AMERICAN INSTITUTES FOR RESEARCH®

Psychometric Advances in Alternate Assessment

Gary W. Phillips, Ph.D.
Vice President & Chief Scientist

Louis Danielson, Ph.D.
Managing Director for
Special Education Projects

Lynnett Wright, M.S.
Senior Alternate Assessment
Development Specialist

American Institutes for Research®
Updated - October 30, 2009

About AIR®

The American Institutes for Research (AIR) has funded and conducted this report as part of our effort to make research relevant to policymakers and practitioners in the field of education. Our mission at AIR is to conduct and apply behavioral and social science research to improve people’s lives and well-being, with a special emphasis on the disadvantaged.

Established in 1946, with headquarters in Washington, DC, and with 42 offices in the United States and around the world, AIR is a nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance both domestically and internationally in the areas of health, education, and workforce productivity.

AMERICAN INSTITUTES FOR RESEARCH®

Psychometric Advances in Alternate Assessments

Gary W. Phillips, Ph.D.
Vice President & Chief Scientist

Louis Danielson, Ph.D.
Managing Director for Special Education Projects

Lynnett Wright, M.S.
Senior Alternate Assessment Development Specialist

American Institutes for Research®

Updated - October 30, 2009

TABLE OF CONTENTS

AMERICAN INSTITUTES FOR RESEARCH®

PSYCHOMETRIC ADVANCES IN ALTERNATE ASSESSMENTS

Executive Summary

This report provides the technical details of an alternate assessment design that has resulted from a long-term research and development effort at the American Institutes for Research (AIR). The methodology is currently in use in two states, South Carolina and New Mexico. The design overcomes many of the inaccuracies and inefficiencies of current methods used by most states. The AIR design is more accurate because: (1) the test is adaptive, i.e., the difficulty of the items is adapted to the ability of the students, and (2) the design puts the psychometrics of alternate assessment on par with regular assessments, allowing states to measure growth across years and grades for special education students. The design is more efficient because it costs less than most other alternate assessments and requires less testing time from students and teachers. The design has been approved under No Child Left Behind (NCLB) for New Mexico and is expected to be approved for South Carolina.

Background

During the early 1990s both Kentucky and Maryland were working to implement assessments designed to measure educational outcomes for children with significant cognitive disabilities. The 1997 amendments to the Individuals with Disabilities Education Act (IDEA) provided the impetus that has led to the widespread development and implementation of alternate assessments with alternate achievement standards (AA-AAS). The 1997 amendments required that all children with disabilities participate in state assessments and that alternate assessments be available for those students who could not meaningfully participate in the regular state assessment even with accommodations. These requirements were to be implemented by July 2000.

This was a challenge for states because little knowledge was available to help develop these assessments. Technical assistance centers were funded to work with states, but they were learning along with the states. There were challenges related to how to assess these students and also challenges concerning what to assess. The initial assessment approaches relied heavily on portfolios and primarily assessed functional skills; these are the skills that help one function independently in a real-world environment. They include self-help skills such as dressing and grooming, social skills such as appropriate interaction with peers, and perhaps also the ability to move about the environment using transportation. The enactment of NCLB in 2001 significantly shifted thinking about the AA-AAS by requiring that these assessments be aligned with grade-level content standards. This forced most states to redesign their assessments to shift from a function-based approach and demonstrate that their assessments were aligned with the grade-level content. In addition, the Peer Review Guidance for NCLB developed by the U.S. Department of Education also elevated standards regarding technical adequacy of these assessments. Discussions began about how one might ensure that these assessments also met the high standards laid out for the general state assessment.

Questions have been raised about the assessment approaches and whether there are ways to measure programmatic improvement for these students. That is, are there year-to-year improvements in the performance of this population of students? This is the fundamental question addressed in this report.

Alternate Assessments with Better Psychometric Properties

One alternate assessment method commonly used in many states is to collect a *Portfolio* or *Body of Evidence* about a student's academic progress toward performance goals in the general curriculum. The assessment is carried out by collecting data over a long period of time from multiple sources (such as student work samples, scores from teacher-made tests, parent interviews, checklists, and audio- and videotapes). Generally this evidence is collected over an entire semester or academic year and could include over 100 pieces of evidence.

The problem with both the Portfolio or Body of Evidence alternate assessments is that they have few psychometric properties; therefore, many claims about them are simply untrue. One significant shortcoming is that claims about test reliability are usually related to inter-rater reliability and not score reliability. Inter-rater reliability can be very high (i.e., two independent raters tend to agree on the student performance) but score reliability can still be very low (i.e., the scores do not reliably assess what the test is intended to measure). Inter-rater reliability is a necessary but not sufficient condition for score reliability.

However, the most serious psychometric shortcoming of Portfolio or Body of Evidence alternate assessments is that they are not scaled. Often only an ad hoc raw score scale is available, which means it is not possible to equate these assessments across years or link them across grades. Although most states claim to measure progress, this is not statistically possible. Without equated tests, progress cannot be measured from one year to the next, and without a vertical scale, progress from one grade to the next cannot be measured.

Adaptive Alternate Assessment of Growth

Over the past several years AIR has had an ongoing research and development effort aimed at improving the state of the art in assessments for special education students. The goal was to develop a design that was *as psychometrically sound as the assessments for the general population*. The design involves the use of Item Response Theory (IRT) psychometric models with an adaptive on-demand alternate assessment. This section describes the AIR design.

ITEMS WRITTEN BY SUBJECT AREA AND SPECIAL EDUCATION EXPERTS

The *Adaptive Alternate Assessment of Growth* is written by an expert team of item writers and special educators. The items are tailored for students with significant cognitive disabilities who have been determined by the Individualized Education Program (IEP) team to be unable to participate in the general state assessments even with appropriate accommodations.

