Article

Hammill Institute on Disabilities

Remedial and Special Education
2017, Vol. 38(4) 195–206

DOI: 10.1177/0741932516678661
rase.sagepub.com
SAGE

Modeling Reading Growth in Grades 3 to 5 With an Alternate Assessment

Dan Farley, MA[1], Daniel Anderson, PhD[1], P. Shawn Irvin, PhD[1], and Gerald Tindal, PhD[1]

Abstract

Modeling growth for students with significant cognitive disabilities (SWSCD) is difficult due to a variety of factors, including, but not limited to, missing data, test scaling, group heterogeneity, and small sample sizes. These challenges may account for the paucity of previous research exploring the academic growth of SWSCD. Our study represents a unique context in which a reading assessment, calibrated to a common scale, was administered statewide to students in consecutive years across Grades 3 to 5. We used a nonlinear latent growth curve pattern-mixture model to estimate students' achievement and growth while accounting for patterns of missing data. While we observed significant intercept differences across disability subgroups, there were no significant slope differences. Incorporating missing data patterns into our models improved model fit. Limitations and directions for future research are discussed.

Keywords

alternate, assessment, students with significant cognitive abilities, growth modeling, testing, special education, high stakes testing, quantitative, research methodology

Modeling students' academic growth can be challenging, even in relatively ideal situations (Raudenbush, 2001; Rogosa, Brandt, & Zimowski, 1982). The challenges in modeling academic growth are amplified for students with significant cognitive disabilities (SWSCD) who are administered alternate assessments based on alternate achievement standards (AA-AAS; see Karvonen, Flowers, & Wakeman, 2013; Saven, Anderson, Nese, Farley, & Tindal, 2016). These challenges can include missing data, test scaling, small sample sizes, and group heterogeneity.

Missing data pose a serious challenge to modeling growth by introducing a potential source of systematic bias that threatens model-based inferences and generalizability (Allison, 2002; McKnight, McKnight, Sidani, & Figueredo, 2007; Schafer & Graham, 2002). Missing data are pervasive in longitudinal AA-AAS data sets, with as much as 75% attrition documented across Grades 3 to 8 (Saven et al., 2016; Tindal, Nese, Farley, Saven, & Elliott, 2016). The missingness can partially be attributed to students *switching* between the general and alternate assessments between years (Saven et al., 2016). When an established relationship between missingness and measured outcomes exists, namely, that students who performed very well on the AA-AAS took the general assessment the following year and were thus missing on the AA-AAS data, the outcomes are missing not at random (MNAR; Allison, 2002; Schafer & Graham, 2002). In situations where data are MNAR, missingness must be accounted for statistically such that the data become missing at random conditional on the modeled covariates (MAR; Hedeker & Gibbons, 1997; Muthén & Muthén, 1998–2007). Note that missingness is ignorable when the data are missing completely at random (MCAR) or MAR conditional on the modeled covariates, while MNAR data bias model estimates to an unknown extent. However, while tests of MCAR exists (Little & Rubin, 2002), it is impossible to know whether the covariates in a model sufficiently account for the missingness and result in MNAR data becoming MAR (Enders, 2011).

The degree of missing data in AA-AAS longitudinal data sets is severe at times. For example, Saven et al. (2016) found that of the 1,182 students who took a state's AA-AAS in Grade 3, only 293 students were administered the measure annually through Grade 8. The authors also documented the extent to which students switched between the alternate and general assessment from year to year and found higher rates than would be expected by chance, with students performing in the top performance level in the AA-AAS or the bottom performance level on the general assessment generally more

[1]University of Oregon, Eugene, USA

Corresponding Author:
Dan Farley, 175 Lokey Education Building, 5262 University of Oregon, Eugene, OR, 97403-5262, USA.
Email: dfarley@uoregon.edu

likely to switch than their peers. These patterns suggested a *non-random* missing data mechanism, yet previous research on academic growth for SWSCD has not directly addressed the missing data problem.

Modeling growth reliably requires test scales that are sensitive to the academic growth of SWSCD and reflective of a comprehensive developmental continuum. A common scale is generally a prerequisite to meaningful interpretation of score changes over time, without which changes in students' performance are confounded by changes in the measurement scale. Although it is possible to model change over time in a normative manner that is not scale dependent, a common scale is required to determine the magnitude of growth (Betebenner & Linn, 2009).

As a group, SWSCD exhibit tremendous heterogeneity in terms of communication levels, support needs, functional skills, academic abilities, and disability classifications (Towles-Reeves, Kearns, Kleinert, & Kleinert, 2009; Ysseldyke & Olsen, 1997). These intrapopulation differences make it difficult to establish reliable measurement scales. The functional skill and academic ability limitations of SWSCD can be confounded with the constructs assessed (Koretz & Hamilton, 2006). Furthermore, there is little evidence on how Individualized Education Program (IEP) decision making affects AA-AAS test participation, which calls into question the constitution of the tested population (Cho & Kingston, 2013a, 2013b). Changes in disability classifications across grades also calls sample stability into question (Schulte & Stevens, 2013, 2015). In addition, small sample sizes present serious methodological limitations that make it difficult to impossible to meaningfully include all SWSCD in growth modeling efforts (Buzick & Laitusis, 2010; Dunn, Roussos, Lonczak, & Sukin, 2012; Karvonen et al., 2013).

There are promising ways of addressing the challenges of modeling growth for SWSCD, but a number of hurdles remain. In the sections that follow, we briefly review the history of AA-AAS, summarize the available literature related to modeling academic growth for SWSCD, and describe the purpose and context of our study before introducing the methods of our study.

A Brief History of Alternate Assessments

Alternate assessments were first federally mandated in 1997 for students whose disabilities precluded meaningful participation in statewide general assessments (Individuals With Disabilities Education Act Amendments [IDEA], 1997). The No Child Left Behind Act of 2001 (NCLB) subsequently defined technical adequacy requirements for AA-AAS, in which SWSCD were first identified (The No Child Left Behind Act of 2001 [NCLB], 2002). The U.S. Department of Education mandated implementation of AA-AAS to involve SWSCD in statewide academic accountability systems, with federal guidance clarifying the definition and technical requirements of AA-AAS multiple times (U.S. Department of Education, 2003, 2005). All states now include annually administered AA-AAS as part of their statewide accountability systems that are reviewed based on federal technical requirements. With adoption of The Every Student Succeeds Act (ESSA; 2015), only the participation-proficiency rates change; otherwise, testing with AA-AAS continues in form and substance as part of states' broader accountability systems.

