Documenting Reading Achievement and Growth for Students Taking Alternate Assessments

Gerald Tindal\(^1\), Joseph F. T. Nese\(^1\), Dan Farley\(^1\), Jessica L. Saven\(^1\), and Stephen N. Elliott\(^2\)

Abstract

Students with disabilities have been included in state accountability systems for more than a decade; however, only in the past few years have alternate assessments of alternate achievement standards (AA-AAS) become stable enough to allow examination of these students’ achievement growth. Using data from Oregon’s AA-AAS in Reading during the period 2008–2009 to 2010–2011, we examined the achievement growth for a sample of 1,061 elementary students using two growth models: a transition matrix and a multilevel linear growth model. The authors found with the transition matrix model that a majority of students remained at the same performance level from one year to the next, whereas with the multilevel linear growth model, students’ scores revealed small, but statistically meaningful, growth year to year. The article concludes by noting advantages and disadvantages of these models to characterize growth and their implications for policy and practice.

Students with significant cognitive disabilities (SWSCDs) have been included in alternate assessments as part of their states’ large-scale testing programs since 2004. During the past decade, a considerable amount of research has been published on various aspects of alternate assessments for students judged against alternate assessments of alternative achievement standards (AA-AAS). However, for a number of technical and conceptual reasons, very little is known about the achievement growth of these students as measured by an AA-AAS.

The purpose of the present study was to examine the reading achievement growth of SWSCDs on Oregon’s AA-AAS using two approaches to determining growth: a transition matrix model and a multilevel linear growth model. This research is critical in that any attention to growth requires technical adequacy in the development and implementation of alternate assessments. We first consider critical issues for alternate assessments that are uniquely defined for this population of students and then review the limited research that has been conducted on growth for SWSCDs.

Policy and Regulatory Context for Alternate Assessments

The 1997 amendments of the Individuals with Disabilities Education Act (IDEA) specified that all students must have access to the general curriculum. This access was reaffirmed in the 2004 reauthorization of IDEA (IDEA, 2006). Also, the No Child Left Behind Act (NCLB; 2006) required that all students be assessed in reading and math for accountability purposes.

\(^1\)University of Oregon  
\(^2\)Arizona State University

Corresponding Author:  
Gerald Tindal, University of Oregon, 175 Education, 5262 University of Oregon, Eugene, OR, 97405-5262, USA.  
E-mail: geraldt@uoregon.edu
and that results be disaggregated by subgroups of students (e.g., race-ethnicity, language, special education status). In the initial regulations, a number of conditions were specified and then later implemented that focused on the design and implementation of alternate assessments (U.S. Department of Education, 2003). These regulations were enacted for including students with the most significant cognitive disabilities and became part of the peer review process (U.S. Department of Education, 2004). Later, nonregulatory guidance was published for states to use in defining the population, describing the testing program to be implemented and the alternate achievement standards to be aligned and achieved, and reporting of results (U.S. Department of Education, 2005).

**Given the complexities of the population and the measurement systems, and the difficulties associated with both aligning to standards and establishing technical adequacy, it is not surprising that little information is available on the achievement growth of this population.**

Students with disabilities need to be allowed to demonstrate achievement, but many cannot do so in the general large-scale assessment even after intensive, evidence-based interventions and appropriate, allowable assessment accommodations. Therefore, the inferences from alternate assessments must take into consideration the supports needed for students essentially taking a modified test. For example, changes are typically allowed in the levels of support or in the breadth, depth, or complexity of the standards being assessed. **Breadth** refers to the “comprehensiveness of the content and skills embodied in the standards, curriculum, or assessments” (Council of Chief State School Officers [CCSSO], 2003, p. 5); **depth** refers to “the level of cognitive processing (i.e., recognition, recall, problem solving, analysis, synthesis, and evaluation) required for success relative to the performance standards” (CCSSO, 2003, p. 6). **Complexity** is considered in the language used to express the item content. Finally, other supports also are allowed in the types of scaffolds, prompts, and assistive technologies used in the administration of the assessment.

In early publications, the focus was on defining alternate assessments and providing the field an introduction (Kleinert, 2001), which has been more recently updated to explicate what alternate assessments were intended to measure (Kettler et al., 2010) and to describe the plethora of options that have been adopted in developing and implementing alternate assessments (Schafer & Lissitz, 2009). Over the decade, considerable research has addressed definitions of the population designed to take the tests as well as the development of the tests.