LINK TO STATE CONTENT STANDARDS

The *Adaptive Alternate Assessment of Growth* is based on each state's content standards, benchmarks or indicators for students in grades 3–8 and high school in Language Arts (reading and writing), Mathematics and Science. The AIR team works with the state department of education's assessment, content area and special education experts to ensure that the tasks are assessing grade-level content at a level accessible for the target population.

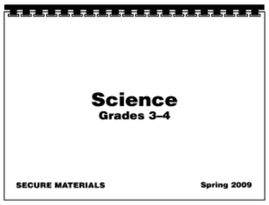
TEST IS ADAPTED TO THE STUDENT'S LEVEL OF ABILITY

The *Adaptive Alternate Assessment of Growth* is an adaptive test of 12 to 15 tasks ordered by difficulty within a grand band. The student's teacher administers the scripted questions orally in a one-on-one setting. The teacher selects the starting task (the one most appropriate for the student's ability level) with the help of a Student Placement Questionnaire (SPQ). This is a brief rating instrument that represents the range of communication levels and cognitive-academic functioning found in the population of alternate assessment examinees. Teachers complete the SPQ in each content area to identify the most appropriate starting task for each student. The process takes approximately 6 minutes. *Research has indicated that the SPQ has worked well in targeting starting tasks, resulting in 88% accuracy in identifying the most appropriate starting task.* This prevents the lowest-achieving students from being embarrassed by tests that are too difficult and prevents the high-achieving students from needing to answer items that we already know they are likely to get correct. Instead, the lowest-ability students are administered an easier set of tasks, and the higher-ability students are administered a more difficult set of tasks.

ON-DEMAND ASSESSMENT

The *Adaptive Alternate Assessment of Growth* administration consists of five major components:

1. **Test booklet:** There is one test booklet per grade band and content area. Each booklet contains 12 to 15 tasks, ordered by difficulty, with 4 to 7 items, also ordered by difficulty, per task. The tasks are placed in the booklet in order of difficulty, and the booklet is spiral bound and divided by task.



Science
Grades 3–4

SECURE MATERIALS Spring 2009

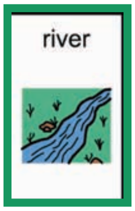
2. **Specific script:** The test contains a specific script that the teacher reads to the student during administration.

   The following is an example of a script:

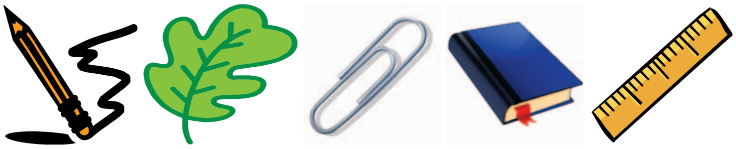
   > **Say:** ***Show (tell) me where fish live: in a river*** (indicate the river card), ***in a desert*** (indicate the desert card***), or on a mountain*** (indicate the mountain card)**.**

3. **Printed manipulatives:** Because students with significant cognitive disabilites communicate at various levels, the alternate assessment allows them to respond orally, through sign

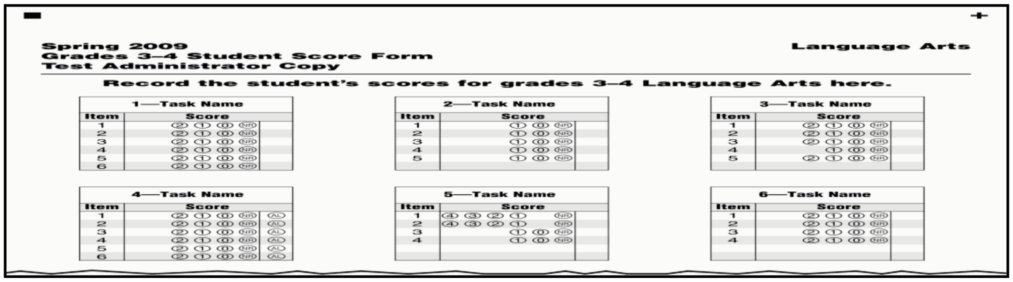
language, through augmentative communication, or by pointing to the correct answer. The following are examples of the printed word cards that are used for student responses. The green box indicates the correct answer.



4. **Physical manipulatives:** Instruction for students with significant cognitive disabilites usually includes the use of concrete objects. Physical manipulatives are used throughout the assessment. The following are examples of physical manipulatives used in the classroom during instruction and in test administration.



5. **Student answer folder:** The student responses are recorded and scored by the teacher. The answer document contains the names of the tasks and the point score earned for each student answer.



Spring 2009
Grades 3–4 Student Score Form
Test Administrator Copy

Language Arts

Record the student's scores for grades 3–4 Language Arts here.

