Article

HAMMILL INSTITUTE ON DISABILITIES

The Journal of Special Education
2016, Vol. 49(4) 209–220

DOI: 10.1177/0022466915582213
journalofspecialeducation.sagepub.com
SAGE

Patterns of Statewide Test Participation for Students With Significant Cognitive Disabilities

Jessica L. Saven, MAT[1], Daniel Anderson, PhD[1], Joseph F. T. Nese, PhD[1], Dan Farley, MA[1], and Gerald Tindal, PhD[1]

Abstract

Students with significant cognitive disabilities are eligible to participate in two statewide testing options for accountability: alternate assessments or general assessments with appropriate accommodations. Participation guidelines are generally quite vague, leading to students "switching" test participation between years. In this study, we tracked test participation for two cohorts of students with a documented disability over 3 years. Results suggested approximately 25% of students who initially took the alternate assessment switched test type at least once, although patterns of switching were not consistent across disabilities. Students on the performance "bubble" were more likely to switch test participation. Test switching poses challenges for monitoring students' academic growth within accountability frameworks.

Keywords

alternate assessment, accountability systems, longitudinal research methods, quantitative research methods

The 1997 reauthorization of the Individuals With Disabilities Education Act (IDEA) included amendments that required the inclusion of all students with disabilities in state- or districtwide testing systems, as well as public reporting of results from such assessments. This legislation required the use of standards-based general assessments (GA) with necessary accommodations for students with disabilities. As students with disabilities participated in systems-level testing, three options for their participation emerged: (a) GA, (b) GA with accommodations, or (c) alternate assessments based on grade-level, modified, or alternate achievement standards (Thurlow, Lazarus, Thompson, & Morse, 2005).

Under this last option in the No Child Left Behind Act (NCLB; 2002), states developed alternate assessments based on alternate achievement standards (AA), as one possible means to assess students with the *most significant* cognitive disabilities ([SWSCDs], U.S. Department of Education [USED], 2005). The USED (2005) estimated that approximately 9% of students with disabilities, or 1% of all students, had a disability significant enough to preclude meaningful participation in the GA. This number became the cap on the number of AA scores that could be counted toward adequate yearly progress (AYP) although specific criteria for participation vary by state (Kleinert, Browder, & Towles-Reeves, 2009). Similar to the requirement for GA, NCLB also required that AA meet several technical adequacy requirements (i.e., reliability, validity) and link to state academic content standards used to determine AYP (Elliott, Compton, & Roach, 2007).

Although the reporting guidelines for AA are clear, participation guidelines vary from state to state (Musson, Thomas, Towles-Reeves, & Kearns, 2010; Roach, 2005; Thurlow, 2004). No federal definition for the term "significant cognitive disability" exists; however, this category was intended to include students within a federal IDEA disability category "whose cognitive impairments may prevent them from attaining grade level achievement standards, even with the very best instruction" (USED, 2005, p. 23). This terminology is subjective, and may contribute to students "switching" test types between years. This switching makes tracking students' academic growth difficult, as the AA and GA test scales are generally not comparable.

[1]Behavioral Research and Teaching, University of Oregon, Eugene, USA

Corresponding Author:
Jessica L. Saven, Behavioral Research and Teaching, 275 Education, 5262 University of Oregon, Eugene, OR 97405-5262, USA.
E-mail: jsaven@uoregon.edu

Alternate Assessment Participation

Three criteria for AA participation are used by most states: (a) the presence of a disability that significantly affects intellectual functioning, (b) an explicit decision made by the students' Individualized Education Program (IEP) team that the student will participate in the AA, and (c) the necessity of substantial adjustments to the general education curriculum to ensure access (Albus & Thurlow, 2012; Musson et al., 2010). Practices used by educators to identify students eligible for the AA vary widely and include checklists, observation protocols, and tests (Musson et al., 2010; Roach, 2005; Thurlow, 2004). The majority of students participating in AA are identified as students with an intellectual disability (ID), autism (AUT), or multiple disabilities (Kearns, Towles-Reeves, Kleinert, Kleinert, & Thomas, 2011).

In interviews with six special education teachers from the same Midwestern state, Cho and Kingston (2012) found teachers based AA participation decisions on students' need for instructional modifications and adaptations, differences in information processing, extremely low academic performance, and cognitive challenges. Some of these traits have also been found in students with mild disabilities, demonstrating the subjectivity of guidelines used to determine test participation. The expansion of the Cho and Kingston (2013) study to 317 teachers across three states indicated that when teachers were presented with three descriptions of students with mild disabilities, they sometimes assigned these students to AA. In particular, students with other health impairments, learning disabilities (LD), and emotional/behavioral disorders were identified due to their need for modifications, intensive instruction, poor academic performance, and other considerations. That is, students were frequently classified as SWSCDs by educators even when they only exhibited mild disabilities (Cho & Kingston, 2013). These trends suggest a prevalent uncertainty about which students qualify for the AA that, in turn, may lead to students being misassigned to either the AA or the GA.

In a later study with eight teachers from a single state, Cho and Kingston (2014) documented test assignment decisions based on subjective and noninstructional factors. Although performance scores on the previous year's state assessment were not specifically included in state eligibility guidelines, some teachers considered them when determining future test participation. Given the federal accountability requirements, including public reporting, IEP teams may feel pressure to switch test assignment to maximize school AYP—particularly students who are either performing highly on the AA or poorly on the GA (i.e., students on the performance "bubble"); there may be additional pressures for IEP teams to switch student test assignment to maximize school AYP. Although in some circumstances a school may benefit from continuing to assign high performing students to AA (e.g., Lemke, Hoerandner, & McMahon, 2006), the 1% cap on AA scores prohibits this group of students from getting too large for AYP reporting purposes. Examining test switching for students on the test bubble is necessary to determine if the proximity of their scores to state cut scores makes them more likely to switch than other students.

