Hammill Institute on Disabilities

The Journal of Special Education
2015, Vol. 49(1) 3–15

DOI: 10.1177/0022466913491834
journalofspecialeducation.sagepub.com
SAGE

Test Design Considerations for Students With Significant Cognitive Disabilities

Daniel Anderson, MS[1], Dan Farley, MA[1], and Gerald Tindal, PhD[1]

Abstract

Students with significant cognitive disabilities present an assessment dilemma that centers on access and validity in large-scale testing programs. Typically, access is improved by eliminating construct-irrelevant barriers, while validity is improved, in part, through test standardization. In this article, one state's alternate assessment data were analyzed to determine the impact of (a) administration supports based on students' level of independence and (b) a scaffold test administration format. Using structural equation modeling, we tested the extent to which students' level of independence mediated the relation between disabilities and latent content knowledge scores. We then tested the invariance of the measurement model across administration formats. The results provide evidence that these supports help students access the test without compromising the validity of test-based inferences.

Keywords

alternate assessment, structural equation modeling, test access, test design, validity

Measuring the academic achievement of students with the most significant cognitive disabilities remains a complex challenge. Two major test designs exist, including individualized body of work portfolios and standardized performance-based assessments (Altman et al., 2010), each with strengths and limitations. Portfolio assessments can be adapted to the unique needs of each student, yet comparisons across students in terms of levels of achievement are difficult given the diminished level of standardization. In contrast, performance-based assessments can be rigorously standardized, providing direct comparisons in achievement across students and over time; performance-based assessments, however, are difficult to individualize. To address these problems, some states have a hybrid approach using standardized portfolios, with standardized procedures for collecting and scoring evidence, but much of the individualized nature of the portfolio—that is, the characteristic that makes the test format so desirable—may be sacrificed (Gong & Marion, 2006).

In Quenemoen's (2008) brief history of alternate assessments based on alternate achievement standards (AA-AAS), policy reporting from the National Center on Educational Outcomes (NCEO) describes 15 years of research and development on alternate assessments. She addresses important characteristics of the student population, noting the need to focus on authentic skills integrated across domains, while continuously documenting outcomes. These issues reflect sensible attention to the inherent difficulty that exists in assessing students with significant cognitive disabilities: "There is more variability in the skill levels and needs of this 1% of the students than there is in the rest of the total student population" (Ysseldyke & Olsen, 1997, p. 16). This variation may come in part from the full range of disabilities usually included within the population taking AA-AAS.

Students eligible for the AA-AAS include only those who qualify for special education services through the Individuals With Disabilities Education Act (IDEA, 1997; Title I—Improving the Academic Achievement of the Disadvantaged, 2003), and have been recommended for participation by an individual education program (IEP) team in relation to state eligibility criteria. The three most common eligibility criteria include the need for (a) students to have a significant cognitive disability, (b) IEP teams to make the eligibility decision, and (c) substantial adjustments to the curriculum for ensuring access to the general education curriculum (Albus & Thurlow, 2012).

Although AA-AAS eligibility decisions do not generally focus on disability category, Towles-Reeves et al. (2012) report that students participating in the AA-AAS are generally identified as having an intellectual disability, autism, or multiple disabilities. Because of their unique abilities and needs, this population provides a measurement challenge for large-scale statewide testing programs, which are

[1]University of Oregon, Eugene, USA

Corresponding Author:
Daniel Anderson, University of Oregon, 175 Lokey Education, 5262 University of Oregon, Eugene, OR 97403-5262, USA.
Email: daniela@uoregon.edu

defined by standardized administration and scoring procedures (Thomas, 2005). While a great deal of policy and practice research has been conducted for AA-AAS (e.g., Albus & Thurlow, 2012; Cameto et al., 2009; Kearns, Towles-Reeves, Kleinert, Kleinert, & Kleine-Kracht, 2011), very little research has been conducted on the actual measures used in standardized testing programs.

The purpose of this article is to explore the accessibility and validity of one state's approach to a standardized performance-based AA-AAS. Two levels of individualization are embedded in this state's AA-AAS, both of which are intended to recognize the unique needs of students with significant cognitive disabilities. While these embedded levels of individualization provide flexibility in the assessment delivery, they also are structured so direct comparisons can be made across students and over time (see Gong & Marion, 2006).

In this state, students' access (e.g., level of independence) is assessed through a "pre-requisite skills" assessment prior to administration. The results of this quick assessment guide administrators in the types of supports he or she is allowed to provide during the administration of the content tasks (e.g., verbal, gestural, etc.). Students with low scores on the prerequisite skills task can be provided greater levels of support (e.g., partial-physical and full-physical), whereas students with high scores on the prerequisite task take the test independently or with minimal levels of support. This differentiation is consistent with many other states that routinely provide differing levels of support based on their level of independence (Cameto et al., 2009).

The test also can be administered in either of two formats: standard and scaffold. In the standard form, teachers follow a script with specific materials. The scaffold format contains the same prompts and scoring expectations as the standard version, but supplemental direction/redirection statements and visual supports focus students' attention on the test materials. These two test formats are designed to provide comparable estimates of student performance while reflecting the same underlying construct.

These two test format options—while potentially helping students access the test—also potentially threaten the validity of comparisons across students if scores are not invariant across test formats. As Geisinger (1994) notes, "one of the fundamental tenets of testing that has held until today is that such tests must be administered under standardized conditions" (p. 121). He expands on the meaning of standardization by referencing Anastasi (1988) to include "exact materials employed, times limits, oral instructions, preliminary demonstrations, ways of handling queries from test takers, and every other detail in the testing situation" (p. 25). Any variation is considered a potential threat to valid inferences, particularly construct misrepresentation (Haladyna & Downing, 2004). Standardization, therefore, is critical in not only test administration but also test planning and development (e.g., test specifications, item development, test design, assembly, and production). Given the variation in test delivery (standard and scaffold formats), the degree to which test scores from each test format are comparable must be empirically tested.

Because AA-AAS are administered individually, flexibility is built into the administration by definition: "The one-on-one nature of the administration, especially with the allowed task or prompting/scaffolding variability, makes it doubtful that truly standardized administration conditions exist" (Gong & Marion, 2006, p. 8). The authors also note that scoring is flexible and depends upon a particular task and "the degree of independence/assistance" (Gong & Marion, 2006, p. 8) associated with getting the student to respond. We focus on two aspects of flexibility in AA-AAS, as discussed by Gong and Marion (2006): (a) variation in the administration conditions, operationalized by the level of support provided by the administrator and (b) variation in the administration format, operationalized through the standard and scaffold test formats.

We first explore the extent to which the prerequisite skills task mediates the relation between students' specific disabilities and their latent content knowledge scores. Ideally, the strength of the relation between students' disability type and content knowledge scores can be reduced once the prerequisite skills task scores are included, indicating that the supports provided by the administrator help students access the test. We then test the invariance of the factor structure across the two formats (standard/scaffold) of the test administration, conditioning on students' level of independence. Thus, our research addresses two primary questions:

Research Question 1: To what extent do prerequisite skills mediate the relation between student disability and total test score?