1—Task Name	
Item	Score
1	② ① ⓪ (NR)
2	② ① ⓪ (NR)
3	② ① ⓪ (NR)
4	② ① ⓪ (NR)
5	② ① ⓪ (NR)
6	② ① ⓪ (NR)

2—Task Name	
Item	Score
1	① ⓪ (NR)
2	① ⓪ (NR)
3	① ⓪ (NR)
4	① ⓪ (NR)
5	① ⓪ (NR)

3—Task Name	
Item	Score
1	② ① ⓪ (NR)
2	② ① ⓪ (NR)
3	② ① ⓪ (NR)
4	① ⓪ (NR)
5	② ① ⓪ (NR)

4—Task Name		
Item	Score	
1	② ① ⓪ (NR)	(AL)
2	② ① ⓪ (NR)	(AL)
3	② ① ⓪ (NR)	(AL)
4	② ① ⓪ (NR)	(AL)
5	② ① ⓪ (NR)	(AL)
6	② ① ⓪ (NR)	(AL)

5—Task Name	
Item	Score
1	④ ③ ② ① (NR)
2	④ ③ ② ① (NR)
3	① ⓪ (NR)
4	① ⓪ (NR)

6—Task Name	
Item	Score
1	② ① ⓪ (NR)
2	② ① ⓪ (NR)
3	② ① ⓪ (NR)
4	② ① ⓪ (NR)

SCORED BY TEACHERS

One of the key features of the *Adaptive Alternate Assessment of Growth* is that the teacher records and scores the student's responses for each item. All scoring directions are on the right side of the script, and possible point scores vary from item to item. The teacher then scores the answer sheet.

A special video study was conducted during the 2008 South Carolina Alternate Assessment administration to confirm that test administrators were following all scoring procedures accurately. For this study, *scoring accuracy* refers to the degree to which teachers follow scaffolding and scoring directions correctly and assign accurate scores to student responses. Scoring accuracy among test administrators was evaluated by having trained raters at AIR review videotapes of the test administrations and assume the role of the test administrator in scoring student responses. The AIR raters did not know the scores assigned by the test administrators at the time of their own assignment of scores. After the raters concluded their scoring of the student responses, the consistency between the test administrators and AIR raters was determined. *The results indicated that there was a high degree of consistency (92% to 95%) between the scores given by the test administrators and those of the AIR raters*, suggesting that test administrators in the state understood the scoring procedures and implemented them accurately when scoring student responses.

Psychometric Characteristics of the Test

Reliability

Classical test theory-based reliability indices such as Cronbach's Alpha are not appropriate for the *Adaptive Alternate Assessment of Growth* design because: (1) the length of the test may differ for each student, and (2) the test uses IRT methods, where the measurement error is a function of student ability (θ). The reliability coefficient for the *Adaptive Alternate Assessment of Growth* uses the *marginal reliability* (Sireci, Thissen, & Wainer, 1991), which is equivalent to internal consistency estimates of reliability.

First we determine the marginal measurement error variance, $\overline{}^{2}_{e}$, across all examinees,

$$\overline{}^{2}_{e} \qquad {}^{2}_{e}p(\quad)d \qquad \frac{{}^{2}_{e}}{N}, \tag{1}$$

where ${}^{2}_{e}$ is the square of the standard error of student ability, θ. Thus, the marginal measurement error variance can be estimated as the average of the squared standard error of θ. Then we estimate the marginal reliability as

$$- \qquad \frac{{}^{2} \quad \overline{}^{2}_{e}}{{}^{2}}, \tag{2}$$

where σ_{θ}^{2} is the variance of observed θ estimates.

The marginal reliability estimate, $\overline{\rho}$, can be interpreted similarly to classical reliability indices such as Cronbach's Alpha. Estimates of the marginal reliability for the test forms corresponding to the five grade bands in New Mexico are provided in Table 1.

Table 1: Marginal Reliability by Grade Band and Subject for 2008 New Mexico Alternate Assessment

	Language Arts		Math		Science		Writing	
Grade Band	***N***	**Reliability**	***N***	**Reliability**	***N***	**Reliability**	***N***	**Reliability**
Grade 3–4	386	0.903	370	0.922	383	0.916	386	0.760
Grade 5–6	384	0.896	382	0.907	399	0.912	384	0.828
Grade 7–8	447	0.887	439	0.898	449	0.924	447	0.841
Grade 9–10	162	0.923	168	0.913	167	0.924	162	0.890
Grade 11–12	270	0.933	261	0.907	273	0.927	270	0.836

The marginal reliability estimates for the South Carolina Alternate Assessment are provided in Table 2.

Table 2: Marginal Reliability by Grade Band and Subject for 2008 South Carolina Alternate Assessment

	English Language Arts		Math		Science		Social Studies	
Grade Band	***N***	**Reliability**	***N***	**Reliability**	***N***	**Reliability**	***N***	**Reliability**
Grade 3–5	1229	0.909	1226	0.903	879	0.901	846	0.891
Grade 6–8	1050	0.908	1045	0.906	761	0.909	724	0.865
Grade 10	373	0.907	372	0.906	370	0.882		

As can be seen from Tables 1 and 2, the *Adaptive Alternate Assessment of Growth* design provides tests in both states that are as reliable as typical tests for the general population.

CLASSIFICATION ACCURACY

In addition to reliable scores, a test needs to classify students correctly into performance levels. In the *Adaptive Alternate Assessment of Growth*, classification accuracy was estimated for each cut score as the average probability of correct performance-level assignments across all examinees (assignments above or below the cut score), given each examinee's estimated proficiency score, θ_i:

$$\mathrm{CA_K} = \frac{\sum_{k \geq K} P(\theta_i > \theta_K^* \mid \theta_i, k_i) + \sum_{k<K} 1 - P(\theta_i > \theta_K^* \mid \theta_i, k_i)}{N}; \qquad (3)$$

where

θ_i is the proficiency (i.e., theta) of student *i*;

k_i is the performance level of student *i*;

θ_K^* is the cut score for the performance level *K* on the theta scale; and

N is the total number of students.

$P(\theta_i \geq \theta_K^* \mid \theta_i, k_i)$ is the probability of a student with θ_i and performance level k_i being at or above the cut score θ_K^*. The classification accuracy is the expected rate of correct classification probability; therefore, the higher value indicates the superior classification accuracy. Table 3 and Table 4 show the classification accuracy by content area, performance level and grade band.