Implications of Switching Test Type

One of the most substantial problems with students switching test participation is that it reduces opportunities to characterize students' growth on a common scale. In practice, GA and AA scores are typically reported on different scales, making it difficult, if not impossible, to include scores from each in a longitudinal growth model (Buzick & Laitusis, 2010; Gong & Marion, 2006; Ho, 2009). Test construction for AA typically relies on either creating the measure using extended content standards with reduced breadth, depth, and complexity, or applying such reductions directly to the items themselves (USED, 2005). As a result, scores from the AA and the GA are not comparable because the targeted skills of the underlying constructs differ. Analyses on students' transitioning between state test performance classifications suffer the same inadequacy as longitudinal growth models because they are based on changes in groups of students in various proficiency categories, which are not equated. That is, the cut scores delineating performance categories generally vary across grades and by assessment type. Although assessment systems for both GA and AA are currently changing, test switching is still a relevant concern as modeling students' growth was a primary goal for national consortia in the development of comprehensive assessment systems (Center for K-12 Assessment & Performance Management, 2012a, 2012b; Sireci, 2012).

Student Growth

States developed AA during a time when the NCLB status model was the metric used to hold states accountable for achievement proficiency, yet in recent years student growth has been explored as a meaningful outcome within accountability frameworks. Educators wanted to recognize students who made substantial annual gains irrespective of their final proficiency levels (Furgol & Helms, 2012).

Growth models in statewide accountability systems typically condition students' current state test performance on a measure of prior achievement (generally the previous year's state test). Teacher- or school-level inferences are then based on the difference between students' expected and observed achievement, with covariates occasionally included in the model to control for teacher or school intake (e.g., Betebenner, 2009; McCaffrey, Lockwood, Koretz, Louis, & Hamiltion, 2004; Sanders & Horn, 1994). Use of growth models with repeated measures of student achievement (across years in statewide testing programs) for the

same cohort of students can help address the lack of comparability between groups present in a status model (e.g., percent proficient) as achievement is referenced to the same students over time. As of 2010, 12 states had adopted growth models to generate estimates of whether or not student achievement would meet state proficiency targets for AYP within 3 years, and 13 states used growth models for state-level accountability systems (Blank, 2010). Although a growth model can account for variance explained by prior student achievement, student mobility, and constant effects of family background (e.g., Blank, 2010), if students switch from one test to a noncomparable test type, it may not be possible to estimate student growth.

Missing Data

When students switch test types, the longitudinal data from successive test administrations become systematically missing by a known mechanism (i.e., missing not at random [MNAR]; see Little & Rubin, 2002). The extent to which test switching biases estimates for the GA population is likely insubstantial, as the proportion of students with missing data is very small relative to the entire GA population. However, missing data within cohorts of students taking the AA may be quite substantial (Saven, Farley, & Tindal, 2013) and such cohort instability may dramatically influence growth models used for accountability purposes.

A number of factors can influence cohort instability and lead to increased bias in estimates of student growth. First, attrition may occur, in which students begin in the first year of the time interval but drop out in later years. Cohort instability also can occur if new groups of students who were previously not present participate (i.e., students move into a state *after* the first data collection). In longitudinal studies, specific mechanisms lead to nonresponse at certain time points in which data are likely MNAR and constitute *nonignorable nonresponses* and in which a dependence exists between the outcome variable and data missingness (Schafer & Graham, 2002). Because the missing data mechanism is known (test switching), it could potentially be modeled and the resulting missingness may then become missing at random conditional on the observed variables. Regardless, test switching introduces substantial challenges to modeling data longitudinally for SWSCDs.

Test Participation and Disability Classification

Exploring test switching patterns for students with low-incidence disabilities is difficult, as the sample sizes are often small (Buzick & Laitusis, 2010). Aggregation across disability types is generally not appropriate (Abedi, Leon, & Kao, 2008) and precludes understanding the unique abilities and individualized needs of students taking the AA (Gong & Marion, 2006). In addition, the phenomenon of students switching between GA and AA from year to year may be related to students' disability type or variability in disability classification (Buzick & Laitusis, 2010; Saven et al., 2013). Concerns regarding variation in participation and disability classification are difficult to address, however, when little evidence exists regarding how consistently state test participation guidelines are applied (Cho & Kingston, 2012, 2013).

The majority of states do not allow disability type to be used as a determination of AA participation (Albus & Thurlow, 2012). That is, students from all IDEA categories are typically eligible to participate in AA (Kleinert et al., 2009). However, the majority of students participating in AA tend to be those with ID, multiple disabilities, and AUT (Kearns et al., 2011; Thurlow, 2004; Towles-Reeves et al., 2012). Further investigation to determine if disability type and previous test performance are predictors of test type switching is necessary, particularly as validity issues arise when the target populations are misrepresented (i.e., students who should take the GA, instead of participating in the AA; Musson et al., 2010; Quenemoen, 2008).

Problem Summary and Research Questions

Most SWSCDs participate in statewide testing programs by taking either the GA with accommodations or one of various forms of an alternate assessment; however, switching test types between years makes the estimation of growth difficult. In this study, we document the degree to which test switching occurred within one state. We then apply multinomial logistic regression to examine the extent to which specific disability types and previous test performance predict whether students will switch test types between years. Outcomes from such analyses can then be used for documenting the potential applicability of growth models to subgroups of students with disabilities, particularly students participating in AA. Such research may also help identify predictors of test switching and increase opportunities to develop techniques to measure the growth of students with disabilities, helping ensure their inclusion in state accountability efforts to improve educational outcomes (e.g., Buzick & Laitusis, 2010; Sireci, 2012).

We investigated the following three research questions:

Research Question 1: What is the likelihood of test switching on the reading portion of the AA and the GA across consecutive years, over a 3-year span for students with ID, AUT, or LD?

Research Question 2: Do students performing highly on the AA or poorly on the GA (i.e., students on the "bubble") have an increased likelihood of switching test type as compared with other students with the same disability?

Research Question 3: Is the observed pattern the same across cohorts of students in middle school as compared with elementary school?

Method

Measures

We analyzed a single state's reading GA and AA data in this study for school years 2009–2012. The GA was aligned to the state's general reading content standards and designed to provide a performance match between the test items and the reading content standards. The state used a computer-adaptive test administered at Grades 3 to 8 and 11. All students, irrespective of disability or language status, were permitted to participate in the GA with or without accommodations, based on individual need. The test was not timed and offered accommodations including typical options for presentation (e.g., adjustable font size, sign language interpretation) and response (e.g., response aids, manipulatives, sign language), in addition to Braille. Test results for the GA were reported at five different performance levels, with the top two levels (*meets* and *exceeds*) counting as proficient for AYP.