**Growth of SWSCDs on Alternate Assessments**

After nearly a decade of development and implementation of alternate assessments, the field is characterized by a variety of assessment adoptions and profiles that vary considerably across states (Cameto et al., 2009; Kohl, McLaughlin, & Nagle, 2006). Nevertheless, and particularly in the presence of this diversity in assessment, the critical issue is whether administration of alternate assessments meaningfully records improvement in the performance of students who qualify for this testing option. Given the complexities of the population and the measurement systems, and the difficulties associated with both aligning to standards and establishing technical adequacy, it is not surprising that little information is available on the achievement growth of this population. This lack of information may also be a result of the limited time frame within which this testing option has been available (roughly 8 years since its inception). In this limited time frame, states have switched systems, partly in response to the initial testing results but also as a function of federal oversight. Nevertheless, a few publications have explicitly addressed growth in achievement for students who participate in large-scale alternate assessments, two focusing on critical issues and two presenting empirical results.
Ahearn (2009) summarized the NCLB growth model pilot program from 2005 to 2008 and reviewed three models: (a) growth-to-proficiency models, (b) value tables and transition models, and (c) projection models. Her focus was on inclusion of students with disabilities in growth models for the 15 participating states. She reported that because alternate assessments were designed for SWSCDs and scored on a different scale from the general assessment, 13 of the states that had growth models approved for use for adequate-yearly-progress purposes did not include students taking the AA-AAS in their accountability growth model. She noted several additional challenges, primarily psychometric issues and changes in the test used as alternate assessments since they were first required in the 1997 amendments to the IDEA.

Buzick and Laitusis (2010) provided an overview of not only the options but also the challenges to using growth models for students with disabilities. They reviewed three models (observed growth score, predicted growth score, and value added) and noted a number of serious issues, such as changes in the testing program, use of accommodations, variation in state policies, inconsistencies with individualized education program (IEP) teams, design of tests not explicitly oriented to document growth, and presence of multiple testing options (including modified achievement standards that had been available). All of these issues introduced difficulty in “tracking of students from year to year and comparability of test forms across the three assessments” (Buzick & Laitusis, 2010, p. 540). Finally, they noted that the number of students in the various disability categories and the changes in student classification across these categories (as well as in and out of special education) limited the use of growth models. Ultimately, Buzick and Laitusis recommended more research on the use of accommodations, the influence of test difficulty, and the characteristics of students with disabilities, all of which had to be addressed before the application of growth models in large-scale test accountability systems.

Using 4 years of mathematics and reading results from a state AA-AAS, Dunn, Rousoss, Lonczak, and Sukin (2012) evaluated the application and impact of a variety of growth models for characterizing students’ performances. Each growth model used 2 or more years of individual student data. A baseline model (Model 1) involved assigning a growth score of 1 to each student who moved up at least one proficiency level from one year to the next, a growth score of −1 assigned to students whose proficiency level dropped one or more levels, and a growth score of 0 assigned to students whose proficiency level was unchanged. For each additional model, Dunn et al. added complexity to account for different characteristics of AA-AAS scores. For example, Model 2 allowed students who had remained at the highest proficiency level from one year to the next to receive a growth score of 1.

Model 3 divided the below-proficient levels into two levels, allowing low-achieving students to demonstrate growth without achieving proficiency. Finally, Models 4 and 5 were based on a change statistic \( Z = \frac{\text{raw score Year 2} - \text{raw score Year 1}}{SD} \) for modeling growth referenced to change in scores instead of a change in proficiency level. Dunn et al. (2012) reported that more statistically sophisticated models (Models 4 and 5) improved on the outcomes of Model 1. A subsequent analysis used modified Model 4 (now referred to as Model 6) by assigning a growth score of 1 to any student who (a) progressed at least one proficiency level or (b) had a \( Z \) score that was greater than 1. Model 6 improved on Model 1 by identifying a small percentage of students (mostly below proficiency) who did not change proficiency levels but had \( Z \) scores greater than 1. Dunn and colleagues found potential added value of detecting growth for students who stay within the same proficiency level across years using the test score difference and comparing it to an appropriate standard error.

Using data from three states, Karvonen, Flowers, and Wakeman (2013) investigated three types of growth models: transition growth, growth to standards, and ordinary least squares (OLS). The transition model results showed a large percentage of students
not changing performance levels and no consistent pattern across the states or across math and reading in the changes that were positive (went up from one year to the next) or negative (went down from one year to the next). For these states, approximately 40%, 60%, and 35% of the students remained in the same achievement level for 3 consecutive years. When the changes across categories were collapsed to show movement from not proficient to proficient (to capture growth to standards), significant differences appeared across the states, with State A having nearly two thirds of the population remain not proficient across successive years and State B showing three fourths of the population moving from not proficient to proficient. In State C, this change occurred more often in reading (with half of the students moving to proficient) than math (with less than a third moving from not proficient to proficient). The final analysis using OLS showed much the same results as the growth to standards. In the end, the three methods for reflecting growth were not correlated.

So far, the analyses tell us that transition tables do not yield stable patterns of growth across consecutive grades. There are patterns of reversals in the transition models, and the regression model used to calculate residuals for the exploratory growth model only accounted for 48% of the variance in 2010 scores. Correlations among the results of the various approaches were inconsistent across subjects and years, even within states. (Karvonen et al., 2013, p. 17)

The Present Study on Growth in Alternate Assessments

Given this limited research base on the growth of students participating in alternate assessments, it is uncertain how much growth can (does) occur. In order to address this question, the population needs to be intact over a period of time. In addition, growth (or lack thereof) needs to be based on an appropriate test designed with reductions in breadth, depth, and complexity. Finally, estimates of growth in a standards-based testing environment depend on technically adequate measures in their linkage to standards as well as in the reliability and validity information supporting the test. The present study embodied these design attributes and addressed fundamental questions about characterizing the achievement growth on alternate assessments by comparing two viable growth models: the transition matrix model and a multilevel linear growth model.