Table 3: Classification Accuracy by Grade Band and Subject for 2008 New Mexico Alternate Assessment

		Grade Band					
Subject	**Performance Level**	**3–4**	**5–6**	**7–8**	**9–10**	**11–12**	**Overall**
Language Arts	Nearing Proficient	0.936	0.945	0.931	0.953	0.968	0.945
	Proficient	0.881	0.898	0.921	0.902	0.881	0.896
	Advanced	0.857	0.864	0.915	0.894	0.869	0.880
Mathematics	Nearing Proficient	0.929	0.918	0.957	0.899	0.895	0.922
	Proficient	0.884	0.890	0.893	0.890	0.887	0.890
	Advanced	0.872	0.809	0.859	0.842	0.853	0.850
Science	Nearing Proficient	0.908	0.924	0.914	0.886	0.941	0.918
	Proficient	0.908	0.889	0.905	0.875	0.893	0.900
	Advanced	0.825	0.805	0.811	0.869	0.863	0.827
Writing	Proficient	0.794	0.829	0.883	0.832	0.838	0.840

Table 4: Classification Accuracy by Grade Band and Subject for 2008 South Carolina Alternate Assessment

		Grade Band			
Subject	**Performance Level**	**3–5**	**6–8**	**10**	**Overall**
English Language Arts	Level 2	0.927	0.928	0.929	0.927
	Level 3	0.846	0.899	0.889	0.873
	Level 4	0.889	0.908	0.898	0.898
Mathematics	Level 2	0.893	0.882	0.889	0.888
	Level 3	0.876	0.901	0.866	0.884
	Level 4	0.868	0.887	0.890	0.879
Science	Level 2	0.875	0.901	0.870	0.884
	Level 3	0.876	0.896	0.837	0.877
	Level 4	0.863	0.883	0.852	0.870
Social Studies	Level 2	0.911	0.910		0.911
	Level 3	0.869	0.879		0.874

These tables indicate that the alternate assessments in both states have very high levels of classification accuracy.

Alternate Achievement Standards

One defining characteristic of alternate assessments is that they have alternate achievement standards. The *Adaptive Alternate Assessment of Growth* uses the same procedure to establish alternate achievement standards as is used for the general education students. The alternate achievement standards are established by a broadly representative group of about 100 panelists from across the state (teachers, administrators, parents, and business and community leaders). The alternate achievement standards are typically determined through the Bookmark method or Item-Descriptor (ID) Matching (which is a simplified Bookmark procedure developed at AIR; see Cizek & Bunch, 2007 and Ferrara, Perie, & Johnson, 2003). With both methods the panelists review test items in an ordered-item booklet (items are ordered in a booklet by difficulty, with the easiest items in the front of the booklet and the hardest toward the end), review corresponding Performance-Level Descriptors (descriptions of what students should know and be able to do for each performance standard) then recommended performance standards such as Basic, Proficient and Advanced. These standards are then translated into cut points on the student proficiency scale using item-response theory methods.

Differential Item Functioning

AIR conducts analyses of differential item functioning (DIF) on all test items to detect potential item bias across demographic groups. The purpose of these analyses is to identify items that may have favored students in one group over students of similar ability in another group. For example, DIF analyses may be conducted to compare Hispanic/White, Native-American/White, and Female/Male student subgroups. Although sample sizes are very small for each subgroup, we calculate DIF statistics wherever possible for the purposes of item review.

All items on the *Adaptive Alternate Assessment* are partial credit items. For partial credit items, we calculate both the Mantel-Haenszel chi-square (MHχ^2; Zwick & Thayer, 1996; Zwick, Donoghue, & Grima, 1993) and the Standardized Mean Difference (SMD) (see Dorans & Kulick, 1986). The classification rules are defined in the table below. Items in the “C” DIF category are flagged for review, indicating evidence of DIF on the items. The DIFF classification rules are displayed in the table below.

Table 5: Summary of DIF Classification Rules for Partial Credit Items

DIF Category	Rule
C	The *p*-value of MHχ^2 is less than .05, and |SMD|/|SD| is greater than 0.25.
B	The *p*-value of MHχ^2 is less than .05, and |SMD|/|SD| is greater than 0.17 and less than 0.25.
A	The *p*-value of MHχ^2 is not significant at the .05 level or |SMD|/|SD| is less than 0.17.

Traditional DIFF analysis requires that students in each group be sorted into comparable raw score categories. That is not possible with the *Adaptive Alternate Assessment* because raw scores are not comparable across students (since each student takes a different set of items). Instead of using raw scores, the *Adaptive Alternate Assessment* uses the Item Response theory theta scale for purposes of sorting groups into comparable categories. Some illustrative examples of differential item functioning (as well as other statistics) are provided in Table 6 and Table 7.

Table 6: Differential Item Function Results (and other statistics) for 5-6 Grade Band 2008 Language Arts in New Mexico Alternate Assessment

					DIF		
Item Position	Adjusted Polyserial Correlation	Average Score	Access Limitation	Omit	Female vs. Male	Hispanic vs. White	Native American vs. White
12	0.55	0.64	0.00	0.02	-A	-A	+A
13	0.55	0.66	0.00	0.01	-A	+A	-B
14	0.52	0.45	0.00	0.01	+A	-A	+B
15	0.76	0.75	0.00	0.00	+A	-A	+A

Table 7: Differential Item Function Results (and other statistics) for 3-5 Grade Band 2008 Social Studies Field Test in South Carolina Alternate Assessment

					DIF	
Item Position	Adjusted Polyserial Correlation	Average Score	% Omit	% Access Limitation	Female vs. Male	Black vs. White
54	0.49	0.38	1.82	0.00	+A	+A
55	0.48	0.43	1.62	1.45	-A	-A
56	0.45	0.40	1.46	0.73	-A	-A
57	0.57	0.40	0.71	0.00	+B	+A

VERTICAL SCALE

The most unique aspect of the *Adaptive Alternate Assessment of Growth* is that it is vertically scaled using the partial-credit Rasch item-response theory model across the grade bands within each subject area. This process provides an alternate assessment that has all the psychometric characteristics of the state's assessment of the general population. These characteristics include the following:

Each item on the test is a partial-credit item which means the partial-credit Rasch model is used for item calibration, scaling and achievement estimation.

