United States Government Accountability Office

GAO

Report to the Chairman, Committee on Health, Education, Labor, and Pensions, U.S. Senate

September 2009

NO CHILD LEFT BEHIND ACT

Enhancements in the Department of Education's Review Process Could Improve State Academic Assessments

GAO-09-911

September 2009

NO CHILD LEFT BEHIND ACT

Enhancements in the Department of Education's Review Process Could Improve State Academic Assessments

GAO Accountability•Integrity•Reliability

Highlights

Highlights of GAO-09-911, a report to the Chairman, Committee on Health, Education, Labor, and Pensions, U.S. Senate

Why GAO Did This Study

The No Child Left Behind Act of 2001 (NCLBA) requires states to develop high-quality academic assessments aligned with state academic standards. Education has provided states with about $400 million for NCLBA assessment implementation every year since 2002. GAO examined (1) changes in reported state expenditures on assessments, and how states have spent funds; (2) factors states have considered in making decisions about question (item) type and assessment content; (3) challenges states have faced in ensuring that their assessments are valid and reliable; and (4) the extent to which Education has supported state efforts to comply with assessment requirements. GAO surveyed state and District of Columbia assessment directors, analyzed Education and state documents, and interviewed assessment officials from Maryland, Rhode Island, South Dakota, and Texas and eight school districts in addition to assessment vendors and experts.

What GAO Recommends

GAO recommends that Education (1) incorporate assessment security best practices into its peer review protocols, (2) improve communication during the review process, and (3) identify for states why its peer review decisions in some cases differed from peer reviewers' written comments. Education indicated that it believes its current practices are sufficient regarding our first recommendation and agreed with GAO's other two recommendations.

View GAO-09-911 or key components.
For more information, contact Cornelia Ashby at (202) 512-7215 or AshbyC@gao.gov.

What GAO Found

States reported their overall annual expenditures for assessments have increased since passage of the No Child Left Behind Act of 2001 (NCLBA), which amended the Elementary and Secondary Education Act of 1965 (ESEA), and assessment development was the largest expense for most states. Forty-eight of 49 states that responded to our survey said that annual expenditures for ESEA assessments have increased since NCLBA was enacted. Over half of the states reported that overall expenditures grew due to development of new assessments. Test and question—also referred to as item—development was most frequently reported by states to be the largest ESEA assessment expense, followed by scoring. State officials in selected states reported that alternate assessments for students with disabilities were more costly than general population assessments. In addition, 19 states reported that assessment budgets had been reduced by state fiscal cutbacks.

Cost and time pressures have influenced state decisions about assessment item type—such as multiple choice or open/constructed response—and content. States most often chose multiple choice items because they can be scored inexpensively within tight time frames resulting from the NCLBA requirement to release results before the next school year. State officials also reported facing trade-offs between efforts to assess highly complex content and to accommodate cost and time pressures. As an alternative to using mostly multiple choice, some states have developed practices, such as pooling resources from multiple states to take advantage of economies of scale, that let them reduce cost and use more open/constructed response items.

Challenges facing states in their efforts to ensure valid and reliable assessments involved staff capacity, alternate assessments, and assessment security. State capacity to provide vendor oversight varied, both in terms of number of state staff and measurement-related expertise. Also, states have been challenged to ensure validity and reliability for alternate assessments. In addition, GAO identified several gaps in assessment security policies that were not addressed in Education's review process for overseeing state assessments that could affect validity and reliability. An Education official said that assessment security was not a focus of its review. The review process was developed before recent efforts to identify assessment security best practices.

Education has provided assistance to states, but issues remain with communication during the review process. Education provided assistance in a variety of ways, and states reported that they most often used written guidance and Education-sponsored meetings and found these helpful. However, Education's review process did not allow states to communicate with reviewers during the process to clarify issues, which led to miscommunication. In addition, state officials were in some cases unclear about what review issues they were required to address because Education did not identify for states why its decisions differed from the reviewers' written comments.

Contents

Letter		1
	Background	4
	States Reported That Assessment Spending Has Increased Since NCLBA Was Enacted and Test Development Has Been the Largest Assessment Cost in Most States	12
	States Have Considered Cost and Time in Making Decisions about Assessment Item Type and Content	18
	States Faced Several Challenges in Their Efforts to Ensure Valid and Reliable ESEA Assessments, including Staff Capacity, Alternate Assessments, and Assessment Security	26
	Education Has Provided Assistance to States, but the Peer Review Process Did Not Allow for Sufficient Communication	34
	Conclusions	37
	Recommendations for Executive Action	39
	Agency Comments and Our Evaluation	39
Appendix I	**Objectives, Scope, and Methodology**	41
Appendix II	**Student Population Assessed on ESEA Assessments in School Year 2007-08**	45
Appendix III	**Validity Requirements for Education's Peer Review**	46
Appendix IV	**Reliability Requirements for Education's Peer Review**	47
Appendix V	**Alignment Requirements for Education's Peer Review**	48
Appendix VI	**Item Types Used Most Frequently by States on General and Alternate Assessments**	49

Appendix VII **Comments from the U.S. Department of Education** 50

Appendix VIII **GAO Contact and Staff Acknowledgments** 53

Table

Table 1: Illustration of Depth of Knowledge Levels 11

Figures

Figure 1: Examples of Item Types 6
Figure 2: State Expenditures for Assessment Vendors, 2007-08 14
Figure 3: ESEA Assessment Activities That Received the Largest Share of States' Total ESEA Assessment Costs, 2007-08 16
Figure 4: The Number of States Reporting Changes in Item Type Use on ESEA Assessments since 2002 19
Figure 5: Number of FTEs Dedicated to ESEA Assessments in States, 2007-08 27

Abbreviations

ARRA	The American Recovery and Reinvestment Act of 2009
AYP	Adequate Yearly Progress
CCSSO	Council of Chief State School Officers
Education	U.S. Department of Education
ESEA	The Elementary and Secondary Education Act
FTE	full-time equivalent
LEP	Limited English Proficiency
NCLBA	The No Child Left Behind Act of 2001
NECAP	The New England Common Assessment Program
SFSF	State Fiscal Stabilization Fund
TAC	Technical Advisory Committee

GAO
Accountability * Integrity * Reliability

United States Government Accountability Office
Washington, DC 20548

September 24, 2009

The Honorable Tom Harkin
Chairman
Committee on Health, Education, Labor, and Pensions
United States Senate

Dear Mr. Chairman:

The No Child Left Behind Act of 2001 (NCLBA), which amended the Elementary and Secondary Education Act of 1965 (ESEA), aims to improve student achievement, particularly among poor and minority students. To reach this goal, the law requires states to develop high-quality academic assessments aligned with challenging state academic standards that measure students' knowledge of reading/language arts, mathematics, and science. Student achievement as measured by these assessments is the basis for school accountability, including corrective actions such as removing principals or implementing new curricula. NCLBA required that states test all students in grades 3 through 8 annually in mathematics and reading/language arts and at least once in one of the high school grades by the 2005-06 school year. It also required that states test students in science at least once in elementary, middle, and high school by 2007-08. Education has provided states with about $400 million for ESEA assessment[1] implementation every year since 2002. To ensure that assessments appropriately measure student achievement, the law requires that assessments be valid and reliable and that they measure higher-order thinking skills and understanding. The U.S. Department of Education's (Education) guidance defines valid assessments as those for which results accurately reflect students' knowledge in a subject, and it defines reliable assessments as those that produce similar results among students with similar levels of knowledge. The law also directs states to assess all students, including those with disabilities. For children with significant cognitive disabilities, Education has directed states to develop alternate assessments that measure achievement on alternate state standards designed for these children.

[1]For purposes of this report, the term "ESEA assessments" refers to assessments currently required under ESEA, as amended. The Improving America's Schools Act of 1994 created some requirements for assessments, and these requirements were later supplemented by the requirements in NCLBA.

States have primary responsibility for developing ESEA assessments and ensuring their technical quality, and can work with private assessment vendors that provide a range of assessment services, such as question (item)[2] development and scoring. Education provides technical assistance and oversees state implementation of ESEA assessment requirements through its standards and assessments peer review process. In Education's peer review process, a group of experts—reviewers—review whether states are complying with ESEA assessment requirements, including requirements for validity and reliability, and that assessments cover the full depth and breadth of academic standards.

NCLBA increased the number of assessments that states are required to develop compared to prior years, and states have reported facing challenges in implementing these new assessments. Little is known about how federal, state, and local funds have been used for assessments, or how states make key decisions as they implement ESEA assessments, such as whether to use multiple choice or open/constructed response items. To shed light on these issues and to assist Congress in its next reauthorization of ESEA, the Chairman of the Senate Committee on Health, Education, Labor, and Pensions requested that GAO provide information on the quality and funding of student assessments. Specifically, you asked GAO to examine the following questions: (1) How have state expenditures on ESEA assessments changed since NCLBA was enacted in 2002, and how have states spent funds? (2) What factors have states considered in making decisions about item type and content of their ESEA assessments? (3) What challenges, if any, have states faced in ensuring the validity and reliability of their ESEA assessments? (4) To what extent has Education supported state efforts to comply with ESEA assessment requirements?

To conduct our work, we used a variety of methods, including reviews of Education and state documents, a 50-state survey, interviews with Education officials, and site visits in 4 states. We also reviewed relevant federal laws and regulations. To learn whether state expenditures for assessments have changed since NCLBA enactment, and if so, how they have changed, and how states have spent these funds, we analyzed responses to our state survey, which was administered to assessment

[2]For purposes of this report, we refer to test questions as "items." The term item can include multiple choice, open/constructed response, and various other types, while the term "question" connotes the usage of a question mark.

directors of the 50 states and the District of Columbia in January 2009. We received responses from 49 states, for a 96 percent response rate.[3]

We also conducted site visits to four states—Maryland, Rhode Island, South Dakota, and Texas—that reflect a range of population size and results from Education's assessment peer review. On these site visits we interviewed state officials, officials from two districts in each state, and technical advisors to each state.

To gather information about factors states consider when making decisions about the item type and content of their assessments, we analyzed our survey and interviewed state officials and state technical advisors from our site visit states. We reviewed studies from our site visit states that evaluated the alignment between state standards and assessments, including the level of cognitive complexity in assessments, and spoke with representatives from four alignment organizations—organizations that evaluate the alignment between state standards and assessments—that states hire to conduct these studies. These alignment organizations included the three organizations that states most frequently hire to conduct alignment studies, and representatives of a fourth alignment organization that was used by one of our site visit states.

In addition, we interviewed four assessment vendors that were selected because they work with a large number of states to obtain their perspectives on ESEA assessments and the assessment industry. We used our survey to collect information about challenges states have faced in ensuring validity and reliability. We also reviewed state documents from our site visit states, such as test security documentation for peer review and assessment security protocols, and interviewed state officials. We asked our site visit states to review a checklist created by the Council of Chief State School Officers (CCSSO), an association of state education agencies. A CCSSO official indicated that this checklist is still valid for state assessment programs.

To address the extent of Education's support and oversight of ESEA assessment implementation, we reviewed Education guidance, summaries of Education assistance, and peer review protocols and training

[3]New York and Rhode Island did not respond to the survey. For the purposes of this report, we refer to the District of Columbia as a state.

documents, and interviewed Education officials in charge of the peer review and assistance efforts.

We conducted this performance audit from August 2008 through September 2009 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

Background

The ESEA was created to improve the academic achievement of disadvantaged children.[4] The Improving America's Schools Act of 1994, which reauthorized ESEA, required states to develop state academic content standards, which specify what all students are expected to know and be able to do, and academic achievement standards, which are explicit definitions of what students must know and be able to do to demonstrate proficiency.[5] In addition, the 1994 reauthorization required assessments aligned to those standards. The most recent reauthorization of the ESEA, the No Child Left Behind Act of 2001, built on the 1994 requirements by, among other things, increasing the number of grades and subject areas in which states were required to assess students.[6] NCLBA also required states to establish goals for the percentage of students attaining proficiency on ESEA assessments that are used to hold schools and districts accountable for the academic performance of students. Schools and districts failing to meet state proficiency goals for 2 or more years must take actions, proscribed by NCLBA, in order to improve student achievement. Every state, district, and school receiving funds under Title I, Part A of ESEA—the federal formula grant program dedicated to improving the academic achievement of the disadvantaged—is required to implement the changes described in NCLBA.

