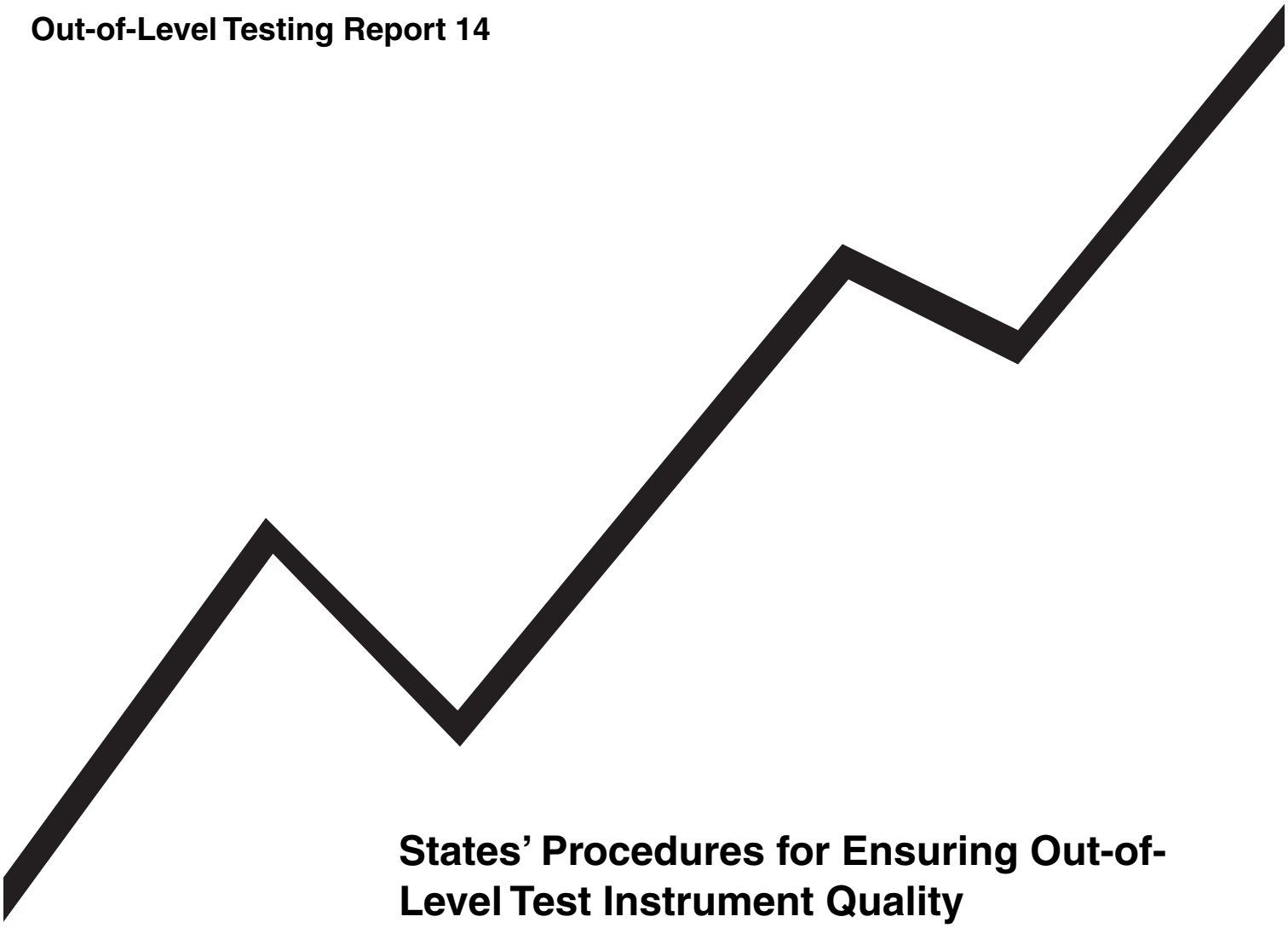
Out-of-Level Testing Report 14

States' Procedures for Ensuring Out-of-Level Test Instrument Quality



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

In collaboration with:

Council of Chief State School Officers (CCSSO)

National Association of State Directors of Special Education (NASDSE)

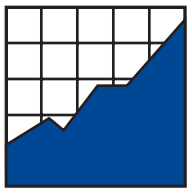
Out-of-Level Testing Report 14

States' Procedures for Ensuring Out-of-Level Test Instrument Quality

Jane E. Minnema • Martha L. Thurlow • Ross E. Moen •
Gretchen R. VanGetson

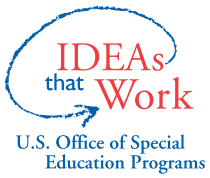
September 2004

Minnema, J. E., Thurlow, M. L., Moen, R. E., & VanGetson, G. R. (2004). *States' procedures for ensuring out-of-level test instrument quality* (Out-of-Level Testing Report 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The Out-of-Level Testing Project is supported by a grant (#H324D990058) from the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.