Method

The transition matrix reflects changes in proficiency categories across successive years. In this state, four categories of proficiency are used: (a) does not yet meet, (b) nearly meets, (c) meets, and (d) exceeds. The cut scores for these categories were based on a standard setting session with teachers from throughout the state convening to use a bookmark method (Mitzel, Lewis, Patz, & Green, 2001). The multilevel model uses a Rasch conversion that extends the lower bounds of the general education test (reflecting Grade 3 achievement centered on 200). The analysis reflects growth over time in relation to various predictors (in this study, student characteristics and proficiency status). Using an extant state test that was accessible following approval from the institutional review board, we determined that the important issues are the sample of students, the measurement of proficiency, and the manner in which the data were analyzed.

Sample

Students in the study were deemed eligible for participating in the alternate assessment if their IEP teams determined that they (a) performed well below grade level, (b) had academic difficulties that were significant and generalized to all subject areas, (c) read significantly below enrolled grade level or did not read, (d) demonstrated performance that was significantly affected by the disability, and (e) had mobility as well as receptive and expressive language difficulties that were significant and generalized. In terms of instruction, the students (a) primarily received instruction
from specialized or functional materials that were significantly reduced in depth, breadth, and complexity and (b) relied on accommodations or modifications to access instructional information (Oregon Department of Education, 2014, p. 9). The state did not have a multiple-disabilities eligibility category; only the primary disability listed within a student’s IEP is reported.

The data were part of a state’s testing program over a 3-year period (2008–2009, 2009–2010, and 2010–2011) with students who were in Grade 3 in 2008–2009 or Grade 4 in 2009–2010 or Grade 5 in 2010–2011. All students eligible to take the alternate assessment were included in the sample. The original sample for this study included all Grade 3 students who took the 2008–2009 Oregon alternate reading assessment ($N = 1,129$). To prepare the sample data, we excluded students who demonstrated an off-track grade-level progression (e.g., retained a grade) and those who did not have complete data available in 2008–2009. The final sample included 1,061 students (approximately 94% of the original sample).

Demographic statistics of the sample at Grade 3 in 2008–2009 are displayed in Table 1. The final sample ($N = 1,061$) was approximately 35% female, 39% non-White, and 70% who qualified for free or reduced-price lunch. All sample students received special education services, with 37% receiving general education classroom instruction less than 40% of the time (the remaining received general education instruction more than 40% of the time). IEP teams classified all sample students displayed in Table 1 as having significant cognitive disabilities, though considerable variation existed in exceptionality categories. For the purpose of this study, we created a variable to represent students with an “intellectual disability” classification, who were those with a classification of intellectual disability or traumatic brain injury. Approximately 19% of the sample had an intellectual disability classification in Grade 3. Finally, approximately 63% of the sample scored above achievement standards on the state alternate reading assessment in Grade 3.

**Measure**

The test blueprint was based on vertical alignment of standards within elementary Grades 3 through 5 and middle school Grades 6 through 8. This alignment allowed item writers to develop test content using a single task across the three grade levels in which items were reduced for breadth, depth, and complexity when vertically aligned to standards, precluding the need for extended standards. The test was separated into a student materials booklet and a teacher administration and scoring protocol. The test development process is described by Tindal et al. (2003) and Carrizales and Tindal (2009). After development of items and creation of the materials and protocols, an alignment study was conducted to determine the relation between items and standards.

In addition to a test blueprint and specifications, a test administration and scoring manual was used to ensure a reliable administration of the tasks and items. All new assessors were trained in regional workshops or by (district) qualified trainers who monitored each district’s qualified assessors. Follow-up training was implemented using a web-based content management system that presented vignettes of administration and scoring using videotaped situations of comparable tasks as presented in the secure test. This training provided teachers extensive options for test administration and a number of different accommodations that could be used. All teachers had to become proficient (as qualified assessors or qualified trainers) on a test of administration and scoring protocols to be allowed access to the website, with the secure tests to be administered during the testing window (approximately end of February through end of April). The test was administered in a one-to-one situation and scores entered into a web-based system later during the test window.

In the test administration setting, the qualified assessors first administered a common set of 10 items to document students’ level of independence in taking tests (scaled from independent to full physical prompts needed); this outcome provided teachers information on how to best administer the test but was not used in calculating a final score. Qualified
assessors also used this information to choose between a standard administration and a scaffold administration with both types comparable on content but with more graphics provided in the scaffold version.

The reading test comprised 10 tasks, each with five associated items that addressed the following skills: decoding, fluency, word recognition and vocabulary, reading to perform a task, general understanding of literary texts and text content and structure, and interpretation. The following task labels were specifically referenced in the test: Task 2, Decoding and Word Recognition; Task 3, Reading Fluency; Task 4, Vocabulary: Root Words and Dictionary; Task 5, Vocabulary: Word Meaning and Context; Task 6, Informational Text: Directions and Procedures; Task 7, Informational Text: Structural Features; Task 8, Informational Text: Main Ideas and Details; Task 9, Informational Text: Cause, Effect, and Inference; Task 10, Literary Text: General Understanding; and Task 11, Literary Text: Develop an Interpretation. These tasks had been aligned with state standards but reduced in breadth, depth, and complexity.