Historically, states have used data from AA-AAS within status-based accountability models. However, interest in assessing students' longitudinal growth is increasing: "Including measures of student growth in accountability is important because it allows schools and teachers to be recognized for student learning not just for student performance at a fixed point in time" (Buzick & Laitusis, 2010, p. 537). As a consequence, state general education systems have been developing growth-based models to supplement NCLB-required status-based models (Blank, 2010; Carey & Manwaring, 2011; Jacobs et al., 2015; O'Malley, Murphy, McClarty, Murphy, & McBride, 2011; U.S. Department of Education, 2015).

In addition, teacher evaluation programs are intensifying the need for reliable growth measures. Thirty-five states currently incorporate measures of student academic growth as a preponderant criterion in such evaluations (Jacobs et al., 2015). ESSA (2015) also allowed for growth-based accountability models.

Academic Growth Literature for SWSCD

Although a number of studies have reviewed the academic growth for students with disabilities using various measures, relatively few have documented growth for SWSCD who participate in AA-AAS (Dunn et al., 2012; Karvonen et al., 2013; Tindal et al., 2016). Dunn et al. (2012) evaluated six modeling approaches for alternate assessments using a transition matrix (TM) approach, comparing a student's performance level (e.g., *above proficient, proficient, nearing proficient, well below proficient*) to successive changes in annual classification. Dunn and colleagues studied students in Grades 3 to 8 in the *below proficient* range to determine the percentages of those who moved up, down, or remained the same in performance level. Dunn et al. determined that combining a 3-year TM model with a method to evaluate gains within a performance level was the most comprehensive of the six studied. Although Dunn and colleagues evaluated changes in the performance of SWSCD across grades, two main limitations constrained the inferences they could draw. First, the authors assumed that the assessments given across Grades 3 to 8 were of comparable difficulty, an assumption that is likely untenable. Second, the authors noted small sample sizes and yearly changes in the analytic sample but did not account for the presence of missing data, which limited their ability to evaluate growth.

Karvonen et al. (2013) applied TM, growth to standards (GTS), and ordinary least squares (OLS) regression models to 3 years of reading and mathematics data for SWSCD in three states. The TM approach appeared too insensitive, as approximately 40% of students never changed performance-level classifications. Karvonen and colleagues reported similar results for the OLS regression analyses. Overall, the researchers observed wide variations across states, due in part to differences in performance-level classifications. Similar to Dunn et al. (2012), small sample sizes and missing data clearly limited the ways in which Karvonen and colleagues could model growth. The researchers collapsed analyses across grades to address small sample size and acknowledged their inability to model growth beyond a 2-year span due to missing data.

Tindal et al. (2016) analyzed reading growth for SWSCD in Grades 3 to 5 over the 2008–2010 school years, comparing a TM with a multi-level linear growth model (MLLGM). When applying the TM method, Tindal and colleagues found that the majority of students remained at the same performance level annually. However, using the MLLGM, the researchers found that students' scores increased in a modest, but statistically significant manner from year to year, on average. They concluded that the MLLGM approach was more sensitive than TM approaches. Tindal and colleagues also reported large variance in intercept and growth estimates, suggesting substantial differences between students in terms of initial achievement and rates of growth. The TM, OLS, and GTS approaches evaluated in previous research, while innovative in that they modeled changes in academic performance of SWSCD across grades, appear too insensitive, whereas the MLLGM applied by Tindal et al. (2016), which was tied to a common scale, demonstrated growth that was small and incremental.

Current Study Context and Purpose

Similar to MLLGM, latent growth curve models (LGCM) make it feasible to estimate annual change across multiple grades, provided a common scale is in place. These models also account for the multivariate structure of the data with a residual error term modeled for each grade, provide statistical avenues for handling missing data, and allow for an examination of the variance around the average initial achievement and rates of growth. The purpose of this study is to present preliminary evidence of reading growth for SWSCD across Grades 3 to 5 in one state, while accounting for patterns of missing data. The study context is relatively unique, in that a common AA-AAS, mapped to a common scale, was administered over 3 consecutive years.

We fit a LGCM across Grades 3 to 5, collected during the 2010–2011, 2011–2012, and 2012–2013 school years. We tested both linear and nonlinear functional forms, using an estimated factor-loading approach (see Kamata, Nese, Patarapichayatham, & Lai, 2013). Similar to previous research on academic growth for SWSCD, we observed substantial missing data, and thus explicitly explored the extent to which modeling missingness improved model fit and changed the parameter estimates via a random-effects pattern-mixture model (Hedeker & Gibbons, 1997; Muthén & Muthén, 1998–2007). We addressed the following three research questions:

Research Question 1: What are typical growth trajectories for SWSCD?

Research Question 2: Do SWSCD with different disability classifications grow at significantly different rates?

Research Question 3: Does incorporating a model for missing data significantly improve model fit or change parameter estimates?

Method

Sample

This study included all students who participated in the reading portion of the AA-AAS in one Pacific Northwest state in each of Grades 3, 4, and 5 (N = 1,612). All third-, fourth-, and fifth-grade students who took the third-, fourth-, and fifth-grade AA-AAS, respectively, from 2011–2013 were included in the sample. This selection process did not include a small number of students who participated in the AA-AAS yet did not make typical grade-level progressions from Grade 3 to 4 (n = 7; .009%) and from Grade 4 to 5 (n = 2; .002%). Demographic variables were treated as fixed, with the subgroup status for the first test administration defining all subsequent grades for individual students. All disability subgroups were retained for analysis; however, subgroups with fewer than 20 cases per classification were combined into one heterogeneous category, labeled *Low-Incidence* (LowInc), to preserve the stability of parameter estimates (Kline, 2016). We removed students with no Statewide Student Identification (SSID) number, as well as students with no recorded disability status. These two exclusion criteria eliminated 32 students, for an analytic sample of 1,580. The overall sample means were virtually unchanged by the deletion.

Approximately 68% of students in the analytic sample were male and 81% were White. Approximately 29% of students were identified with a specific learning disability (SLD), 18% with a communication disorder (CD), which is defined federally as a speech–language impairment (SLI), 17% with autism spectrum disorder (ASD), 15% with an intellectual disability (ID), 14% with other health impairments (OHI), 4% with emotional disturbance (ED), and 2% with an orthopedic impairment (OI). An additional 2% of students were classified as low-incidence, which included students with a hearing impairment (HI), visual impairment (VI), deaf-blindness (DB), and traumatic brain injury (TBI). These demographics were consistent with AA-AAS population expectations in the state.