NCEO Core Staff

Deb A. Albus
Ann T. Clapper
Christopher J. Johnstone
Jane L. Krentz
Sheryl Lazarus
Kristi K. Liu
Jane E. Minnema
Ross E. Moen
Michael L. Moore
Rachel F. Quenemoen
Dorene L. Scott
Sandra J. Thompson
Martha L. Thurlow, Director

National Center on Educational Outcomes
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
Phone 612/624-8561 • Fax 612/624-0879
http://nceo.info

Executive Summary

The present study was initiated to gather information about states' out-of-level testing practices. Specifically, we wanted to see whether states attempted to align out-of-level tests with grade of enrollment content standards, what processes are used to complete this task, and what psychometric information states offer as evidence of that alignment. We collected narrative data in this study from two sources: technical information about states' large-scale statewide assessment and information gleaned from interviews with state assessment directors or other individuals knowledgeable about the state's large-scale assessment. We used the technical information gathered prior to the interviews to provide context and support for our interview data analysis process. We then compiled the results thematically.

Five critical issues were highlighted in the results.

1. There was an increasing need for states to provide easily accessible technical information that includes out-of-level testing information.
2. States' arguments supporting their decision to not equate out-of-level test scores with on-level test scores were stronger than arguments supporting this practice.
3. States provided incomplete and inconclusive information about the psychometric properties of out-of-level test scores.
4. States were not consistent in their opinions about the use of out-of-level tests.
5. States made questionable assumptions about out-of-level testing.

Overall, the wide variability of out-of-level testing practices among the states raises many concerns about the practice of out-of-level testing.

Background

Out-of-level testing, "the administration of a test at a level above or below the level generally recommended for students based on their age-grade level" (Study Group on Alternate Assessment, 1999), first arose in the 1960s as a way to measure Title I program effectiveness (Cleland & Idstein, 1980; Crowder & Gallas, 1978; Jones, Barnette, & Callahan, 1983; Long, Schaffran, & Kellogg, 1977). The logic behind this was that out-of-level testing would yield more reliable and valid test results for students who were not achieving at grade level (Ayrer & McNamara, 1973). With the advent of standards-based reform and large-scale statewide assessments ushered in by recent legislation, including the No Child Left Behind (NCLB) Act of 2001, out-of-level testing seemed like an appealing option for fulfilling the legal requirement to include all students in statewide testing. Thus, students achieving below-grade level, typically students with disabilities, who were historically omitted from large-scale assessment practices, were thought to show with increased participation and performance.

The increased popularity of out-of-level testing has occurred amidst controversy within highly politicized settings (Thurlow & Minnema, 2001). States differed greatly in their use of out-of-level testing, which was reflected in their preferred terms for below grade level testing. Multiple terms were used, from instructional level testing to alternate assessment. There was variability in other areas that are of concern, such as the alignment of the test content with state content standards and the psychometric properties of out-of-level test scores.

Standards-based large-scale assessments are used as indicators that, under NCLB, all students are working toward proficiency on rigorous on-grade level curriculum guided by states' academic content standards (Linn, Baker, & Betebenner, 2002). Curriculum and assessment are two of the three elements of education that, when joined by the third element (instruction) comprise a triad of core educational elements (Pellegrino, 2002). Alignment is the process of ensuring the agreement between these three elements, defined as "the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do" (Webb, 2002, p. 3). The purpose of content standards is to provide clear and concise guidelines for instructional and curricular development serving as the foundation of the alignment process.

Many state assessments have been custom built to align with the state's content standards even though this process has its shortcomings. Many states' content standards are too broad to provide clear and concise guidelines for alignment (Popham, 2001). There are multiple processes for aligning large-scale assessments with content standards. Some are less rigorous than others, which introduces a lack of consistency across states about their thoroughness (Council of Chief State School Officers, 2002). There is continuing concern that failing to ensure proper alignment between assessments and content standards will result in students being "taught to

the tests" and not participating in a rigorous on-grade level curriculum based on challenging content standards (Rothman, Slattery, Vranek, & Resnick, 2002).

Alignment causes even more of a concern when considering out-of-level testing. In addition to the challenges of alignment faced by on-grade level assessments, alignment of out-of-level tests also begs the question as to which grade level content standards the out-of-level test should align. Out-of-level tests should then be aligned with alternate achievement standards that are "clearly different from the achievement standards in the target grade" (Federal Register, 2003). It is important to note that data collection for this report was conducted prior to the release of these regulations; therefore, the results of this study may not necessarily reflect this federal mandate.

Along with test alignment, another issue in developing and demonstrating the quality of the test instrument is psychometric soundness. Two aspects of psychometric properties that have been emphasized in the area of out-of-level testing are the concepts of precision and accuracy. Precision is concerned with random error and accuracy is concerned with systematic error or bias. These two concepts can be thought of as roughly comparable to reliability and validity. Validity (accuracy) speaks to whether you are hitting the right target and reliability (precision) speaks to how consistently you are hitting one target.