Content items were scored using a partial-credit model with 0 = incorrect, 1 = partially correct, and 2 = exactly correct (Masters, 1982). Each year, field test items were embedded into the test and then scaled for use as an operational item the following year. The possible total score for each task in Tasks 2 through 11 of the test is 10 points per task for a total of 100 points for the assessment. The score was converted to a Rasch score so it could be placed on the same scale as the general education summative test (extending downward from about 165 points and lower). Students averaged between 6 and 7 points on each task, with a total of about 60 raw points overall (and a standard deviation of approximately 25 points). Once converted to the Rasch score, the average was about 100 points with a standard deviation of 5 points.

After the initial test development, standards were set to establish cut scores using the bookmark method (Mitzel et al., 2001). In this form of standard setting, items were arranged from low to high on the Rasch scale, and teams of teachers met in grade levels to identify the item that best represented the difference between

<table>
<thead>
<tr>
<th>Table 1. Sample Demographic Characteristics.</th>
</tr>
</thead>
<tbody>
<tr>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Non-White</td>
</tr>
<tr>
<td>Free/reduced lunch</td>
</tr>
<tr>
<td>General education &lt;40%</td>
</tr>
<tr>
<td>Primary exceptionality classification</td>
</tr>
<tr>
<td>No disability classification</td>
</tr>
<tr>
<td>Intellectual disability</td>
</tr>
<tr>
<td>Hearing impairment</td>
</tr>
<tr>
<td>Visual impairment</td>
</tr>
<tr>
<td>Communication disorder</td>
</tr>
<tr>
<td>Emotional disturbance</td>
</tr>
<tr>
<td>Orthopedic impairment</td>
</tr>
<tr>
<td>Traumatic brain injury</td>
</tr>
<tr>
<td>Other health impairments</td>
</tr>
<tr>
<td>Autism spectrum disorder</td>
</tr>
<tr>
<td>Specific learning disability</td>
</tr>
<tr>
<td>Met achievement proficiency</td>
</tr>
</tbody>
</table>

Content items were scored using a partial-credit model with 0 = incorrect, 1 = partially correct, and 2 = exactly correct (Masters, 1982). Each year, field test items were embedded into the test and then scaled for use as an operational item the following year. The possible total score for each task in Tasks 2 through 11 of the test is 10 points per task for a total of 100 points for the assessment. The score was converted to a Rasch score so it could be placed on the same scale as the general education summative test (extending downward from about 165 points and lower). Students averaged between 6 and 7 points on each task, with a total of about 60 raw points overall (and a standard deviation of approximately 25 points). Once converted to the Rasch score, the average was about 100 points with a standard deviation of 5 points.

After the initial test development, standards were set to establish cut scores using the bookmark method (Mitzel et al., 2001). In this form of standard setting, items were arranged from low to high on the Rasch scale, and teams of teachers met in grade levels to identify the item that best represented the difference between

Content items were scored using a partial-credit model with 0 = incorrect, 1 = partially correct, and 2 = exactly correct (Masters, 1982). Each year, field test items were embedded into the test and then scaled for use as an operational item the following year. The possible total score for each task in Tasks 2 through 11 of the test is 10 points per task for a total of 100 points for the assessment. The score was converted to a Rasch score so it could be placed on the same scale as the general education summative test (extending downward from about 165 points and lower). Students averaged between 6 and 7 points on each task, with a total of about 60 raw points overall (and a standard deviation of approximately 25 points). Once converted to the Rasch score, the average was about 100 points with a standard deviation of 5 points.

After the initial test development, standards were set to establish cut scores using the bookmark method (Mitzel et al., 2001). In this form of standard setting, items were arranged from low to high on the Rasch scale, and teams of teachers met in grade levels to identify the item that best represented the difference between

Content items were scored using a partial-credit model with 0 = incorrect, 1 = partially correct, and 2 = exactly correct (Masters, 1982). Each year, field test items were embedded into the test and then scaled for use as an operational item the following year. The possible total score for each task in Tasks 2 through 11 of the test is 10 points per task for a total of 100 points for the assessment. The score was converted to a Rasch score so it could be placed on the same scale as the general education summative test (extending downward from about 165 points and lower). Students averaged between 6 and 7 points on each task, with a total of about 60 raw points overall (and a standard deviation of approximately 25 points). Once converted to the Rasch score, the average was about 100 points with a standard deviation of 5 points.

After the initial test development, standards were set to establish cut scores using the bookmark method (Mitzel et al., 2001). In this form of standard setting, items were arranged from low to high on the Rasch scale, and teams of teachers met in grade levels to identify the item that best represented the difference between
successive proficiency levels in each grade. In 2007, a standard setting proposal was adopted by the state board of education in which these cut scores were delineated in Grades 3 through 8 and 11 using four categories (with accompanying achievement level descriptors): does not yet meet, nearly meets, meets, and exceeds. Generally, these values ranged from 95 and below for third graders who “did not meet” to 122 and above for fifth graders who “exceeded,” with a linear trend showing minimal overlap within a category across grades. In the elementary grade levels for this study, approximately one fifth to one quarter of the students either did not yet meet or nearly met proficiency, about one third met proficiency, and one fifth to one quarter exceeded proficiency.