Research Question 2: Is the alternate assessment factor structure invariant across test formats (standard/scaffold)?

For Research Question 1, evidence of mediation would suggest that the supports provided by the administrators are functioning as intended, providing access and allowing students to demonstrate their content knowledge net of their specific disabilities. Given that the population of students tested all have significant cognitive disabilities, it is unreasonable to expect that the supports would fully eliminate (i.e., fully mediate; see Baron & Kenny, 1986) the effect of their disabilities on their total test scores. However, the supports may reduce the effect in terms of magnitude (i.e., partially mediate), which would provide empirical evidence that the test functions as intended. Research Question 2 addresses the validity of inferences for comparisons across students: Poor model fit—or lack of invariance across test

formats—would suggest that the validity of inferences is threatened and a reconceptualization or revision of the test design may be necessary. By contrast, adequate model fit and test format invariance would provide empirical evidence for the validity of test-based inferences, and suggest that the standard and scaffold formats function equivalently.

Method

Measures

Data for this study came from one state's alternate assessment during the 2011-2012 school year. We limited our analysis to only data from the reading and mathematics measures in Grades 4 and 7. All assessments contained 40 operational items linked to the state's content standards, but reduced in depth, breadth, and complexity. Eight field test items in mathematics and 10 field test items in reading were also embedded within the test forms, but were not used for this study.

In reading, a series of testlets were used, and in mathematics, items were bundled into related (but separate) tasks, based on state achievement standards. Reading testlets consisted of a common stimulus followed by a set of five items (with three response options each) related to the stimulus. In math, tasks were organized by state achievement standards with each item containing a unique stimulus (and three response options). All testlets (in reading) and tasks (in math) contained five items addressing the state standards corresponding to the respective grade level and subject area. All items were scored with a partial credit model on a 0-2 scale, with 0 representing an incorrect response, 1 representing a partially correct response (as determined by the administrator), and 2 representing a fully correct response. Universal design components were embedded in all item and test structures in an effort to ensure that all students could access the assessments regardless of exceptionality (see Thompson, Johnstone, & Thurlow, 2002).

At any point during testing, administrators could use professional judgment to determine that the test had become too difficult for the student. In these instances, the test was discontinued and the student received a score of zero on all subsequent items for the purposes of accountability reporting. Students also may have refused to respond to an item, in which case the item was also scored 0 and the administrator continued testing. However, a sufficient number of items must be administered to meet the minimum participation rule prior to discontinuing the test. Students were required to take all of the prerequisite skills and attempt at least two complete content tasks to meet the state's minimum participation requirement. In our data set, we treated nonresponses as missing, rather than scoring them as 0 and assuming the student had no competence on the targeted construct. Missing values were handled during estimation with full-information maximum likelihood. Overall, the amount of missing data at the task level ranged from 0% to 17% across grades and subject areas, with a median of 12%.

All test administrators completed training and proficiency requirements to become certified as a "Qualified Assessor" prior to administering the test to students. As part of the certification process, test administrators were trained in providing different levels of support based on students' individual needs. The training was intended to ensure that supports were provided in a consistent manner without distorting the underlying construct.

Levels of independence. The level of independence score was determined by the prerequisite skills task that preceded all content prompts. During the administration of the prerequisite task, administrators worked with students to bring them to success (i.e., correctly respond to the item). Each item was scored based on the level of support the administrator needed to provide for the student to correctly respond to the item (0 = *Refusal/Inappropriate-Inaccessible*, 1 = *Full Physical Support*, 2 = *Partial Physical Support*, 3 = *Verbal/Gestural Support*, and 4 = *Independent*). Administrators also had the option of not administering any item and awarding full points if the student already had the skill being assessed. The level of independence score was determined by the mode of the item responses. The corresponding level of support then became the maximal allowable support the administrator could provide during the administration of the content tasks.

Prerequisite items denoted a range of performance, from responding to the administrator when cued (e.g., "Wave hi") to identifying content-related symbols and terms. For example, in mathematics, the assessor may prompt the student by asking, "Which is the minus sign?" or "Tell me or show me when I touch the minus sign." Students responded from three answer options. If the student initially responded incorrectly, the administrator continued working with and redirecting the student until he or she responded correctly. Students were rated not on the accuracy of their response, but on the level of support the administrator provided to bring the student to success. The goal of the prerequisite skills task was thus to determine the level of support necessary to bring the student to success, content and difficulty notwithstanding. As of 2009, the majority of states (76%) used level of independence as part of their approach to scoring AA-AAS (Cameto et al., 2009).

The prerequisite task total was not included in accountability reporting. Generally, students scored quite high on the prerequisite task. It was rare for students to receive a score of 0 and quite common to receive a perfect score. Descriptive statistics for the prerequisite skills task are reported for our sample in Table 1 by grade, subject, and test administration type. As can be seen, students receiving the standard administration format scored near perfect, on

Table 1. Descriptive Statistics by Disability Group.

	Grade 4				Grade 7			
	Standard		Scaffold		Standard		Scaffold	
Variable	*n*	Mean (*SD*)	*n*	Mean (*SD*)	*n*	Mean (*SD*)	*n*	Mean (*SD*)
Math								
Pre-req[a]								
ID	105	39.06 (1.42)	104	32.59 (9.64)	104	39.47 (1.43)	133	34.40 (8.45)
CD	130	39.11 (2.10)	19	38.68 (1.25)	35	39.74 (0.44)	7	39.14 (1.07)
OHI	106	39.60 (0.79)	44	34.25 (8.79)	62	39.68 (0.74)	23	31.65 (11.68)
ASD	96	38.58 (2.68)	133	31.81 (9.24)	59	39.19 (2.99)	76	31.53 (10.95)
Total Raw[a]								
ID	105	45.74 (12.81)	101	21.48 (18.55)	104	41.80 (13.42)	132	23.58 (18.86)
CD	129	53.60 (11.22)	19	42.79 (17.66)	35	46.20 (10.00)	7	34.86 (16.04)
OHI	106	53.87 (13.01)	44	27.43 (22.86)	62	45.84 (15.29)	23	20.61 (17.36)
ASD	96	47.73 (16.48)	133	21.34 (20.08)	59	42.75 (16.22)	73	21.37 (19.13)
Reading								
Pre-req[a]								
ID	102	38.35 (1.73)	102	32.52 (8.46)	107	38.93 (2.15)	124	33.23 (8.20)
CD	155	38.83 (2.07)	18	37.89 (2.83)	32	39.25 (1.22)	6	38.83 (0.75)
OHI	123	39.14 (1.53)	41	32.83 (8.99)	57	39.51 (1.00)	20	30.25 (12.28)
ASD	95	38.52 (2.06)	130	30.66 (9.46)	59	38.58 (4.16)	72	30.22 (10.87)
Total Raw[a]								
ID	101	59.62 (13.95)	90	38.86 (22.98)	107	64.24 (12.91)	109	39.79 (23.97)
CD	154	66.36 (10.18)	17	61.35 (15.99)	32	67.91 (6.76)	6	47.67 (19.32)
OHI	123	66.60 (12.57)	36	42.53 (22.74)	57	65.42 (12.58)	17	41.53 (26.07)
ASD	95	58.91 (15.15)	107	33.51 (25.39)	59	62.15 (13.64)	56	35.04 (24.34)

Note. Pre-req score was not included in students' content score; prerequisite skills scored on a 0 to 40 scale. Total Raw ranged from 0 to 80. ID = intellectual disability; CD = communication disorder; OHI = other health impairment; ASD = autism spectrum disorder.
[a]Prerequisite skills variable was centered around the mean prior to analysis.

average, across grades and subjects, while students receiving the scaffold version scored comparatively lower. The data suggested that, on average, students receiving the scaffold test format had lower levels of independence than students receiving the standard test format.