The high percentage of students with SLD and CD included in the AA-AAS is complicated by the fact that the state does not have a multiple disabilities category and because our data included only students' primary disability category. It is likely that students labeled as either SLD or CD within the data set had more complex needs than a primary label might suggest, and, therefore, reflected students with multiple secondary disabilities. It is also possible that the assessment was administered to students who were misclassified as eligible for participation in the AA-AAS.

Measures

The reading portion of the AA-AAS was used for all analyses. The measure was composed of two versions, *Standard* and *Scaffold*. The *Scaffold* version had the same item prompts as the *Standard* version, but included supportive directive statements and visual supports. Each version included *Prerequisite Skills* and *Content* tasks. The *Prerequisite Skills* task determined the level of support students required during the administration of the *Content* tasks, which were designed to assess students' academic knowledge and skills linked to the statewide content standards. Students' scores on the *Prerequisite Skills* task did not count toward the students' reported scores for accountability. A total of 40 *Content* task items were included.

Anderson, Farley, and Tindal (2013) determined that the *Prerequisite Skills* task functioned as a mediator of student disability on latent content knowledge; in addition, the researchers determined that the factor structures for the *Standard* and *Scaffold* versions were equivalent. Anderson and colleagues' study provided empirical evidence that the test design was working as intended and provided comparable results across versions. The internal consistency of the measures, reported by the statewide technical reports, was quite high, with coefficient alpha reported at .92, .95, and .96 for the Grades 3 to 5 test administrations, respectively.

From 2011 to 2013, the state's AA-AAS was a paper/pencil assessment distributed via a secure website. Trained district staff members served as qualified assessors and individually administered the reading portion of the AA-AAS annually each spring. Students were directed to select the most appropriate answer from three choices, with items scored as 0 = *incorrect*, 1 = *partially correct*, and 2 = *correct*. Alternate forms of the same test were administered over the 3 consecutive years, with scores calibrated to a common scale. *Rasch Unit Scale* scores were used for all analyses, estimated via a partial credit Rasch model (Masters, 1982), where RIT = Rasch Unit = $(\theta * 10) + 100$, with θ representing the Rasch person ability estimates.

The scores ranged from approximately 43 to 159 RIT points, across grades. The means for Grades 3 to 5 were 105.56 (n = 1,062, *Mdn* = 111.0, *SD* = 23.61), 110.74 (n = 1,195, *Mdn* = 110.7, *SD* = 24.39), and 109.94 (n = 1,046, *Mdn* = 117.1, *SD* = 23.59), for Grades 3 to 5, respectively. When the scale was initially created with the partial credit Rasch model, it was scaled to have a mean of 100 and a standard deviation of 10. The item difficulty parameters for items included in the initial scale creation were anchored at the initial calibration values for future test administrations, making scores comparable across years. The mean scores for our sample exceeded the initial scale mean of 100 due to new cohorts of students performing at a higher level than the cohort used to establish the initial scale. Approximately 33% of students were missing data from Grade 3, while 24% and 34% of students were missing data in Grades 4 and 5, respectively. The measures correlated at .62 between Grades 3 and 4, .61 between Grades 3 and 5, and .64 between Grades 4 and 5.

Analyses

Our baseline model included a linear LGCM with reading scores from Grade 3, 4, and 5, modeled by latent *Intercept* and *Slope* variables. Factor loadings for the *Intercept* variable were fixed at 1.0, while the factor loadings for *Slope* were initially fixed at 0, 1, and 2, for Grades 3 to 5, respectively. The baseline model also included estimation of the variance around the mean intercept and slope, and the covariance between the two. This model was compared with a nonlinear model, with the factor loading for Grade 5 freely estimated for the latent *Slope* variable (0, 1, and λ; see Kamata et al., 2013). The nonlinear model did not converge without errors. However, given the apparent nonlinearity in the data, we explored a more parsimonious model with slope variance fixed at 0 (implying all students progressed, statistically, at the same rate) for both linear and nonlinear models. Information criteria (described below) indicated the nonlinear model with fixed slope variance fit the data best (see Figure 1), among the parameterizations of the unconditional model. The nonlinear model therefore became our new baseline model, from which all subsequent models were built and compared. Given the non-normality of the score distributions, we used maximum likelihood with robust standard errors (MLR) estimation within the *Mplus 7.3* software for all LGCM analyses (Muthén & Muthén, 1998–2007), which is robust to deviations from multivariate normality.

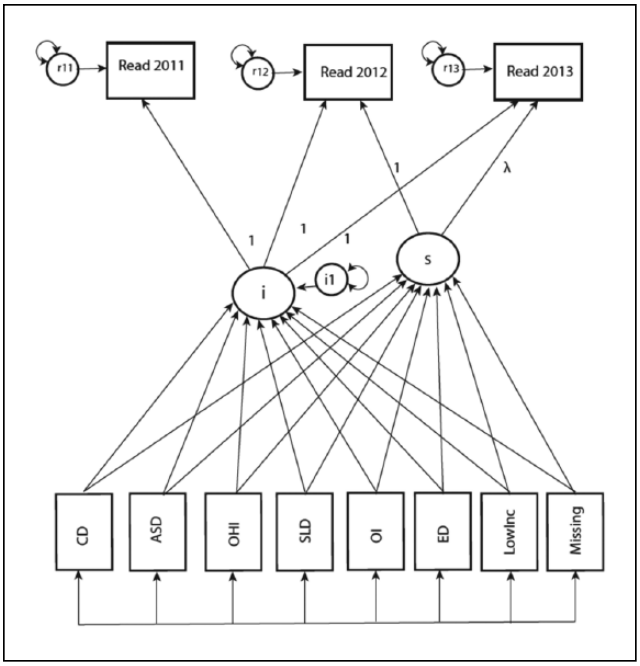

Figure 1. Model 6: Our best-fitting model of reading achievement and growth conditional on students' disability status and patterns of missing data (with fixed slope variance).
Note. CD = communication disorder; ASD = autism spectrum disorder; OHI = other health impairment; SLD = specific learning disability; OI = orthopedic impairment; ED = emotional disturbance; LowInc = hearing impairment, visual impairment, deaf-blindness, or traumatic brain injury; Missing = the missing data patterns, with students with no missing data as the reference group.

Model building. Following estimation of the baseline model, we specified sex (male/female) and race/ethnicity (White/non-White) as predictors of students' initial achievement (Model 2) and then as predictors of student initial achievement and rate of growth (Model 3). These variables did not improve the fit of the model over the unconditional model and were not variables of theoretical interest; they were therefore removed from subsequent models. We next included disability as a predictor of students' initial achievement (Model 4), followed by disability as a predictor of students' initial achievement and rate of growth (Model 5). Disability status was entered with a set of dummy-coded vectors, representing whether the student was identified as CD, ASD, OHI, SLD, OI, ED, or as part of the LowInc category. Students' with ID represented the reference group.