Annually, a technical report has been published, summarizing reliability and validity information, including content-related evidence (alignment), test outcomes by grade level and grade band, response and scoring process evidence (focusing on teacher training and administration), score distributions, comparability of test administration (standard vs. scaffold), and internal (factor) structure of the test. In all grade levels, over 90% of the reading items were rated as strongly linked to the associated standards. Administration proficiency has averaged 91%, with the vast majority of teachers achieving this in the first attempt. Reliability has been reported for tasks and the entire test (with Cronbach’s alpha based on standardized items typically .95 or greater). Item statistics and intercorrelations have been calculated to summarize the range of performance and outcomes. These results have been stable each year.

Analysis

In analyzing two different strategies for summarizing growth, we highlight critical features of each analysis in the context of the nascent literature in the field. Importantly, these two analyses represent important choices for state education agency (SEA) personnel and policy makers. Without appropriate clarity of purpose and outcomes, it is not possible to make valid interpretations of an accountability system.

The transition matrix was based upon movement across performance levels used by the state for monitoring AYP. In the analysis, we simply summed the number of students at each performance level transition for each pair of years (Grade 3 to 4 and Grade 4 to 5). For example, students improved by moving from “did not meet” to any of the other three categories, from “nearly meets” to either of the other two categories, or from “meets” to “exceeds”; of course, the reverse trend also was possible and transitions were possible across more than a single performance level. No weights or other values were added to this transition as done by Dunn et al. (2012). In the table of results, we also reported the percentage at each of these transitions and summarized the number (and percentage) of performance-level changes. This transition matrix has been used to document the effects from implementing an accountability system that (a) addressed the critical issue of all students reaching proficiency as targeted in NCLB and (b) improved the status model in which each year’s outcomes is based on a unique cohort. This analysis required complete data from students for both years of transition (from Grade 3 to 4 and from Grade 4 to 5).

For the multilevel model, we used HLM Version 7.0 with full maximum likelihood to estimate all models. We first examined unconditional growth models (without predictors but with a time variable) to examine mean and variance of between-subject reading proficiency and to provide baseline statistics for evaluating subsequent conditional models (Raudenbush & Bryk, 2002). The two parameters of interest in Level 1 represented intercept and slope, which was centered on the Grade 3 performance and expressed as the linear increase from one year to the next.

We ran conditional models with predictors at Level 2 to determine the influence of student characteristics and measurement conditions on both the intercept and slope. The four student characteristics were as follows: (a) sex (male = 0, female = 1), (b) ethnicity-race (ethnicity; White = 0 and non-White = 1), (c) economic disadvantage (EconDis; 0 = no free or reduced-price lunch and 1 = free or reduced-price lunch),
and (d) intellectual disability (non–intellectual disability and non–traumatic brain injury = 0 and intellectual disability and traumatic brain injury = 1). The two measurement conditions were (a) program placement (GenEd40; general education instruction >40% of time = 0 and general education instruction <40% of time = 1) and (b) performance level (PerfLevel; below proficiency standards in Grade 3 = 0 and above proficiency standards in Grade 3 = 1). After documenting descriptive statistics for the student characteristics and measurement conditions, we ran a two-level hierarchical linear growth model. This model took advantage of using a vertical scale that is potentially more sensitive than a four-category proficiency scale. Using the transformed Rasch score, we were able to determine the estimated amount of change from year to year, even with missing data. Further, the analysis has been run to determine the effects of various student characteristics.

**Results**

Transition matrices are presented in Tables 2 and 4 with level change in Tables 3 and 5 indicating the number of students who changed performance levels from one year to the next. The majority of students remained in the same performance level in each of the two transitions (i.e., approximately 60% at Grade 4 and 64% at Grade 5). In addition, approximately 28% and 15% of the sample increased performance levels, and 13% and 21% decreased levels at Grades 4 and 5, respectively.

The observed means of state alternate assessment scores for each year were as follows: Grade 3, \( M = 101.84 \) (\( SD = 19.70; n = 1061 \)); Grade 4, \( M = 104.68 \) (\( SD = 21.07; n = 771 \)); and Grade 5, \( M = 106.14 \) (\( SD = 20.68; n = 653 \)). The results of the unconditional growth model are presented in Tables 6 and 7. Across all students, the estimated mean alternate assessment reading scale score in Grade 3 was 102.15, and the average linear growth was 4.64; both were statistically significantly different from zero. The model also indicated that there was statistically significant variance between students for both the intercept and slope parameters.