Asserting that on-level tests yield imprecise measures for students who are instructed at levels below the grade in which they are enrolled in school, proponents of out-of-level testing claim that testing students at the level of instruction is a more precise measure of what they know and can do. Psychometric theory and research agree that there is more random measurement error when students take tests that are much too hard for them (Bielinski, Thurlow, Minnema, & Scott, 2000). It follows then that giving students a test closer to their achievement level will produce less random error in measuring students' ability. But the picture increases in complexity when inferences are made beyond the test results, such as when below grade level tests are used to infer achievement on content standards set for grade of enrollment. When out-of-level test results are used to infer how a student would perform on an on-level test, the measurement error of both tests must be taken into account. *It is imperative that states consider whether the precision gained by using an assessment closer to a student's achievement level (out-of-level test) outweighs the added error introduced by inferring on-level test performance from out-of-level test results.*

The issue of psychometric test score accuracy in aligning out-of-level tests with grade of enrollment content standards falls in the realm of equating, and more particularly, vertically equating test scores. In the area of general equating tests, the *Standards for Educational and Psychological Testing* state that "the fundamental concern is to show that equated scores measure essentially the same construct, with very similar levels of reliability and conditional standard errors of

measurement" (AERA/APA/NCME, 1999, p. 57). Measurement specialists are able to address this concern when they create parallel test forms intended to measure the same difficulty level on the same constructs for the same population. Annually produced versions of college entrance examinations such as the SAT or ACT tests are examples of successfully meeting these conditions. Moving away from any of these three conditions—comparable difficulty, construct, and population—complicates the process. For example, a National Research Council (NRC) committee was charged by Congress to find a common scale to equate tests such as a variety of 4th grade reading tests. After reviewing a variety of equating issues, methods, and studies, the committee concluded that "comparing the full array of currently administered commercial and state achievement tests to one another, through the development of a single equivalency or linking scale, is not feasible" (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999, p. 4).

Even reducing random measurement error to acceptable limits, as might be accomplished through methods such as item response theory (IRT) scaling, would not settle the issue of whether the same construct is being measured by different levels of testing. The NRC committee spoke clearly to this issue by describing an extreme case of creating a formula that links the scores from a reading test with the scores from a mathematics test (Feuer et al., 1999). Although reading and mathematics are obviously different constructs, it is arithmetically possible to link the two scores together, resulting in a deceptive representation of student performance in one of the two content areas. Developers of out-of-level tests need to show whether, for example, 5th grade students who receive a 3rd grade reading test out of level are being assessed on the same constructs as the majority of 5th graders who receive the on-level reading test. If the same constructs are not being measured, inferring anything about proficiency on 5th grade content standards from performance on a 3rd grade test is compromising. *Of particular concern is that some students who are tested on lower level standards are prevented from demonstrating proficiency on grade of enrollment content standards because they had no opportunity to show what they could do on grade level standards.*

The present study was initiated to gather information about states' out-of-level testing practices. Specifically, we wanted to see whether they attempted to align out-of-level tests with grade of enrollment content standards, what processes were used to complete this task, and what psychometric information they offer as evidence of that alignment. The specific research questions that guided this study were:

(1) What processes do states use to ensure alignment of out-of-level tests with state standards?

(2) What is the grade level of the standards with which the out-of-level tests are aligned?

(3) What evidence do states offer that their assessments are psychometrically sound?

(4) Are scores from out-of-level tests equated with scores from on-grade level tests? And what is the rationale for this process?

Method

We collected narrative data in this study from two different sources: technical reports that contained information about states' large-scale test instruments and information gleaned from interviews with state assessment directors or other individuals knowledgeable about the state's large-scale assessment program.

Procedure—Task One

The first task in this study was to gather the technical information. At the beginning of this study in 2002, 14 states were identified as using out-of-level testing in their statewide assessment programs (Thurlow & Minnema, 2003). We attempted to obtain technical reports, as well as any other technical information about the state's assessment program, from each of these states. To accomplish this task, we searched each state education agency's Web site and downloaded any technical information or test development information, including test blueprints. If detailed technical information was not posted online, we contacted each state or the state's test contractor directly in an attempt to obtain a hard copy of the technical report. We received technical reports from two states; two other states indicated that the report was not available at that time, but would be available within the year. Unfortunately, the window of time for collecting technical reports for our study closed before these reports were completed. Two test publishers sent a copy of the technical report for two norm-referenced tests used in some states that test out of level. We reviewed all of the technical information before conducting the telephone interview. For states from which technical information was not received, we reviewed the information we had found on the states' Web site.

Procedure—Task Two

Assessment directors in states that tested out of level were selected as prospective participants for the telephone interviews. An NCEO researcher contacted each state assessment director by e-mail. We attached a copy of the interview questions and the study's research proposal to the e-mail for advanced information about the study. Participants were given the option of a telephone interview or responding to the interview questions via e-mail. They could designate another individual to participate instead if someone else was more knowledgeable on out-of-level testing in that state. A follow-up e-mail identical to the initial recruitment e-mail with a brief reminder note was sent to non-respondents after one month.

Nine states agreed to participate in the interview (California, Connecticut, Delaware, Iowa,

Mississippi, Oregon, South Carolina, Utah, Vermont); three states declined participation, and two states failed to respond. One assessment director completed the interview via e-mail while eight states participated in the telephone interview. Of the eight states that participated in the telephone interviews, two assessment directors, one program associate of an educational testing company, one university professor, and two assessment specialists participated individually, and two states requested group interviews. Of these group interviews, one state included the assessment director, two assessment consultants, and one manager of communications. The other state included the assessment director, one assessment consultant, and one assessment coordinator. The telephone interviews typically lasted 20 to 30 minutes and were tape recorded for transcription and qualitative data analysis.