Missing data were analyzed with Little's MCAR test, using the MissMech *R* software package (Jamshidian, Jalal, & Jansen, 2014). The test was rejected ($p < .05$), indicating probable divergence from MCAR, and a possible bias around estimation (Little & Rubin, 2002). Following Model 5, we fit random-effects pattern-mixture models to account for missingness. Pattern-mixture models represent one method of accounting for data that are potentially MNAR by including patterns of missing data as predictors of the estimated parameters. Pattern-mixture models separate the analysis into strata by the observed missingness patterns, and growth models are fit within each stratum (Hedeker & Gibbons, 1997; Muthén & Muthén, 1998–2007). The difference in parameter estimates between missing data patterns can then be evaluated with the main effects (i.e., average initial achievement and rate of growth). In other words, the data become MAR, conditional on the observed covariates (see Little & Rubin, 2002). Observed missing data patterns and percentages are displayed in Table 1.

Table 1. Missing Data Patterns, Population Sizes, and Percentages.

	Grade level				
Pattern	Grade 3	Grade 4	Grade 5	*n*	%
Missing G34	0	0	1	135	8.54
Missing G35	0	1	0	142	8.99
Missing G45	1	0	0	213	13.48
Missing G3	0	1	1	241	15.25
Missing G4	1	0	1	37	2.34
Missing G5	1	1	0	179	11.33
No missing data	1	1	1	633	40.06

Note. A zero indicates missing data, while a one indicates a valid score. Missing G3 = missing data in third grade only; Missing G4 = missing data in fourth grade only; Missing G5 = missing data in fifth grade only; Missing G34 = missing data in third and fourth grade; Missing G45 = missing data in fourth and fifth grade. Missing G35 = missing data in third and fifth grade. No missing data = student had complete data over the 3-year period.

Model fit evaluation. When evaluating model fit, we used cut off criteria recommended by Hu and Bentler (1999) a priori, targeting a comparative fit index (CFI) ≥ 0.95, standardized root mean square residual (SRMR) ≤ 0.08, and root mean square error of approximation (RMSEA) ≤ 0.06. When comparing competing models, we used multimodel inference (Burnham & Anderson, 2004), with Akaike information criterion (AIC; Akaike, 1973) and Bayesian information criterion (BIC; also referred to as Schwartz criterion; Schwartz, 1978). AIC and BIC are transformations of the log likelihood that balance model fit with parsimony (i.e., both fit indices include penalties for the number of estimated parameters), in which lower values indicate better fitting models. We also calculated Akaike weights (using both AIC and BIC; see Burnham & Anderson, 2004), which transform raw AIC and BIC values into conditional probabilities. The weights can be interpreted as the evidence in favor of one model over others in a set of models.

Magnitude of growth. To facilitate interpretation of the magnitude of observed growth with test results that are likely on

Table 2. Model Parameter Estimates: Reference Group Is Students With ID and No Missing Data.

	Unconditional (1)		Disability on intercept (4)		Disability on intercept and slope (5)		Disability and missing data on intercept (6)		Fully conditional (7)	
Parameter	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*
Intercept	106.83**	0.62	94.77**	1.48	94.89**	1.62	92.30**	1.53	92.97**	1.68
CD			19.69**	1.60	19.45**	1.82	17.50**	1.59	16.91**	1.83
ASD			−2.66	2.19	−2.27	2.40	−3.46	2.17	−3.44	2.37
OHI			14.44**	2.05	14.55**	2.32	12.65**	1.54	12.43**	2.30
SLD			24.43**	1.51	23.77**	1.75	21.31**	1.54	20.09**	1.79
OI			−24.25**	6.04	−26.45**	6.32	−25.34**	6.14	−28.33**	6.56
ED			19.54**	1.99	19.03**	2.39	16.65**	1.99	15.36**	2.42
LowInc			3.47	4.89	4.47	5.34	2.39	4.67	3.37	4.87
MissG34							3.68*	1.78	3.10	1.79
MissG35							8.52**	1.59	9.17**	1.63
MissG45							12.38**	1.47	12.40**	1.48
MissG3							3.61*	1.46	8.27**	2.50
MissG4							−1.11	3.66	0.23	3.85
MissG5							8.19**	1.47	7.00**	1.58
Slope	5.49**	0.47	5.33**	0.47	5.12	0.92	5.67**	0.51	4.29**	0.89
G5 factor loading	1.27	0.10	1.30	0.10	1.32	0.11	1.35	0.10	1.64	0.19
CD					0.32	1.16			0.60	0.96
ASD					−0.54	1.37			−0.12	1.14
OHI					−0.15	1.23			0.18	1.02
SLD					0.86	1.05			1.34	0.89
OI					2.72	3.40			3.22	2.88
ED					0.67	1.87			1.43	1.64
LowInc					−1.35	2.74			−1.28	2.37
MissG3									−3.43*	1.58
MissG4									−1.99	1.98
MissG5									3.17*	1.28

Note. ID = intellectual disability; CD = communication disorder; ASD = autism spectrum disorder; OHI = other health impairment; SLD = specific learning disability; OI = orthopedic impairment; ED = emotional disturbance; LowInc = hearing impairment, visual impairment, deaf-blind, and traumatic brain injury. MissG3, G4, G5 = missing data in third, fourth, or fifth grade. MissG34 = missing data in third and fourth grade, and so on.
**p* < .05. ***p* < .001.

an unfamiliar scale and to provide a comparative source for future studies of SWSCD academic growth, we calculated effect sizes between Grades 3, 4, and 5 for each disability group. These effect size calculations convey the mean differences between each grade divided by the pooled standard deviation and provide a normative expectation of annual reading growth for SWSCD. We also calculated between-group effect size gaps for each grade level. Although it is not feasible at present to compare these effect size gaps with other studies with SWSCD, they may nonetheless provide a relevant benchmark for researchers and statewide accountability system policy makers interested in performance gaps when defining appropriate growth expectations for SWSCD (Hill, Bloom, Black, & Lipsey, 2008).

Results

Parameter estimates are displayed for Models 1, 4, 5, 6, and 7 in Table 2. These models were most relevant to our research questions and also conveyed the best model fit statistics. Intercept and residual variances, as well as model fit statistics, are provided in Table 3. The baseline model, Model 1, indicated that students scored, on average, 106.83 RIT points in Grade 3. The estimated factor loading for Grade 5 was 1.27, while the estimated slope was 5.49. These results imply that students gained, on average, 5.49 RIT points from Grades 3 to 4 and $5.49 \times 1.27 = 6.97$ points from Grades 3 to 5, without accounting for student disability or controlling for missing data patterns. The average difference in gains from Grades 3 to 4 (5.49) to Grades 4 to 5 (1.48) was substantial and nonlinear.