The difficulty of the items is *adapted* to the ability level of the student. This results in better measurement of student proficiency.

Interval level scaled scores (instead of raw scores) are used to summarize student performance.

Embedded field tests can be used to field test new items (permitting the release of previously used items) while maintaining the same scale year after year.

Scales are linked across grades so growth can be measured from year to year.

- The same quality of statistical analysis is provided as in the general education population (reliability and validity indices, bias and fairness reviews, classification consistency indices, generalizability analyses, etc.).

The vertical linking design is presented below for the mathematics and science for both New Mexico and South Carolina. These tables show how tasks are distributed across grades to permit the vertical linking.

Table 8: Vertical Linking Design for Tasks in Mathematics and Science in 2008 New Mexico Alternate Assessment

Grade	3-4	3-4/ 5-6	5-6	5-6/ 7-8	7-8	7-8/ 9-12	9-12	
Math	**Unique**	**Linking**	**Unique**	**Linking**	**Unique**	**Linking**	**Unique**	**Total**
3-4	7	5						12
5-6			2	5				12
7-8					2	5		12
9-12							7	12
Science	**Unique**	**Linking**	**Unique**	**Linking**	**Unique**	**Linking**	**Unique**	**Total**
3-4	7	5						12
5-6			2	5				12
7-8					2	5		12
9-12							7	12

Table 9: Vertical Linking Design for Tasks in Mathematics and Science in 2008 South Carolina Alternate Assessment

Mathematics Grade	Unique Tasks	Linking Tasks			Total
10	6	2	—	4	12
6–8	2		4		12
3–5	4	—			12
Science Grade	**Unique Tasks**	**Linking Tasks**			**Total**
10	8	4	—		12
6–8	3		5		12
3–5	7	—			12

MEASUREMENT OF GROWTH

A major benefit of the psychometric properties of the *Adaptive Alternate Assessment of Growth* is that the state has the capacity to measure growth for individual students.

The *Adaptive Alternate Assessment of Growth* uses the Rasch model as the underlying psychometric model. Most users of the Rasch model use number-correct scoring. However, since the test is adaptive each student receives a different set of items, so number-correct scoring cannot be used with the *Adaptive Alternate Assessment of Growth*. Instead, AIR uses pattern scoring with the Rasch model. This is accomplished as follows.

Indexing items by *i*, the likelihood function based on the *j*th person's score pattern for k_i items is

$$L_j(\theta \mid \mathrm{z}_\mathrm{j}, \mathbf{b}'^j_1, \ldots \mathbf{b}'^j_{k_j}) = \prod_{i=1}^{k_i} p_i(z_{ji} \mid \theta, b^j_{i,1}, \ldots, b^j_{i,m^j_i}), \tag{4}$$

where $\mathbf{b}'^j_i = (b^j_{i,1}, \ldots, b^j_{i,m^j_i})$ is the *i*th item's parameters and m^j_i is the possible score of this item. Depending on the item type, the probability $p_i(z_{ji} \mid \theta, b^j_{i,1}, \ldots, b^j_{i,m^j_i})$ takes either the form of a dichotomously scored item (in which case, we only have $b^j_{i,1}$, which can be simply written as b^j_i), or the form based on Master's partial credit model for the polytomous items.

In case of dichotomously scored items, we have

$$p_i(z_{ji} \mid \theta, b^j_i) = \begin{cases} \dfrac{\exp(\theta - b^j_i)}{1 + \exp(\theta - b^j_i)} = p_i & \text{if } z_{ji} = 1 \\ \dfrac{1}{1 + \exp(\theta - b^j_i)} = 1 - p_i & \text{if } z_{ji} = 0 \end{cases},$$