After the transcribed interviews were read, the primary participant was contacted once again via e-mail for any additional follow-up questions or clarifications. We began data analysis by revisiting the technical information gathered prior to the interviews. This information provided context and support for our analysis process. Next, we reviewed each interview transcript and coded the narrative data into subcategories of information. We then compiled the results thematically. Each states' final set of results were e-mailed to the primary interview participant for final review prior to drafting the report.

Results

States' alignment and technical information for their large-scale statewide assessments was generally available online or in hard copy. Table 1 provides details on the availability of states' alignment and technical information, including what information was available online and what information was available in hard copy for those who request it. Some states (California, Connecticut, Mississippi, Oregon, South Carolina, Utah) posted online their test blueprints or a similar form of test specifications that explained the alignment between the state's content standards and test items. Five states (Delaware, Mississippi, Oregon, South Carolina, Utah) provided more detailed alignment and test development information. This information was presented as a stand-alone document with information such as phases of test instrument development or descriptions of test development committees and panels (Oregon, Utah). Other stand-alone documents were in the form of technical manuals (Delaware, South Carolina) or summaries of technical information (Connecticut, Mississippi, South Carolina, Utah, Vermont).

Five states (Connecticut, Delaware, Iowa, Utah, South Carolina) indicated that test blueprints were available in hard copy, while five states (Connecticut, Iowa, Mississippi, Oregon, South Carolina) indicated that some form of more detailed technical information was also available upon request. The majority of the states (California, Connecticut, Iowa, Mississippi, South Carolina, Utah, Vermont) responded that the complete technical manual for the state's assessment was available in hard copy from either the state educational agency or the test publisher.

Table 1. Availability of States' Large-Scale Assessment Information

	CA	CT	DE	IA	MS	OR	SC	UT	VT
Online information									
Blueprints or test specifications	X	X			X	X	X	X	
Technical manual			X				X		
Test development information			X		X	X	X	X	
Some technical information		X			X		X	X	X
Hard copy materials on request									
Blueprints		X	X	X			X	X	
Technical manual	X	X		X	X		X	X	X
Detailed technical information		X		X	X	X	X		

States made technical information available to the public. Overall, states had responded to their consumers' need for technical information by making the information available in consumer-friendly formats, either online or in hard copy. Three states (Connecticut, Oregon, South Carolina) attempted to put information online that was useful for a broad audience while providing instructions on how to obtain more specific information available in hard copy. While one state (Delaware) preferred to respond to the public need for such information by posting the entire technical manual online as a way to answer common questions, another state (Iowa) commented that consumers (i.e., district test coordinators) could more readily access the information in hard copy format. Making technical information, and especially test development information and blueprints, available to public consumers offered these consumers the opportunity to access states' procedures for test alignment with state content standards.

We gathered information about states' processes for aligning on-level tests to content standards to provide the context for a discussion of aligning out-of-level tests with content standards. Table 2 displays states' information regarding on-level test alignment. All of the states involved in this study, with the exception of one (Iowa), had developed statewide content standards. Every state with statewide content standards also had some type of documented link between those standards and the state's large-scale statewide assessment. These states had blueprints or test specifications based on the content standards that guided test item development. In these docu-

ments, each content standard was divided into testable portions, and each portion was assigned one or more test items to assess this section of the content standard.

States organized groups to conduct alignment procedures in a variety of ways. Many states (California, Delaware, Oregon, South Carolina, Utah) deployed special panels or committees to review test alignment as shown in Table 2. For example, Oregon used both content and sensitivity panels to check for alignment, South Carolina developed an Education Oversight Committee to, in part, ensure test alignment to content standards, and Utah used an advisory committee for this purpose. Two states (Connecticut, Vermont) relied on stakeholders in the assessment development process to review alignment during and at the conclusion of the actual test development process. Further, one state (Iowa), because it does not have statewide standards, provided training programs to each district in the state to teach about aligning the district assessment to district standards. One state (Mississippi) did not cite a specific review process to ensure alignment of assessments with content standards.

Each state included a unique combination of individuals in their alignment and test development processes. In particular, content specialists were designated as having a role in these processes (California, Delaware, Iowa, Oregon, South Carolina, Utah) as well as state educational agency staff members (Connecticut, Delaware, South Carolina, Utah, Vermont). Additionally, test publisher staff members contributed to these processes (California, Connecticut, Delaware, Iowa, South Carolina, Utah) along with educators (California, Connecticut, Delaware, Iowa, Mississippi, Oregon, South Carolina, Utah, Vermont) and administrators (South Carolina, Utah, Vermont).

States assumed that out-of-level tests were aligned with students' grade level of instruction. As indicated in Table 3, all the states responded that the tests used for out-of-level assessment were the same tests that were used for on-level assessment, just presented at a grade level below a student's enrollment grade. For instance, an 8th grade student administered a 5th grade out-of-level test would take the same test as 5th grade students who were participating in the general assessment on grade level. In most cases, states indicated that students were tested out of level at their instructional level, meaning that an 8th grade student would be administered a 5th grade test out-of-level because instruction was delivered at the 5th grade level.