The reading AA, designed for SWSCDs, was linked to grade-level content standards that had been reduced in depth and breadth, pursuant to the nonregulatory guidance published by the USED (2005). The AA was administered in grade bands (i.e., Grades 3–5, 6–8, and 11) using a paper–pencil format. Universal design concepts were built into the test design to ensure access for students with a wide range of sensory and cognitive limitations (Thompson, Johnston, & Thurlow, 2002). Students' levels of independence were tested prior to the administration of content tasks to provide appropriate supports during administration, but performance was based only on content task items. Furthermore, the AA was designed for flexible, yet standardized administration, to meet the unique needs of all students while producing scores directly comparable across students (Anderson, Farley, & Tindal, 2013). All accommodations afforded to students taking the GA were likewise made available for students taking the AA. Although the same labels were used, the performance levels of various proficiency categories were not comparable between the GA and AA. Test results for the AA were reported at four levels (*does not yet meet, nearly meets, meets, exceeds*). For both types of tests, two categories in the state data set counted for proficiency toward AYP (*meets* and *exceeds*).

Participants

We used reading test data from students with normal grade progression in Grade 3 and Grade 6 at the beginning of the study (2009–2010), excluding students who switched disability classification during the 3 years to examine whether or not disability type influenced test switching. Analyses were conducted with two cohorts to examine if the effects observed with an elementary cohort replicated with a middle school cohort. We included all students with at least two test scores whose primary disability was coded as ID, LD, or AUT, as the sample size for these groups was sufficient for adequate statistical power (i.e., ≥.80 with an effect size of 0.40 and $\alpha = .05$).

Generally, students with LD are not intended to be included in AA participation. They are also not to be categorically excluded given that the AA participation option is available to all students with IEPs (Kleinert et al., 2009). This state had a large percentage of students with LD taking the AA, compared with nationally representative survey results (Kearns et al., 2011). This can be attributed in part to the fact that multiple disabilities is not an eligibility category in this state; students with complex disability profiles have primary, secondary, tertiary, and so forth, disability categories. The students labeled as LD in this state who participated in AA are likely more complex than the mere label conveys. However, we hypothesized that students with LD would be more likely to have been misclassified into the AA due to their achievement levels generally being higher than ID or AUT students. We therefore hypothesized that the probability of switching from AA to GA would be higher for LD students than ID or AUT students.

Within each disability category, we identified students on the performance "bubble" who scored either in the highest performance category (*exceeds*) on the AA or the lowest performance category (*very low*) on the GA in the previous year. Students on the bubble were hypothesized to have a higher likelihood of switching test participation than other students.

The final sample included 3,048 students in the Grade 3 Cohort, and 3,911 students in the Grade 6 Cohort. For the Grade 3 Cohort, approximately 65% of students were male, 64% were White, and 22% were of Hispanic/Latino descent. Of the 217 students (7%) with ID, approximately 13% were identified on the bubble after Year 1 and 12% after Year 2. Of the 561 (18%) students with AUT, 9% were on the bubble after each year. Finally, of 2,270 students (75%) with LD, 18% were on the bubble after Year 1, and 15% after Year 2.

The Grade 6 Cohort consisted of approximately 3,911 students with approximately 63% male, 64% White, and 24% of Hispanic/Latino descent. Of the 273 students (7%) with ID, 21% were identified on the bubble after Year 1 and 19% after Year 2. The group of 587 (15%) students with AUT included 8% on the bubble after both Years 1 and 2. Of the 3,051 (78%) with LD, 11% were on the bubble after Year 1, and 7% after Year 2.

Analyses

Descriptive analyses of the test type switching patterns were conducted to characterize the patterns of test switching for the two cohorts over the 3-year time span. Eight possible participation patterns for the 3-year period were investigated, as displayed in Table 1.

Table 1. Grade 3 (*n* = 3,048) and Grade 6 (*n* = 3,911) Cohort Test Patterns 2009–2010 Through 2011–2012.

Testing pattern	Grade 3 cohort (%)	Grade 6 cohort (%)
AA to AA to AA	446 (14.63)	341 (8.69)
AA to AA to GA	72 (2.36)	39 (1.00)
AA to GA to AA	7 (0.23)	6 (0.15)
AA to GA to GA	0	0
GA to GA to GA	2,226 (73.03)	3,221 (82.36)
GA to GA to AA	52 (1.71)	20 (0.51)
GA to AA to GA	36 (1.18)	19 (0.49)
GA to AA to AA	92 (3.02)	37 (0.95)
Missing a time point	187 (6.14)	282 (7.21)

In addition to exploring test switching descriptively, we used multinomial logistic regression to explore how the likelihood of students represented in each testing pattern changed by the predictor variables of interest. All analyses were conducted with the Mplus software, Version 7.11 (Muthén & Muthén, 1998–2007); plots were produced with the ggplot2 package (Wickham, 2009) within the R statistical software (R Core Team, 2013). Each cohort and each potential switch (Years 1–2 and Years 2–3) were analyzed separately (four total analyses; Grade 3 Years 1–2 and Years 2–3, as well as Grade 6 Years 1–2 and 2–3). The multinomial outcome had four levels, as follows

$$y_i = \begin{cases} 0 & \text{if AA Year 1 \& AA Year 2,} \\ 1 & \text{if AA Year 1 \& GA Year 2,} \\ 2 & \text{if GA Year 1 \& AA Year 2,} \\ 3 & \text{if GA Year 1 \& GA Year 2.} \end{cases} \quad (1)$$

Equation 1 displays the levels for the outcome for the analysis between Year 1 and Year 2. The levels were defined equivalently for the analysis between Year 2 and Year 3.