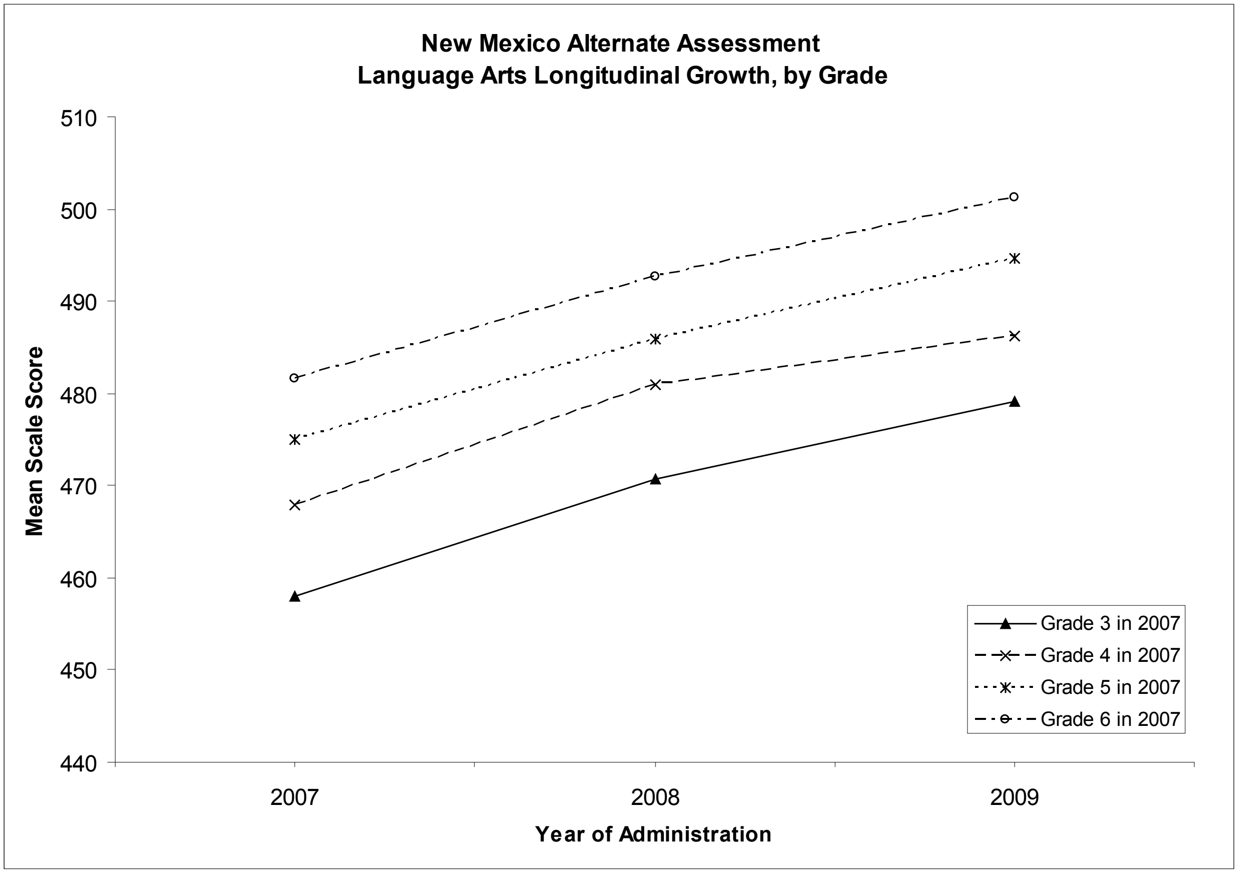
and in case of polytomous items, $p_i(z_{ji} \mid \theta, b^j_{i,1}, \ldots, b^j_{i,m^j_i}) = \begin{cases} \dfrac{\exp(\sum_{r=1}^{z_{ji}} (\theta - b^j_{i,r}))}{s_i(\theta, b^j_{i,1}, \ldots, b^j_{i,m^j_i})} & \textit{if } z_{ji} > 0 \\ \dfrac{1}{s_i(\theta, b^j_{i,1}, \ldots, b^j_{i,m^j_i})} & \textit{if } z_{ji} = 0 \end{cases}$, where

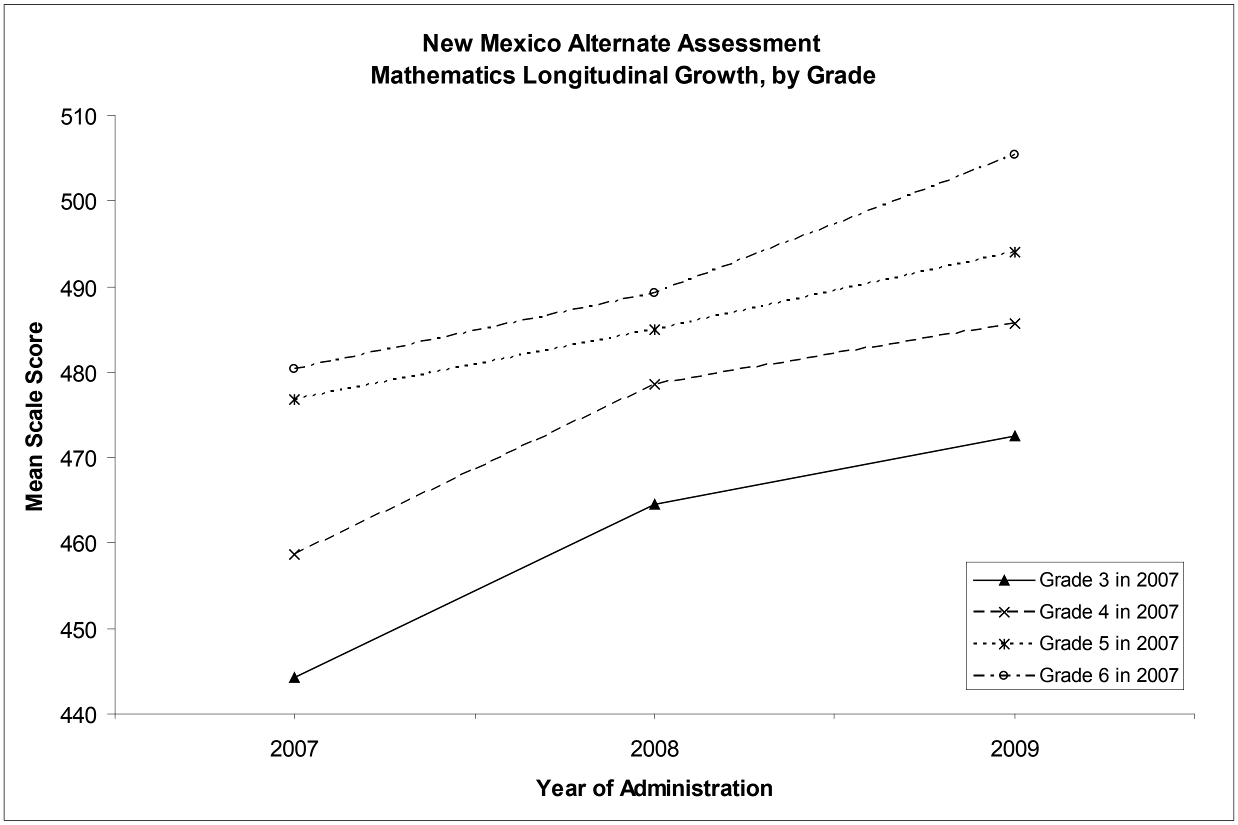
$s_i(\theta, b^j_{i,1}, \ldots, b^j_{i,m^j_i}) = 1 + \sum_{l=1}^{m^j_i} \exp(\sum_{r=1}^{l} (\theta - b^j_{i,r}))$ and the loglikelihood function is