Student disability covariates were added (Model 4) to determine the relationship among the disability groups in terms of initial status. The addition of disability covariates resulted in a substantial drop in AIC and BIC relative to Model 1 (ΔAIC = −517.66, ΔBIC = −480.11), suggesting improved model fit to the data. Student disability subgroups performed in the following ranked order (highest to

Table 3. Variances and Model Fit Statistics.

Statistic	Unconditional Model 1		Disability on intercept Model 4		Disability on intercept and slope Model 5		Disability and missing pattern on intercept Model 6		Fully conditional Model 7	
Variance	Variance	*SD*	Variance	*SD*	Variance	*SD*	Variance	*SD*	Variance	*SD*
Intercept	444.34**	21.08	302.42**	17.39	301.88**	17.37	286.17**	16.92	285.42**	16.89
Residual 2011	104.00**	10.20	104.12**	10.20	104.38**	10.22	102.45*	10.12	102.37**	10.12
Residual 2012	123.98**	11.13	122.37**	11.06	122.49**	11.07	121.16**	11.01	120.44**	10.97
Residual 2013	101.08**	10.05	100.60**	110.03	100.09**	10.00	102.28*	10.11	101.06*	10.05
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Information criteria	28,345.25	28,382.81	27,827.59	27,902.70	27,838.35	27,951.02	27,776.83	27,884.14	27,780.87	27,941.83
Akaike weights	0.0	0.0	0.0	0.0	0.0	0.0	0.88	1.0	0.12	0.0
Fit indices										
CFI	0.999		1.000		0.998		0.988		0.988	
SRMR	0.044		0.017		0.016		0.033		0.034	
RMSEA	0.014		0.000		0.014		0.025		0.033	

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion. CFI = comparative fit index; SRMR = standardized root mean square residual; and, RMSEA = root mean square error of approximation.
*p < .01. **p < .001.

lowest): SLD, CD, ED, OHI, LowInc, ID, ASD, and OI. Disability covariates were included as predictors of the slope in Model 5, but model fit was not improved compared with Model 4 (ΔAIC = +10.76, ΔBIC = +48.32). None of the individual parameter estimates indicated that students identified in any disability category progressed at a significantly different rate, on average, compared with students with ID.

Missing data patterns were included as predictors of students' initial achievement in Model 6, along with the previously included disability covariates. The inclusion of missing data patterns resulted in another drop in model AIC and BIC compared with Model 4, the best-fitting prior model (ΔAIC = −50.76, ΔBIC = −18.56). The Model 6 intercept values for each disability decreased compared with Model 5 after accounting for missing data patterns. However, the ranked-order pattern of initial student achievement by disability grouping from Models 5 to 6 remained consistent. Finally, we fit our fully conditional model, with disability and missing data patterns entered as predictors of students' initial achievement and rate of growth (Model 7; Note: We were not able to include all missing data patterns as predictors of the slope in Model 7, as three of the patterns had only 1 year of data). This did not result in improvement for AIC or BIC compared with our best-fitting prior model, Model 6 (ΔAIC = +4.04, ΔBIC = +57.69). Akaike weights for Model 6 show that it is our best-fitting model, rounding to 1.0 for BIC and 0.88 for AIC.

Analysis of model fit statistics suggested that our models fit the data quite well, given our a priori criteria. The CFI ranged from 0.988 to 1.000 (Model 6 = .988), the SRMR ranged from 0.016 to 0.044 (Model 6 = 0.033), and the RMSEA ranged from 0.000 to 0.033 (Model 6 = 0.025). Our best-fitting model, Model 6, indicated that the initial achievement of the reference group (ID) was 92.30 RIT scale points, and that they gained 5.67 RIT points between Grades 3 and 4 and 1.98 RIT points between Grades 4 and 5 ($\lambda = 1.35$).

We observed broad mean differences across models in students' initial achievement, depending on disability and patterns of missing data (see Table 2). Model 6, for instance, which displayed the best evidence of fit, indicated that, on average, students with SLD scored 21.31 points higher initially than ID students with no missing data, while students with CD, ED, and OHI scored 17.50, 16.65, and 12.65 points higher, respectively. Students with ASD or in the LowInc category did not have an initial achievement significantly different from students with ID. All model observed means and nonlinear growth trajectories by disability are presented in Figure 2.

Tables 4 and 5 convey effect sizes for the average growth between grades and effect size gaps among the disability groups, respectively. Overall, patterns of growth are in accord with our models, where greater average growth occurred between Grades 3 and 4 than between Grades 4 and 5. Students with SLD, OI, and ED outperformed their peers in this cohort, with effect sizes ranging from 0.56 to 0.65 over the studied grades. The lowest performing group in terms of growth was LowInc (i.e., VI, HI, DB, and TBI), which appeared to marginally regress, on average. The effect size gaps between disability subgroups, reported in Table 5, indicated that the size of the achievement gaps