Multinomial logistic regression represents an extension of binary logistic regression for cases in which the outcome has more than two unordered categories. Conceptually, the model is similar to a set of binary logistic regression models fit to each of a series of contrasts comparing the odds of membership in a reference category (Category 0) to the corresponding category. An unconditional model can then be fit, producing $j - 1$ intercept coefficients (where j represents the number of levels in the multinomial outcome). The intercepts represent the baseline odds of membership in each category as compared with the reference category. Covariates can then be included to predict group membership, with a set of $j - 1$ coefficients produced for each covariate. Formally, the model is specified as a system of $j - 1$ equations, as follows:

$$\ln\left(\frac{Pr(y_i = 1)\,|\,\mathbf{X}}{Pr(y_i = 0)\,|\,\mathbf{X}}\right) = \mathbf{X}_i\beta_k^1, \quad (2.1)$$

$$\ln\left(\frac{Pr(y_i = 2)\,|\,\mathbf{X}}{Pr(y_i = 0)\,|\,\mathbf{X}}\right) = \mathbf{X}_i\beta_k^2, \quad (2.2)$$

$$\ln\left(\frac{Pr(y_i = 3)\,|\,\mathbf{X}}{Pr(y_i = 0)\,|\,\mathbf{X}}\right) = \mathbf{X}_i\beta_k^3, \quad (2.3)$$

where $Pr(y_i = j)$ represents the probability of student i being represented in category j, $\mathbf{X}$ represents an n (number of observations) by k (number of predictor variables) matrix of covariates, and $\boldsymbol{\beta}$ represents a vector of coefficients of length k. Equation 2.1 models the odds of membership in Category 1, given the set of covariates $\mathbf{X}$, relative to the odds of membership in Category 0, while Equations 2.2 and 2.3 model the odds of membership in Category 2 or 3 relative to Category 0, respectively. The covariate matrix $\mathbf{X}$ also included a leading column of 1s to define the intercept for each corresponding category. The superscript on the beta coefficients represents the corresponding category for which the coefficient was estimated. It is important to note that Equations 2.1 to 2.3 were estimated simultaneously, with maximum likelihood.

Covariates in this study included dummy vectors representing whether the students were classified as LD (0 = *non-LD*, 1 = *LD*) or AUT (0 = *non-AUT*, 1 = *AUT*), along with a third dummy vector representing whether the student was on the performance bubble (0 = *student was not on the performance bubble*, 1 = *student was on the performance bubble*). The intercept for each category represented the likelihood of students with ID (reference group) who were not on the bubble being represented in each of Categories 1 to 3, as opposed to Category 0 (reference category). The relative importance of these variables in predicting group membership was then evaluated.

Following the multinomial analysis, the probability of each student group being represented in each testing pattern was calculated, along with a 95% confidence interval (CI) of each estimate. Student groups included each of the three student disability types, as well as students within each disability type who were also on the bubble (six total student

groups). The group membership probabilities were the primary purpose of the analysis, and the CIs provided an indication of whether student groups differed significantly in their likelihood of being represented in one testing pattern over another.

Results

Our descriptive analysis illustrating patterns of student test taking by cohort, examined all students as one group. As illustrated in Table 1, approximately 13% of the Grade 3 Cohort and approximately 6% of the Grade 6 Cohort switched test type at least once in the two transitions. Furthermore, 43 students (1.41%) in the Grade 3 Cohort and 25 students (0.64%) in the Grade 6 Cohort switched in both transitions. Overall, the Grade 3 Cohort had a higher percentage of students switching test types than students in the Grade 6 Cohort.

Within the Grade 3 Cohort, about 29% of the students who started by taking the AA switched to the GA at least once, approximately 18% after the first year and 12% after the second year. Nearly 4% of students switched back to the AA after a year of taking the GA. Overall, about 8% of students in third grade who started by taking GA switched to AA; approximately 5% after the first administration, and 2% after the second administration. Of students who switched from GA to AA, nearly 20% switched back to the GA after only 1 year.

About 23% of the students in the Grade 6 Cohort whose first test was AA switched to the GA at least once. Approximately 16% of students tested with AA switched to GA after the first year, and nearly 8% after the second administration. Five percent of students who switched to GA from AA switched back to taking AA after only 1 year. Of students in sixth grade taking GA in 2009–2010, about 2% switched to taking AA in the subsequent 2 years: 1.73% after the first administration and 0.59% after the second administration. Approximately, 24% of students who switched to take the AA switched back to take the GA the following year.

Multinomial Logistic Regression

Table 2 reports the multinomial logistic regression odds ratios, coefficients, and a 95% CI for each coefficient for the Grade 3 and 6 Cohorts, respectively. When interpreting this table, it is important to keep the reference group in mind: students with ID who took the AA both years. For example, during the first change period in Grade 3, students with AUT were approximately 3 times more likely than students with ID to be represented in the AA to GA pattern, as opposed to the AA to AA pattern, which was statistically significant. However, for the second change period, students with AUT were no more likely than students with ID to be represented in the AA to GA pattern, as opposed to the AA to AA pattern.

The primary purpose of the multinomial logistic regression models was to calculate the predicted probability of each student group being represented in each assessment pattern. These results are displayed in Table 3, with a 95% CI for each probability. Across cohorts, transition years, and student groups, students were more likely to participate in the same test than to switch participation. These results conformed to our descriptive analysis findings. The probability of students being represented in one switching pattern (AA to GA) was generally not significantly different than the probability of being represented in the other switching pattern (GA to AA). The exception was students with LD who were on the bubble. Those students were significantly more likely to be represented by the AA to GA pattern than the GA to AA pattern (with students in Grade 3 for the first transition being the exception).

Students with ID were most likely to be represented in the AA to AA pattern for both transition years in each cohort, regardless of whether the student was on the performance bubble or not. The probabilities of students with ID being represented in any of the remaining three test patterns was quite low across all models, with the next most likely pattern varying by cohort and transition. Students with ID who were not on the bubble were more likely to only take the GA than to switch test types. Students with AUT were most likely to be represented in the GA to GA pattern across cohorts and transition years. However, students with AUT who were on the performance bubble were much more likely to be represented in other patterns, with the AA to AA pattern generally being the most likely, and with slightly greater likelihood of switching out of rather than into AA. Students with LD were overwhelmingly likely to be represented in the GA to GA pattern across cohorts and transitions (.90 – .98 probabilities). These probabilities dropped considerably if the student was on the test performance bubble. For the Grade 3 Cohort, students with LD on the bubble had just more than a 50% probability of being represented in the GA to GA pattern, with the other three patterns being roughly equally probable. For the Grade 6 Cohort, the probabilities for students with LD on the bubble also dropped relative to LD students not on the bubble, but the reduction was not as dramatic (dropping from .97 and .98 probability to .73 and .66 probabilities for Transitions 1 and 2, respectively).