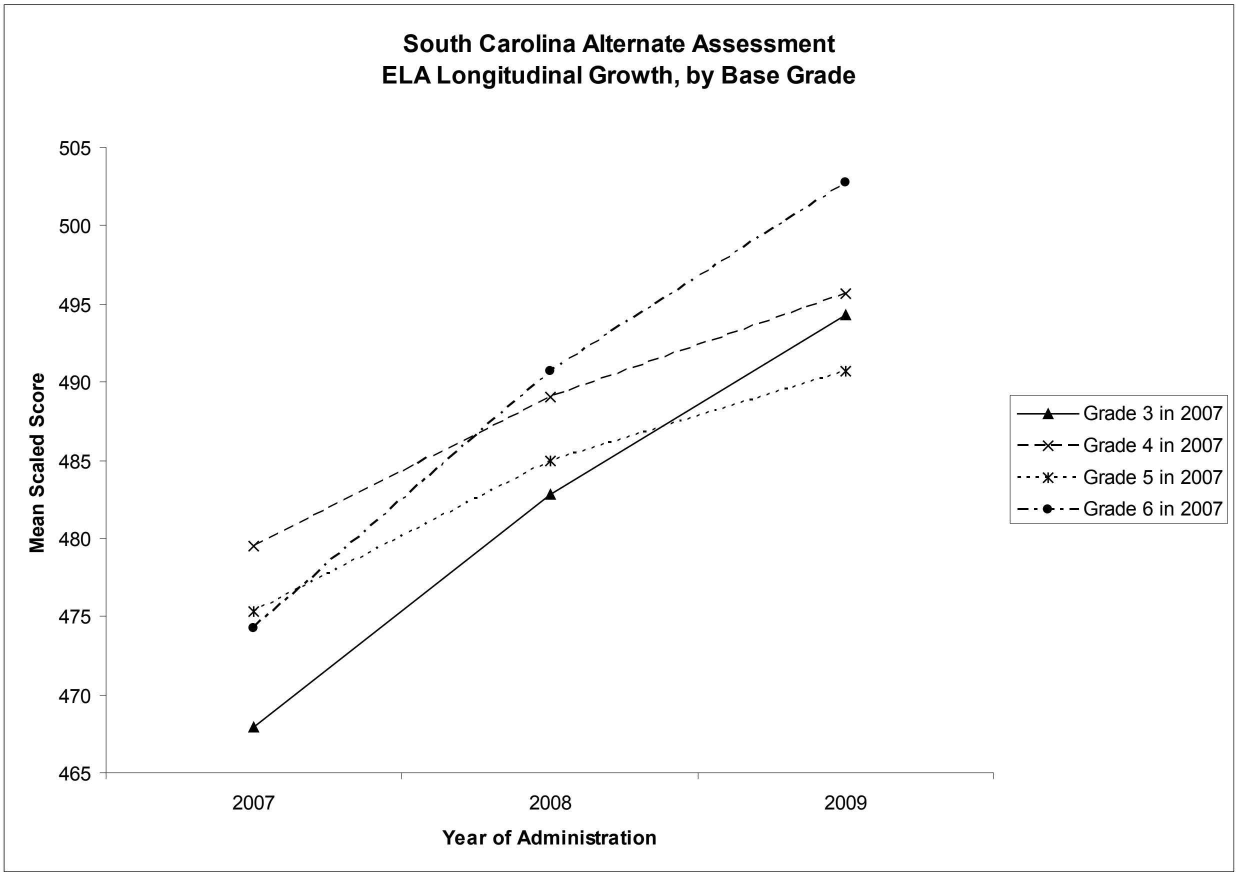
$$\begin{aligned} l_j(\theta \mid \mathrm{z}_\mathrm{j}, b^j_{i,1}, \ldots, b^j_{i,m^j_i}) &= \log(l_j(\theta \mid \mathrm{z}_\mathrm{j}, b^j_{i,1}, \ldots, b^j_{i,m^j_i})) \\ &= \sum_{i=1}^{k} \log(p_i(z_{ji} \mid \theta, b^j_{i,1}, \ldots, b^j_{i,m^j_i})) \end{aligned} \tag{5}$$