The results of the conditional growth model are presented in Tables 6 and 7. For the reference group (i.e., males, White, non–economically disadvantaged, non–intellectual disability, received general education instruction more than 40% of time, and scored below proficiency at Grade 3), the average Grade 3 alternate assessment score was 86.30, and the average linear growth was 6.30. Students’ sex, ethnicity, and disability were not significant

### Table 2. Transition Matrix From Grade 3 to Grade 4.

<table>
<thead>
<tr>
<th>Grade 3 (2008–2009)</th>
<th>Low</th>
<th>Nearly meets</th>
<th>Meets</th>
<th>Exceeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>161 (20.3)</td>
<td>37 (4.7)</td>
<td>9 (1.1)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Nearly meets</td>
<td>34 (4.3)</td>
<td>54 (6.8)</td>
<td>49 (6.2)</td>
<td>8 (1.0)</td>
</tr>
<tr>
<td>Meets</td>
<td>5 (0.6)</td>
<td>40 (5.0)</td>
<td>145 (18.2)</td>
<td>116 (14.6)</td>
</tr>
<tr>
<td>Exceeds</td>
<td>0 (0)</td>
<td>1 (0.1)</td>
<td>21 (2.6)</td>
<td>115 (14.5)</td>
</tr>
</tbody>
</table>

*Note. \( N = 795 \). Frequencies shown with percentages in parentheses. There were no students in the lowest level (very low).*

### Table 3. Level Change From Grade 3 to Grade 4.

<table>
<thead>
<tr>
<th>–3</th>
<th>–2</th>
<th>–1</th>
<th>0</th>
<th>+1</th>
<th>+2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (0)</td>
<td>6 (0.8)</td>
<td>95 (11.9)</td>
<td>475 (59.7)</td>
<td>202 (25.4)</td>
<td>17 (2.1)</td>
</tr>
</tbody>
</table>

*Note. \( N = 795 \). Frequencies shown with percentages in parentheses.*
predictors of either intercept or slope. EconDis was significant for intercept but not slope; all else constant, the average Grade 3 alternate assessment reading score for students who received free/reduced lunch was approximately 3 scaled points greater than the reference group. GenEd40 and PerfLevel were significant for both intercept and slope. All else constant, the average Grade 3 alternate assessment reading score for students who received general education instruction less than 40% of time was approximately 5 scaled points less than the reference group, and their growth rate was about 2 scaled points less than the reference group. All else constant, the average Grade 3 alternate assessment reading score for students who were at or above Grade 3 proficiency was approximately 26 scaled points greater than the reference group, but their slope was approximately 2 scaled points less than the reference group.

This set of predictors explained approximately 58% of the variance between students; specifically, the conditional model explained 60% of the variance in the intercept and 10% of variance in the slope. The final model results, however, indicated that even with the inclusion of these predictors, there remained statistically significant variance across students for both the intercept and slope parameters.

Discussion

Documenting the achievement growth of all students is central to conceptualization of learning and development. Yet for students participating in alternate assessments, who by definition experience serious difficulties learning, it is particularly challenging to meaningfully document their achievement growth against states’ content standards. Two growth models in particular, however, hold promise for characterizing the achievement growth on alternate assessments: transition matrix models and multilevel linear growth models.

For students participating in alternate assessments, who by definition experience serious difficulties learning, it is particularly challenging to meaningfully document their achievement growth against states’ content standards.

Comparison of Models for Expressing Growth

Both strategies for expressing growth, of course, require the usual technical adequacy of the test (reliability and evidence of
construct validity with appropriate test design, alignment to standards, and appropriate criterion-related evidence). When used longitudinally, they both have the advantage of documenting growth with the same cohort, in contrast to the status model of NCLB. However, they also offer unique advantages and disadvantages that need to be understood by SEA personnel and policy makers.

**Documenting growth using a transition matrix.** A transition matrix reveals whether students are making adequate progress across performance levels. The primary advantage of a transition matrix to connote growth is that categories of proficiency represent the “coin of the realm” for accountability. That is, growth is expressed as an increase in proficiency (from “does not yet meet” to “meets” or from various other categories). This outcome is ultimately what is important and most interpretable for administrators, teachers, parents, and the general public. The transition matrix also presents observed growth information in a format that is easily accessed by the user with visual inspection.

Of course, this metric is at the end of a validation logic chain. Such movement across proficiency categories also depends on the standard-setting process (i.e., assignment of cut scores to proficiency categories). The transition matrix relies on the assumption that the performance-level categories are well defined within and across grades, based on well-articulated standards, such that categories have consistent meaning across grades (Castellano & Ho, 2013). However, the scaling within and across categories as well as the cut score assignments may not produce equal-interval scales, making it difficult to traverse the categories consistently. Further, a transition

---

**Table 6.** Unconditional and Final Conditional Model Parameters With Robust Standard Errors.

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>Unconditional model</th>
<th>Final model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>SE</td>
</tr>
<tr>
<td>Intercept, $\beta_{00}$</td>
<td>102.15</td>
<td>0.61</td>
</tr>
<tr>
<td>Sex, $\beta_{01}$</td>
<td>-1.72</td>
<td>0.90</td>
</tr>
<tr>
<td>Ethnicity, $\beta_{02}$</td>
<td>-1.08</td>
<td>0.86</td>
</tr>
<tr>
<td>EconDis, $\beta_{03}$</td>
<td>3.04</td>
<td>1.04</td>
</tr>
<tr>
<td>Intellectual disability, $\beta_{04}$</td>
<td>0.13</td>
<td>1.30</td>
</tr>
<tr>
<td>GenEd40, $\beta_{05}$</td>
<td>-5.34</td>
<td>1.03</td>
</tr>
<tr>
<td>PerfLevel, $\beta_{06}$</td>
<td>26.35</td>
<td>1.12</td>
</tr>
<tr>
<td>Slope, $\beta_{10}$</td>
<td>4.64</td>
<td>0.24</td>
</tr>
<tr>
<td>Sex, $\beta_{11}$</td>
<td>-0.50</td>
<td>0.48</td>
</tr>
<tr>
<td>Ethnicity, $\beta_{12}$</td>
<td>-0.09</td>
<td>0.50</td>
</tr>
<tr>
<td>EconDis, $\beta_{13}$</td>
<td>0.93</td>
<td>0.53</td>
</tr>
<tr>
<td>Intellectual disability, $\beta_{14}$</td>
<td>-0.49</td>
<td>0.58</td>
</tr>
<tr>
<td>GenEd40, $\beta_{15}$</td>
<td>-1.86</td>
<td>0.52</td>
</tr>
<tr>
<td>PerfLevel, $\beta_{16}$</td>
<td>-1.89</td>
<td>0.52</td>
</tr>
</tbody>
</table>