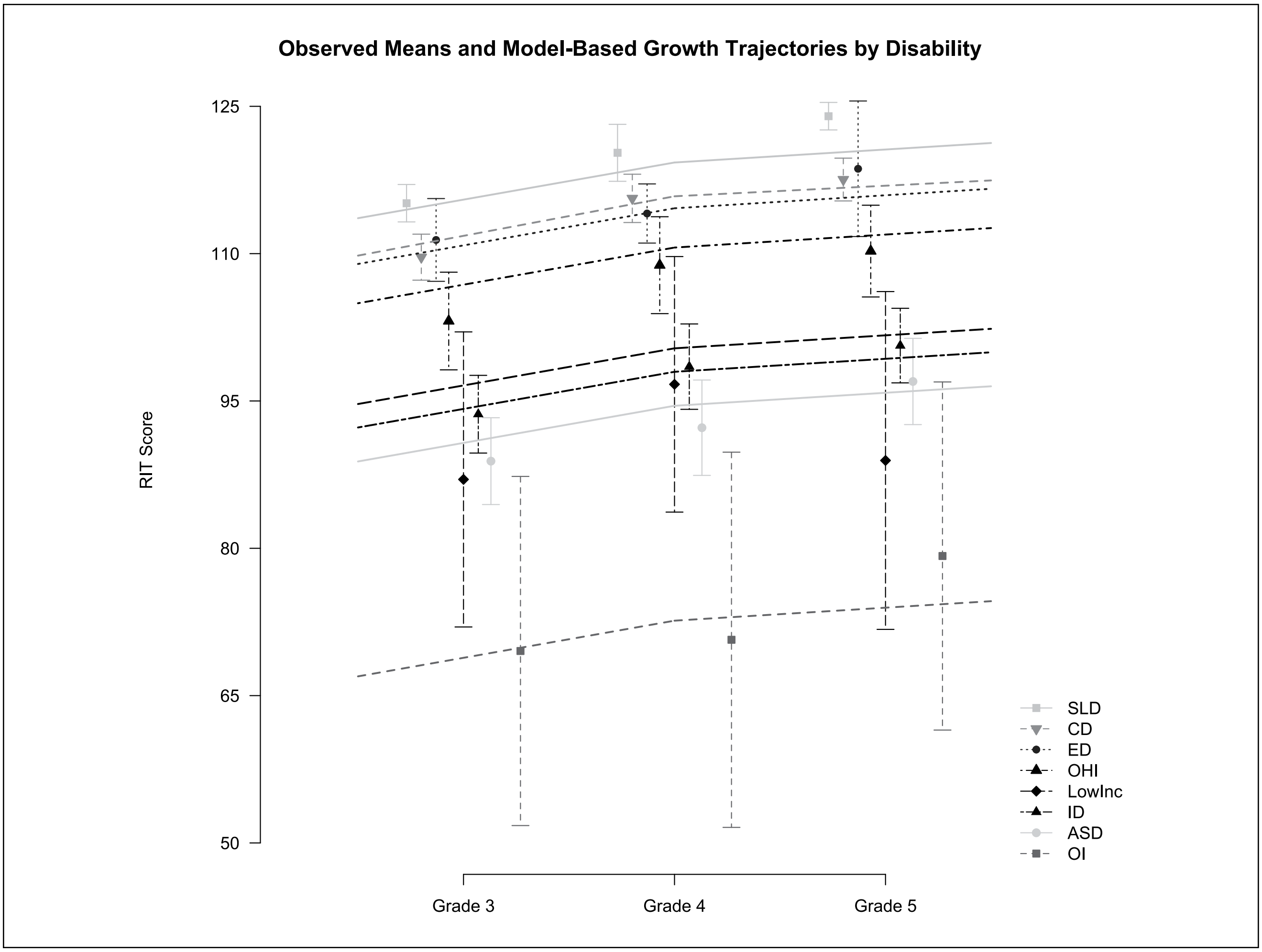

Figure 2. Observed means are displayed with two standard errors around each mean.
Note. Observed means are jittered to match the hierarchical order presented in the legend, from highest performing to lowest performing. Nonlinear reading growth trajectories for each of the eight disability groups included in our study resultant from Model 6, our best-fitting model, are also presented. RIT = Rasch unit; SLD = specific learning disability; CD = communication disorder; ED = emotional disturbance; OHI = other health impairment; LowInc = hearing impairment, visual impairment, deaf-blindness, or traumatic brain injury; ID = intellectual disability (reference group); ASD = autism spectrum disorder; OI = orthopedic impairment.

Table 4. Effect Size by Disability.

Variable	Grade 3 to 4 Δ	Grade 4 to 5 Δ	Grade 3 to 5 Δ
ID	0.22	0.02	0.26
CD	0.45	−0.07	0.37
ASD	0.09	0.02	0.11
OHI	0.19	0.06	0.27
SLD	0.45	0.02	0.56
OI	0.27	0.25	0.56
ED	0.54	0.20	0.65
LowInc	0.21	−0.31	−0.09
Total	0.22	−0.03	0.19

Note. Δ = change in effect size between grades. ID = intellectual disability; CD = communication disorder; ASD = autism spectrum disorder; OHI = other health impairment; SLD = specific learning disability; OI = orthopedic impairment; ED = emotional disturbance; and LowInc = hearing impairment, visual impairment, deaf-blind, and traumatic brain injury as one combined category.

between student disability groups remained stable across grades. This is consistent with our modeling results, suggesting that students exhibit significant intercept differences, but progressed at similar average rates.

Discussion

Little previous research has modeled the academic growth of SWSCD using AA-AAS, which inherently presents complex challenges. The results of this study add to the growing body of literature (Dunn et al., 2012; Karvonen et al., 2013; Tindal et al., 2016), using a sophisticated LGCM approach while accounting for the presence of missing data patterns. Our research attempted to answer three primary questions.

Our first research question addressed the typical academic reading growth trajectory for SWSCD in Grades 3 to 5. The model that displayed the best fit to the data, Model

Table 5. Effect Size Gaps Between Disability Subgroups.

Variable	Grade	ID	CD	ASD	OHI	SLD	OI	ED	LowInc
CD	3	0.81	—	—	—	—	—	—	—
	4	0.78							
	5	0.75							
ASD	3	−0.06	−0.87	—	—	—	—	—	—
	4	−0.19	−0.97						
	5	−0.19	−0.95						
OHI	3	0.56	−0.25	0.62	—	—	—	—	—
	4	0.49	−0.29	0.68					
	5	0.54	−0.20	0.74					
SLD	3	1.02	0.21	1.08	0.47	—	—	—	—
	4	0.97	0.19	1.16	0.48				
	5	0.99	0.24	1.18	0.44				
OI	3	−1.26	−2.07	−1.19	−1.81	−2.28	—	—	—
	4	−1.05	−1.84	−0.87	−1.55	−2.03			
	5	−0.74	1.49	−0.54	−1.29	−1.73			
ED	3	0.78	−0.03	0.84	0.22	−0.24	2.03	—	—
	4	0.75	−0.03	0.94	0.26	−0.22	1.82		
	5	0.85	0.10	1.04	0.30	−0.14	1.59		
LowInc	3	0.14	−0.67	0.21	−0.41	−0.88	1.40	−0.63	—
	4	0.15	−0.63	0.34	−0.34	−0.82	1.21	−0.60	
	5	−0.24	−0.99	−0.05	−0.79	−1.23	0.50	−1.09	

Note. Effect size gaps are listed by year: 2011, 2012, and 2013. The column variables were subtracted from the row variables; therefore, positive values mean that the row variable had a higher mean, where negative values mean that the column variable had a higher mean. For example, students with CD outperformed students with ID across all 3 years, but performance gaps remained consistent (0.81, 0.78, and 0.75). ID = intellectual disability; CD = communication disorder; ASD = autism spectrum disorder; OHI = other health impairment; SLD = specific learning disability; OI = orthopedic impairment; ED = emotional disturbance; LowInc = hearing impairment, visual impairment, deaf-blind, and traumatic brain injury.