The probabilities for each cohort and transition year are plotted in Figure 1, with student groups plotted along the *x*-axis, probability plotted along the *y*-axis, testing patterns represented by different shapes, and test switching indicated by shapes filled in black. To compare between groups, all estimates are displayed with a 95% CI. Note the similarity in probabilities across cohorts and transition years modeled.

Table 2. Multinomial Regression Results: Grade 3 and 6 Cohorts.

		Grade 3 Cohort								Grade 6 Cohort							
		Model 1: Years 1–2				Model 2: Years 2–3				Model 1: Years 1–2				Model 2: Years 2–3			
				95% CI				95% CI				95% CI				95% CI	
Assessment pattern	Group	OR	*b*	Lower	Upper	OR	*b*	Lower	Upper	OR	*b*	Lower	Upper	OR	*b*	Lower	Upper
AA to GA	ID	0.02	−3.79	−4.63	−2.94	0.04	−3.17	−3.80	−2.54	0.02	−3.72	−4.42	−3.01	0.03	−3.68	−4.51	−2.85
	AUT	3.01	1.10	0.17	2.03	1.26	0.23†	−0.66	1.11	1.46	0.38†	−0.53	1.29	0.94	−0.06†	−1.09	0.97
	LD	7.02	1.95	1.12	2.76	4.91	1.59	0.82	2.37	10.93	2.39	1.67	3.12	7.57	2.03	1.26	2.79
	Bubble	6.69	1.90	1.41	2.39	3.60	1.28	0.76	1.80	8.34	2.12	1.52	2.73	6.67	1.90	1.21	2.58
GA to AA	ID	0.03	−3.41	−4.21	−2.61	0.03	−3.43	−4.32	−2.54	0.05	−3.03	−3.56	−2.51	0.02	−3.93	−4.87	−2.99
	AUT	3.10	1.13	0.17	2.09	2.42	0.88†	−0.20	1.97	0.72	−0.33†	−1.30	0.63	2.45	0.90†	−0.28	2.07
	LD	17.97	2.89	2.00	3.77	7.72	2.04	1.06	3.03	5.91	1.78	1.10	2.45	6.94	1.94	0.86	3.01
	Bubble	1.13	0.12†	−0.38	0.62	3.60	−0.20†	−0.82	0.42	3.70	1.31	0.68	1.93	2.34	0.85†	0.00	1.70
GA to GA	ID	0.14	−1.95	−2.39	−1.50	0.17	−1.76	−2.20	−1.31	0.20	−1.62	−1.99	−1.26	0.23	−1.46	−1.83	−1.09
	AUT	18.64	2.93	2.43	3.42	17.15	2.84	2.35	3.34	19.36	2.96	2.54	3.39	20.27	3.01	2.57	3.45
	LD	105.44	4.66	4.16	5.16	98.07	4.59	4.09	5.08	219.85	5.39	4.95	5.84	269.48	5.60	5.11	6.09
	Bubble	0.25	−1.39	−1.71	−1.08	0.13	−2.08	−2.38	−1.77	0.27	−1.32	−1.77	−0.86	0.08	−2.52	−3.01	−2.03
	ID	0.02	−3.79	−4.63	−2.94	0.04	−3.17	−3.80	−2.54	0.02	−3.72	−4.42	−3.01	0.03	−3.68	−4.51	−2.85

Note. Students with an intellectual disability who were administered the AA for two consecutive years were the reference group in each analysis. Bubble = students scoring in the top performance category of the AA *or* the bottom category of the GA. OR = odds ratio; CI = confidence interval; AA = alternate assessments based on alternate achievement standards; GA = general assessments; ID = intellectual disability; AUT = autism spectrum disorder; LD = learning disability.

†$p > .05$; all other values significant, $p < .05$.

Table 3. Group Probabilities: Grades 3 and 6 Cohorts.

		Grade 3 Cohort						Grade 6 Cohort					
		Model 1: Years 1–2			Model 2: Years 2–3			Model 1: Years 1–2			Model 2: Years 2–3		
			95% CI			95% CI			95% CI			95% CI	
Disability	Pattern	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper
ID	GA only	.12	.07	.17	.14	.09	.19	.16	.11	.20	.18	.13	.24
	AA to GA	.02	.00	.04	.03	.01	.05	.02	.01	.03	.02	.00	.04
	GA to AA	.03	.01	.05	.03	.00	.05	.04	.02	.06	.02	.00	.03
	AA only	.83	.78	.89	.80	.74	.86	.79	.74	.84	.78	.73	.84
ID bubble	GA only	.03	.01	.05	.02	.01	.03	.04	.02	.06	.02	.01	.03
	AA to GA	.12	.04	.21	.13	.04	.21	.14	.07	.22	.14	.06	.21
	GA to AA	.03	.00	.06	.02	.00	.05	.12	.05	.20	.04	.00	.07
	AA only	.82	.73	.90	.83	.74	.92	.70	.60	.79	.81	.73	.89
AUT	GA only	.70	.66	.73	.72	.68	.76	.78	.75	.82	.82	.78	.85
	AA to GA	.02	.01	.03	.01	.01	.02	.01	.00	.01	.00	.00	.01
	GA to AA	.03	.01	.04	.02	.01	.03	.01	.00	.01	.01	.00	.02
	AA only	.26	.22	.30	.24	.21	.28	.21	.17	.24	.17	.14	.21
AUT bubble	GA only	.30	.23	.36	.23	.17	.29	.42	.32	.52	.23	.15	.31
	AA to GA	.21	.12	.29	.12	.05	.19	.12	.04	.20	.10	.02	.17
	GA to AA	.05	.02	.09	.04	.01	.07	.05	.01	.10	.07	.01	.13
	AA only	.45	.36	.54	.62	.53	.70	.41	.30	.52	.61	.49	.72
LD	GA only	.90	.88	.91	.92	.91	.93	.97	.96	.97	.98	.97	.98
	AA to GA	.01	.01	.01	.01	.01	.02	.01	.00	.01	.00	.00	.01
	GA to AA	.04	.03	.04	.01	.01	.02	.01	.00	.01	.00	.00	.00
	AA only	.06	.05	.07	.05	.05	.06	.02	.02	.03	.02	.01	.02
LD bubble	GA only	.58	.53	.62	.52	.47	.57	.73	.68	.78	.66	.59	.73
	AA to GA	.17	.13	.20	.18	.15	.22	.14	.10	.18	.17	.12	.22
	GA to AA	.10	.08	.13	.05	.03	.07	.07	.04	.09	.04	.02	.07
	AA only	.15	.12	.19	.25	.20	.29	.06	.04	.09	.13	.08	.18

Note. CI = confidence interval; ID = intellectual disability; AA = alternate assessments based on alternate achievement standards; GA = general assessments; AUT = autism spectrum disorder; LD = learning disability.