While most likely true for more states, two states (Delaware, Iowa) indicated that the test level may not always match the instructional level. The respondent from Delaware expressed hesitation in claiming that all students were tested out of level at their instructional level. This expectation was articulated in the state's policy language, but actual implementation of the policy could vary from teacher to teacher. The respondent from Iowa explained that Iowa has no large-scale statewide assessment program, leaving assessment policy to be determined by each individual school district. Given that, he did not assume that every student tested out of

Table 2. State Practices In Aligning On-Level Tests to Content Standards

	CA	CT	DE	IA	MS	OR	SC	UT	VT
Adopted Statewide Content Standards									
Yes	X	X	X		X	X	X	X	X
No				X					
Documented Link Between Standards and Assessment									
Yes	X	X	X		X	X	X	X	X
No				X					
Alignment Review Process in Place									
Yes	X	X	X	X		X	X	X	X
No					X				
Individuals Involved in Test Development and Alignment Process									
Content specialists	X		X	X		X	X	X	
State department of education staff		X	X				X	X	X
Test publisher staff	X	X	X	X			X	X	
Educators	X	X	X	X	X	X	X	X	X
Administrators							X	X	X

level was tested at the student's instructional level because out-of-level testing practices could vary across districts.

Table 3 also shows states' responses about the grade level of the content standard to which out-of-level tests are aligned, meaning the grade at which the student was tested or grade at which the student was enrolled in school. For all but one of the states, the out-of-level test was aligned with content standards for a grade one or more levels lower than the student's grade of enrollment. For example, the out-of-level test administered to an 8th grade student was aligned with the 5th grade content standards so that an 8th grade student was being tested on proficiency for 5th grade content standards. The one state (Vermont) that differed indicated that students tested out of level were taking tests that were aligned with the content standards specific to their grade of enrollment but at a significantly lower difficulty level than their same-grade peers.

States differed in deeming out-of-level test scores as more precise and accurate than on-level test scores. The psychometric properties of precision and accuracy for out-of-level tests are displayed in Table 4. One state (California) stated explicitly that out-of-level tests were considered accurate only when assessing proficiency on instructional level content standards, but not accurate when reporting achievement relative to the content standards of a student's enrollment grade.

Table 3. Out-of-Level Test Characteristics

	CA	CT	DE	IA	MS	OR	SC	UT	VT
Test Used									
Same as on-level	X	X	X	X	X	X	X	X	X
Tested at Instructional Level									
Yes	X	X			X	X	X	X	X
Unsure			X	X					
Content Standards Tested									
Test grade level	X	X	X	X	X	X	X	X	
Enrollment grade level									X

Four states (Connecticut, Iowa, Oregon, South Carolina) assumed that out-of-level test scores were as accurate as on-level test scores. Further, one state (Vermont) refrained from making assertions about the accuracy of out-of-level test results until currently planned validation and reliability studies were completed. Three states (Delaware, Mississippi, Utah) did not address the concept of accuracy in the interview.

Table 4. Psychometric Properties of Out-of-Level Tests

	CA	CT	DE	IA	MS	OR	SC	UT	VT
Assume same accuracy as on-level tests									
Yes		X		X		X	X		
No									X
Distinguish the standard	X								
Not addressed			X		X			X	
Assume same precision as on-level tests									
Yes	X	X		X	X	X			X
No			X				X	X	

In terms of test score precision, some states (California, Connecticut, Iowa, Mississippi, Oregon, Vermont) believed that out-of-level test scores contained degrees of precision comparable to the on-level test, although few states had data to support this assumption. Of those states that did

not assume equality between the precision of on-level and out-of-level tests, one state (Utah) said that they "guessed" that out-of-level test scores might have more measurement error than on-level test scores. Limited department resources prevented them from investigating this issue further. Another state (Delaware) pointed to the small number of students tested below their grade of enrollment making it impossible to evaluate the precision of out-of-level test scores. A final state (South Carolina) stated that they did not have the proper statistics to support this claim.

States provided a variety of rationales for not equating out-of-level test scores to on-level test scores. Given previous responses, it is not surprising that when states were asked about methods used to equate out-of-level test scores to on-level test scores, most of them indicated that they did not attempt to do such equating. As depicted in Table 5, the seven states (California, Connecticut, Delaware, Mississippi, Oregon, South Carolina, Utah) that did not attempt to equate out-of-level test scores to on-level test scores cited several rationales for this decision. Some states (California, Utah) thought that equating these types of test scores was not statistically possible. Other states indicated that equating to on-level test scores was unnecessary because out-of-level test results were best used to only guide instruction (Delaware) or that these results should only be used in reference to the grade level of the test (South Carolina). Three states (Mississippi, Connecticut, Oregon) suggested that equating was inappropriate due to differences between out-of-level tests and on-grade level tests. For instance, one state (Mississippi) named a test validity issue since out-of-level and on-level tests measure different constructs. Another state (Connecticut) did not equate these test scores because out-of-level and on-level tests are the same measurement instrument while another state (Oregon) used no equating procedures for reasons that were "self evident."