6, indicated that SWSCD involved in our study grew 5.67 RIT scale points between Grades 3 and 4 and 1.98 RIT scale points between Grades 4 and 5 (λ = 1.35). Growth was nonlinear, with more growth occurring between Grades 3 and 4 than between Grades 4 and 5, which underscores the importance of first determining functional form in growth modeling.

We present effect sizes in Table 4 to provide scale-independent analyses of our findings. Effect sizes indicated that some disability subgroups demonstrated moderate growth over the 2-year period (ED, SLD, and OI), while others displayed moderate to small (CD) or small growth (ID, ASD, OHI) based on Cohen's (1988) thresholds. However, Hill et al. (2008) argued that effect sizes should be interpreted in context and cannot be aptly understood with effect size rules of thumb. Although not a perfect comparison due to the fact that their meta-analyses included general education students, Hill et al. conducted a comprehensive literature review and found that annual reading effect sizes gains were 0.36 from Grades 3 to 4 and 0.40 from Grades 4 to 5 (0.76 across Grades 3–5). Averaging across disabilities, the effect size gain from Grades 3 to 5 in our study was 0.19, more than .5 standard deviations lower than that reported by Hill et al. for general education students. Performance differences, perhaps expectedly dissimilar between general education peers and SWSCD, provide a comparative context for future research involving AA-AAS. They are also important for statewide accountability policy makers to consider, particularly for those systems that incorporate growth expectations across diverse student populations. Additional research should help hone growth expectations for SWSCD.

Our second research question considered whether students with different disability classifications progressed at significantly different rates. In our best-fitting model, we found that students with different disability classifications differed in initial average achievement, but did not grow at significantly different rates (see Figure 2). Our finding was in contrast to Tindal et al. (2015) who found substantial differences in the rates of growth among SWSCD and merits further study. Overall, our results indicated that disability category was not a significant predictor of slope. However, it is important to recognize that this finding might have been due to limitations in the areas of power or scale sensitivity. For example, Model 7 showed growth differences between student subgroups that are of theoretical interest and deserve further study (see Figure 2). Students with SLD, CD, ED, and OHI began with higher initial status and grew at faster rates, while students classified as LowInc, ID, and ASD began with lower initial status and grew at slower rates. Students with OI, by contrast, began with a low initial status, yet grew at a substantial rate.

These findings, however, should be interpreted cautiously, given that the inclusion of the disability subgroup parameters reduced model fit and parallel slopes fit the data better.

Our final research question addressed whether including missing data patterns improved model fit and/or affected parameter estimates. We found that incorporating missing data patterns improved model fit (see Table 2). The observed missing data patterns are consistent with those found by Saven et al. (2016) and suggest that some students switched between the general and alternate assessment. Saven and colleagues (2016) found that approximately 25% of students switched between the general assessment and the AA-AAS in the state reviewed over a 3-year period. Our annual missingness rates varied from 24% to 34%. It is feasible, therefore, that the missing data mechanism in our models may be largely accounted for by test switching. Determining the reasons for observed patterns of missingness in AA-AAS is an important avenue of future study, as it may afford inclusion of additional meaningful predictors of growth.

Statistically addressing observed missingness has not been part of previous research on the estimation of growth for SWSCD, despite being a prevalent and substantial hurdle. Had we included only students with complete data, the analytic sample would have included only 633 students—a loss of about 60% of the overall sample. The pattern-mixture model represents one method for accounting for meaningful patterns of missing data, and future research should explore the sensitivity of this and alternative methods. The pattern of results suggests that missingness affected parameter estimates and should be accounted for when modeling growth.

Beyond evaluating the academic growth of SWSCD for substantive reasons (e.g., exploring typical rates of improvement for select disability subgroups), the educational policy environment has placed considerable emphasis on evaluating students' growth as part of accountability frameworks. This might be a difficult goal to achieve for SWSCD, given (a) missing data are pervasive and often make up a substantial proportion of testing samples, (b) common scales are generally not available even within AA-AAS testing frameworks, (c) the population of students has high variability in performance (even within subgroups), and (d) subgroup sample sizes are often very small (even at the state level). Furthermore, to date, content domains have generally not been constructed to define a developmental continuum that supports meaningful interpretations of score changes across grades with respect to actual student learning. Drawing valid inferences about growth requires a set of academic content standards and achievement-level descriptors (ALDs) that map vertically and outline expected gains in knowledge and skills as assessment scores increase (Karvonen et al., 2013). We used a common scale for our study, but our ALDs were not mapped vertically by design.

Limitations

The purpose of this study was to document the academic reading growth of SWSCD in Grades 3 to 5, while accounting for disability and missingness. Our results must be qualified in light of their limitations. First, we assumed that students' disability classifications were singular in nature and that they did not change over the course of the study. However, classification can be multifaceted (e.g., a student might be identified as having multiple disabilities) and change from grade to grade. Evidence suggests that measuring and accounting for this variability may improve model fit and affect study results (Schulte & Stevens, 2013).

Second, as noted in our discussion of the sample, high percentages of students with SLD and CD were included in this state's AA-AAS. Although it is likely that students labeled as either SLD or CD had more complex needs than a primary label might suggest, we expect that their inclusion may have inflated grand mean and slope estimates, though these were accounted for in subsequent models. These findings suggest that some students might have been misclassified as eligible for participation in this state's AA-AAS. The state should thus reconsider the rigor of its AA-AAS eligibility criteria.

Third, although we attempted to account for the observed patterns of missing data with a pattern-mixture model, in which data can be considered MAR conditional on the observed covariates (Little & Rubin, 2002), it is clear that missing data in our study were MNAR. Unfortunately, it is impossible to know whether the pattern-mixture model adequately accounted for the bias introduced by the MNAR process (Enders, 2011). We simply argue that missing data are a consistent complexity in AA-AAS modeling contexts and that the effect of missing data on growth estimates cannot be ignored. We also recommend further exploration of the systematic factors that are generating missing data, such as test switching (Saven et al., 2016) and IEP team eligibility decision making (Cho & Kingston, 2013a, 2013b).

Finally, our growth modeling assumed that the common assessment scale was sufficient, including an assumption that it aptly quantified growth across the identified achievement continuum. These assumptions are questionable and worthy of substantiation (Betebenner & Linn, 2009). Having a common scale, while a prerequisite for modeling progress across grades, may be insufficient to model growth if it is not matched to a developmental scope and coherent sequence of academic content and performance expectations (Karvonen et al., 2013). An important design component of the AA-AAS used for this study was that the content standards were initially vertically aligned and then all items linked to them; however, the standards and ALDs have not yet been evaluated through the lens of modeling growth.