Standard and scaffold test formats. The standard and scaffold formats both included the same specific item prompts and response options. The scaffold format, however, contained additional verbal redirection, prompting, and occasionally, additional clarifying graphics. For example, the standard format of a math item might have asked the student, "Which bar on this graph shows how many apples the students picked?" The scaffold version would provide an additional introductory statement, or statements, to focus the student's attention on the test materials before the prompt is read, such as "Here is a graph with four bars showing the kinds of fruit that students picked (point to each bar)." The prompt would then be read to the student. Note that the parentheses of this prompt instruct the administrator to point to specific aspects of the test materials, highlighting important elements of the item. These additional supports were intended to help students access the test materials by focusing their attention on the appropriate stimuli, while not changing the underlying construct. Students were assigned to the scaffold or standard format based on IEP teams determining which format better met the student's unique needs. Determinations were based on criteria published by the state, which included an assessment of curricular modifications required for the student, content area skill levels, and generalized deficits resultant from the student's disability.

Participants

Participants in this study included a subsample of students taking the math and reading portions of the alternate assessment in the cooperating state in Grades 4 and 7. The state's AA-AAS eligibility criteria required that students have an IEP and were determined eligible for the test by an IEP team. These guidelines led to a total population of test takers that included many students with a learning disability, and a few other students with ostensibly higher functioning disabilities (e.g., emotional disturbance). Given that the

tests were designed for students with the most significant disabilities, however, we limited our analyses to a subsample of the total population of test takers. This subsample included only students whose *primary* disability was autism spectrum disorder, intellectual disability, communication disorder, or other health impairment. These students better represented the population of students for whom the test was originally designed, and all groups were of sufficient size for analysis.

It is important to note that student disability classifications for the cooperating state were limited to a single, primary disability (which did not include a "multiple disabilities" category). In addition, there were occasionally other student groups who aligned with the target population of test takers, but were simply not of sufficient size for analysis (e.g., Traumatic Brain Injury, Grade 4 $n = 6$). Because we were interested in the mediating effect of the prerequisite skills tasks for specific groups of students, the analysis had to be constrained to those groups who fit the target population and were of sufficient size.

Descriptive statistics for the sample used in all analyses are displayed by grade and testing condition in Table 1. Across disability groups, the Grade 4 sample size was 819, of which approximately 70% were male, 59% identified as White, and 35% identified as Multiethnic. The Grade 7 sample size was 521, of which approximately 64% were male, 64% identified as White, and 27% identified as Multiethnic. In both grades, the largest portion of Multiethnic students identified as White and Hispanic.

Analyses

Two families of models were tested to address the primary research questions. For Research Question 1 (access), we used a full structural equation model (SEM) that included a series of mediation models with student disability types predicting prerequisite skills total and latent content knowledge. For Research Question 2 (validity), we used a confirmatory factor analysis (CFA) model for invariance across test formats (standard and scaffold). In our sequence of analyses, we first fit a standard CFA model, as displayed in Figure 1. Following adequate model fit, we then tested the mediation model, as displayed in Figure 2. Finally, we tested the CFA model in Figure 1 for invariance. All analyses were conducted with the *Mplus* software (Muthén & Muthén, 2007) using maximum likelihood with robust standard errors (MLR) estimation, which is robust to departures from multivariate normality, and produces Satorra–Bentler scaled chi-square values.

CFA. The theoretical CFA model is displayed in Figure 1. The structure of the test was equivalent across grades and subject areas. That is, all tests targeted a single latent "content knowledge" factor, which was measured by a series of eight testlets or tasks, as described previously. In all models, the measurement weight for Testlet/Task 8 was fixed at 1.0 to identify the model. We first fit the model for the full sample of students in each grade and subject, then, provided adequate model fit, proceeded with further model testing (i.e., mediation or invariance testing). Across all models, we evaluated model fit with the comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean residual (SRMR) using the cutoff criteria outlined by Hu and Bentler (1999) of .95 and above, .06 and below, and .08 and below, respectively.

SEM disability mediation models. The full theoretical SEM mediation model for all grades and subject areas is displayed in Figure 2. Note that direct paths from the exogenous variables to the latent factor are displayed with hatched lines. This is because in testing for mediation, we followed a two-staged process, as recommended by Baron and Kenny (1986), who suggested first establishing all direct effects before testing for indirect effects with the theoretical mediating variable. In other words, the model was estimated twice: first with the prerequisite skills task excluded from the model and second with the full model displayed. Differences in the magnitude of the direct (hatched) paths were then evaluated between models. For the purposes of our study, complete or full mediation was defined as cases in which the direct effect became nonsignificant after accounting for the mediating variable. Partial mediation was defined as cases in which the direct effect remained a significant predictor in both models but was reduced in magnitude in the model including prerequisite skills *and* the indirect effect was significant. Indirect effects were tested for significance with standard errors obtained from the multivariate delta method (also known as the Sobel test). For all analyses, students with a communication disorder served as the reference group because they generally represented the highest performing group of students.