ESEA assessments may contain one or more of various item types, including multiple choice, open/constructed response, checklists, rating scales, and work samples or portfolios. GAO's prior work has found that

[4]Pub. L. No. 89-10.

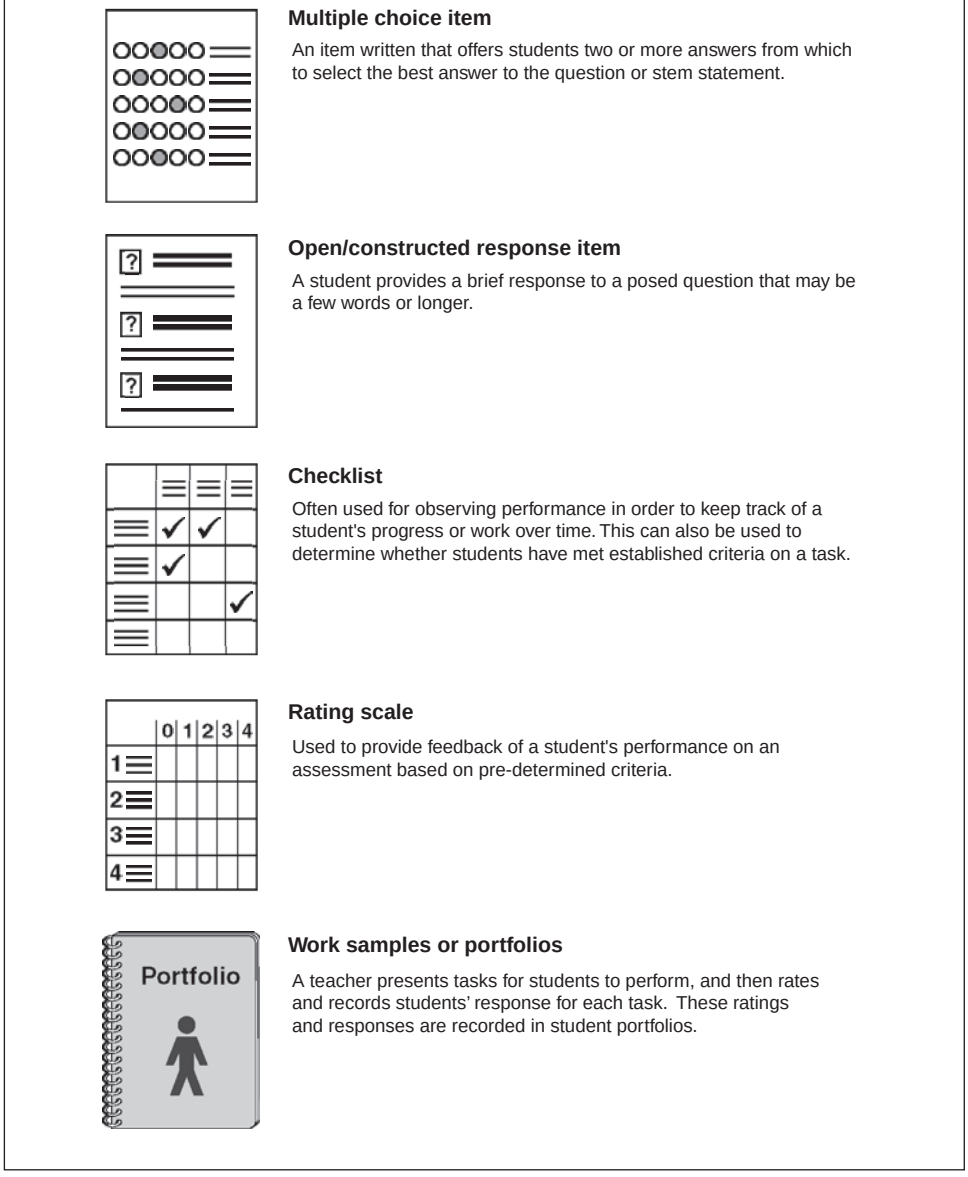
[5]Pub. L. No. 103-382.

[6]Pub. L. No. 107-110.

item type is a major factor influencing the overall cost of state assessments and that multiple choice items are less expensive to score than open/constructed response items.[7] Figure 1 describes several item types states use to assess student knowledge.

[7]GAO, *Title I: Characteristics of Tests Will Influence Expenses; Information Sharing May Help States Realize Efficiencies*, GAO-03-389 (Washington, D.C.: May 2003).

Figure 1: Examples of Item Types



Source: GAO; images, Art Explosion.

NCLBA authorized additional funding to states for these assessments under the Grants for State Assessments program. Each year states have received a $3 million base amount regardless of its size, plus an additional

amount based on its share of the nation's school age population. States must first use the funds to pay the cost of developing the additional state standards and assessments. If a state has already developed the required standards and assessments, NCLBA allows these funds to be used to administer assessments or for other activities, such as developing challenging state academic standards in subject areas other than those required by NCLBA and ensuring that state assessments remain valid and reliable. In years that the grants have been awarded, the Grants for Enhanced Assessment Instruments program (Enhanced Assessment grants) has provided between $4 million and $17 million to several states. Applicants for Enhanced Assessment grants receive preference if they plan to fund assessments for students with disabilities, for Limited English Proficiency (LEP) students or are part of a collaborative effort between states. States may also use other federal funds for assessment-related activities, such as funds for students with disabilities, and funds provided under the American Recovery and Reinvestment Act of 2009 (ARRA).[8] ARRA provides about $100 billion for education through a number of different programs, including the State Fiscal Stabilization Fund (SFSF). In order to receive SFSF funds, states must provide certain assurances, including that the state is committed to improving the quality of state academic standards and assessments. In addition, Education recently announced plans to make $4.35 billion in incentive grants available to states through SFSF on a competitive basis. These grants—referred to by Education as the Race to the Top program—can be used by states for, among other things, improving the quality of assessments.

Like other students, those with disabilities must be included in statewide ESEA assessments. This is accomplished in different ways, depending on the effects of a student's disability. Most students with disabilities participate in the regular statewide assessment either without accommodations or with appropriate accommodations, such as having unlimited time to complete the assessments, using large print or Braille editions of the assessments, or being provided individualized or small group administration of the assessments. States are permitted to use alternate academic achievement standards to evaluate the performance of students with the most significant cognitive disabilities. Alternate achievement standards must be linked to the state's grade-level academic content standards but may include prerequisite skills within the continuum of skills culminating in grade-level proficiency. For these

[8]Pub. L. No. 111-5.

students, a state must offer alternate assessments that measure students' performance. For example, the alternate assessment might assess students' knowledge of fractions by splitting groups of objects into two, three, or more equal parts. While alternate assessments can be administered to all eligible children, the number of proficient and advanced scores from alternate assessments based on alternate achievement standards included in Adequate Yearly Progress (AYP)[9] decisions generally is limited to 1 percent of the total tested population at the state and district levels.[10] In addition, states may develop modified academic achievement standards—achievement standards that define proficiency at a lower level than the achievement standards used for the general assessment population, but are still aligned with grade-level content standards—and use alternate assessments based on those standards for eligible students whose disabilities preclude them from achieving grade-level proficiency within the same period of time as other students. States may include scores from such assessments in making AYP decisions but those scores generally are capped at 2 percent of the total tested population.[11]

States are also required to include LEP students in their ESEA assessments. To assess these students, states have the option of developing assessments in students' native languages. These assessments are designed to cover the content in state academic content standards at the same level of difficulty and complexity as the general assessments.[12] In the absence of native language assessments, states are required to provide testing accommodations for LEP students, such as providing additional time to complete the test, allowing the use of a dictionary, administering assessments in small groups, or simplified instructions.

[9]Adequate Yearly Progress is a measure of year-to-year student achievement under ESEA. AYP is used to make determinations about whether or not schools or school districts have met state academic proficiency targets. All schools and districts are expected to reach 100 percent proficiency by the 2013-14 school year.

[10]For the total number of students tested on each of the different types of assessment in 2007-08, see appendix II.

[11]The 2 percent of the scores being included in AYP using the alternate assessment based on modified academic achievement standards is in addition to the one percent of the student population included with the alternate assessment based on alternate academic achievement standards.

[12]LEP students may only take assessments in their native language for a limited number of years.

By law, Education is responsible for determining whether or not states' assessments comply with statutory requirements. The standards and assessments peer review process used by Education to determine state compliance began under the 1994 reauthorization of ESEA and is an ongoing process that states go through whenever they develop new assessments. In the first step of the peer review process, a group of at least three experts—peer reviewers—examines evidence submitted by the state to demonstrate compliance with NCLBA requirements, identifies areas for which additional state evidence is needed, and summarizes their comments. The reviewers are state assessment directors, researchers, and others selected for their expertise in assessments. After the peer reviewers complete their review, an Education official assigned to the state reviews the peer reviewers' comments and the state's evidence and, using the same guidelines as the peer reviewers, makes a recommendation on whether the state meets, partially meets, or does not meet each assessment system critical element and on whether the state's assessment system should be approved. A group of Education officials from the relevant Education offices—including a representative from the Office of the Assistant Secretary of Elementary and Secondary Education—meet as a panel to discuss the findings. The panel makes a recommendation about whether to approve the state and the Assistant Secretary makes the final approval decision. Afterwards a letter is sent to the state notifying them of whether they have been approved, and—if the state was not approved—Education's letter identifies why the state was not approved. States also receive a copy of the peer reviewers' written comments as a technical assistance tool to support improvement.

Education has the authority to withhold federal funds provided for state administration until it determines that the state has fulfilled ESEA assessment requirements and has taken this step with several states since NCLBA was enacted. Education also provides states with technical assistance in meeting the academic assessment requirements.

ESEA assessments must be valid and reliable for the purposes for which they are intended and aligned to challenging state academic standards. Education has interpreted these requirements in its peer review guidance to mean that states must show evidence of technical quality—including validity and reliability—and alignment with academic standards. According to Education's peer review guidance, the main consideration in determining validity is whether states have evidence that their assessment results can be interpreted in a manner consistent with their intended purposes. See appendix III for a complete description of the evidence used by Education to determine validity.

A reliable assessment, according to the peer review guidance, minimizes the many sources of unwanted variation in assessment results. To show evidence of consistency of assessment results, states are required to (1) make a reasonable effort to determine the types of error that may distort interpretations of the findings, (2) estimate the likely magnitude of these distortions, and (3) make every possible effort to alert the users to this lack of certainty. As part of this requirement, states are required to demonstrate that assessment security guidelines are clearly specified and followed. See appendix IV for a full description of the reliability requirements.

Alignment, according to Education's peer review guidance, means that states' assessment systems adequately measure the knowledge and skills specified in state academic content standards. If a state's assessments do not adequately measure the knowledge and skills specified in its content standards or if they measure something other than what these standards specify, it will be difficult to determine whether students have achieved the intended knowledge and skills. See appendix V for details about the characteristics states need to consider to ensure that its standards and assessments are aligned.

In its guidance and peer review process, Education requires that—as one component of demonstrating alignment between state assessments and academic standards—states must demonstrate that their assessments are as cognitively challenging as their standards. To demonstrate this, states have contracted with organizations to assess the alignment of their ESEA assessments with the states' standards. These organizations have developed similar models of measuring the cognitive challenge of assessment items. For example, the Webb model categorizes items into four levels—depths of knowledge—ranging in complexity from level 1—recall, which is the least difficult for students to answer, to level 4—extended thinking, which is the most difficult for students to answer. Table 1 provides an illustration, using the Webb model, of how depth of knowledge levels may be measured.

Table 1: Illustration of Depth of Knowledge Levels

Depth of knowledge level	Description
Level 1 – Recall	Includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. Other key words that signify a Level 1 activity include "identify," "recall," "recognize," "use," and "measure."
Level 2 – Skill/ Concept	Includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. Other Level 2 activities include noticing and describing non-trivial patterns; explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.
Level 3 – Strategic Thinking	Requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is a Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. Other Level 3 activities include drawing conclusions from observations, citing evidence and developing a logical argument for concepts, explaining phenomena in terms of concepts, and using concepts to solve problems.
Level 4 – Extended Thinking	Requires complex reasoning, planning, developing, and thinking most likely over an extended period of time. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas within the content area or among content areas—and would have to select one approach among many alternatives on how the situation should be solved, in order to be at this highest level. Level 4 activities include developing and proving conjectures; designing and conducting experiments; making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs.