Discussion

Understanding what SWSCDs know and how they progress is critical to the development of educational practices that better support them. Yet, understanding this population in terms of levels of proficiency and then modeling their growth is challenging, given the design and scaling of most AA and the relative frequency with which they switch tests (GA vs. AA). We documented and empirically investigated the likelihood of students with ID, AUT, and LD switching reading test types over a 3-year span. Each potential transition period was modeled separately for two cohorts of students (elementary and middle school), using proximity to proficiency (bubble membership) and disability type as predictors of their test-taking patterns.

Overall, students were most likely to stay within the test type they had been administered the previous year in either of the two potential transition periods. However, a large portion of students who began by taking the AA did switch at some point during the study (29% Grade 3 Cohort, 24% Grade 6 Cohort). These instances of switching test type are not entirely surprising, given the variation in practices used by IEP teams to determine eligibility for AA (Musson et al., 2010; Roach, 2005; Thurlow, 2004). This documented test switching makes it difficult to develop accountability systems using growth. Furthermore, cohort instability limits the representativeness of the students who are stable over occasions. Finally, because the data of students who switch are not missing at random (Little & Rubin, 2002), growth analyses of the subgroup of nonswitching students may lead to biased estimates.

One of the more surprising findings of this study was the relatively high frequency of students who switched tests multiple times. We found that students not only switch tests, but 18% of students in this study who switched test type in 2010–2011 switched back to their previous test type in 2011–2012. This variation indicates a high potential for

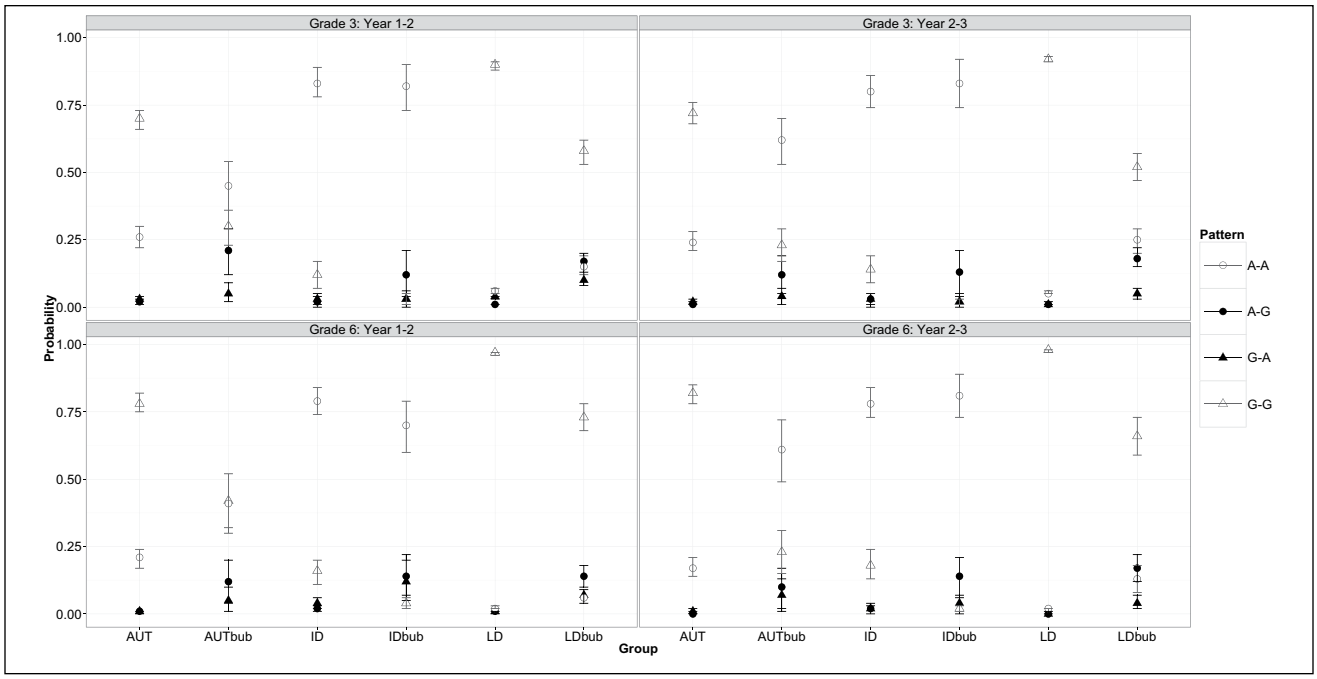

Figure 1. Probability of each student group being represented in each change category.
Note. Students beginning in the AA are represented with circles, while those beginning in the GA are represented with triangles. Students who remained in the same assessment across years are filled in white, while those changing test types are filled in black. All estimates are displayed with a 95% confidence interval (CI), so comparisons between groups can be more readily made (e.g., ID and IDbub students are equally likely to be in the A-A pattern, as indicated by the overlapping error bands). AUT = autism; ID = intellectual disability; LD = learning disability; bub = performance-level "bubble" during first year of transition; A = alternate assessment; G = general assessment.

variability in the methods used to assess the academic achievement of SWSCDs. This finding would be less surprising in a recently developed state AA system in which appropriate identification of participating students was still being fine-tuned. However, this assessment system studied here was well established, having been in operation and undergoing continuous improvement since the spring of 2000.