Invariance testing. The CFA model described above, and displayed in Figure 1, was tested for invariance across the standard and scaffold formats of the test. For basic CFA models, Kline (2013) suggests testing for invariance by first freely estimating the model for both groups, then placing a series of increasingly restrictive equality constraints on the models. Specifically, Kline suggests testing for configural invariance, followed by metric, scalar, and strict factorial invariance. We followed these recommendations, but included two additional constraints—structural weights and structural residuals—given that our theoretical CFA model included students' prerequisite skills total as a predictor of their latent content knowledge score. We tested for configural invariance by freely estimating the model for both groups and evaluating model fit. The configural model served as the baseline for subsequent comparisons. We then

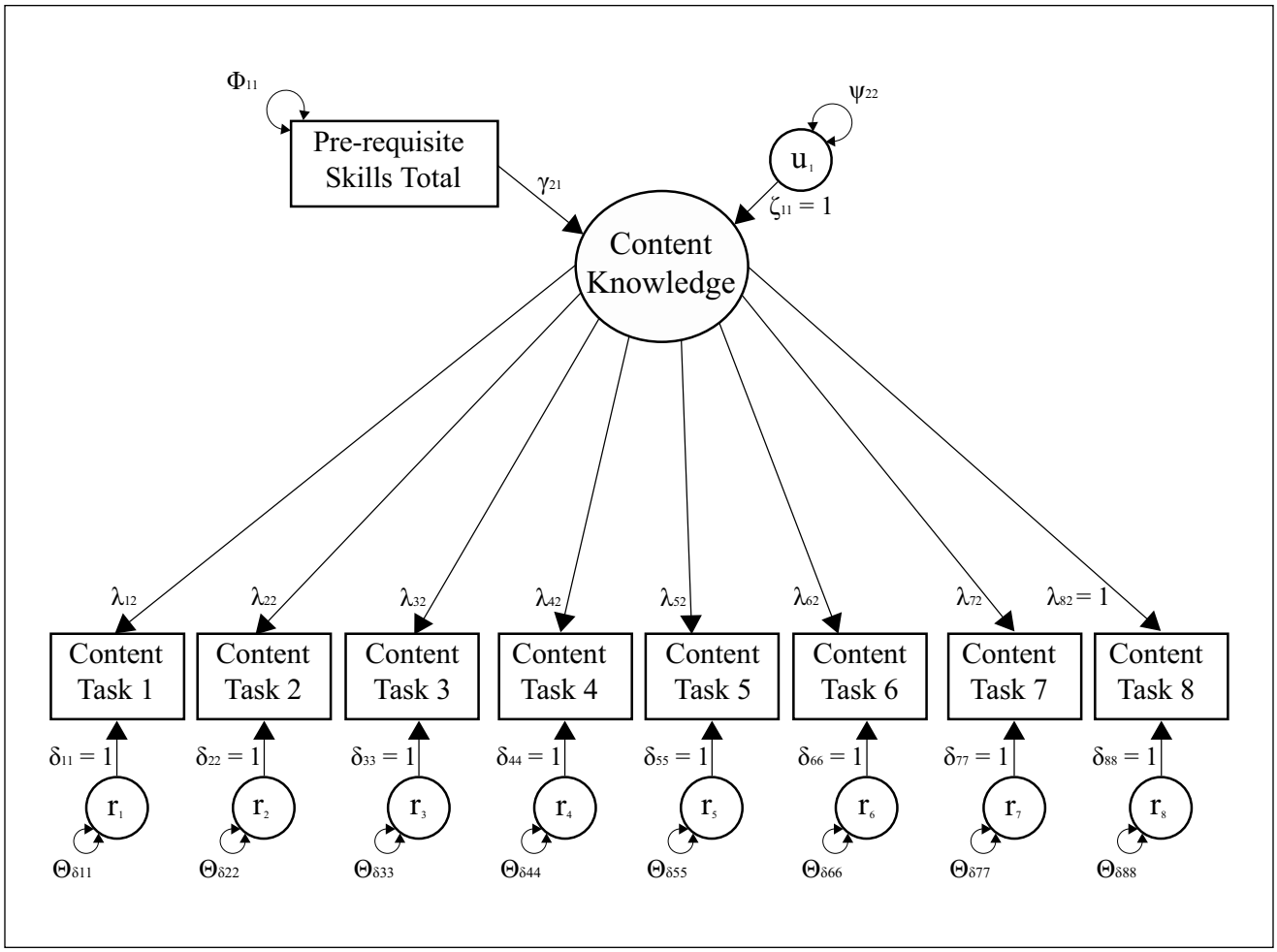

Figure 1. Theoretical confirmatory factor analysis model.

Note. Prerequisite skills total represented students score on a prerequisite task administered prior to content tasks. Students' median score on prerequisite items determined the level of support the administrator provided to the student during administration of the content tasks. Model was tested for invariance at Grades 4 and 7 in reading and math. The mathematics test contained tasks of items adressing related standards, while the reading test contained testlets (rather than tasks) based on a common stimulus. The overarching structure in reading and math tests, however, was equivalent, as displayed above.

constrained the measurement weights (λ) to be equal across groups to test for metric invariance, and constrained the intercepts of the factor indicators (τ) to be equal across groups to test for scalar invariance. Next, we tested for structural invariance by further constraining the model so the structural weights (γ) were equal, followed by the structural residuals (ζ). Finally, we tested for strict factorial invariance by fully constraining the model to be equal across groups by constraining the factor indicator residuals (Θ_{δ}).

When testing for structural invariance, we did not constrain the intercept, mean, or variance of the prerequisite skills total to be equal across groups because we would not theoretically expect them to be equal, given the test administration procedures. As previously mentioned, the administrator can award full points to any student who already has the skill being assessed. Generally, students provided the standard format of the test are higher functioning than those who participate in the scaffold format, and received full credit across all or the majority of items more often (see Table 1). We would therefore not expect—nor do the administration procedures intend—for the groups to have equivalent intercepts, means, or variances in their prerequisite skill scores. Thus, we allowed the parameters to be freely estimated at all times given that, theoretically, they should not be equivalent, and nonequivalent parameters do not threaten the validity of test-based inferences.

We also tested for differences in model fit with the chi-square difference test ($\chi^2\Delta$). However, because we used the Satorra–Bentler scaled chi-square, the values were not directly comparable across the nested models (Satorra, 1999). Thus, chi-square difference tests were calculated