Source: Norman L. Webb, Issues Related to Judging the Alignment of Curriculum Standards and Assessments, April 2005.

States Reported That Assessment Spending Has Increased Since NCLBA Was Enacted and Test Development Has Been the Largest Assessment Cost in Most States

Assessment Expenditures Have Grown in Nearly Every State since 2002, and Most States Reported Spending More for Vendors than State Staff

State ESEA assessment expenditures have increased in nearly every state since the enactment of NCLBA in 2002, and the majority of these states reported that adding assessments was a major reason for the increased expenditures. Forty-eight of 49 states that responded to our survey said their states' overall annual expenditures for ESEA assessments have increased, and over half of these 48 states indicated that adding assessments to their state assessment systems was a major reason for increased expenditures.[13] In other cases, even states that were testing students in reading/language arts and mathematics in all of the grades that were required when NCLBA was enacted reported that assessment expenditures increased due to additional assessments. For example, officials in Texas—which was assessing general population students in all of the required grades at the time NCLBA was enacted—told us that they created additional assessments for students with disabilities.

In addition to the cost of adding new assessments, states reported that increased vendor costs have also contributed to the increased cost of assessments. On our survey, increasing vendor costs was the second most frequent reason that states cited for increased ESEA assessment costs. One vendor official told us that shortly after the 2002 enactment of NCLBA, states benefited from increased competition because many new vendors entered the market and wanted to gain market share, which drove down prices. In addition, vendors were still learning about the level of effort and costs required to complete this type of work. Consequently, as

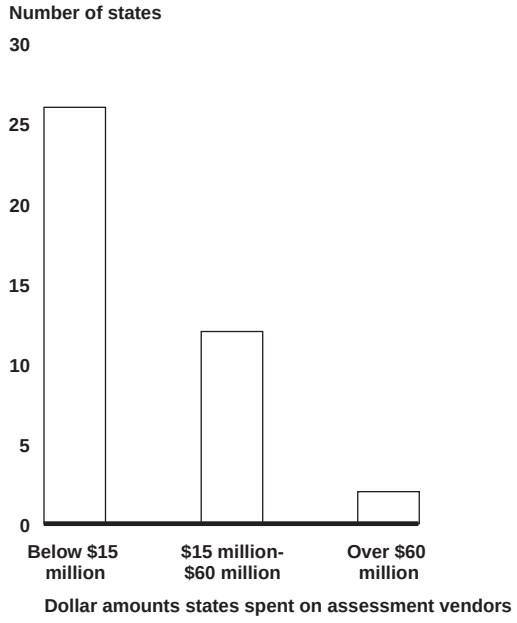
[13]GAO's 2003 report (GAO-03-389) found that item type has a major influence on overall state expenditures for assessments. However, regarding the changes to state expenditures for assessments since the enactment of NCLBA—which our survey examined—few states reported that item type was a major factor.

the ESEA assessment market has stabilized and vendors have gained experience pricing assessments, the cost of ESEA assessment contracts have increased to reflect the true cost of vendor assessment work. One assessment vendor that works with over half of the states on ESEA assessments told us that vendor costs have also been increasing as states have been moving toward more sophisticated and costly procedures and reporting.

Nearly all states reported higher expenditures for assessment vendors than for state assessment staff. According to our survey responses, 44 out of the 46 states that responded said that of the total cost of ESEA assessments, much more was paid to vendors than to state employees. For example, one state reported it paid approximately $83 million to vendors and approximately $1 million to state employees in the 2007-08 school year. The 20 states that provided information for the costs of both vendors and state employees in 2007-08 reported spending more than $350 million for vendors to develop, administer, score, and report the results of ESEA assessments—more than 10 times the amount they spent on state employees.

State expenditures for ESEA assessment vendors, which were far larger than expenditures for state staff, varied. Spending for vendors on ESEA assessments in the 40 states that reported spending figures on our survey ranged from $500,000 to $83 million, and in total all 40 states spent more than $640 million for vendors to develop, administer, score, and report results of the ESEA assessments in 2007-08. The average cost in these 40 states was about $16 million. See figure 2 for the distribution of state expenditures for vendors in 2007-08.

Figure 2: State Expenditures for Assessment Vendors, 2007-08



Source: GAO survey.

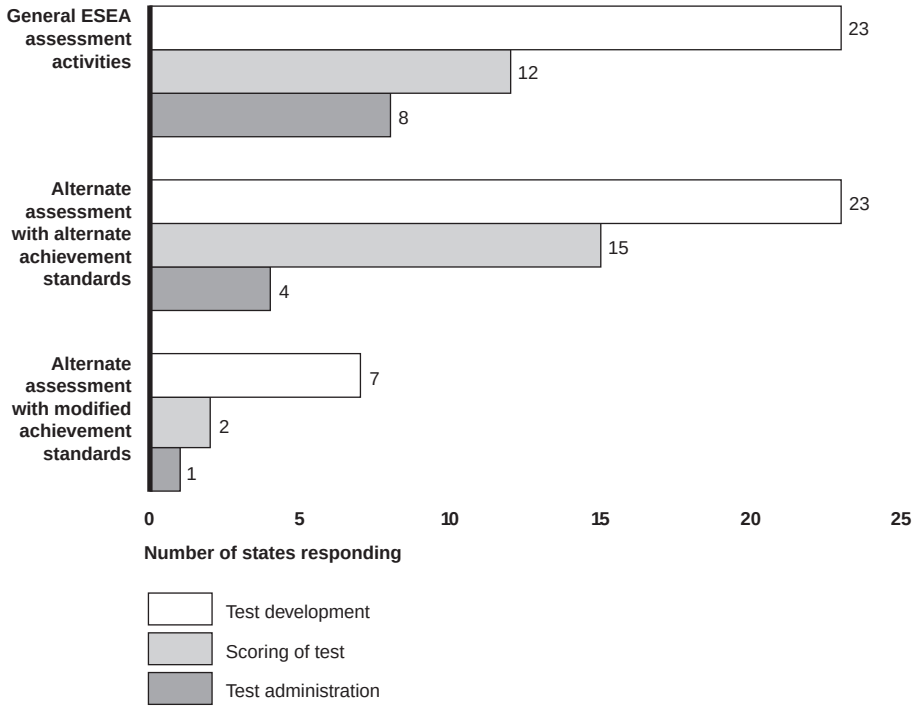
Over half of the states reported that the majority of their funding for ESEA assessments—including funding for expenses other than vendors—came from their state governments. Of the 44 states that responded to the survey question, 26 reported that the majority of their state's total funding for ESEA assessments came from state government funds for 2007-08, and 18 reported that less than half came from state funds. For example, officials from one state that we visited, Maryland, reported that 84 percent of their total funding for ESEA assessments came from state government funds and that 16 percent of the state's funding for ESEA assessments came from the federal Grants for State Assessments program in 2007-08. In addition to state funds, all states reported using Education's Grants for State Assessments for ESEA assessments, and 17 of 45 states responding to the survey question reported using other federal funds for assessments. One state reported that all of its funding for ESEA assessments came from the Grants for State Assessments program. The other federal funds used by states for assessments included Enhanced Assessment grants.

The Majority of States Reported That Assessment Development Was the Most Expensive Component of the Assessment Process; Development Has Been More Challenging for Small States

More than half of the states reported that assessment development costs were more expensive than any other component of the student assessment process, such as administering or scoring assessments.[14] Twenty-three of 43 states that responded to the question in our survey told us that test and item development and revision was the largest assessment cost for 2007-08. For example, Texas officials said that the cost of developing tests is higher than the costs associated with any other component of the assessment process. After test and item development costs, scoring was most frequently cited as the most costly activity, with 12 states reporting it as their largest assessment cost. Similarly, states reported that test and item development was the largest assessment cost for alternate assessments, followed by scoring. See figure 3 for more information.

[14]We asked states to rank the cost of test/item development, scoring, administration, reporting test results, data management, and all other assessment activities.

Figure 3: ESEA Assessment Activities That Received the Largest Share of States' Total ESEA Assessment Costs, 2007-08



Source: GAO survey data.

The cost of developing assessments was affected by whether states release assessment items to the public.[15] According to state and vendor officials, development costs are related to the percentage of items states release to the public every year because new items must be developed to replace released items. According to vendor officials, nearly all states release at least some test items to the public, but they vary in the percentage of items that they release. In states that release 100 percent of their test items each year, assessment costs are generally high and steady over time because states must develop additional items every year. However, some states release only a portion of items. For example, Rhode Island state officials told us that they release 20 to 50 percent of their reading and math

[15]Although GAO-03-389 found that item type was a key factor in determining the overall cost of state ESEA assessments, these differences were related to the cost of scoring assessments rather than developing assessments. Our research did not find that item type affected the cost of development.

assessment items every year. State and vendor officials told us that despite the costs associated with the release of ESEA assessment items, releasing assessment items builds credibility with parents and helps policymakers and the public understand how assessment items relate to state content standards.

The cost of development has been particularly challenging for smaller states.[16] Assessment vendors and Education officials said that the price of developing an assessment is fixed regardless of state size and that, as a result smaller states with fewer students usually have higher per pupil costs for development. For example, state assessment officials from South Dakota told us that their state and other states with small student populations have the same development costs as states with large assessment populations, regardless of the number of students being assessed. In contrast to development costs, administration and scoring costs vary based on the number of students being assessed and the item types used. Although large and small states face similar costs for development, each has control over some factors—such as item type and releasing test items—that can increase or decrease costs.

Selected States Are Concerned about Costs of Developing and Administering Alternate Assessments for Students with Disabilities and Budget Cuts

State officials from the four states we visited told us that alternate assessments based on alternate achievement standards were far more expensive on a per pupil basis than general assessments. In Maryland, state officials told us that general assessments cost $30 per pupil, and alternate assessments cost between $300 and $400 per pupil. Rhode Island state officials also reported that alternate assessments cost much more than general assessments. These officials also said that, in addition to direct costs, the administration of alternate assessments has resulted in significant indirect costs, such as professional development for teachers. Technical advisors and district and state officials told us that developing alternate assessments is costly on a per pupil basis because the number of students taking these assessments is small. See appendix VI for more information about states' use of various item types for alternate assessments.

In light of recent economic conditions, many states have experienced fiscal reductions, including within ESEA assessment budgets. As of

[16]We defined small states as those states administering 500,000 or fewer ESEA assessments in 2007-08. Reading/language arts and mathematics assessments were counted separately.

January 2009, 19 states said their state's total ESEA assessment budget had been reduced as a result of state fiscal cutbacks. Fourteen states said their state's total ESEA assessment budgets had not been reduced, but 10 of these states also said they anticipated future reductions. Half of the 46 states that responded to the question told us that in developing their budget proposals for the next fiscal year they anticipated a reduction in state funds for ESEA assessments. For example, one state that responded to our survey said it had been asked to prepare for a 15 percent reduction in state funds.