Some trends emerged from analysis of test patterns for students by disability type across cohort and transition. By and large, students with different disability types exhibited different transition patterns. For students with AUT and LD, these patterns also differed by performance bubble membership, whereas students with ID were, overall, most likely to stay in AA. These observed differences between students grouped by disability type and, in some cases, bubble membership, indicate the importance of avoiding a one-size-fits-all approach to determining which test a student should take. The differences across disability are yet another reminder of the need to make decisions about test participation based on individual student needs.

Students with LD on the performance bubble were the only group with a distinct directionality to documented test switching. Students with LD are not commonly represented as SWSCDs and generally not administered the AA (Kearns et al., 2011; Thurlow, 2004; Towles-Reeves et al., 2012). Therefore, it is encouraging that students with LD on the bubble who switched test type were generally more likely to switch to the GA than to AA, indicating perhaps a better understanding by teachers of not only the population of students with disabilities, but also the test demands. However, schools can also "game" AYP results by manipulating the group of students with disabilities being tested (Wong, 2011). For some instances of switching, school accountability pressures might drive test switching.

Limitations and Future Research

The large number of students with LD taking the AA investigated in this study is likely due, in part, to the state's lack of a multiple disabilities classification. This trend illuminates one limitation of this research, which is the availability of only the students' primary disability code. We used students' disability codes to classify them into groups, yet some students included may have had comorbid disabilities not accounted for here. As documented by Schulte and Stevens (2015), it is already difficult to summarize or generalize findings within specific disability groups using growth models given that students may change classifications over time. In the results we report, students with an LD classification could have been (or could be) classified with another disability category in prior or subsequent

years. However, the extent to which this was the case is unknown.

In addition, the grouping of students into the performance bubble limits the generalizability of our findings, as the bubble variable included students exceeding expectations on the AA *and* performing very low on the GA. Findings for these students have limited generalizability given the use of state-specific cut scores. This state's student population, special education identification policies, assessment system (both GA and AA), and associated testing policies and practices may not generalize well to other states. However, it is worth noting that the bubble variable was created as a proxy for test performance, which may have greater generalizability (e.g., students performing well on the AA are more likely to switch to the GA).

This study demonstrated both that test switching takes place and that patterns of switching vary by primary disability category. Investigating patterns of switching in other states that use instruments such as the Learner Characteristics Inventory (Kearns, Kleinert, Kleinert, & Towles-Reeves, 2006) to describe characteristics of students who participate in AA could help to determine if use of such systems to monitor student test participation helps to reduce the proportion of students switching, or if patterns are similar when such systems are in place. A more detailed understanding of SWSCD test participation, test switching, and resulting outcomes is necessary both to help educators develop practices to better support their needs and to inform policy makers on their progress. Research into patterns of test switching could be used to create professional development to inform educators about the impact of test switching on understanding student proficiency levels and measuring student growth. Such training could assist educators in making more informed decisions about student test participation.

Regardless of training, some students will necessarily continue to switch test participation based on their observed needs. Students who acquire tools to help them express what they know more effectively will move from the AA to the GA with accommodations, and students with a documented need will switch from GA to AA. Future methodological research around methods to incorporate students who switch test type into models of student growth is necessary. In this manner, switching students could be included in accountability systems more accurately, without compromising their needs for access to assessment systems.

Implications

As states investigate growth modeling as a means to represent student academic achievement, test switching becomes a salient concern. If one goal of an accountability system is to document the progress of all students using growth models, mechanisms must be found to include SWSCDs and ensure appropriate participation in the testing program over time. Otherwise, high percentages of students switching test types limit the accuracy of estimates of growth for these students. That being said, students' needs and classifications change over the course of their schooling (Buzick & Laitusis, 2010), sometimes necessitating switching from one test type to another. If a high percentage of students switch test types from one year to another, interpretations of students' levels of proficiency and progress over time is greatly complicated. The utility of student growth models being developed by alternate assessment consortia such as Dynamic Learning Maps and the National Center and State Collaborative (Center for K-12 Assessment & Performance Management, 2012a, 2012b) for characterizing student growth may be reduced without proper attention to the phenomenon of student test switching.

Documentation of test switching provides insight into observed patterns of missingness for students taking AA when constructing longitudinal cohorts, which could perhaps be included in statistical models to reduce the overall bias introduced by the missingness. It is difficult, if not impossible, to make accurate estimations of student growth based on such cohort instability. For example, Saven et al. (2013) found that within a single cohort across Grades 3 to 8, only 3% took the AA across all 6 years. While advanced methods for handling the missingness may help (e.g., multiple imputation, pattern mixture modeling), they are unlikely to fully resolve the issue, as essentially an entirely different set of students from those initially included would be represented by the conclusion of the study. Furthermore, the overall extent to which cohort instability biases the estimates (i.e., the effect size of the missingness) would remain unknown.

Establishing more effective practices to help educators determine test participation for SWSCDs so that their academic progress can be accurately reported is critical. Reexamining switching based on such informed decision making may reflect more accurate insights into the characteristics of the population (Cho & Kingston, 2012; Thurlow, 2004). Future efforts to help educators more accurately identify students as eligible for AA are necessary, given the substantial number of students who switch test type. It is important to ensure that educators align students with the appropriate test type that fits their needs, which should minimize test switching from year to year. The current transition to new assessment systems provides an opportunity to build in appropriate training to ensure the needs of SWSCDs are met.

Authors' Note

The findings and conclusions expressed do not necessarily represent the views or opinions of the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded in part by a Cooperative Service Agreement from the Institute of Education Sciences (IES) establishing the National Center on Assessment and Accountability for Special Education—NCAASE (PR/Award No. R324C110004). Oregon Department of Education funded the development and implementation of the statewide assessments used in this research.

References

Abedi, J., Leon, S., & Kao, J. C. (2008). *Examining differential item functioning in reading assessments for students with disabilities* (CRESST Report No. 744). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Albus, D., & Thurlow, M. L. (2012). *Alternate assessments based on alternate achievement standards (AA-AAAS) participation policies* (Synthesis Report No. 88). Minneapolis, MN: National Center on Educational Outcomes.