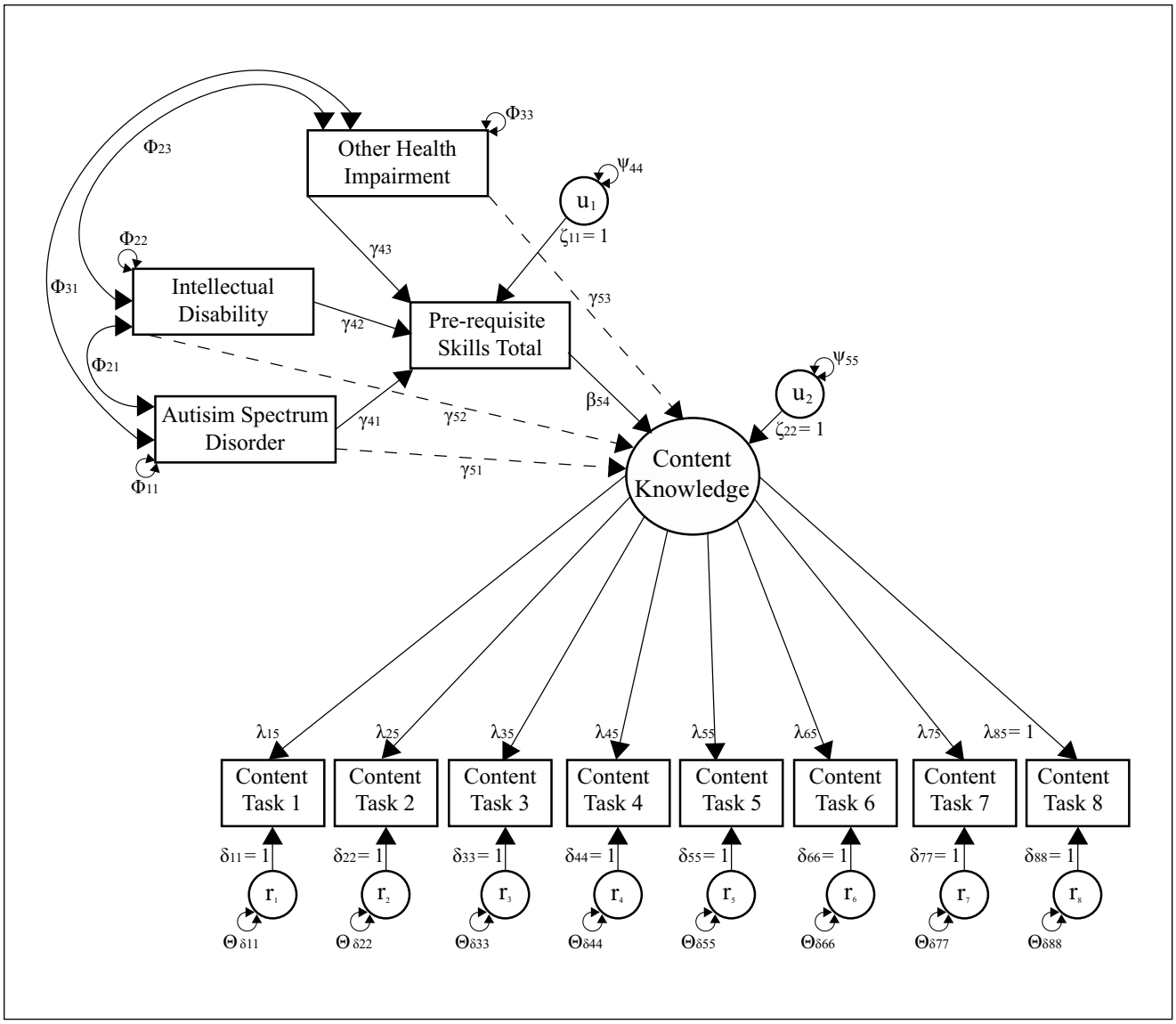

Figure 2. Theoretical structural equation model with disability mediation models.
Note. Model evaluated the extent to which the prerequisite skills total mediated the relationship between students disability and their latent content knowledge score. Differences in the magnitude of direct effects (hatched lines) were evaluated with the prerequisite skills total variable included, versus excluded. Measurement model was equivalent to the model displayed in Figure 1. Students with a communication disorder served as the reference group.

after first applying a model-based scaling correction factor (see Muthén & Muthén, n.d.). However, similar to critiques of the chi-square test with standard model evaluation (see Bentler & Bonett, 1980), chi-square difference tests are sensitive to sample size, with larger samples leading to a higher likelihood of detecting significant differences in model fit (Kline, 2013). Cheung and Rensvold (2002) thus recommend evaluating differences in CFI, which is less influenced by sample size, and recommend rejecting model equivalence when CFI differences exceed .01. Given our relatively large sample size, we used differences in CFI (CFIΔ) as our *primary* indicator of model equivalence, but used a family of model fit indicators to determine model adequacy. Inadequate model fit at any step in the process would suggest the structure of the test was no longer equivalent across groups.

According to Kline (2013), models with configural, metric, and scalar invariance are referred to as having *strong* factorial invariance. However, only models that still fit the data well after having the factor indicator residual terms constrained to be equal across groups should be referred to as having *strict* factorial invariance.

Table 2. Parameter Estimates for Freely Estimated Models.

Parameters	Grade 4		Grade 7	
	Reading	Math	Reading	Math
Weights[a]				
Task 1, λ_{12}	.88 (.01)	.77 (.02)	.83 (.02)	.74 (.02)
Task 2, λ_{22}	.82 (.02)	.65 (.02)	.91 (.01)	.67 (.03)
Task 3, λ_{32}	.91 (.01)	.77 (.02)	.92 (.01)	.81 (.02)
Task 4, λ_{42}	.89 (.01)	.84 (.02)	.89 (.02)	.78 (.03)
Task 5, λ_{52}	.91 (.01)	.75 (.02)	.93 (.01)	.83 (.02)
Task 6, λ_{62}	.91 (.01)	.84 (.02)	.91 (.01)	.64 (.04)
Task 7, λ_{72}	.91 (.01)	.82 (.02)	.91 (.01)	.74 (.03)
Task 8, λ_{82}	.88 (.01)	.77 (.02)	.88 (.02)	.81 (.02)
Pre-req, γ_{21}	.81 (.02)	.68 (.03)	.90 (.03)	.73 (.03)
Intercepts[b]				
Task 1, τ_1	7.44 (.08)	5.59 (.09)	5.83 (.13)	4.64 (.12)
Task 2, τ_2	7.14 (.09)	4.36 (.10)	6.72 (.12)	4.35 (.11)
Task 3, τ_3	7.52 (.08)	4.95 (.10)	7.42 (.12)	5.17 (.12)
Task 4, τ_4	7.41 (.09)	5.92 (.11)	6.72 (.12)	4.08 (.14)
Task 5, τ_5	6.73 (.10)	4.38 (.11)	6.81 (.12)	5.03 (.14)
Task 6, τ_6	6.87 (.10)	6.26 (.11)	7.02 (.13)	3.27 (.12)
Task 7, τ_7	6.16 (.10)	5.40 (.11)	6.63 (.13)	3.87 (.12)
Task 8, τ_8	6.46 (.09)	4.77 (.12)	6.10 (.12)	4.53 (.13)
Residual variances[a]				
r1, $\Theta_{\delta 11}$	.22 (.02)	.40 (.03)	.31 (.03)	.46 (.03)
r2, $\Theta_{\delta 22}$	.33 (.03)	.57 (.03)	.17 (.02)	.55 (.04)
r3, $\Theta_{\delta 33}$	.17 (.02)	.41 (.03)	.15 (.02)	.34 (.03)
r4, $\Theta_{\delta 44}$	.21 (.02)	.29 (.02)	.22 (.03)	.40 (.04)
r5, $\Theta_{\delta 55}$	.17 (.02)	.43 (.03)	.14 (.02)	.32 (.03)
r6, $\Theta_{\delta 66}$	.18 (.02)	.29 (.03)	.17 (.02)	.59 (.05)
r7, $\Theta_{\delta 77}$	.31 (.03)	.32 (.03)	.17 (.02)	.45 (.04)
r8, $\Theta_{\delta 88}$	.22 (.02)	.41 (.03)	.23 (.03)	.35 (.04)
u1, Ψ_{22}	.35 (.03)	.54 (.05)	.37 (.05)	.47 (.04)

Note. Model results for full model, not separated by test administration type. Standard errors are displayed in parentheses. Tasks scored on a 0 to 10 scale. All parameters are significant at the $p < .05$ level. Fit statistics reported in Tables 4 and 5.
[a]Reported in standardized form. [b]Reported in unstandardized form.