Table 5. Out-of-level Test Score Equating

	CA	CT	DE	IA	MS	OR	SC	UT	VT
Equating Out-of-Level Test Scores									
No equating performed	X	X	X		X	X	X	X	
Standard developmental growth scale				X					
Score transformation rules									X

Two states (Iowa, Vermont) did discuss a process for aligning out-of-level and on-level test results, although neither state labeled the process test equating. Iowa used a norm-referenced assessment with a developmental growth scale that was developed so that an on-level equivalency could be derived from out-of-level test scores when the scores were only one grade apart.

Vermont described a process by which judges used rubrics that yielded scores, which could be subjected to transformation rules to infer on-level test results from out-of-level test results.

Discussion

Examining out-of-level testing in the current context has been like trying to walk on shifting sand. Although this study was intended to be an investigation of state practices at one point in time, these practices and the surrounding issues have continued to fluctuate throughout the course of the study. For example, some participants from states in the process of changing their out-of-level testing policy or practices found it difficult to describe out-of-level testing in their state. In all states, unforeseen changes may have occurred since the conclusion of this study. Thus, the instability of out-of-level testing throughout the nation and the difficulties that arise due to this controversial approach to testing should be held in mind when considering the results of this study. Given this disclaimer, we have discerned five critical issues through the interpretation of our results.

The first issue is that there is an increasing need for states to provide easily accessible technical information that includes out-of-level testing information. Detailed large-scale assessment technical information, such as a technical manual, was typically only available in hard copy to individuals who requested it. Those individuals may be deterred from obtaining these reports due to the length of the report (i.e., hundreds of pages long), having to order the report directly from the test publisher, or even having to purchase the technical manual. Because this information should be readily available to the public, states should strive toward including all technical information on their Web sites as well as in hard copy to accommodate non-state or district personnel.

Specific to out-of-level testing, we found it difficult to access this information at all. Typically, the only information available online about out-of-level testing was policy information or participation criteria. Very few states made information about their out-of-level testing practices or test results available online, and out-of-level testing information was seldom mentioned in states' assessment program technical manuals. *As states begin to incorporate more assessment information in their Web sites, they need to include the same detailed information regarding their out-of-level assessment options as they do for other assessment options in their large-scale assessment programs.*

The second issue concerns equating. States' arguments supporting their decision to not equate out-of-level test scores with on-level test scores were stronger than arguments supporting this practice. Very few states used out-of-level testing results to indicate whether students were proficient on grade of enrollment content standards. In other words, most states did not attempt to

prove that out-of-level tests were aligned with on-level standards. Thus, they did not attempt to prove that out-of-level tests could be used to assess on-level proficiency. These states thought that equating out-of-level test scores with on-level test scores was either statistically impossible, unnecessary, or simply inappropriate.

Of the two states that did attempt to equate out-of-level test scores, neither developed evidence to support the psychometric soundness of testing students with disabilities two or more grade levels below their grade of enrollment. One state (Iowa) asserted that its series of tests permitted inferring grade of enrollment proficiency from the results of an out-of-level test that was one grade level lower than the grade of enrollment. Nevertheless, it is critical to consider whether a single grade difference is enough to make out-of-level testing worthwhile. Other matters such as a student's opportunity to learn or the need for changes in instructional delivery or curricular format could resolve these assessment issues. It is important to consider these matters for all students who are being tested out of level, which would include those students who are tested many grade levels below their grades of enrollment in school.

The other state (Vermont), also the only state that indicated that out-of-level tests were aligned with grade of enrollment content standards, had yet to conduct empirical studies evaluating the psychometric soundness of its score transformation rules. This evaluation was planned for the future. *The lack of explicit evidence that supports equating out-of-level test scores to on-level test scores confirms that this equating practice should not occur for test results that demonstrate academic progress toward meeting a grade-level criterion.*

The third issue is that states provided incomplete and inconclusive information about the psychometric properties of out-of-level test scores. Two states (Iowa, South Carolina) indicated that out-of-level test scores were as accurate as on-level test scores, which was as close as any interview respondent came to commenting on the psychometric properties of out-of-level test scores. Later, when states were reviewing our results before publishing this report, two more states (Connecticut, Oregon) indicated that the accuracy of out-of-level and on-grade level test results were equivalent. Only one state (California) did not make this claim either during the interview or the review of our results.

It should be noted that the four states that claimed out-of-level and on-grade level test score equivalency did so without specifying the grade level at which the out-of-level test scores were accurate for measuring proficiency—the grade level of the test or the grade level of a student's enrollment in school. Perhaps these states believed that since their out-of-level tests measured the content standards for the grade level of the test, and not the grade level of enrollment, their assumption was clear. Out-of-level test scores are as accurate in measuring proficiency on test grade content standards, but *only* for the grade level at which the test is administered.

Questions remain unanswered regarding the psychometric properties of out-of-level test scores.

Did states have sufficient empirical evidence to support their responses? Was this evidence derived from research on out-of-level tests, or was it simply an interpretive extension of research conducted on on-level tests? Three states acknowledged that they did not have the statistics available to support any claim of precision or accuracy on out-of-level tests and because of that, refrained from doing so. *If states are using out-of-level tests as part of their large-scale assessment program, they should conduct the same studies of score accuracy and precision as they do for on-level tests.*

A fourth issue is that states are not consistent in their opinions about the use of out-of-level tests; states commented on the benefits as well as the limitations of out-of-level testing. Benefits included supporting the participation of more students in states' large-scale assessment and accountability programs (Mississippi, South Carolina, Vermont), matching or exceeding the integrity of on-grade level test scores (Oregon, South Carolina, Vermont), and asserting that out-of-level test results provided more instructionally useful information for each individual student because out-of-level tests are instructionally appropriate for students who are achieving below their grade of enrollment (California, Connecticut, Delaware, Utah, Vermont).