Anderson, D., Farley, D., & Tindal, G. (2013). Test design considerations for students with significant cognitive disabilities. *The Journal of Special Education*. Advance online publication. doi:10.1177/0022466913491834

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*, 42–51. doi:10.1111/j.1745-3992.2009.00161.x

Blank, R. K. (2010). *State growth models for school accountability: Progress on development and reporting measures of student growth*. Washington, DC: Council of Chief State School Officers.

Buzick, H. M., & Laitusis, C. C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher*, *39*, 537–544. doi:10.3102/0013189X10383560

Center for K-12 Assessment & Performance Management. (2012a). *The alternate assessment consortia: Dynamic learning* maps *(DLM)*. Austin, TX: Educational Testing Service.

Center for K-12 Assessment & Performance Management. (2012b). *The alternate assessment consortia: National Center and State Collaborative (NCSC)*. Austin, TX: Educational Testing Service.

Cho, H. J., & Kingston, N. (2012). Why IEP teams assign low performers with mild disabilities to the alternate assessment based on alternate achievement standards. *The Journal of Special Education*, *47*, 162–174. doi:10.1177/0022466911435416

Cho, H. J., & Kingston, N. (2013). Examining teachers' decisions on test-type assignment for statewide assessments. *The Journal of Special Education*. Advance online publication. doi:10.1177/0022466913498772

Cho, H. J., & Kingston, N. (2014). Understanding test-type assignment: Why do special educators make unexpected test-type assignments? *Psychology in the Schools*, *51*, 866–878. doi:10.1002/pits.21783

Elliott, S. N., Compton, E., & Roach, A. T. (2007). Building validity evidence for scores on a state-wide alternate assessment: A contrasting groups multi-method approach. *Educational Measurement: Issues and Practice*, *26*, 30–43. doi:10.1111/j.1745-3992.2007.00092.x

Furgol, K. E., & Helms, L. B. (2012). Lessons in leveraging implementation: Rulemaking, growth models, and policy dynamics under NCLB. *Educational Policy*, *26*, 777–812. doi:10.1177/0895904811417588

Gong, B., & Marion, S. (2006). *Dealing with flexibility in assessments for students with significant cognitive disabilities* (Synthesis Report No. 60). Minneapolis, MN: National Center on Educational Outcomes.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics*, *34*, 201–228. doi:10.3102/1076998609332755

Individuals With Disabilities Education Act 20 U.S.C., Pub. L. No. 105-117 § 1400 et seq. (1997).

Kearns, J., Kleinert, H., Kleinert, J., & Towles-Reeves, E. (2006). *Learner Characteristics Inventory*. Lexington: University of Kentucky, National Alternate Assessment Center.

Kearns, J., Towles-Reeves, E., Kleinert, H. L., Kleinert, J. O., & Thomas, M. K.-K. (2011). Characteristics of and implications for students participating in alternate assessments based on alternate academic achievement standards. *The Journal of Special Education*, *45*, 3–14. doi:10.1177/0022466909344223

Kleinert, H. L., Browder, D. M., & Towles-Reeves, E. A. (2009). Models of cognition for students with significant cognitive disabilities: Implications for assessment. *Review of Educational Research*, *79*, 301–326. doi:10.3102/0034654308326160

Lemke, R. J., Hoerandner, C. M., & McMahon, R. E. (2006). Student assessments, non-test-takers, and school accountability. *Education Economics*, *14*, 235–250.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York, NY: Wiley.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamiltion, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 67–101. doi:10.3102/10769986029001067

Musson, J. E., Thomas, M. K., Towles-Reeves, E., & Kearns, J. (2010). An analysis of state alternate assessment participation guidelines. *The Journal of Special Education*, *44*, 67–78. doi:10.1177/0022466909333515

Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Author.

No Child Left Behind Act, 20 U.S.C., Pub. L. No. 107-110 § 1424 et seq., 6301 Stat. (2002).

Quenemoen, R. (2008). *A brief history of alternate assessments based on alternate achievement standards* (Synthesis Report No. 68). Minneapolis, MN: National Center on Educational Outcomes.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org/

Roach, A. T. (2005). Alternate assessment as the "ultimate accommodation": Four challenges for policy and practice. *Assessment for Effective Intervention*, *31*, 73–78. doi:10.1177/073724770503100107

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*, 299–311. doi:10.1007/BF00973726

Saven, J. L., Farley, D., & Tindal, G. (2013). *Constructing alternate assessment cohorts: An Oregon perspective*. Retrieved from http://ncaase.com/publications/in-briefs

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. doi:10.1037//1082-989X.7.2.147

Schulte, A. C., & Stevens, J. (2015). Once, sometimes, or always in special education: Mathematics growth and achievement gaps. *Exceptional Children*, *81*(3), 370–387. doi:10.1177/0014402914563695

Sireci, S. G. (2012). *Smarter balanced assessment consortium: Comprehensive research agenda*. Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/08/Smarter-Balanced-Research-Agenda_Recommendations-2012-12-31.pdf

Thompson, S., Johnston, C., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Vol. *44*). Minneapolis, MN: National Center on Educational Outcomes.

Thurlow, M. L. (2004). *How state policies and practices for alternate assessment impact who is included in NAEP State Assessments*. Paper presented at the NAGB Conference on Increasing the Participation of SD and LEP Students in NAEP. Retrieved from http://www.nagb.org/publications/conferences/thurlow.pdf

Thurlow, M. L., Lazarus, S. S., Thompson, S. J., & Morse, A. B. (2005). State policies on assessment participation and accommodations for students with disabilities. *The Journal of Special Education*, *38*, 232–240.

Towles-Reeves, E., Kearns, J., Flowers, C., Hart, L., Kerbel, A., Kleinert, H., . . . Thurlow, M. L. (2012). *Learner characteristics inventory project report: A product of the NCSC validity evaluation*. Minneapolis, MN: National Center and State Collaborative.

U.S. Department of Education. (2005). *Alternate achievement standards for students with the most significant cognitive disabilities: Non-regulatory guidance*. Washington, DC: Author.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer.

Wong, V. C. (2011, March). Games schools play: How schools near the proficiency threshold respond to accountability pressures under No Child Left Behind. In A. Gamoran (Chair), *Education policy.* Symposium conducted at the Spring meeting of the Society for Research on Educational Effectiveness. Retrieved from http://files.eric.ed.gov/fulltext/ED518224.pdf