Results

Prior to conducting analyses addressing our research questions, we first fit a standard CFA model for the full sample of students in each grade, as displayed in Figure 1. The purpose was to first establish that the theoretical measurement model fit the data before including predictors of the latent factor (i.e., testing for mediation) or constraining parameters to be equal across student groups (i.e., testing for invariance).

Parameter estimates from these analyses are displayed in Table 2. All model fit statistics are presented as "full sample" along with the models tested for invariance (displayed for Grade 4 in Table 4 and Grade 7 in Table 5). Across both subjects and grades, the data fit the models quite well. The Satorra–Bentler chi-square statistic was universally significant but was likely due to the large sample size, ranging from 477 to 767. All other fit indices met a priori cutoff criteria, with CFI approximately .99 across models, RMSEA in the .03 to .04 range, and SRMR ranging from .02 to .03.

For the remainder of this section, we present results from our primary analyses addressing our two research questions. First, we present the SEM mediation model in which student prerequisite skills total (levels of independence) served as a mediator between disability and the latent content knowledge factor score. Second, we present two tables that summarize the results from our invariance testing of the scaffold and standard test administrations.

Table 3. Standardized Parameter Estimates for the Mediation Model With All Direct and Indirect Effects.

Parameters	Grade 4				Grade 7			
	Reading		Math		Reading		Math	
Direct								
ASD, γ_{41}	−.21	(.03)	−.21	(.03)	−.14	(.04)	−.16	(.05)
ID, γ_{42}	−.18	(.03)	−.28	(.03)	−.17	(.02)	−.19	(.05)
OHI, γ_{43}	−.05	(.02)	−.08	(.03)	−.01†	(.03)	−.04†	(.04)
Indirect								
ASD, $\gamma_{41} \times \beta_{54}$	−.25	(.03)	−.19	(.02)	−.24	(.03)	−.20	(.03)
ID, $\gamma_{42} \times \beta_{54}$	−.17	(.02)	−.13	(.02)	−.11	(.04)	−.14	(.02)
OHI, $\gamma_{43} \times \beta_{54}$	−.06	(.02)	−.04	(.02)	−.08	(.03)	−.08	(.03)

Note. Students with a communication disorder served as the reference group. Standard errors displayed in parentheses. In the model not including prerequisite skills, conducted prior to testing for mediation, the direct effect for Grade 7 OHI in reading was not significantly different from students with a communication disorder. All other direct effects were significant in the model not including prerequisite skills. ASD = autism spectrum disorder; ID = intellectual disability; OHI = other health impairment.
†$p > .05$, all other values significant at $p < .05$.

SEM Mediation Model

The model displayed in Figure 2 was fit for both grades and subject areas to test the extent to which the prerequisite skills task mediated the relation between disability and latent content knowledge score. The chi-square test of model fit was significant across all models, suggesting lack of fit. However, all other fit indices suggested the data fit the model quite well, with CFI approximately .99 for all models, RMSEA ranging from .02 to .04, and SRMR ranging from .02 to .03. The significant chi-square value was thus, again, likely due to the large sample size. Estimates of direct and indirect effects are displayed in Table 3, with students with a communication disorder serving as the reference group. The disability variables accounted for approximately 3% to 8% of the variance in the prerequisite skills total, depending on the specific model tested.

Table 4. Invariance Testing: Model Fit: Grade 4.

Model	S-B χ^2	*df*	$\chi^2\Delta$	*df*Δ	CFI	RMSEA	SRMR
Math (n = 737)[a]							
Full sample	65.11*	27	—	—	.99	.04	.03
Configural	97.60*	54	—	—	.98	.05	.03
Metric (λ, constrained)	120.00*	61	24.03*	7	.97	.05	.05
Scalar (λ, τ, constrained)	151.73*	68	30.27*	7	.95	.06	.07
Structural weight (λ, τ, γ, constrained)	172.87*	69	8.65*	1	.94	.06	.09
Reading (n = 767)[b]							
Full sample	61.27*	27	—	—	.99	.04	.02
Configural	81.14*	54	—	—	.99	.04	.02
Metric (λ, constrained)	118.85*	61	41.20*	7	.98	.05	.07
Scalar (λ, τ, constrained)	149.44*	68	30.99*	7	.97	.06	.09
Structural weight (λ, τ, γ, constrained)	149.70*	69	1.25	1	.97	.06	.09

Note. Chi-square difference tests were computed with a scaling correction factor applied. Each $\chi^2\Delta$ value represents a comparison between the corresponding model and the preceding model. Models displayed in bold represent the final model meeting invariance fit criteria. CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean residual; S-B χ^2 = Satorra–Bentler scaled chi-square, which is not directly comparable across models.
[a]437 Standard, 300 Scaffold. [b]475 Standard, 292 Scaffold.
*$p < .05$.

In testing for mediation, the model was first analyzed with the prerequisite skills variable excluded from the model. Direct paths from disability to latent content knowledge were then evaluated. All paths were significant, with the exception of Grade 7 reading for students with other health impairments, suggesting these students did not function significantly different from students with a communication disorder (reference group). The model was then reanalyzed with the prerequisite skills variable included and indirect paths specified, which resulted in the direct paths on content knowledge being, universally, reduced in magnitude. All indirect effects were also significant.

For Grade 4, the reduction in magnitude of the direct effects on the latent content knowledge factor, and the significant indirect effects, suggested partial mediation in both reading and math for all disability groups. For Grade 7, the prerequisite skills task also partially mediated the relation for students with autism spectrum disorder or an intellectual disability. For students with other health impairments, however, there was no evidence of mediation in reading, given that the direct effect on content knowledge was not significant in the model without prerequisite skills. Contrarily, there was evidence of *full* mediation in math because the indirect path was significant, and the direct effect on content knowledge became nonsignificant after including prerequisite skills in the model. Overall, the reduction in magnitude for the direct effects on content knowledge was most pronounced for students with autism spectrum disorder or an intellectual disability—generally the two student groups requiring the most intensive supports.

Invariance Testing

Given that all CFA models fit the data well with the full sample of students, as displayed in Figure 1, we began testing for invariance across testing conditions (standard vs. scaffold). All model fit statistics for our sequence of models are displayed in Table 4 for Grade 4 and Table 5 for Grade 7. We first estimated the same CFA model displayed in Figure 1 simultaneously for students taking the standard and scaffold version of the test (i.e., configural invariance), which served as the baseline model for all comparisons. The model resulted in satisfactory model fit across both grades and subject areas, with CFI ranging from .98 to .99, RMSEA from .03 to .05, and SRMR from .02 to .04.