States Have Considered Cost and Time in Making Decisions about Assessment Item Type and Content

States Used Primarily Multiple Choice Items in Their ESEA Assessments Because They Are Cost-Effective and Can Be Scored within Tight Time Frames for Reporting Results

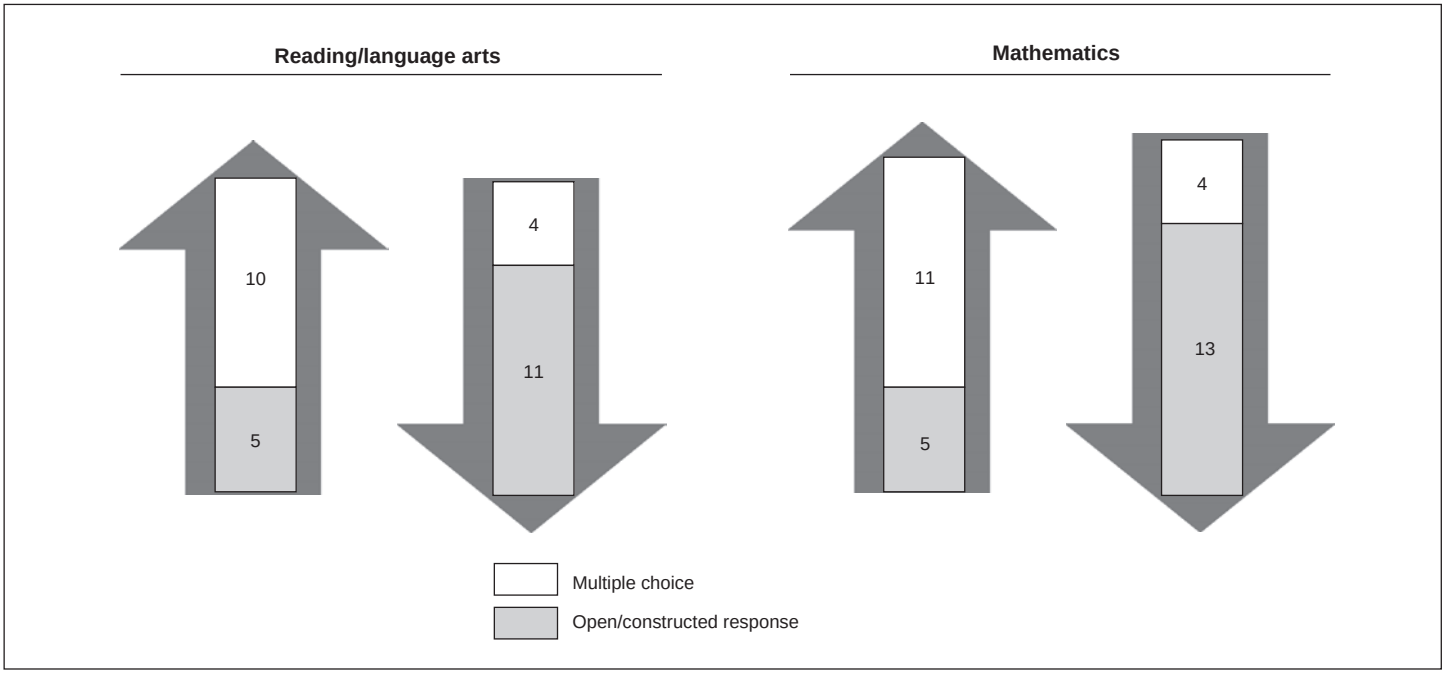
States have most often chosen multiple choice items over other item types on assessments. In 2003, we reported that the majority of states used a combination of multiple choice and a limited number of open-ended items for their assessments.[17] According to our survey, multiple choice items comprise the majority of unweighted score points (points)—the number of points that can be earned based on the number of items answered correctly—for ESEA reading/language arts and mathematics general assessments administered by most responding states. Specifically, 38 of 48 states that responded said that multiple choice items comprise all or most of the points for their reading/ language arts assessments, and 39 states said that multiple choice items comprise all or most of the points for mathematics assessments. Open/constructed response items are the second most frequently used item type for reading/language arts or mathematics general assessments. All states that responded to our survey reported using multiple choice items on their general reading/language arts and mathematics assessments, and most used some open/constructed

[17]GAO, *Title I: Characteristics of Tests Will Influence Expenses; Information Sharing May Help States Realize Efficiencies*, GAO-03-389 (Washington, D.C.: May 8, 2003).

response items. See appendix VI for more information about the types of items used by states on assessments.

Some states also reported on our survey that, since 2002, they have increased their use of multiple choice items and decreased their use of other item types. Of the 47 states that responded to our survey question, 10 reported increasing the use of multiple choice items on reading/language arts general assessments, and 11 reported increasing their use of multiple choice items on mathematics assessments. For example, prior to the enactment of NCLBA, Maryland administered an assessment that was fully comprised of open/constructed response items, but state assessment officials told us that they have moved to an assessment that is primarily multiple choice and plan to eliminate open/constructed response items from assessments. However, several states reported that they have decreased the use of multiple choice items and/or increased the use of open/constructed response items. For more information about how states reported changing the mix of items on their assessments, see figure 4.

Figure 4: The Number of States Reporting Changes in Item Type Use on ESEA Assessments since 2002



Source: GAO.

States reported that total cost of use and the ability to score assessments quickly were key considerations in choosing multiple choice item types. In response to our survey, most states reported considering the cost of different item types and the ability to score the tests quickly when making decisions about item types for ESEA assessments. Officials from the states we visited reported choosing multiple choice items because they can be scored inexpensively within challenging time frames. State officials, assessment experts, and vendors told us that multiple choice item types are scored electronically, which is inexpensive, but that open/constructed response items are usually scored manually, making them more expensive to score. Multiple scorers of open/constructed response items are sometimes involved to ensure consistency, but this also increases costs. In addition, state officials said that training scorers of open/constructed response items is costly. For example, assessment officials in Texas told us that the state has a costly 3-week long training process for teachers to become qualified to assess the open-ended responses. State assessment officials also told us that they used multiple choice items because they can be scored quickly, and assessment vendors reported that states were under pressure to release assessment results to the public before the beginning of the next school year in accordance with NCLBA requirements. For example, assessment officials from South Dakota told us that they explored using open/constructed response items on their assessments but that they ultimately determined it would not be feasible to return results in the required period of time. States also reported considering whether item types would meet certain technical considerations, such as validity and reliability. Texas assessment officials said that using multiple choice items allows the state more time to check test scores for reliability.

States Reported That the Use of Multiple Choice Items in Assessments Has Limited the Content and Complexity of What They Test

Despite the cost- and time-saving benefits to states, the use of multiple choice items on assessments has limited the content included in the assessments. Many state assessment officials, alignment experts, and vendor officials told us that items possess different characteristics that affect how amenable they are to testing various types of content. State officials and their technical advisors told us that they have faced significant trade-offs between their efforts to assess highly cognitively complex content and their efforts to accommodate cost and time pressures. All four of the states that we visited reported separating at least a minor portion of standards into those that are used for ESEA assessment and those that are for instructional purposes only. Three of the four states reported that standards for instructional purposes only included highly cognitively complex material that could not be assessed using multiple

choice items. For example, a South Dakota assessment official told us that a cognitively complex portion of the state's new reading standards could not be tested by multiple choice; therefore, the state identified these standards as for instructional purposes only and did not include them in ESEA assessments. In addition to these three states, officials from the fourth state—Maryland—told us that they do not include certain content in their standards because it is difficult to assess. Many state officials and experts we spoke with told us that multiple choice items limit states from assessing highly cognitively complex content. For example, Texas assessment officials told us that some aspects of state standards, such as a student's ability to conduct scientific research, cannot be assessed using multiple choice.

Representatives of the alignment organizations told us that it is difficult, and in some cases not possible, to measure highly cognitively complex content with multiple choice items. Three of the four main groups that conduct alignment studies, including alignment studies for all of our site visit states, told us that states cannot measure content of the highest complexity with multiple choice and that ESEA assessments should include greater cognitive complexity. Maryland state officials said that before NCLBA was enacted the state administered an assessment that was fully comprised of open/constructed response items. Maryland technical advisors told us that because the state faced pressure to return assessment results quickly, the state changed its test to include mostly multiple choice items, but that this had limited the content assessed in the test. According to an analysis performed in 2002 after the enactment of NCLBA, of 36 scorable items on one Maryland high school mathematics assessment, about 94 percent of the items were rated at the two lowest levels of cognitive demand, out of four levels based on an independent alignment review.[18, 19] Representatives of all four alignment groups told us that multiple choice items can measure intermediate levels of cognitive complexity, but it is difficult and costly to develop these items. These alignment experts said that developing multiple choice items that measure

[18]This does not necessarily indicate that state assessments were not aligned to state standards. For example, if the content in standards does not include the highest cognitive level, assessments that do not address the highest cognitive level could be aligned to standards.

[19]The alignment review was conducted by Achieve, Inc., which was one of the four alignment organizations that we interviewed.

cognitively challenging content is more expensive and time-consuming than for less challenging multiple choice items.

Vendor officials had differing views about whether multiple choice items assess cognitively complex content. For example, officials from three vendors said that multiple choice items can address cognitively complex content. However, officials from another vendor told us that it is not possible to measure certain highly cognitively complex content with multiple choice items. Moreover, two other vendors told us that there are certain content and testing purposes that are more amenable to assessment with item types other than with multiple choice items. Several of the vendors reported that there are some standards that, because of practical limitations faced by states, cannot be assessed on standardized, paper-and-pencil assessments. For example, one vendor official told us that performance-based tasks enabled states to assess a wider variety of content but that the limited funds and quick turnaround times required under the law require states to eliminate these item types.

Although most state officials, state technical advisors, and alignment experts said that ESEA assessments should include more open/constructed response items and other item types, they also said that multiple choice items have strengths and that there are challenges with other types of items. For example, in 2008 a national panel of assessment experts appointed and overseen by Education reported that multiple choice items do not measure different aspects of mathematics competency than open/constructed response items. Also, alignment experts said that multiple choice items can quickly and effectively assess lower level content, which is also important to assess. Moreover, open/constructed response items do not always assess highly complex content, according to an alignment expert. This point has been corroborated by several researchers who have found that performance tasks, which are usually intended to assess higher-level cognitive content may inadvertently measure low-level content.[20] For example, one study describes a project in which students were given a collection of insects and asked to organize them for display. High-scoring students were supposed to demonstrate complex thinking skills by sorting insects based on scientific classification systems, rather than less complex criteria, such as whether or not insects

[20]Committee on the Foundations of Assessment, James W. Pellegrino, Naomi Chudowsky, and Robert Glaser, editors, *Knowing What Students Know: The Science and Design of Educational Assessment* (Washington, D.C.: National Academy Press, 2001) 194.

are able to fly. However, analysis of student responses showed that high scorers could not be distinguished from low scorers in terms of their knowledge of the insects' features or of the scientific classification system.[21]

The presence or absence of highly complex content in assessments can impact classroom curriculum. Several research studies have found that content contained in assessments influences what teachers teach in the classroom. One study found that including open-ended items on an assessment prompted teachers to ask students to explain their thinking and emphasize problem solving more often.[22] Assessment experts told us that the particular content that is tested impacts classroom curriculum. For example, one assessment expert told us that the focus on student results, combined with the focus on multiple choice items, has led to teachers teaching a narrow curriculum that is focused on basic skills.

Under the federal peer review process, Education and peer reviewers examined evidence that ESEA assessments are aligned with the state's academic standards. Specifically, peer reviewers examined state evidence that assessments cover the full depth and breadth of the state academic standards in terms of cognitive complexity and level of difficulty. However, consistent with federal law, it is Education's policy not to directly examine a state's academic standards, assessments, or specific test items.[23] Education officials told us that it is not the department's role to evaluate standards and assessments themselves and that few at Education have the expertise that would be required to do so. Instead, they explained that Education's role is to evaluate the evidence provided by states to determine whether the necessary requirements are met.

[21]Gail P. Baxter and Robert Glaser, "Investigating the Cognitive Complexity of Science Assessments," *Educational Measurement: Issues and Practice*, vol. 17, no. 3 (1998).

[22]Helen S. Apthorp, et al., "Standards in Classroom Practice Research Synthesis," Mid-Continent Research for Education and Learning (October 2001).

[23]For example, see 20 U.S.C. § 7907(c)(1) and 20 U.S.C. § 6575.

States Used Alternative Practices to Reduce Cost and Meet Quick Turnaround Times while Attempting to Assess Complex Material

As an alternative to using mostly multiple choice items on ESEA assessments, states used a variety of practices to reduce costs and meet quick turnaround times while also attempting to assess cognitively complex material. For example, some states have developed and administered ESEA assessments in collaboration with other states, which has allowed these states to pool resources and use a greater diversity of item types. In addition, some states administered assessments at the beginning of the year that test students on material taught during the prior year to allow additional time for scoring of open-response items, or administered assessments online to decrease turnaround time for reporting results. States have reported advantages and disadvantages associated with each of these practices:

- **Collaboration among states:** All four states that we visited—Maryland, Texas, South Dakota, and Rhode Island—indicated interest in collaborating with other states in the development of ESEA reading/language arts or mathematics assessments, as of March 2009, but only Rhode Island was. Under the New England Common Assessments Program (NECAP), Rhode Island, Vermont, New Hampshire, and Maine share a vendor, a common set of standards, and item development costs. Under this agreement, the cost of administration and scoring are based on per pupil rates. NECAP states use a combination of multiple choice, short answer, and open/constructed response items. According to Rhode Island assessment officials, more rigorous items, including half of their math items, are typically embedded within open/constructed response items.