On the other hand, respondents cited various concerns about the limitations of testing students with disabilities out of level. Limitations included providing no information about a student's performance on-grade level (California, Connecticut, Delaware, Oregon, South Carolina), assuming that the student fails to meet on-grade level proficiency (California, Connecticut, Delaware, Mississippi, Utah, Vermont), including age inappropriate test content for some students (Iowa, South Carolina), and handling the test results differently from the results of on-level tests in reporting and interpretation (Connecticut, Delaware, Utah). *The variability in responses indicates that states are not in agreement on the benefits and limitations of out-of-level testing, nor thoroughly understand the purpose of statewide testing under NCLB— that of system accountability rather than student accountability.*

The final issue, and arguably most important, is that states make dangerous assumptions about out-of-level testing. Three examples of this issue emerged from our findings. The first example surfaced when examining out-of-level test alignment. States assumed that out-of-level tests were aligned with the student's grade level of instruction. Yet, two states (Delaware, Iowa) made a critical point; that it is impossible to ensure that all students tested out of level are assessed at their instructional level. In fact, case study research conducted in local schools has demonstrated that students with disabilities who are tested out of level are not consistently tested at the grade level on which core content instruction is delivered (Minnema, Thurlow, & Warren, 2004a). Some students with disabilities were tested out of level below *both* the grade level of instruction and the grade level of enrollment. Even if the state's policy indicates that an out-of-level test must be administered at a student's instructional level, it cannot be assumed that every out-of-level test across a state is implemented in accordance with the intent of the policy. *States need to*

take measures to monitor the consistency with which out-of-level tests are used, and not simply assume that the out-of-level testing policy is consistently implemented in practice.

The second example appeared when considering states' opinions about out-of-level testing. States commented that one of the limitations of out-of-level testing is that it is assumed that students with disabilities who are tested out of level cannot achieve proficiency on grade of enrollment content standards. By assessing these students with out-of-level tests, they never receive the opportunity to demonstrate on-grade level proficiency. Students with disabilities—as is true of all students—deserve the chance to surprise their teachers and parents. Numerous stories have emerged in practice that testify to the possibilities of grade level proficiency when students are given the opportunity to learn from a challenging, rich curriculum that is delivered by high-quality, standards-based instruction (Minnema, Thurlow, & Warren, 2004b). When a subgroup of students is not afforded the opportunity to strive for grade-level standard proficiency, an incomplete picture is provided on which school improvement plans are derived. *Both legally and morally, it is inappropriate to systematically exclude some students from the full benefits of standards-based reform—the ultimate result of testing students with disabilities below the grade in which they are enrolled in school.*

The third example emerged in discussing psychometric properties of out-of-level test instruments, which again is disconcerting. States are assuming the psychometric soundness of both out-of-level tests and the resulting test scores without supporting this assertion with empirical evidence. Although this issue is repetitive, its importance merits the extra emphasis. *Again, states that are testing students with disabilities out of level need to statistically validate out-of-level tests with the population of students who are being assessed below grade level, which considers test score interpretation as part these critical validation studies.*

Concluding Remarks

In addition to the five key points discussed above, there is an additional theme that emerged from our findings that brings us back to the place at which we began discussing our findings. Across the states that participated in this study, there is wide variability in how states use out-of-level testing as an option for statewide testing. The unpredictability and volatility of out-of-level testing across the nation was noted in the disclaimer that opened our discussion. Our findings substantiate this concern—that the uneven practice of out-of-level testing promotes fluctuating circumstances that surround the testing of students with disabilities below the grade in which they are enrolled in school. In conclusion, it behooves states to examine both the instructional and assessment needs of students with disabilities who are perceived as a poor fit for either the general assessment or the alternate assessment. By incorporating principles of universally designed tests, using accommodations more appropriately, or improving instructional delivery,

these students will be given a better opportunity to demonstrate academic proficiency, which in turn will reflect enhanced quality in testing instruments.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Ayrer, J. E., & McNamara, T. C. (1973). Survey testing on an out-of-level basis. *Journal of Educational Measurement, 10*(2), 79-84.

Bielinksi, J., Thurlow, M., Minnema, J, & Scott., J (2000). *How out-of-level testing affects the psychometric quality of test scores* (Out-of-Level Testing Report 2). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at http://education.umn.edu/NCEO/OnlinePubs/OOLT2.html

Cleland, W. E., & Idstein, P. M. (1980, April). *In-level versus out-of-level testing of sixth grade special education students*. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Council of Chief State School Officers. (2002). *Model for alignment analysis and assistance to states*. Retrieved August 24, 2004, from http://www.ccsso.org/content/pdfs/AlignmentModels.pdf

Crowder, C. R., & Gallas, E. J. (1978, March). *Relation of out-of-level testing to ceiling and floor effects on third and 5th grade students*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada.