*Note. EconDis = economic disadvantage; GenEd40 = program placement; PerfLevel = performance level.*

**Table 7.** Variances From Final Unconditional and Conditional Models.

<table>
<thead>
<tr>
<th>Random effects</th>
<th>Unconditional model</th>
<th>Final model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Variance</td>
<td>p value</td>
</tr>
<tr>
<td>Intercept, $r_{0i}$</td>
<td>342.18</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Slope, $r_{1i}$</td>
<td>12.54</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Level 1 residual, $e_8$</td>
<td>54.22</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Documenting growth using multilevel modeling. Any summary of outcomes using a four-category scale (reflecting various proficiencies) is considerably less sensitive than the raw or converted scale using many points in a distribution. In using multilevel models to document growth, the primary advantage is the use of the vertical scale with scores in a distribution. Multilevel growth models also can accommodate parsing variance at the student, classroom, or school level and with relevant characteristics to reflect important dimensions of influence. Finally, as transition models express change across 2 years, multilevel growth models can express change across 3 or more years and offer flexibility in handling missing data and accommodating different functional forms of growth.

Perhaps the most significant disadvantage to a multilevel model is the complexity of design and analysis used for public reporting. Further, growth estimates, though based on a more sensitive metric, may be difficult to interpret (e.g., whether such growth is adequate or not). For example, although the results may reflect statistically significant growth, the actual amount may not translate into improvement across proficiency categories.

Findings and Interpretations in Comparing the Two Strategies

With the scores from Oregon’s AA-AAS, we found that the transition matrix revealed that the majority of students in this sample remained at the same performance level from one year to the next. Of all possible performance-level combinations in transition matrices for Grades 3 to 4 and Grades 4 to 5, the largest proportion of students in both matrices were those students who remained in the does not yet meet level in consecutive years. Fewer than 9% of students in both matrices went from below proficiency to above. These findings from our transition matrix approach complement the findings from the two previous empirical studies. Similar to Buzick and Laitusis (2013), most of the students were stable in their levels of proficiency, with few students moving up or down more than one performance level. From Grade 3 to Grade 4, the proportion of students moving up was greater than that of students moving down, but the majority of the former was movement from meets to exceeds. The sample for the transition matrix analyses decreased 19% from Grade 4 to 5. This was likely due to the requirement that students needed valid scores from 2 consecutive years of tests to be included in the transition matrix analysis.

Growth models, using methods such as multilevel or structural equation modeling and the maximum likelihood estimator, however, do not require outcome data to be listwise complete. Thus, alternate assessment scores from any available year can be included even when data from other years are missing (Singer & Willet, 2003). The results of our growth modeling of state alternate assessment reading scores indicated that on average, the mean scores are quite close to the proficiency cut scores. Figure 1 shows that the unconditional grand-mean growth trajectory generally follows the proficiency cut score at each grade, so that the mean is just below the cut score at Grade 3, right at the cut score at Grade 4, and just above the cut score at Grade 5. This would indicate small but meaningful average gains across years.

Figure 1 also shows the estimated trajectories for several groups of interest. The results of the conditional growth model displays the estimated trajectories for several groups and suggests that on average, controlling for other predictors, students above proficiency at Grade 3 (113) score significantly higher than those below proficiency (86), reflecting a difference of approximately 1.42 standard deviations. Students who score below proficiency at Grade 3, however, have significantly greater average growth than their counterparts.
(controlling for other predictors), a difference of approximately 0.53 growth standard deviations. Another interesting finding is that on average and controlling for all other covariates, students receiving free or reduced-price lunch had small, but statistically significant, higher scores (3 points) at Grade 3 than did students not receiving free or reduced-price lunch, but there is no difference in their average estimated growth rates. Finally, a noteworthy finding is that the effect of intellectual disability on both the intercept and slope was not statistically significantly different from the reference group, suggesting that although a variation of disabilities may be represented in the SWSCD sample, the average achievement trajectory for students with intellectual disability is commensurate with the average trajectory for students with any other disability classification.