  When asked about the benefits of working in collaboration with other states to develop ESEA assessments, assessment officials for Rhode Island told us that the fiscal savings are very apparent. Specifically, they stated that Rhode Island will save approximately $250,000 per year with the addition of Maine to the NECAP consortium because, as Rhode Island assessment officials noted, Maine will take on an additional share of item development costs. Also, officials said that with a multi-state partnership, Rhode Island is able to pay more for highly skilled people who share a common vision. Finally, they said that higher standards are easier to defend politically as part of collaboration because there are more stakeholders in favor of them. An assessment expert from New Hampshire said that the consortium has been a “lifesaver” because it has saved the state considerable funding and allowed it to meet ESEA assessment requirements.

  Assessment experts from Rhode Island and New Hampshire told us that there are some challenges to working in collaboration with other states to develop ESEA assessments. Because decisions are made by consensus and

the NECAP states have philosophical differences in areas such as item development, scoring, and use of item types, decision-making is a lengthy process. In addition, a Rhode Island official said that assessment leadership in the states changes frequently, which also makes decision-making difficult.

- **Beginning of year test administration:** NECAP states currently administer assessments in the beginning of the year, which eases time pressures associated with the scoring of open/constructed response items. As a result, the inclusion of open/constructed response items on the assessment has been easier because there is enough time to meet NCLBA deadlines for reporting results. However, Rhode Island officials said that there are challenges to administering tests at the beginning of the year. For example, one official stated that coordinating testing with the already challenging start of school is daunting. For example, she said that state assessment officials are required to use school enrollment lists to print school labels for individual tests, but because enrollment lists often change in the beginning of the year, officials are required to correct a lot of data. District assessment officials also cited this as a major problem.

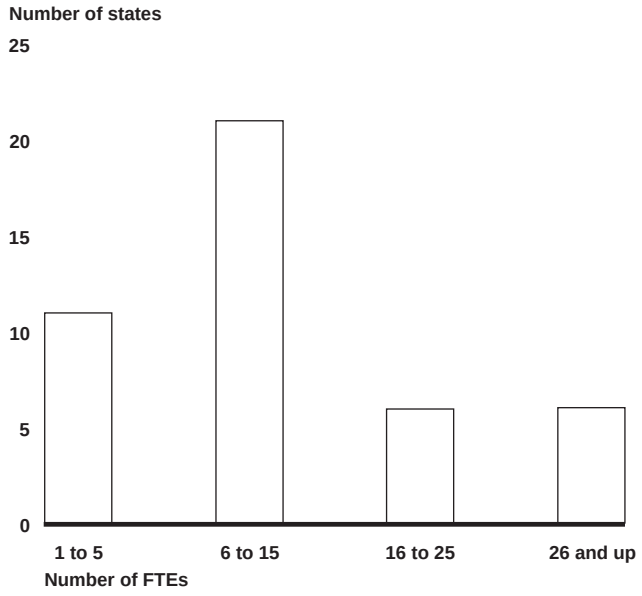
- **Computerized testing:** Of the states we visited, Texas was the only one administering a portion of its ESEA assessments online, but Maryland and Rhode Island were moving toward this goal. One assessment vendor with whom we spoke said that many states are anticipating this change in the not-too-distant future. Assessment vendors and state assessment officials cited some major benefits of online assessment. For example, one vendor told us that online test administration reduces costs by using technology for automated scoring. They also told us that states are using online assessments to address cognitively complex content in standards that are difficult to assess, such as scientific knowledge that is best demonstrated through experiments. In addition, assessment officials told us that online assessments are less cumbersome and easier than paper tests to manage at the school level if schools have the required technology and that they enable quicker turnaround on scores. State and district assessment officials and a vendor with whom we spoke also cited several challenges associated with administering tests online, including security of the tests; variability in students' computer literacy; strain on school computer resources, computer classrooms/labs, and interruption of classroom/lab instruction; and lack of necessary computer infrastructure.

States Faced Several Challenges in Their Efforts to Ensure Valid and Reliable ESEA Assessments, including Staff Capacity, Alternate Assessments, and Assessment Security

States Varied in Their Capacity to Guide and Oversee Vendors

State officials are responsible for guiding the development of the state assessment program and overseeing vendors, but states varied in their capacity to fulfill these roles. State officials reported that they are responsible for making key decisions about the direction of their states' assessment programs, such as whether to develop alternate assessments based on modified achievement standards, or online assessments. In addition, state officials said that they are responsible for overseeing the assessment vendors used by their states. However, state assessment offices varied based on the measurement expertise of their staff. About three-quarters of the 48 responding states had at least one state assessment staff member with a Ph.D. in psychometrics or another measurement-related field. Three states—North Carolina, South Carolina, and Texas—each reported having five staff with this expertise. However, 13 states did not have any staff with this expertise. In addition, states varied in the number of full-time equivalent professional staff (FTE) dedicated to ESEA assessments from 55 professional staff in Texas to 1 professional staff in Idaho and the District of Columbia. See figure 5 for more information about the number of FTEs dedicated to ESEA in the states.

Figure 5: Number of FTEs Dedicated to ESEA Assessments in States, 2007-08



Source: GAO survey.

Small states had less assessment staff capacity than larger states. The capacity of state assessment offices was related to the amount of funding spent on state assessment programs in different states, according to state officials. For example, South Dakota officials told us that they had tried to hire someone with psychometric expertise but that they would need to quadruple the salary that they could offer to compete with the salaries being offered by other organizations. State officials said that assessment vendors can often pay higher salaries than states and that it is difficult to hire and retain staff with measurement-related expertise.

State officials and assessment experts told us that the capacity of state assessment offices was the key challenge for states implementing NCLBA. Greater state capacity allows states to be more thoughtful in developing their state assessment systems, and provide greater oversight of their assessment vendors, according to state officials. Officials in Texas and other states said that having high assessment staff capacity—both in terms of number of staff and measurement-related expertise—allows them to research and implement practices that improve student assessment. For example, Texas state officials said that they conduct research regarding how LEP students and students with disabilities can best be included in ESEA assessments, which state officials said helped them improve the

state's assessments for these students. In contrast, officials in lower capacity states said that they struggled to meet ESEA assessment requirements and did not have the capacity to conduct research or implement additional strategies. For example, officials in South Dakota told us that they had not developed alternate assessments based on modified achievement standards because they did not have the staff capacity or funding to implement these assessments.

Also, of three states we visited that completed a checklist of important assessment quality control steps,[24] those with fewer assessment staff addressed fewer key quality control steps. Specifically, Rhode Island, South Dakota, and Texas reviewed and completed a CCSSO[25] checklist on student assessment, the Quality Control Checklist for Processing, Scoring, and Reporting. These states varied with regard to fulfilling the steps outlined by this checklist. For example, state officials in Texas, which has 55 full-time professional staff working on ESEA assessments, including multiple staff with measurement-related expertise, reported that they fulfill 31 of the 33 steps described in the checklist and address the 2 other steps in certain circumstances. Officials in Rhode Island, who told us that they have six assessment staff and work in conjunction with other states in its assessment consortium, said that they fulfill 27 of the 33 steps. South Dakota, which had three professional full-time staff working on ESEA assessments—and no staff with measurement-related expertise—addressed nine of the steps, according to state officials. For example, South Dakota officials said that the state does not verify the accuracy of answer keys in the data file provided by the vendor using actual student responses, which increases the risk of incorrectly scoring assessments. Because South Dakota does not have staff with measurement-related expertise and has fewer state assessment staff, there are fewer individuals to fulfill these quality control steps than in a state with greater capacity, according to state officials.

[24]Maryland did not complete this checklist.

[25]CCSSO is an association of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five extra-state jurisdictions. It provides advocacy and technical assistance to its members. The CCSSO checklist describes 33 steps that state officials should take to ensure quality control in assessment programs that are used to make decisions with consequences for students or schools. The checklist can be found at http://www.ccsso.org.

Having staff with psychometric or other measurement-related expertise improved states' ability to oversee the work of vendors. For example, the CCSSO checklist recommends that states have psychometric or other research expertise for nearly all of the 33 steps. Having staff with measurement-related expertise allows states to know what key technical questions or data to ask of vendors, according to state officials, and without this expertise they would be more dependent on vendors. State advisors from technical advisory committees (TAC)—panels of assessment experts that states convene to assist them with technical oversight—said that TACs are useful, but that they generally only meet every 6 months. For example, one South Dakota TAC member said that TACs can provide guidance and expertise, but that ensuring the validity and reliability of a state assessment system is a full-time job. The TAC member said that questions arise on a regular basis for which it would be helpful to bring measurement-related expertise to bear. Officials from assessment vendors varied in what they told us. Several told us that states do not need measurement-related expertise, but others said that states needed this expertise on staff.

Education's Inspector General (OIG) found reliability issues with management controls over state ESEA assessments.[26] Specifically, the OIG found that Tennessee did not have sufficient monitoring of contractor activities for the state assessments such as ensuring that individuals scoring open/constructed response items had proper qualifications. In addition, the OIG found that the state lacked written policies and procedures describing internal controls for scoring and reporting.

States Have Faced Challenges in Ensuring the Validity and Reliability of Alternate Assessments for Students with Disabilities

Although most states have met peer review expectations for validity and reliability of their general assessments, ensuring the validity of alternate assessments for students with disabilities is still a challenge. For example, our review of Education documents as of July 15, 2009, showed that 12 states' reading/language arts and mathematics standards and assessment systems—which include general assessments and alternate assessments based on alternate achievement standards—had not received full approval

[26]U.S. Department of Education, Office of the Inspector General, *Tennessee Department of Education Controls Over State Assessment Scoring*, ED-OIG/A02I0034 (New York, N.Y.: May 2009).

under Education's peer review process and that alternate assessments were a factor preventing approval in 11 of these states.[27]

In the four states[28] where alternate assessments were the only issue preventing full approval, technical quality (which includes validity and reliability) or alignment was a problem. For example, in a letter to Hawaii education officials dated October 30, 2007, documenting steps the state must take to gain full approval of its standards and assessments system, Education officials wrote that Hawaii officials needed to document the validity and alignment of the state alternate assessment.

States had more difficulty assessing the validity and reliability of alternate assessments using alternate achievement standards than ESEA assessments for the general student population. In our survey, nearly two-thirds of the states reported that assessing the validity and reliability of alternate assessments with alternate achievement standards was either moderately or very difficult. In contrast, few states reported that either validity or reliability were moderately or very difficult for general assessments.

We identified two specific challenges to the development of valid and reliable alternate assessments with alternate achievement standards. First, ensuring the validity and reliability of these alternate assessments has been challenging because of the highly diverse population of students being assessed. Alternate assessments are administered to students with a wide range of significant cognitive disabilities. For example, some students may only be able to communicate by moving their eyes and blinking. As a result, measuring the achievement of these students often requires greater individualization. In addition, because these assessments are administered to relatively small student populations, it can be difficult for states to gather the evidence needed to demonstrate their validity and reliability.

In addition, developing valid and reliable alternate assessments with alternate achievement standards has been challenging for states because

[27]The 12 states that had not received full approval were California, the District of Columbia, Florida, Hawaii, Michigan, Mississippi, Nebraska, Nevada, New Hampshire, New Jersey, Vermont, and Wyoming. In all of these states except California the alternate assessments based on alternate achievement standards were a factor preventing full approval.

[28]The four states were Florida, New Hampshire, New Jersey, and Vermont.

there is a lack of research about the development of these assessments, according to state officials and assessment experts. States have been challenged to design alternate assessments that appropriately measure what eligible students know and provide similar scores for similar levels of performance. Experts and state officials told us that more research would help them ensure validity and reliability. An Education official agreed that alternate assessments are still a challenge for states and said that there is little consensus about what types of alternate assessments are psychometrically appropriate. Although there is currently a lack of research, Education is providing assistance to states with alternate assessments and has funded a number of grants to help states implement alternate assessments.

