assessments and how they figured in fixed costs, but the testing company was unwilling to share this information.

Using the total national cost figures discussed in the previous section (\$6.1-8.5 billion), and 3-8 percent fixed costs, this implies a potential cost savings from common SAA of \$160-\$680 million per year. This range is relatively wide because it reflects uncertainty in both the total cost figure and in the percentage of those costs that are fixed.

VIII. Just Beyond Compliance: The Costs of Add-Ons

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 three general categories of costs that are just outside the compliance definition: (a) costs that are natural responses to accountability, in order to reach the stated policy objectives; and (b) costs, such as maintaining alignment between standards and other education policies, that are arguably not new costs.³⁷ In the first category, we include the following add-ons: benchmarking systems, teacher merit pay and school rewards, support funding for low-performing schools, high-quality tests, and social promotion rules.

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 benchmark systems, the reasons districts adopt these systems, the timing of adoption, and, most

³⁷ We thank Margaret Goertz for pointing out this category.

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 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.³⁸ 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.

In addition to testing more often, there is growing interest in improving the quality of the tests. In terms of cost, there are two main issues: (1) the costs of creating the test standards, creating and testing items, and maintaining item banks; and (2) the costs of scoring tests, which are lowest with machine scored, multiple-choice tests and higher with constructed-response tests. For information on the first of these points we looked to the Advanced Placement tests, which are widely considered to have high standards and high-quality test items. The College Board, the owner of the AP, charges \$84 per test. Using this market rate as the cost measure, it appears evident that such high-quality tests are considerably more expensive than the costs of the tests being used in our sample states; recall that the contracts between states and testing companies ranged from \$15-\$34 per student for tests in multiple subjects across all grades.

³⁸ 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).

One testing company we contacted indicated that the cost of scoring an exam can run from as low as \$5 per exam to as high as \$28 per exam. Because the AP is machine scored, it is possible to estimate that the costs of most sophisticated state assessments—high-quality content tested through constructed response items—would be \$107 or more per test.³⁹ For tests in three subjects across grades 3-8 and one high school grade (as required by NCLB), the costs of AP-like tests could be orders of magnitude larger than current assessments.

Another type of add-on program is the provision of supplemental funds to schools and districts where students are struggling to pass high stakes tests. Texas' Student Success Initiative (SSI) is an example of one such program. The NCLB menu of options for schools in need of improvement is one example. Texas spent \$175 million or \$38 per pupil for the SSI in 2007. Those funds went to extra instruction for students identified as being at risk of failing math or reading in the third and fifth grades, and for students who did not pass those tests on the first attempt. Study guides were also provided to any student who did not pass TAKS, in both high-stakes and benchmarking grades.

As indicated earlier, Florida has programs for teacher and administrator merit pay, funded at a level of \$129 million in 2007. Florida is one of the only states to have such a system on a statewide basis, though interest is growing at the local, state and federal levels. Texas' Educator Excellence Grants Program, which is targeted at schools with a high concentration low income students, provides \$100 million per year for incentive pay linked to student performance. As indicated earlier, these costs only count as real resource costs if the funds attract more talented teachers (with higher opportunity costs)

³⁹ This figure comes from the \$84 for the AP plus the difference in cost between machine scoring (\$5) and the costs of tests with constructed response (\$28): $84+(28-5)=\$107$.

or induce greater time and effort from existing teachers. As budgetary costs, the resources going toward merit pay are potentially quite expensive add-ons because, as a political matter, it is quite difficult to get the programs adopted without adding new funds to pay for it. Also, for the programs to be credible, they have to involve meaningfully large percentage increases in compensation and that awards must be given to a substantial percentage of educators.

The high stakes tests in Florida and Texas are used to prevent “social promotion” by keeping students from going on to the next grade without having passed a minimum bar on the state exam. The potential cost of these programs is that students stay in schools an additional year, which is equivalent to having to educating more students, which requires hiring more teachers, and so on. For example, suppose that 10 percent of students are held back once during their school careers. For each cohort of students, this means that are, in effect, 13.1 grades for the average student (instead of 13 grades, K-12). If this required a proportional increase in costs, then this would amount to a 0.7 percent increase over current pupil expenditures in Florida, or \$55 per student on average.