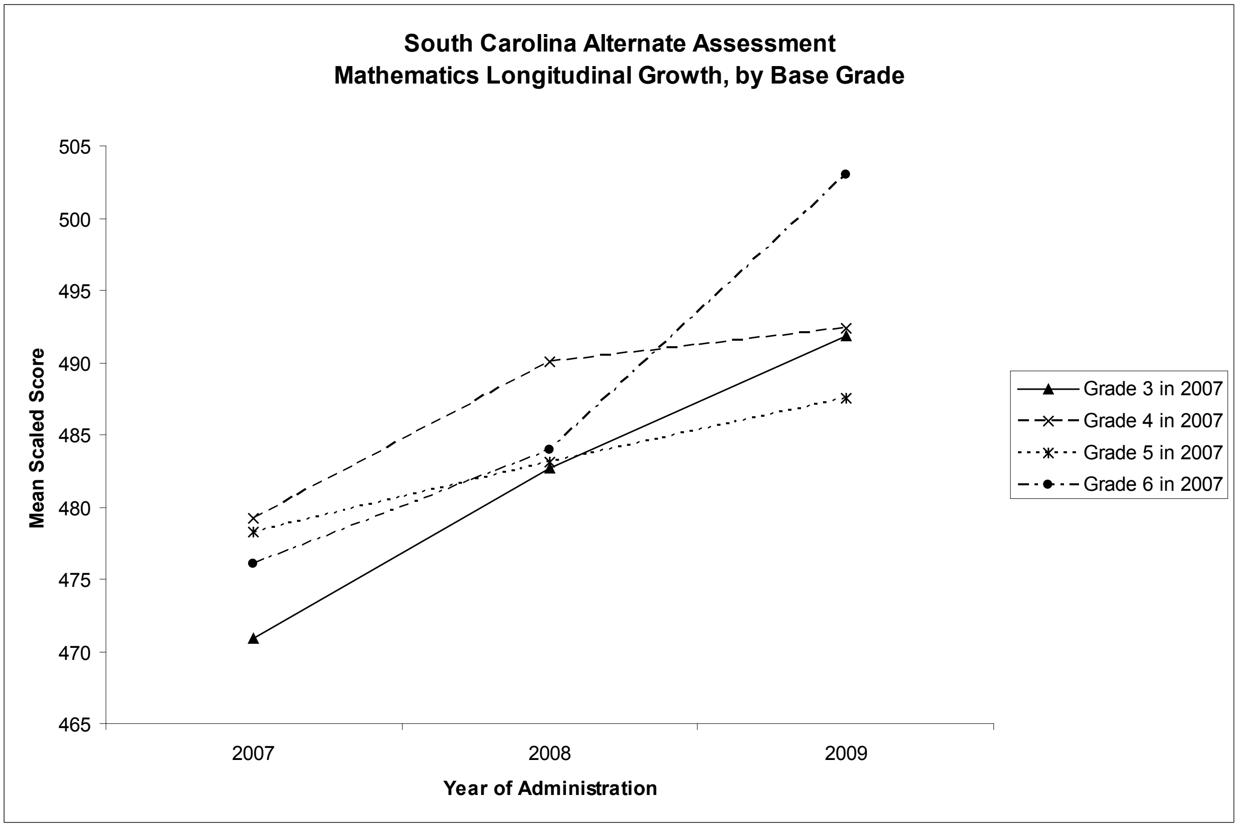
The ability θ is estimated by maximizing the log likelihood function defined in equation 5, and the standard error of measurement (SEM) is approximated by the square root of the inverse of the Fisher information evaluated at the maximum likelihood estimate (MLE) of θ.

Below are four examples of growth. The first two (Language Arts and Mathematics) are from New Mexico, and the second two (English Language Arts and Mathematics) are from South Carolina. In all four graphs we have tracked the cohort of students from 2007 to 2009. The 2007 year is the base test grade for the cohort and 2008 and 2009 are the subsequent assessments for the same students. As can be seen from the graphs, the *Adaptive Alternate Assessment of Growth* generally shows growth in student proficiency from year to year.




New Mexico Alternate Assessment
Mathematics Longitudinal Growth, by Grade
Mean Scale Score
510
500
490
480
470
460
450
440
2007
2008
2009
Year of Administration
Grade 3 in 2007
Grade 4 in 2007
Grade 5 in 2007
Grade 6 in 2007


South Carolina Alternate Assessment
ELA Longitudinal Growth, by Base Grade
505
500
495
490
485
480
475
470
465
Mean Scaled Score
2007
2008
2009
Year of Administration
Grade 3 in 2007
Grade 4 in 2007
Grade 5 in 2007
Grade 6 in 2007



Discussion

Getting an accurate measurement of the achievement of students with disabilities is one of the most important responsibilities of our schools. It is only after we know what they have learned that we are able to target instruction that helps them make progress. Unfortunately, alternate assessments have historically not been developed with the same level of psychometric rigor as the regular criterion-referenced assessments developed for the general education population. As a general rule they are not scaled in any sophisticated way, are not equated from year to year, and therefore cannot measure growth across grades or over time. Even though vendors and state departments of education claim they can measure growth, in fact they cannot.

The long-term research and development effort at the American Institutes for Research was undertaken in order to develop a technically sound, criterion-referenced assessment for students with disabilities that was on par with the current assessments for the general population. This paper reports the results of this research and demonstrates that it is possible to provide students with disabilities an assessment that is cost effective, is less burdensome to students and teachers, measures growth in student achievement across grades and from one year to the next, and is technically comparable to assessments with the general population.

References

Cizek, G. J., & Bunch, M. B. (2007). Standard setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks, CA: Sage Publications.

Dorans, N., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368.

Ferrara, S., Perie, M., & Johnson, E. (2002, September). *Matching the judgmental task with standard setting panelist expertise: The Item-Descriptor (ID) Matching procedure*. Invited colloquium for the Board on Testing and Assessment of the National Research Council, Washington, DC.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 234–247.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233–251.

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics*, *21*(3), 187–201.

AIR®

Headquarters ■ 1000 Thomas Jefferson Street, NW ■ Washington, DC 20007-3835 ■ www.air.org