States that have chosen to implement alternate assessments with modified achievement standards and native language assessments have faced similar challenges, but relatively few states are implementing these assessments. On our survey, 8 of the 47 states responding to this question reported that in 2007-08 they administered alternate assessments based on modified achievement standards, which are optional for states, and several more reported being in the process of developing these assessments. Fifteen states reported administering native language assessments, which are also optional. States reported mixed results regarding the difficulty of assessing the validity and reliability of these assessments, with about two-thirds indicating that each of these tasks was moderately or very difficult for both the alternate assessments with modified achievement standards and native language assessments. Officials in states that are not offering these assessments reported that they lacked the funds necessary to develop these assessments or that they lacked the staff or time.

States Have Taken Measures to Ensure Assessment Security, but Gaps Exist

The four states that we visited and districts in those states had taken steps to ensure the security of ESEA assessments. Each of the four states had a test administration manual that is intended to establish controls over the processes and procedures used by school districts when they administer the assessments. For example, the Texas test administration manual covered procedures for keeping assessment materials secure prior to administration, ensuring proper administration, returning student answer forms for scoring, and notifying administrators in the event of assessment irregularities. States also required teachers administering the assessments to sign forms saying that they would ensure security and had penalties for teachers or administrators who violated the rules. For example, South Dakota officials told us that teachers who breach the state's security measures could lose their teaching licenses.

Despite these efforts, there have been a number of documented instances of teachers and administrators cheating in recent years. For example, researchers in one major city examined the frequency of cheating by test administrators.[29] They estimated that at least 4 to 5 percent of the teachers and administrators cheated on student assessments by changing student responses on answer sheets, providing correct answers to students, or illegitimately obtaining copies of exams prior to the test date and teaching students using knowledge of the precise exam items. Further, the study found that teachers' and administrators' decisions about whether to cheat responded to incentives. For example, when schools faced the possibility of being sanctioned for low assessment scores, teachers were more likely to cheat. In addition, the study found that teachers in low-performing classrooms were more likely to cheat.

In our work, we identified several gaps in state assessment security policies. For example, assessment security experts said that many states do not conduct any statistical analyses of assessment results to detect indications of cheating. Among our site visit states, one state—Rhode Island—reported analyzing test results for unexpected gains in schools' performance. Another state, Texas, had conducted an erasure analysis to determine whether schools or classrooms had an unusually high number of erased responses that were changed to correct responses, possibly indicating cheating. These types of analysis were described as a key component of assessment security by security experts. In addition, we identified one specific state assessment policy where teachers had an opportunity to change test answers. South Dakota's assessment administration manual required classroom teachers to inspect all student answers to multiple choice items and darken any marks that were too light for scanners to read. Further, teachers were instructed to erase any stray marks, and ensure that, when a student had changed an answer, the unwanted response was completely erased. This policy provided teachers an opportunity to change the answers, and improve assessment results. South Dakota officials told us that they had considered taking steps to mitigate the potential for cheating, such as contracting for an analysis that would identify patterns of similar erasure marks that could indicate cheating, but that it was too expensive for the state.

[29]Brian A. Jacob and Steven D. Levitt, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *The Quarterly Journal of Economics* (August 2003).

States' assessment security policies and procedures were examined during Education's standards and assessments peer review process. According to Education's peer review guidance, which Education officials told us were the criteria used by peer reviewers to examine state assessment systems, states must demonstrate the establishment of clear criteria for the administration, scoring, analysis, and reporting components of state assessment systems. One example of evidence of adequate security procedures listed in the peer review guidance was that the state uses training and monitoring to ensure that people responsible for handling or administering state assessments properly protect the security of the assessments. Education indicated that a state could submit as evidence documentation that the state's test security policy and consequences for violating the policy are communicated to educators, and documentation of the state's plan for training and monitoring assessment administration. According to Education officials, similar indicators are included in Education's ongoing efforts to monitor state administration and implementation of ESEA assessment requirements.

Although test security was included as a component in the peer review process, we identified several gaps in how the process evaluated assessment security. The peer reviewers did not examine whether states used any type of data analysis to review student assessment results for irregularities. When we spoke with Education's director of student achievement and school accountability programs—who manages the standards and assessments peer review process—about how assessment security was examined in the peer review process, he told us that security was not a focus of peer review. The official indicated that the review already required a great deal of time and effort by reviewers and state officials and that Education had given a higher priority to other assessment issues. In addition, the state policy described above in which teachers darken marks or erase unwanted responses was approved through the peer review process.

The Education official who manages the standards and assessments peer review process told us that the peer review requirements, including the assessment security portion, were based on the Standards for Educational and Psychological Testing[30] when they were developed in 1999. The

[30]American Educational Research Association, American Psychological Association, National Council on Measurement in Education, *Standards for Education and Psychological Testing* (1999).

Standards provide general guidelines for assessment security, such as that test users have the responsibility of protecting the security of test materials at all times. However, they do not provide comprehensive best practices for assessment security issues. The Association of Test Publishers developed draft assessment security guidelines in 2007. In addition, in the spring 2010, the Association of Test Publishers and CCSSO plan to release a best practices guide for state departments of education that is expected to offer best practices for test security.

Education has made certain modifications to the peer review process but does not plan to update the assessment security requirements. Education updated the peer review protocols to address issues with the alternate assessment using modified achievement standards after those regulations were released. In addition, Education has made certain modifications to the process that were requested by states. However, Education officials indicated that they do not have plans to update the peer review assessment security requirements.

Education Has Provided Assistance to States, but the Peer Review Process Did Not Allow for Sufficient Communication

Education Provided Technical Assistance with Assessments, including Those for Students with Disabilities and LEP Students

Education provided technical assistance to states in a variety of ways. Education provided technical assistance through meetings, written guidance, user guides, contact with Education staff, and assistance from its Comprehensive Centers and Clearinghouses. In our survey, states reported they most often used written guidance and Education-sponsored meetings and found these helpful. States reported mixed results in obtaining assistance from Education staff. Some reported receiving consistent helpful support while others reported staff were not helpful or responsive. Relevant program offices within Education provided additional assistance as needed. For example, the Office of Special Education Programs provided assistance to states in developing alternate assessments for students with disabilities and the Office of English

Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students assisted states in developing their assessments for LEP students. In addition, beginning in 2002, Education awarded competitive Enhanced Assessment Grants to state collaboratives working on a variety of assessment topics such as developing valid and reliable assessments for students with disabilities and LEP students. For example, one consortium of 14 states and jurisdictions was awarded about $836,000 to investigate and provide information on the validity of accommodations for future assessments for LEP students with disabilities, a group of students with dual challenges. States awarded grants are required to share the outcomes of their projects with other states at national conferences; however, since these are multi-year projects, the results of many of them are not yet available.

Education's Peer Review Process Did Not Allow Direct Communication between States and Reviewers to Quickly Resolve Problems

Education's peer review process did not allow for direct communication between states and peer reviewers that could have more quickly resolved questions or problems that arose throughout the peer review process. After states submitted evidence of compliance with ESEA assessment requirements to Education, groups of three reviewers examined the materials and made recommendations to Education. To ensure the anonymity of the peer reviewers, Education did not permit communication between reviewers and state officials. Instead, Education liaisons periodically relayed peer reviewers' questions and comments to the states and then relayed answers back to the peer reviewers. Education officials told us the assurance of anonymity was an important factor in their ability to recruit peer reviewers who may not have felt comfortable making substantive comments on states' assessment systems if their identity was known.

However, the lack of direct communication resulted in miscommunication and prevented quick resolutions to questions arising during the peer review process. State officials and reviewers told us that there was not enough communication between states and reviewers during the process, preventing the quick resolution of questions that arose during the review process. For example, one state official reported on our survey that the lack of direct communication with peer reviewers led to misunderstandings that could have been readily resolved with a conversation with peer reviewers. A number of the peer reviewers who we surveyed provided similar information. For example, one said that the process was missing direct communication, which would allow state officials to provide immediate responses to the reviewers' questions. The Education official who manages the standards and assessments peer

review process recognized that the lack of communication, such as a state not understanding how to interpret peer reviewers' comments, created confusion. Two experts we interviewed about peer review processes in general said that communication between reviewers and state officials is critical to having an efficient process that avoids miscommunication and unnecessary work. State officials said that the peer review process was extensive and that miscommunication made it more challenging.

In response to states' concerns, Education has taken steps to improve the peer review process by offering states the option of having greater communication with reviewers after the peer review process is complete. However, the department has not taken action to allow direct communication between states and peer reviewers during the process to ensure a quick resolution to questions or issues that arise, preferring to continue its reliance on Education staff to relay information between states and peer reviewers and protecting the anonymity of the peer reviewers.

Reasons for Key Decisions Stemming from Education's Peer Review Process Were Not Communicated to States

In some cases, the final approval decisions made by Education, which has final decision-making authority, differed from the peer reviewers' written comments, but Education could not tell us how often this occurred. Education's panels assessed each state's assessment system using the same guidelines used by the peer reviewers, and agency officials told us that peer reviewers' comments carried considerable weight in the agency's final decisions. However, Education officials said that—in addition to peer reviewers' comments—they also considered other factors in determining whether a state should receive full approval, including the time needed by the state to come into compliance and the scope of the outstanding issues. Education and state officials told us that, in some cases, Education reached different decisions than the peer reviewers. For example, the Education official who manages the standards and assessments peer review process described a situation in which the state was changing its content standards and frequently submitting new documentation for its mathematics assessment as the new content standards were incorporated. Education officials told us the peer reviewers got confused by the documentation, but Education officials gave the state credit for the most recent documentation. However, Education could not tell us how often the agency's final decisions matched the written comments of the peer reviewers because it did not track this information.

In cases in which Education's final decisions differed from the peer reviewers' comments, Education did not explain to states why it reached

its decisions. Although Education released the official decision letters describing reasons that states had not been approved through peer review, the letters did not document whether their decisions differed from the peer reviewers' comments or why their decisions were different. Because Education did not communicate this to states, it was unclear to states how written peer reviewer comments related to Education's decisions about peer review approval. For example, in our survey, one state reported that the comments provided to the state by peer reviewers and the letters sent to the state by Education describing their final decisions about approval status did not match.

State officials we interviewed reported confusion about what issues needed to be addressed to receive full approval of their assessment system. For example, some state officials reported confusion about how to receive final peer review approval when the written summary of the peer review comments differed from the steps necessary to receive full approval that were outlined in the official decision letters from Education. The Education official who manages the standards and assessments peer review process said that in some cases the differences between decision letters and peer reviewers' written comments led to state officials being unclear about whether they were required to address the issues in Education's decision letters, comments from peer reviewers, or both.

Conclusions

NCLBA set lofty goals for states to work toward having all students reach academic proficiency by 2013-2014, and Congress has provided significant funding to assist states. NCLBA required a major expansion in the use of student assessments, and states must measure higher order thinking skills and understanding with these assessments. Education currently reviews states' adherence to NCLBA standards and assessment requirements through its peer review process in which the agency examines evidence submitted by each state that is intended to show that state standards and assessment systems meet NCLBA requirements. However, ESEA, as amended, prohibits federal approval or certification of state standards. Education reviews the procedures that states use to develop their standards, but does not review the state standards on which ESEA assessments are based or evaluate whether state assessments cover highly cognitively complex content. As a result, there is no assurance that states include highly cognitively complex content in their assessments.

Although Education does not assess whether state assessments cover highly complex content, Education's peer review process does examine state assessment security procedures, which are critical to ensuring that

assessments are valid and reliable. In addition, the security of ESEA assessments is critical because these assessments are the key tool used to hold schools accountable for student performance. However, Education has not made assessment security a focus of its peer review process and has not incorporated best practices in assessment security into its peer review protocols. Unless Education takes advantage of forthcoming best practices that include assessment security issues, incorporates them into the peer review process, and places proper emphasis on this important issue, some states may continue to rely on inadequate security procedures that could affect the reliability and validity of their assessment systems.

State ESEA assessment systems are complex and require a great deal of time and effort from state officials to develop and maintain. Due to the size of these systems, the peer review process is an extensive process that also took a great deal of time and effort on the part of state officials. However, because Education, in an attempt to maintain peer reviewer confidentiality, does not permit direct communication between state officials and peer reviewers, miscommunication may have resulted in some states spending more time than necessary clarifying issues and providing additional documentation. While Education officials told us the assurance of anonymity was an important factor in their ability to recruit peer reviewers, anonymity should not automatically preclude communications between state officials and peer reviewers during the peer review process. For example, technological solutions could be used to retain anonymity while still allowing for direct communications. Direct communication between reviewers and state officials during the peer review process could reduce the amount of time and effort required of both peer reviewers and state officials.