Federal Register (2003, December 9). *Title I -- Improving the academic achievement of the disadvantaged, Volume 68* (236). Retrieved December 9, 2003 from http://www.ed.gov/legislation/FedRegister/finrule/2003-4/120903a.html

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Jones, E. D., Barnette, J. J., & Callahan, C. M. (1983, April). *Out-of-level testing for special education students with mild learning handicaps*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the no child left behind act of 2001. *Educational Researcher, 31*(6), 3-16.

Long, J. V., Schaffran, J. A., & Kellogg, T. M. (1977). Effects of out-of-level survey testing on reading achievement scores of Title I, ESEA students. *Journal of Educational Measurement, 14*(3), 203-213.

Minnema, J.E., Thurlow, M.L., & Warren, S. (2004a). *Understanding out-of-level testing in local schools: A first case study of policy implementation and effects*. (Out-of-Level Testing Report 11). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Minnema, J.E., Thurlow, M.L., & Warren, S. (2004b). *Understanding out-of-level testing in local schools: A second case study of policy implementation and effects*. (Out-of-Level Testing Report 12). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Pellegrino, J. W. (2002, February 1). *Understanding how students learn and inferring what they know: Implications for the design of curriculum, instruction and assessment*. Paper presented at the NSF Instructional Materials Development Conference, Washington, DC.

Popham, W. J. (2001, April). *Standards-based assessment: Solution or charade*? Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002, May). *Benchmarking and alignment of standards and testing* (Technical Report 566). Los Angeles, CA: University of California, National Center for Research on Evaluation.

Study Group on Alternate Assessment. (1999). *Alternate assessment resource matrix: Considerations, options, and implications* (ASES SCASS Report). Washington, DC: Council of Chief State School Officers.

Thurlow, M., & Minnema, J. (2001). *States' out-of-level testing policies* (Out-of-Level Testing Report 4). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M.L., & Minnema, J.E. (2003). *Reporting out-of-level test scores: Are these students included in accountability programs?* (Out-of-Level Testing Report 10). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Webb, N. L. (2002, December). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Madison, WI: University of Wisconsin, Wisconsin Center for Educational Research.

Appendix A

Interview Protocol

Hi_____, first of all I want to thank you for taking the time to participate in this study. We think that it will yield some valuable information about out-of-level test alignment with academic content standards and technical information. Did you have any questions about the information provided to you in the e-mail (or fax)?

Yes: What are those questions?

No: Wonderful.

I would like to tape record this interview so that I do not lose any of the information you provide. If that is OK with you, I will begin tape recording the interview now. OK? [start tape recorder] OK, the tape recorder is on.

We realize that you may feel that some questions may be better answered by someone you work with. If so, please let me know and I will contact that person. Also, I want to emphasize that anytime standards are mentioned in the interview, we are referring to your state academic content standards. Do you have your questions in front of you at this time? Do you have any questions before we being the interview?

Ok:

1) Is there a clear statement or outline of the skills measured by the assessment? What research was conducted to arrive at these skills?

2) What processes do you use to ensure alignment of your statewide large-scale assessment(s) with standards?

3) If the test publisher performed this process of alignment, how was this process explained to you? Who was involved? [may we contact this group?]

4) If the alignment process was performed by another individual or group, what process did they use? How was it explained to you? Who was involved? [may we contact this group?]

5) Do you have a process to ensure alignment of out-of-level tests with standards at each grade level? If so, what is this process?

6) What grade level content standards are out-of-level tests aligned with: grade at which the student is tested or grade at which the student is enrolled in school? Are those the standards that the student is working towards in the classroom?

7) Do you have information that demonstrates that the knowledge and skills measured by on-level and out-of-level scores are the same? (probe: what information?)

8) Where could the general public find information on the alignment of statewide assessments with standards?

Thank you. We're at the half-way point! The second part of the interview will focus on the psychometric properties of scores derived from out-of-level testing procedures.

9) Where could consumers find on-line information that describes the test development and measurement characteristics of the tests? Does this information include references to out-of-level testing?

10) Is there a reason why the test technical information is (not) online?

11) Is the test technical manual/report available?

12) Could you please describe the equating process you used in the construction of your out-of-level procedures?

13) Could you please describe the rationale for your out-of-level equating (or your decision to not equate out-of-level scores)?

14) What are the intended interpretations and limitations of scores obtained from your out-of-level tests, if any? Are these the same interpretations and limitations as scores obtained from on-level tests?

15) Can you describe the accuracy of the equating functions used in your out-of-level equating procedures? Is information available that supports this statement?

16) Are scores obtained from out-of-level tests as precise as scores from on-level tests? Is information available that supports this statement, such as conditional standard errors of measurement for scores obtained from out-of-level testing?

17) Do you have information/data available from your testing program that reports the percent of students scoring at or below chance level on both on-level and out-of-level tests?

Once again, thank you very much for participating in this interview. We are conducting interviews with 13 other states. We will submit to you the portion of the final report where you state's information is included for you to review. Additionally, we will send you a copy of our final report. Feel free to contact me, the grant coordinator, or the principal investigator if you have any further questions. Have a great rest of your day!