Conclusions and Future Directions

Although progress is being made in efforts to estimate reading growth for SWSCD taking AA-AAS, the process gives rise to many unanswered questions. We found that most SWSCD demonstrated growth when the assessment results were anchored to a common scale; yet, there were no statistically significant differences in growth rates between student disability subgroups. We also found that growth is quite gradual and decreases over time, on average, for many students.

The impact of missing data for AA-AAS growth modeling cannot be ignored, given its degree and ostensible pervasiveness. Missing data impact growth estimates by potentially limiting the types of available analyses and threatening the validity of interpretations by introducing additional sources of bias. Including missing data patterns within growth models is one promising way of handling this issue, yet further study is needed. Our study accounted for missing data patterns in estimates of intercept and growth, but investigation into the contextual factors that contribute to the missing data could lead to more advanced methods (i.e., embedding models for the missing data mechanisms within growth models). Irrespective of the methods used, procedures should account for missing data when the rates are as high as we and others have observed (Karvonen et al., 2013; Saven et al., 2016).

AA-AAS growth modeling efforts are contingent on the development of common scales that are sensitive to detecting changes in students' ability. To improve the approach taken to model growth further, thoughtful and sequential design of academic standards and ALDs that link vertically should also be effected, which would help make changes in students' scores across grades more interpretable. The focus of assessment development could then aptly shift to creating a range of items reflecting performance that is tied to an interpretable vertical scale, to allow for more reliable growth estimation and comparisons across grades, and the academic profiles of students taking the AA-AAS to be better understood longitudinally.

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 R324C110004).

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: SAGE.

Anderson, D., Farley, D., & Tindal, G. (2013). Test design considerations for students with significant cognitive disabilities. *The Journal of Special Education*, *49*, 3–15. doi:10.1177/0022466913491834

Betebenner, D. W., & Linn, R. L. (2009). *Growth in student achievement: Issues of measurement, longitudinal data analysis, and accountability*. Retrieved from http://www.ets.org/Media/Research/pdf/BetebennerandLinnPresenterSession1.pdf

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.

Burnham, K. P., & Anderson, D. R. (2004). Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–304. doi:10.1177/0049124104268644

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

Carey, K., & Manwaring, R. (2011). *Growth models and accountability: A recipe for remaking ESEA*. Washington, DC: Education Sector.

Cho, H.-J., & Kingston, N. (2013a). Examining teachers' decisions on test-type assignment for statewide assessments. *The Journal of Special Education*, *49*, 16–27. doi:10.1177/0022466913498772

Cho, H.-J., & Kingston, N. (2013b). 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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Dunn, J., Roussos, L., Lonczak, T., & Sukin, T. (2012, April). *An exploration of growth models for an alternate assessment*. Paper presented at the American Educational Research Association Annual Conference, Vancouver, British Columbia, Canada.

Enders, C. K. (2011). Missing not at random models for latent growth curve analysis. *Psychological Methods*, *16*, 1–16.

The Every Student Succeeds Act of 2015, Pub. L. No. 114–195. § 1111. (2015).

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*, 64–78.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55.

Individuals with Disabilities Education Act Amendments, Pub. L. No.105–117. (1997).

Jacobs, S., Doherty, K. M., Joseph, N., Nithya, J., Lakis, K., Staresina, L., & Wasbotten, C. (2015). *2015 state teacher policy yearbook: National summary*. Washington, DC: National Council on Teacher Quality.

Jamshidian, M., Jalal, S., & Jansen, C. (2014). MissMech: An R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR). *Journal of Statistical Software*, *56*(6), 1–31.

Kamata, A., Nese, J. F. T., Patarapichayatham, C., & Lai, C.-F. (2013). Modeling nonlinear growth with three data points: Illustration with benchmarking data. *Assessment for Effective Intervention*, *38*, 105–116.

Karvonen, M., Flowers, C., & Wakeman, S. Y. (2013, April). *An exploration of methods for measuring academic growth for students with significant cognitive disabilities*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Kline, R. B. (2016). *Principles and practices of structural equation modeling* (4th ed.). New York, NY: The Guilford Press.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport. CT: Praeger.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analyses with missing data*. Hoboken, NJ: John Wiley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi:10.1007/BF02296272

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. New York, NY: The Guilford Press.

Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

The No Child Left Behind Act of 2001, Pub. L. No. 107-110 § 115, Stat 1425. (2002).

O'Malley, K. J., Murphy, S., McClarty, K. L., Murphy, D., & McBride, Y. (2011). *Overview of student growth models* (Test, Measurement, & Research Services White Paper). Iowa City, IA: Pearson. Retrieved from http://researchnetwork.pearson.com/wp-content/uploads/StudentGrowthWP083111.pdf

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, *52*, 501–525. doi:10.1146/annurev.psych.52.1.501

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*, 726–748. doi:10.1037/0033-2909.92.3.726

Saven, J. L., Anderson, D., Nese, J. F. T., Farley, D., & Tindal, G. (2016). Patterns of state assessment participation for students with significant cognitive disabilities. *The Journal of Special Education*, *49*, 209–220. doi:10.1177/0022466915582213

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of state of the art. *Psychological Methods*, *7*, 147–177.

Schulte, A., & Stevens, J. J. (2013, April). *Special education growth: Contrasting stable and variable identification of special education student status across grades*. Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco, CA.

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

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Tindal, G., Nese, J. F. T., Farley, D., Saven, J. L., & Elliott, S. N. (2016). Documenting reading achievement and growth for students taking alternate assessments. *Exceptional Children*, *82*, 321–336. doi:10.1177/0014402915585492

Towles-Reeves, E., Kearns, J., Kleinert, H., & Kleinert, J. O. (2009). An analysis of the learning characteristics of students taking alternate assessments based on alternate achievement standards. *The Journal of Special Education*, *42*, 241–254.

U.S. Department of Education. (2003). Title 1 – Improving the academic achievement of the disadvantaged; final rule, 34 CFR § 200.

U.S. Department of Education. (2005). *Alternate achievement standards for students with the most significant cognitive disabilities: Non-regulatory guidance*. Washington, DC. Retrieved from https://www2.ed.gov/policy/elsec/guid/altguidance.doc

U.S. Department of Education. (2015). *ESEA flexibility*. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

Ysseldyke, J. E., & Olsen, K. R. (1997). *Putting alternate assessments into practice: What to measure and possible sources of data* (Synthesis Report No. 28). Minneapolis: University of Minnesota, National Center on Educational Outcomes.