The standards and assessments peer review is a high-stakes decision-making process for states. States that do not meet ESEA requirements for their standards and assessments systems can ultimately lose federal Title I, Part A funds. Transparency is a critical element for ensuring that decisions are fully understood and peer review issues are addressed by states. However, because critical Education decisions about state standards and assessments systems sometimes differed from peer reviewers' written comments, but the reasons behind these differences were not communicated to states, states were confused about the issues they needed to address.

Recommendations for Executive Action

To help ensure the validity and reliability of ESEA assessments, we recommend that the Secretary of Education update Education's peer review protocols to incorporate best practices in assessment security when they become available in spring 2010.

To improve the efficiency of Education's peer review process, the Secretary of Education should develop methods for peer reviewers and states to communicate directly during the peer review process so questions that arise can be addressed quickly. For example, peer reviewers could be assigned a generic e-mail address that would allow them to remain anonymous but still allow them to communicate directly with states.

To improve the transparency of its approval decisions pertaining to states' standards and assessment systems and help states understand what they need to do to improve their systems, in cases where the Secretary of Education's peer review decisions differed from those of the reviewers, the Secretary should explain why they differed.

Agency Comments and Our Evaluation

We provided a draft of this report to the Secretary of Education for review and comment. Education's comments are reproduced in appendix VII. In its comments, Education recognizes the value of test security practices in maintaining the validity and reliability of states' assessment systems. However, regarding our recommendation to incorporate test security best practices into the peer review protocols, Education indicated that it believes that its current practices are sufficient to ensure that appropriate test security policies and procedures are implemented. Education officials indicated that states currently provide the agency with evidence of state statutes, rules of professional conduct, administrative manuals, and memoranda that address test security and reporting of test irregularities. Education officials also stated that additional procedures and requirements, such as security methods and techniques to uncover testing irregularities, are typically included in contractual agreements with test publishers or collective bargaining agreements and that details on these additional provisions are best handled locally based on the considerations of risk and cost. Furthermore, Education stated that it plans to continue to monitor test security practices and to require corrective action by states they find to have weak or incomplete test security practices. As stated in our conclusions, we continue to believe that Education should incorporate forthcoming best practices, including assessment security issues into the peer review process. Otherwise, some states may continue to rely on

inadequate security procedures, which could ultimately affect the reliability and validity of their assessment systems.

Education agreed with our recommendations to develop methods to improve communication during the review process and to identify for states why its peer review decisions in some cases differed from peer reviewers' written comments. Education officials noted that the agency is considering the use of a secure server as a means for state officials to submit questions, documents, and other evidence to strengthen communication during the review process. Education also indicated that it will conduct a conference call prior to upcoming peer reviews to clarify why the agency's approval decisions in some cases differ from peer reviewers' written comments. Education also provided technical comments that we incorporated into the report as appropriate.

We are sending copies of this report to appropriate congressional committees, the Secretary of Education, and other interested parties. In addition, the report will be available at no charge on GAO's Web site at http://www.gao.gov. Please contact me at (202) 512-7215 if you or your staff have any questions about this report. Contact points for our Offices of Congressional Relations and Public Affairs may be found on the last page of this report. Other major contributors to this report are listed in appendix VIII.

Sincerely yours,

Cornelia M. Ashby
Director, Education, Workforce,
and Income Security Issues

Appendix I: Objectives, Scope, and Methodology

The objectives of this study were to answer the following questions: (1) How have state expenditures on assessments required by the Elementary and Secondary Education Act of 1965 (ESEA) changed since the No Child Left Behind Act of 2001 (NCLBA) was enacted in 2002, and how have states spent funds? (2) What factors have states considered in making decisions about question (item) type and content of their ESEA assessments? (3) What challenges, if any, have states faced in ensuring the validity and reliability of their ESEA assessments? (4) To what extent has the U.S. Department of Education (Education) supported and overseen state efforts to comply with ESEA assessment requirements?

To meet these objectives, we used a variety of methods, including document reviews of Education and state documents, a Web-based survey of the 50 states and the District of Columbia, interviews with Education officials and assessment experts, site visits in four states, and a review of the relevant federal laws and regulations. The survey we used was reviewed by several external reviewers, and we incorporated their comments as appropriate.

We conducted this performance audit from August 2008 through September 2009 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

Providing Information on How State Expenditures on Assessments Have Changed Since the Enactment of NCLBA and How States Have Spent Funds

To learn how state expenditures for ESEA assessments have changed since NCLBA was enacted in 2002 and how states spent these funds, we analyzed responses to our state survey, which was administered to state assessment directors in January 2009. In the survey, we asked states to provide information about the percentage of their funding from federal and state sources, their use of contractors, cost and availability of human resources, and rank order cost of assessment activities. The survey used self-administered, electronic questionnaires that were posted on the Internet. We received responses from 49 states,[1] for a 96 percent response rate. We did not receive responses from New York and Rhode Island. We reviewed state responses and followed up by telephone and e-mail with

[1]In this report, we refer to the District of Columbia as a state.

states for additional clarification and obtained corrected information for our final survey analysis.

Nonresponse is one type of nonsampling error that could affect data quality. Other types of nonsampling error include variations in how respondents interpret questions, respondents' willingness to offer accurate responses, and data collection and processing errors. We included steps in developing the survey, and collecting, editing, and analyzing survey data to minimize such nonsampling error. In developing the Web survey, we pretested draft versions of the instrument with state officials and assessment experts in various states to check the clarity of the questions and the flow and layout of the survey. On the basis of the pretests, we made slight to moderate revisions of the survey. Using a Web-based survey also helped remove error in our data collection effort. By allowing state assessment directors to enter their responses directly into an electronic instrument, this method automatically created a record for each assessment director in a data file and eliminated the need for and the errors (and costs) associated with a manual data entry process. In addition, the program used to analyze the survey data was independently verified to ensure the accuracy of this work.

We also conducted site visits to four states—Maryland, Rhode Island, South Dakota, and Texas—that reflect a range of population size and results on Education's assessment peer review. On these site visits we interviewed state officials, officials from two districts in each state—selected in consultation with state officials to cover heavily- and sparsely-populated areas—and technical advisors to each state.

Identifying Factors That States Have Considered in Making Decisions about Item Type and Content of Their Assessments

To gather information about factors states consider when making decisions about the item type and content of their assessments, we analyzed survey results. We asked states to provide information about their use of item types, including the types of items they use for each of their assessments (e.g., general, alternate, modified achievement standards, or native language), and changes in their relative use of multiple choice and open/constructed response items and factors influencing their decisions on which item types to use for reading/language arts and mathematics general assessments. We interviewed selected state officials and state technical advisors. We also interviewed officials from other states that had policies that helped address the challenge of including cognitively-complex content in state assessments. We interviewed four major assessment vendors to provide us a broad perspective of the views of the assessment industry. Vendors were

selected in consultation with the Association of American Publishers because its members include the major assessment vendors states have contracted with for ESEA assessment work. We reviewed studies that our site visit states submitted as evidence for Education's peer review approval process to document whether assessments are aligned with academic content standards, including the level of cognitive complexity in standards and assessments. We also spoke with representatives from three alignment organizations that states most frequently hire to conduct this type of study, and representatives of a fourth alignment organization that was used by one of our site visit states, who provided a national perspective on the cognitive complexity of assessment content. In addition, we reviewed selected academic research studies that examined the relationship between assessments and classroom curricula using GAO's data reliability tests. We determined that the results of these research studies were sufficiently valid and reliable for the purposes of our work.

Describing Challenges, If Any, That States Have Faced in Ensuring the Validity and Reliability of Their ESEA Assessments

To gather information about challenges states have faced in ensuring validity and reliability, we used our survey to collect information about state capacity and technical quality issues associated with assessments. We conducted reviews of state documents, such as assessment security protocols, and interviewed state officials. We asked state officials from the states we visited to complete a CCSSO checklist on student assessment—the Quality Control Checklist for Processing, Scoring, and Reporting—to show which steps they took to ensure quality control in high-stakes assessment programs. We used this specific document created by CCSSO because, as an association of public education officials, the organization provides considerable technical assistance to states on assessment. We confirmed with CCSSO that the document is still valid for state assessment programs and has not been updated. We also interviewed four assessment vendors and assessment security experts that were selected based on the extent of their involvement in statewide assessments. We also reviewed summaries of the peer review issues for states that have not yet been approved through the peer review process, the portion of peer review protocols that address assessment security, and the assessment security documents used to obtain approval in our four site visit states.

Describing the Extent to Which Education Has Supported State Efforts to Comply with ESEA Assessment Requirements

To address the extent of Education's support of ESEA assessment implementation, we reviewed Education guidance, summaries of Education assistance, peer review training documents, and previous GAO work on peer review processes. In addition, we analyzed survey results. We asked states to provide information on the federal role in state assessments, including their perspectives on technical assistance offered by Education and Education's peer review process. We also asked peer reviewers to provide their perspectives on Education's peer review process. Of the 76 peer reviewers Education provided us, we randomly sampled 20 and sent them a short questionnaire asking about their perspectives on the peer review process. We obtained responses from nine peer reviewers. In addition, we interviewed Education officials in charge of the peer review and assistance efforts.

Appendix II: Student Population Assessed on ESEA Assessments in School Year 2007-08

	Approximate number of students assessed
General Reading/Language Arts Assessment	25 million in each of reading/language arts and mathematics in 49 states reporting
Alternate Reading/Language Arts Assessment Using Alternate Achievement Standards	250,000 in each of reading/language arts and mathematics in 48 states reporting
Alternate Reading/Language Arts Assessment Using Modified Achievement Standards	200,000 in each of reading/language arts and mathematics in 46 states reporting

Source: GAO.

Appendix III: Validity Requirements for Education's Peer Review

Education's guidance describes the evidence states needed to provide during the peer review process. These are:

1.	**Evidence based on test content (content validity).** Content validity is the alignment of the standards and the assessment.
2.	**Evidence of the assessment's relationship with other variables.** This means documenting the validity of an assessment by confirming its positive relationship with other assessments or evidence that is known or assumed to be valid. For example, if students who do well on the assessment in question also do well on some trusted assessment or rating, such as teachers' judgments, it might be said to be valid. It is also useful to gather evidence about what a test does not measure. For example, a test of mathematical reasoning should be more highly correlated with another math test, or perhaps with grades in math, than with a test of scientific reasoning or a reading comprehension test.
3.	**Evidence based on student response processes.** The best opportunity for detecting and eliminating sources of test invalidity occurs during the test development process. Items need to be reviewed for ambiguity, irrelevant clues, and inaccuracy. More direct evidence bearing on the meaning of the scores can be gathered during the development process by asking students to "think-aloud" and describe the processes they "think" they are using as they struggle with the task. Many states now use this "assessment lab" approach to validating and refining assessment items and tasks.
4.	**Evidence based on internal structure**. A variety of statistical techniques have been developed to study the structure of a test. These are used to study both the validity and the reliability of an assessment. The well-known technique of item analysis used during test development is actually a measure of how well a given item correlates with the other items on the test. A combination of several statistical techniques can help to ensure a balanced assessment, avoiding, on the one hand, the assessment of a narrow range of knowledge and skills but one that shows very high reliability, and on the other hand, the assessment of a very wide range of content and skills, triggering a decrease in the consistency of the results.
In validating an assessment, the state must also consider the consequences of its interpretation and use. States must attend not only to the intended effects, but also to unintended effects. The disproportional placement of certain categories of students in special education as a result of accountability considerations rather than appropriate diagnosis is an example of an unintended—and negative—consequence of what had been considered proper use of instruments that were considered valid.	

Source: NCLB Standards and Assessments Peer Review Guidance.