Perhaps the most critical finding is that on average and controlling for all else, students taking alternate assessments who received more than 40% general education instruction scored on average significantly higher at Grade 3 (about 5 points, or 0.29 standard deviations) and grew significantly greater (1.86, or 0.61 standard deviations) than those receiving less than 40% general education instruction. This may occur because students with disabilities spending more time in general education instruction are better able to access that instruction than their more disabled peers. More likely, this result is a function of

---

Figure 1. The estimated growth trajectories of the unconditional growth model (i.e., grand mean), three student subgroups of interest based on results, and the proficiency cut scores at each of Grades 3 through 5. The Grade 3 Below Proficient group represents the reference group (scored below proficiency at Grade 3, White, males, non–economically disadvantaged, nondisability, received general education instruction more than 40% of time). The Grade 3 Proficient group represents students that scored above proficiency at Grade 3 (all other predictors equal to the reference). The Grade 3 GenEd < 40% group represents students that received general education instruction less than 40% of the time (all other predictors equal to the reference).
students with less interfering behavior or severity of disability being placed in the general education classroom more of the time (however, it is worthwhile to note that we did control for intellectual disability).

These results are complementary to those from Dunn and colleagues (2012), who used an intriguing initial application of a transition matrix to document growth on alternate assessments. In their work, however, nine performance levels were used (which is unusual as most states have three or four), and the instrument had the same range of total score points, raw scores at the proficiency cuts, and interpretation of proficiency across grade levels (also an unusual feature of most statewide alternate assessments). This is a relatively atypical scoring format for an AA-AAS and limits the generalizability of the study’s results to other states’ alternate assessments. Nevertheless, such a fine-grained scoring system may well enhance understanding of movement in a way that is comparable to the use of score changes in a multilevel model.

Probably the most significant contribution of the present study is the interpretation of growth, an issue that has not previously been documented with empirical results along with information on test development, administration, and technical adequacy, extending the findings of previous research on this population, particularly its definition (Kearns, Towles-Reeves, Kleinert, Kleinert, & Thomas, 2009; Quenemoen, 2009; Towles-Reeves, Kearns, Kleinert, & Kleinert, 2008). The real question about growth is consideration of the context of that growth. Without knowing the test content (and alignment to standards) and its technical adequacy (reliability and validity of score meaning), growth cannot be accurately defined.

What is not known, however, and what remains for future researchers to examine, is the consequential validity or the effect of the growth on teachers or systems. To a limited extent, we know about impact on training (given our extensive training using a content management system with proficiency). However, we do not know the impact on perceptions of administrators (Towles-Reeves, Kleinert, & Anderman, 2008) and teachers (Kleinert, Kennedy, & Farmer Kearns, 1999), impact in classrooms (Restorff & Abery, 2013) or consideration of curriculum (Browder, Flowers, et al., 2002; Lowrey, Drasgow, Renzagila, & Chezan, 2007). All of these considerations are likely paramount for SEA personnel and policy makers.

**Limitations**

Although this research clearly and meaningfully applied both a transition matrix model and a multilevel linear growth model to AA-AAS test scores, the generalizations that can be drawn are limited by several factors. First, the number and percentage of students with intellectual disability was quite low, and therefore the results may not generalize to other states where the alternate assessment is used more conservatively. Further, the study focused on elementary students from Oregon and their reading test scores only. Finally, although a scaffold administration was available for teachers to assist the student in accessing the text, it is important to acknowledge the difficulty of showing growth for this population of students.

*Although we found more growth for students spending more time in the general education classroom, this variable is likely confounded with severity of disability.*

In this study, we described an alternate assessment, its development, and technical adequacy and then analyzed growth in both changes of proficiency and the test score. Both growth models, also, are entirely contingent on the quality of the tests scores. The Oregon AA-AAS is a good example, however, of a test that allows for the use of more sensitive growth models and thus gives educators a more accurate picture of student improvement.

Although we found more growth for students spending more time in the general education classroom, this variable is likely confounded with severity of disability (e.g.,
more seriously involved students spend less
time in the general education environment). More difficult to explain was that students from economically disadvantaged back-
grounds performed higher than those from less disadvantaged backgrounds; it may be that this predictor also is confounded with another student characteristic.

**Conclusion**

Achievement growth can meaningfully be documented on alternate assessments, and multilevel growth models are likely more sen-
sitive than transition matrices. Consistent with previous research (Buzick & Laitusis, 2010), we found little growth in proficiency levels, which might be expected given that the resolution of the basic scale for documenting growth is far more coarse than the use of a test score. Further, we documented growth in relation to specific predictors, an outcome not reported in prior research.

Future investigations are needed to both replicate and extend the research related to the achievement growth on AA-AAS. Growth models in this context are often applied for the purpose of making inferences about account-
ability, and although that was beyond the scope of this article, the two growth models applied in this study are a small sample of the possibilities that can be explored and perhaps used by SEA personnel to either develop policy or document outcomes as part of an accountable

**References**


Authors’ Note

This research was funded in part by a cooperative service agreement from the Institute of Education Sciences establishing the National Center on
Assessment and Accountability for Special Education (PR/Award No. R324C110004); the findings and conclusions expressed do not necessarily represent the views or opinions of the U.S. Department of Education.

The Oregon Department of Education funded the development and implementation of the alternate assessments judged against alternate assessment standards.

Manuscript received September 7, 2014; accepted April 12, 2015.