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 system, in which each teacher could choose the content to be taught, there would still be a curriculum and textbooks would still a necessary part of the instructional process. One could even argue that the costs of textbooks under SAA are lower because the range of the curriculum is narrowed and therefore a smaller number of textbooks can supply the 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

This study provides a comprehensive analysis of the costs of standards, assessments and accountability (SAA) systems and the potential costs savings of a common SAA system. Specifically, we have provided answers to three important questions:

First, 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? We find that the real resource costs are in the range of \$125-174

per pupil, or \$6.1-8.5 billion total per year throughout the nation. This represents 1.7-1.9 percent of total public education expenditures. Compared with previous studies, our estimates include a wider range of cost categories and the additional costs arising from NCLB, which became law after the data were collected for previous studies. The differences in results are dramatic—we find that SAA costs are nearly six times larger than those found in earlier analyses (ECS 2001; GAO 1993; Hoxby 2002).

We do not assert that the costs of SAA are large in an absolute sense. At 1-2 percent of total expenditures, the costs still seem somewhat small, but judging the size of the costs requires accurate estimates both the costs *and benefits* of SAA systems, as well as the costs and benefits of alternative policy options. Perceptions about the size of the costs also depend on the perspective of the reader. We have focused on costs to the nation but, within the nation, are various stakeholders who bear higher shares of the costs and who vary in their ability to provide the necessary resources. State education agencies, for example, have no control over the vast majority of public education spending, so our estimates of SAA “costs as a percentage of total education spending” do not accurately represent the pressures and budget constraints these agencies face.

Second, what costs would the nation incur if the current state-based system of standards, assessments, and accountability required of each individual state by NCLB were implemented as common system? The potential savings from a common system come from the fact that some SAA system costs incurred in every state—the “fixed costs”—would be incurred only once in a common system. We find that fixed costs are a 3-8 percent of total costs. Combined with the above total national cost, this implies potential cost savings of \$160-\$680 million per year.

Third, what are the costs of some of the specific “add-ons” to SAA systems that are used in some states and districts, but not required by state or federal law? In answering this question, we consider the following specific add-ons: benchmarking systems, high-quality assessments, teacher merit pay and school performance rewards, funding support for low-performing schools and social promotion. In many cases, the costs of the add-ons exceed the entire cost of the prototype system. The larger point, however, is that SAA can be implemented in widely varying ways—and each way involves its own unique costs.

These findings inform both the general issue of standards, assessments and accountability and the specific design of the systems, including the degree to which systems should be centralized. The costs are of course not the only issue that needs to be considered in these decisions, but it is one of many that will be important as the nation considers changes in its policies to improve the nation’s schools.

References

- Baker, B.D., Taylor, L.L., & Vedlitz, A. (2005). *Measuring educational adequacy in public schools," Bush School Working Paper #580*. Retrieved December 20, 2007 from <http://bush.tamu.edu/research/workingpapers/>
- Betts, J. R., & Costrell, R. M. (2001). Incentives and equity under standards-based reform, in *Brookings papers on education policy*, (ed.) Diane Ravitch. 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 in *Handbook of the economics of education, Volume 2*, (eds.) Eric Hanushek & Finis Welch. Amsterdam: Elsevier.
- Burch, P., & 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.
- 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://fcad.fldoe.org/pdf/BriefingBook07web.pdf>
- 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.N. (2001). What caused the effects of the Florida A+ Program: Ratings or vouchers? In *School Vouchers*, (ed.) Martin Carnoy. Washington, DC: Economic Policy Institute.

- Harris, D.N. (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.N., Herrington, C.D., & 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 *Florida 2001: Educational Policy Alternatives*, (eds.) Carolyn D. Herrington & Katherine Kasten. Jacksonville: Florida Institute of Education, University of North Florida.
- Hoxby, C. (2002) The cost of accountability. In *School accountability*, (eds.) Williamson M. Evers & Herbert J. Walberg. Stanford, CA: Hoover Institution Press.
- Levin, H. & P. McEwan (2001). *Cost-effectiveness analysis*, 2nd Edition. London: Sage Publications.
- 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>
- No Child Left Behind (NCLB) Act of 2001, 20 U.S.C.S § 6301 et seq. (2008).
- Phelps, R. P. (2000). Estimating the cost of standardized student testing in the United States. *Journal of Education Finance*, 25(3), 343-80.
- Podgursky, M. (2003). Fringe benefits. *Education Next*, 3, 71-76.

- St. John, E., Hill, J., & 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*, 3, 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). *Annual administrative and programs strategic budget: Budget year 2008*. Austin, TX: author.
- Texas Education Agency. (2004). *Technical digest for the academic year 2003-2004*. Austin TX: author.
- Texas Education Agency (2002). *A standard-setting plan for the State Board of Education*. Austin, TX: author.
- U.S. Census Bureau. (2005). *Public education finances*. Washington, DC: author.
- 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*, (eds.) Chester E. Finn, Jr, & Diane Ravitch. Washington, D.C.: Thomas B. Fordham Institute.