Appendix IV: Reliability Requirements for Education's Peer Review

The traditional methods of portraying the consistency of test results, including reliability coefficients and standard errors of measurement, should be augmented by techniques that more accurately and visibly portray the actual level of accuracy. Most of these methods focus on error in terms of the probability that a student with a given score, or pattern of scores, is properly classified at a given performance level, such as "proficient." For school-level or district-level results, the report should indicate the estimated amount of error associated with the percent of students classified at each achievement level. For example, if a school reported that 47 percent of its students were proficient, the report might say that the reader could be confident at the 95 percent level that the school's true percent of students at the proficient level is between 33 percent and 61 percent. Furthermore, since the focus on results in a Title I context is on improvement over time, the report should also indicate the accuracy of the year-to-year changes in scores.

Source: NCLB Standards and Assessments Peer Review Guidance.

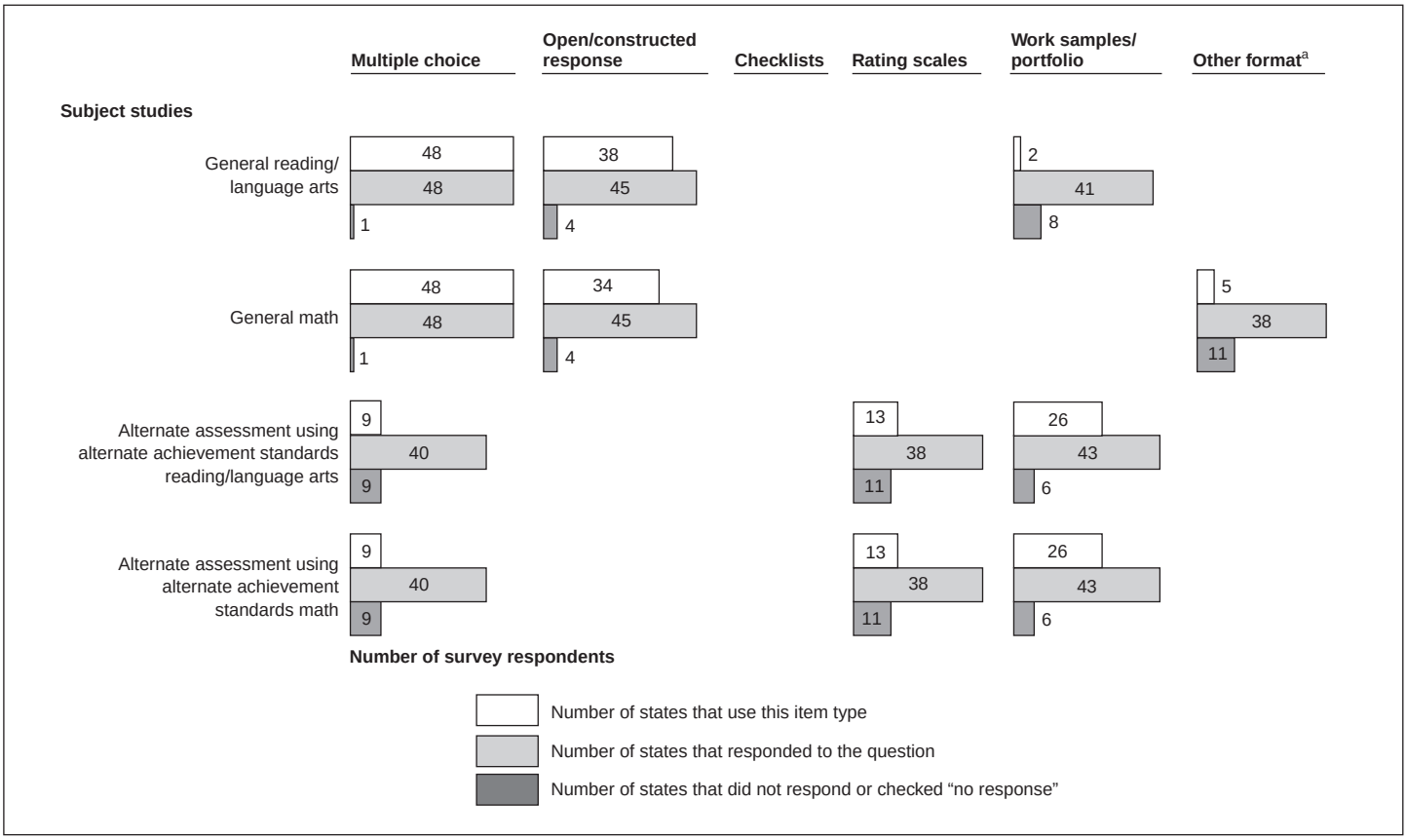
Appendix V: Alignment Requirements for Education's Peer Review

To ensure that its standards and assessments are aligned, states need to consider whether the assessments:

- Cover the full range of content specified in the state's academic content standards, meaning that all of the standards are represented legitimately in the assessments.
- Measure both the content (what students know) and the process (what students can do) aspects of the academic content standards.
- Reflect the same degree and pattern of emphasis apparent in the academic content standards (e.g., if the academic content standards place a lot of emphasis on operations, then so too should the assessments).
- Reflect the full range of cognitive complexity and level of difficulty of the concepts and processes described, and depth represented, in the state's academic content standards, meaning that the assessments are as demanding as the standards.
- Yield results that represent all achievement levels specified in the state's academic achievement standards.

Source: NCLB Standards and Assessments Peer Review Guidance.

Appendix VI: Item Types Used Most Frequently by States on General and Alternate Assessments



Source: GAO survey.

[a]Other format includes gridded response, performance event, scaffolded multiple choice and performance events, and locally-developed formats.

Appendix VII: Comments from the U.S. Department of Education

UNITED STATES DEPARTMENT OF EDUCATION

OFFICE OF ELEMENTARY AND SECONDARY EDUCATION

September 8, 2009 THE ASSISTANT SECRETARY

Ms. Cornelia M. Ashby
Director
Education, Workforce, and
Income Security Issues
U.S. Government Accountability Office
441 G Street, NW
Washington, DC 20548

Dear Ms. Ashby:

I am writing in response to your request for comments on the draft Government Accountability Office (GAO) report, "No Child Left Behind Act: Enhancements in the Department of Education's Review Process Could Improve State Academic Assessments" (GAO-09-911).

This report has three recommendations for the Secretary of Education. Following is the Department's response.

Recommendation: Incorporate test security best practices into the peer review protocols.

Response: The Department recognizes the value of this recommendation and the importance of test security practices in maintaining the validity and reliability of each State's assessment system. Currently, as part of the peer review process, States do provide us with evidence of State statutes, rules of professional conduct, administrative manuals, and memoranda that address test security and reporting of test irregularities. Other procedures and requirements (e.g., remedies for teacher misconduct) are typically included in contractual agreements with test publishers and other parties, or collective bargaining agreements. The Department does not examine those additional provisions because we believe that our current practices are sufficient to ensure that appropriate test security policies and procedures are promulgated and implemented at the State level. Details on these additional provisions such as security methods and techniques to discover testing irregularities are best handled locally based on consideration of risk and cost factors. As the report mentions, the Department monitors the implementation of State test security policies in its regularly scheduled Title I monitoring visits to State and local educational agencies. Department staff will continue to monitor test security practices during the monitoring visits, issue findings to States with weak or incomplete test security practices, and require corrective action by States with monitoring findings.

400 MARYLAND AVE., S.W., WASHINGTON, DC 20202
www.ed.gov

Our mission is to ensure equal access to education and to promote educational excellence throughout the nation.

Recommendation: Develop methods to improve communication during the review process.

Response: The Department has made the following improvements over the last year to improve communications with States during the peer review process. First, peers and the Department staff member assigned to review the State's assessment system typically call State assessment officials and discuss the submission and the peers' concerns. This occurs prior to the conclusion of the peer review, giving peers time to correct any misconceptions before they complete their review. Through this process, State officials have opportunities to ask questions and obtain clarification regarding the peers' and Department's concerns. Second, during the Technical Assistance Peer Review (May 2008), State assessment professionals (individuals or teams) met directly with the peers or peer team leader and the Department staff member assigned to the State to thoroughly discuss the peers' comments and concerns. A technical assistance review is conducted to help States understand where further development is required before the system is ready for review. The Department will continue this process.

Furthermore, the Department is looking into the possibility of using a secure server as a means for State officials to submit questions, documents, and other evidence that would only be viewed by the reviewers, State officials, and Department staff. We believe that the use of a secure server, in combination with the procedures already in place, would strengthen the communication that takes place during the peer review process.

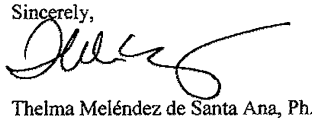
Recommendation: Identify for States why its peer review decisions in some cases differed from peer reviewers' written comments.

Response: Peer notes sometimes address areas outside of the Department's purview, offer recommendations to improve elements of the system beyond the requirements of the law and regulations, or offer opinions on technical matters. We do not use those recommendations in judging the merits of the assessment system, but, as a professional courtesy, we include them as technical assistance in the peer notes provided to the States. Peer notes, and the deliberations they document, are recommendations to the Assistant Secretary for Elementary and Secondary Education, and on occasion, Department staff may disagree with the peers' summary comments. The Assistant Secretary is presented with these discrepancies after they have been discussed internally among Department staff. These discrepancies usually deal with limits on the range of evidence that is required to be provided to demonstrate compliance with the applicable statutory and regulatory provisions and the extent to which the Department has authority in judging the quality of certain features of a State assessment system. For example, the Department has no prerogative to deny approval of an assessment system based on the substance of content standards nor is the State required to submit evidence on that issue. The Department and peers review only the process used to develop a State's content standards, ensure broad participation of stakeholders in the process, and ensure that a State demonstrates the rigor of the standards. Hence, there are no peer-review "decisions," only peer recommendations reflecting the professional experience and perspectives of the reviewers. The Assistant Secretary takes these recommendations under consideration, along with those of Department staff, in making a decision regarding the approval of a State's assessment system.

However, in response to this recommendation, Department staff will conduct a conference call in advance of upcoming peer reviews to clarify why the Department's decisions in some cases differ from peer reviewers' written comments.

I appreciate the opportunity to share our comments on the draft report. I hope that these comments are useful to you. In addition, we have provided some suggested technical edits that should be considered to add clarity to the report.

Sincerely,



Thelma Meléndez de Santa Ana, Ph.D.

Appendix VIII: GAO Contact and Staff Acknowledgments

GAO Contact

Cornelia M. Ashby (202) 512-7215 or ashbyc@gao.gov

Staff Acknowledgments

Bryon Gordon, Assistant Director, and Scott Spicer, Analyst-in-Charge, managed this assignment and made significant contributions to all aspects of this report. Jaime Allentuck, Karen Brown, and Alysia Darjean also made significant contributions. Additionally, Carolyn Boyce, Doreen Feldman, Cynthia Grant, Sheila R. McCoy, Luann Moy, and Charlie Willson aided in this assignment.

GAO's Mission

The Government Accountability Office, the audit, evaluation, and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.

Obtaining Copies of GAO Reports and Testimony

The fastest and easiest way to obtain copies of GAO documents at no cost is through GAO's Web site (www.gao.gov). Each weekday afternoon, GAO posts on its Web site newly released reports, testimony, and correspondence. To have GAO e-mail you a list of newly posted products, go to www.gao.gov and select "E-mail Updates."

Order by Phone

The price of each GAO publication reflects GAO's actual cost of production and distribution and depends on the number of pages in the publication and whether the publication is printed in color or black and white. Pricing and ordering information is posted on GAO's Web site, http://www.gao.gov/ordering.htm.

Place orders by calling (202) 512-6000, toll free (866) 801-7077, or TDD (202) 512-2537.

Orders may be paid for using American Express, Discover Card, MasterCard, Visa, check, or money order. Call for additional information.

To Report Fraud, Waste, and Abuse in Federal Programs

Contact:

Web site: www.gao.gov/fraudnet/fraudnet.htm
E-mail: fraudnet@gao.gov
Automated answering system: (800) 424-5454 or (202) 512-7470

Congressional Relations

Ralph Dawn, Managing Director, dawnr@gao.gov, (202) 512-4400
U.S. Government Accountability Office, 441 G Street NW, Room 7125
Washington, DC 20548

Public Affairs

Chuck Young, Managing Director, youngc1@gao.gov, (202) 512-4800
U.S. Government Accountability Office, 441 G Street NW, Room 7149
Washington, DC 20548