Following adequate configural invariance, we began constraining model parameters to be equal across groups, beginning with the measurement weights (i.e., metric invariance). For Grade 7 math, constraining the measurement weights to be equal resulted in a nonsignificant difference in model fit, as indicated by the scaling corrected chi-square difference test. Across all other models the chi-square difference test was significant, suggesting the model was no longer equivalent. However, the difference in CFI was .01 across all models, meeting the Cheung and Rensvold (2002) criteria for adequate model fit with large samples. Furthermore, all other fit statistics remained adequate, with RMSEA ranging from .04 to .05 and SRMR ranging from .05 to .07.

We then tested for scalar invariance by constraining the intercepts of the factor indicators to be equal across groups, which resulted in a significantly worse fitting model, as

Table 5. Invariance Testing: Model Fit: Grade 7.

Model	S-Bχ^2	*df*	$\chi^2\Delta$	*df*Δ	CFI	RMSEA	SRMR
Math (n = 499)[a]							
Full sample	41.53*	27	—	—	.99	.03	.02
Configural	74.05*	54	—	—	.98	.04	.04
Metric (λ, constrained)	87.62*	61	14.04	7	.98	.04	.06
Scalar (λ, τ, constrained)	102.33*	68	14.16*	7	.97	.05	.07
Structural weight (λ, τ, γ, constrained)	105.94*	69	4.18*	1	.97	.05	.08
Structural residual (λ, τ, γ, ζ, constrained)	108.02*	70	1.98	1	.97	.05	.07
Strict (λ, τ, γ, ζ, Θ_δ, constrained)	118.66*	78	10.47	8	.97	.05	.08
Reading (n = 477)							
Full sample	52.22*	27	—	—	.99	.04	.02
Configural	66.45*	54	—	—	.99	.03	.03
Metric (λ, constrained)	90.20*	61	24.35*	7	.98	.05	.07
Scalar (λ, τ, constrained)	100.46*	68	10.24	7	.98	.05	.07
Structural weight (λ, τ, γ, constrained)	98.65*	69	0.04	1	.98	.04	.07
Structural residual (λ, τ, γ, ζ, constrained)	159.74*	70	39.67*	1	.95	.07	.27

Note. Chi-square difference tests were computed with a scaling correction factor applied. Each $\chi^2\Delta$ value represents a comparison between the corresponding model and the preceding model. Models displayed in bold represent the final model meeting invariance fit criteria; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean residual; S-B χ^2 = Satorra–Bentler scaled chi-square, which is not directly comparable across models.
[a]260 Standard, 239 Scaffold. [b]255 Standard, 222 Scaffold.
*$p < .05$.

indicated by the scaling corrected chi-square difference test, for all groups with the exception of reading in Grade 7. For Grade 4, both models in reading and math fell outside of the a priori fit criteria, suggesting the model was no longer equivalent across groups (CFIΔ ≥ .02). The Grade 4 models thus displayed metric, but not scalar invariance. The Grade 7 models both met the a priori CFI difference criteria.

When constraining the structural weight, the scaling corrected chi-square difference test was significant for Grade 7 math, but not Grade 7 reading. The difference in CFI for Grade 7 math, however, remained at .01 and, again, all other fit statistics met a priori fit criteria.

Thus, at Grade 7, the "structural" models fit the data similar to the "scalar" models. Constraining the structural residuals resulted in a significant scaling corrected chi-square difference test for Grade 7 reading, but not for Grade 7 math. The difference in CFI exceeded .01 for Grade 7 reading, but remained at .01 for Grade 7 math. Thus, the Grade 7 reading model with the structural residual constrained was no longer equivalent to the configural model. Constraining the measurement residuals for Grade 7 math again did not result in a significantly worse fitting model, and the difference in CFI remained at .01. Furthermore, all other fit statistics met a priori cutoff criteria.

Overall, the results were mixed. At Grade 4, the models for reading and math only met the metric invariance criteria, suggesting that only the measurement weights were equivalent between test formats. At Grade 7, the results were quite a bit better, with the model for math displaying strict factorial invariance, and the model for reading displaying structural invariance (although not structural residual invariance). However, note that in Tables 4 and 5, we report statistics from further invariance testing, even after lack of invariance was found. In Grade 4 math, CFI, RMSEA, and SRMR all meet a priori fit criteria for adequate model fit up through scalar invariance. Thus, while the model was statistically different from the configural model (CFIΔ = .03), it still fit the data quite well, and the statistics were reasonable through the structural weight step. Similarly, the fit statistics for the model in Grade 4 reading were quite reasonable through the structural weight step, despite being statistically different from the configural model (CFIΔ = .02). Although *statistically* different, then, the *practical* difference in inferences based on test scores for students under each format are likely minimal. However, outside of Grade 7 math, it would be difficult to argue that any model was invariant beyond the structural weight step, even from a practical perspective.

Discussion

The two major findings from our analyses were that (a) the prerequisite skills task nearly universally functioned as a mediator of the relationship between students' disabilities and their latent content knowledge score and (b) the factor structure was equivalent for both test administration formats to varying degrees (depending on grade/subject area). Both analyses provide a degree of empirical support for the

test design used in this state. For the mediation models, the analyses provided empirical evidence that the supports provided to students based on their level of independence helped students access the test. For the invariance testing, the measurement weights were equivalent in Grade 4, while the Grade 7 models met Kline's (2013) criteria for *strong* and *strict* factorial invariance for reading and math, respectively. Future research should investigate the models for Grade 4 in more depth to better understand why they did not meet the strong factorial invariance criteria.

Role of Standardization

Clearly, standardization is a significant issue in large-scale statewide testing programs. Although standardized tests were developed in this state, two accessibility supports, consistent with the principles of universal design, were built into the manner in which administrators interacted with students, both of which have been noted by Gong and Marion (2006). First, the items were administered differently with various levels of assistance provided so students could access the items. Second, items were designed to have various prompts embedded so students could display their content skill. These supports, however, had not previously been empirically validated, despite widespread implementation in the field (Cameto et al., 2009). Our results provide initial confirmatory evidence that the supports function as intended in helping students with significant cognitive disabilities access standardized performance-based alternate assessments without interfering with the intended construct.

When teachers provide supports to students, they become part of the testing routine. As Gerber and Semmel (1984) argue, the teacher *is* the test. However, the variation in test administration (through different levels of support) was designed to reduce, if not remove, construct-irrelevant variance (Haladyna & Downing, 2004). In addition, implementation of these supports has been shown to not compromise the reliability of the test. In a study by Tindal, Yovanoff, and Gellar (2010), generalizability theory analyses suggested that the test administrator was a negligible facet in the reliability of the test. These results and the Tindal et al. results suggest that well-defined administrator-provided supports in standardized, performance-based alternate assessments can help students access the test while not compromising reliability and the validity of test-based inferences.

Limitations of the Study

It is important to note that all analyses in the current study were conducted under a basic fidelity of implementation assumption relative to the prerequisite skills total score. That is, we assumed that a student's prerequisite skills total score was a reasonably valid indicator of the support the student received during the assessment. Theoretically, administrators used the results of the prerequisite skills total to determine the amount of support they provide the student during administration of the content tasks. However, whether the appropriate level of support was provided, given the prerequisite skill total, was unknown. As noted earlier, however, teachers went through a rigorous training and certification process annually to become a "Qualified Assessor" or recertify before they were allowed to administer the assessment. Thus, the fidelity of implementation assumption was likely borne out in the majority of cases.

As mentioned in the participant portion of the methods section, a substantial number of students with learning disabilities were part of the original AA-AAS dataset used here. Although categorical exclusion from AA-AAS of any IDEA subgroup is generally avoided by most states, it is clearly the intent of the No Child Left Behind Act (2001) to provide access to statewide accountability systems for students with the most significant cognitive disabilities, including, but not limited to, intellectual disabilities, autism, and multiple disabilities. All students with a learning disability were excluded from the analyses reported here, to be consistent with the population for which the test was designed. It is also important to note that the research reported here investigated tests used in reading and math in Grades 4 and 7. Whether the results generalize to other grades and/or subject areas is unknown.

Finally, the invariance testing approach we took was an "all or nothing" approach, by which the model either passed or failed the invariance test. However, it is quite possible that the Grade 4 models had *partial* scalar invariance, which was not investigated here. In other words, the model may have displayed scalar invariance if all but a single intercept, or measurement weight was constrained. In this case, the single discrepant parameter would have caused the model as a whole to display lack of invariance, despite the majority of parameters functioning equivalently between groups. Future research should investigate partial invariance for both Grade 4 measures by systematically fixing and freeing each measurement weight in isolation, as well as each intercept in isolation, to identify the specific parameter or parameters causing the lack of invariance. Were the specific parameters identified, revisions to the tasks or testlets could be made so the model functioned equivalently between groups.

Conclusion—Looking Forward

Within accountability contexts, students' academic growth has become an increasingly present component of the overall metric used to judge the "effectiveness" of the school and, in some cases, the teacher (Buzick & Laitusis, 2010). Yet, discussions of how students with significant cognitive

disabilities fit within these systems have been largely absent. Measuring growth requires a common scale across time, which essentially eliminates the possibility of personalized portfolio-type indicators of student achievement. While it may be possible to construct a common portfolio scale with standardized procedures for collecting and scoring evidence, much of the personalized nature of the portfolio would be sacrificed. Standardized performance-based assessments, though also potentially low in personalization, can be high in efficiency, and a common scale can readily be constructed for modeling growth.

To be clear, our purpose was not to model growth with alternate assessment data, but rather to provide evidence on the validity and accessibility of a test format readily amenable to standard item response theory (IRT) scaling and equating practices for the construction of a common, vertical scale—a prerequisite for modeling growth (Rogosa, Brandt, & Zimowski, 1982). However, some students may not show growth if they cannot access the test and display their content proficiency. Ideally, evidence from this cross-sectional study can serve as an initial step to longitudinal studies that investigate the validity of score changes over time.

As the field shifts from local statewide assessments toward AA-AAS designed by national consortia (see Dynamic Learning Maps, 2010; National Center and State Collaborative, 2010), we must be careful to ensure that flexibility is built into the test design in a manner that does not compromise reliability and validity. States moving toward growth models based on AA-AAS data must seriously consider the impact of test design and standardization. The test formats selected must be amenable to scaling and equating procedures, while items and test structures must be invariant across student subgroups and testing conditions.

Acknowledgment

The authors would like to thank Dr. Joseph F. T. Nese for his consultation during the preparation of this manuscript.

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) received no financial support for the research, authorship, and/or publication of this article.

References

Albus, D., & Thurlow, M. (2012). *Alternate assessments based on alternate achievement standards (AA-AAS) participation policies*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Altman, J. R., Lazarus, S. S., Quenemoen, R. F., Kearns, J., Quenemoen, M., & Thurlow, M. L. (2010). *2009 survey of states: Accomplishments and new issues at the end of a decade of change*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York, NY: Macmillan.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182. doi:10.1037/0022-3514.51.6.1173

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606. doi:10.1037/0033-2909.88.3.588

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.

Cameto, R., Knokey, A.-M., Nagle, K., Sanford, C., Blackorby, J., Sinclair, B., & Riley, D. (2009). *National profile on alternate assessments based on alternate achievement standards. A report from the national study on alternate assessments* (NCSER 2009-3014). Menlo Park, CA: SRI International.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255. doi:10.1207/S15328007SEM0902_5

Dynamic Learning Maps Alternate Assessment System Consortium. (2010). *IDEA General Supervision Enhancement Grant—Alternate Academic Achievement Standards*. U.S. Department of Education.

Geisinger, K. F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education*, *7*, 121–140. doi:10.1207/s15324818ame0702_2

Gerber, M. M., & Semmel, M. I. (1984). Teacher as imperfect test: Reconceptualizing the referral process. *Educational Psychologist, 19*, 1–12. doi:10.1080/00461528409529290

Gong, B., & Marion, S. (2006). *Dealing with flexibility in assessments for students with significant cognitive disabilities* (Synthesis Report 60). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Haladyna, T. M., & Downing, S. M. (2004). Construct irrelevant variance in high stakes testing. *Educational Measurement, Issues and Practice*, *23*(1), 17–27.

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. 105–17 (1997).

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

Kline, R. B. (2013). Assessing statistical aspects of test fairness in structural equation modeling. *Educational Research and Evaluation*, *19*, 204–222. doi:10.1080/13803611.2013.767624.

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

Muthén, L. K., & Muthén, B. O. (n.d.). *Chi-square difference testing using the Satorra-Bentler scaled chi-square*. Retrieved from http://www.statmodel.com/chidiff.shtml

National Center and State Collaborative. (2010). *National Center and State Collaborative General Supervision Enhancement Grant (NCSC GSEG)*. Washington, DC: U.S. Department of Education.

The No Child Left Behind Act, Pub. L. No. 107–110 (2001).

Quenemoen, R. (2008). *A brief history of alternate assessments based on alternate achievement standards* (Synthesis Report 68). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Rogosa, D., 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

Satorra, A. (1999, July). *Scaled and adjusted restricted tests in multi-sample analysis of moment structures* (UPF Economic & Business Working Paper No. 395). Retrieved from http://www.econ.upf.edu/docs/papers/downloads/395.pdf

Thomas, F. M. (2005). *High stakes testing: Coping with collateral damage*. Mahwah, NJ: Lawrence Erlbaum.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). Universal design applied to large scale assessments (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Tindal, G., Yovanoff, P., & Geller, J. P. (2010). Generalizability theory applied to reading assessments for students with significant cognitive disabilities. *Journal of Special Education, 44*, 3–17. doi:10.1177/0022466908323008

Title I—Improving the Academic Achievement of the Disadvantaged. (2003). *Final Rule, 34 CFR Part 200* (*December 2003*). Retrieved from http://www2.ed.gov/legislation/FedRegister/finrule/2003-4/120903a.pdf

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

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, MN: University of Minnesota, National Center on Educational Outcomes.