Examining the Validity and Fairness of a State Standards-Based Assessment of English-Language Arts for Deaf or Hard of Hearing Students

Jonathan Steinberg
Frederick Cline
Guangming Ling
Linda Cook
Namrata Tognatta

Educational Testing Service
Princeton, NJ

Abstract

This study examines the appropriateness of a large-scale state standards-based English-Language Arts (ELA) assessment for students who are deaf or hard of hearing by comparing the internal test structures for these students to students without disabilities. The Grade 4 and 8 ELA assessments were analyzed via a series of parcel-level exploratory and confirmatory factor analyses, where both groups were further split based on English language learner (ELL) status. Differential item functioning (DIF) analyses were also conducted for these groups of students, and where sample sizes were sufficient, the groups were additionally split based on test accommodation status. Results showed similar factor structures across the groups of students studied and minimal DIF, which could be interpreted as lending support for aggregating scores for Annual Yearly Progress (AYP) purposes from students who are deaf or hard of hearing.

Examining the Validity and Fairness of a State Standards-Based Assessment of English-Language Arts for Deaf or Hard of Hearing Students

Introduction

Over the last decade several federal laws have been passed to ensure that students with disabilities in the K-12 system receive appropriate accommodations during standardized testing (IDEA, 1997), and that all students between Grades 3-8, regardless of disability, be tested to measure annual yearly progress (AYP) in the areas of reading and math (NCLB, 2002). These mandates have led to a greater emphasis in studying the appropriateness of standardized tests and the accommodations provided when administered to students with disabilities. Elbaum, Arguelles, Campbell, and Saleh (2004) state that, "…the increasing participation of students with disabilities in statewide assessments has stimulated considerable research and discussion concerning the appropriate assignment of testing accommodations, the impact of accommodations on test performance of students with and without disabilities, and the validity of interpretations of test performance when students are awarded particular accommodations."

However, analyzing the appropriateness of an assessment when administered to students with disabilities can be a complex task. The main issue is whether the reported test scores have the same meaning for students with disabilities as they do for students without disabilities. Simply comparing results such as average test scores or percent correct on individual items cannot separate the appropriateness of the test for the population from the population's actual achievement level. More sophisticated statistical procedures are available but the low incidence of many disabilities means that sufficient numbers of test takers may not be available to produce results that are stable and statistically valid. Sample sizes can be further reduced when the lack of homogeneity of students within a specific disability category is taken into account, for

instance the severity of the disability, the language background, and the variety and impact of appropriate test accommodations.

Nevertheless, studying the characteristics of tests when given to students with disabilities is needed as the underlying assumption of placing the same expectations on all students based on a single standardized test is that inferences made from the test scores are equally valid for, and fair to, all groups. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) chapter on Fairness in Testing and Test Use (Chapter 7) points out that while there are many interpretations of test fairness, the core issue is for there to be no systematic differences in test-taker performance specifically due to group membership. Any difference in performance caused by group membership alone could represent construct-irrelevant variance (Messick, 1989), meaning the test results could be impacted by factors other than an examinee's ability to answer the questions. The *Standards* make explicit that this concept of fairness is the goal when testing students with disabilities as well. Standard 10.1 states: "In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement."

Test accommodations are one of the methods used to ensure compliance with *Standard 10.1* (*Standards*, p. 106). Accommodations change the conditions under which the test is administered to minimize any difficulty an examinee may have in interacting with the test due to a disability in order to accurately measure their performance, in effect removing construct-irrelevant variance while maintaining construct representation. More extreme changes to the test conditions that also are considered to change the construct being measured are considered to be modifications (Thurlow & Wiener, 2000).

This study examines the appropriateness of a large-scale state standards-based English-Language Arts (ELA) assessment for students who are deaf or hard of hearing by comparing the results for these students to students without disabilities, based on the assumption that if the internal structure of the exam is consistent across non-disabled and disabled groups, the resulting scores are equally fair and valid for these groups. A relatively large sample of non-disabled and deaf or hard of hearing students, further sub-divided based on English language learner (ELL) status, allowed us to study the internal structure of the Grade 4 and 8 ELA assessments via a series of parcel-level exploratory and confirmatory factor analyses as well as differential item functioning (DIF) analysis. Factor analysis based on whether the test was taken with or without accommodations was not possible as the majority of the deaf or hard of hearing students took the test without accommodations, and sample sizes for this kind of analysis were too small. However, in some instances sample sizes did permit DIF analyses based on accommodation status.

DIF analysis identifies items that are much harder or easier for a subgroup of examinees when compared to a reference group of examinees of the same ability. By matching on ability, with total test score commonly used as a proxy for ability, differences in performance on specific items are isolated apart from any overall differences in ability between the groups. The assumption is that items should be equally difficult for students of the same overall ability regardless of group membership; if not, some aspect of the item may be interacting with group membership to make it less appropriate for that group. Most large-scale testing programs use DIF procedures on items during both pre-testing and scoring to identify items that are differentially difficult based on race, sex and sometimes native language. Items identified during pre-testing are reviewed and are either revised or not included in an operational test form. Items

identified after testing are reviewed and may be excluded from the calculations of the reported score. Due to small sample sizes it is uncommon to perform DIF analysis on items based on disability status.

Factor analysis, when used on test data, attempts to identify the relative cohesiveness of individual items with each other in forming a stable underlying construct, or factor. Often it is simply used to determine whether all test items are part of the same general factor, indicating the test is unidimensional, or if the test reports subscores, that the subset of items used to compute those subscores group together appropriately on unique factors. When used to compare the internal structure of a test across groups, the first step is sometimes an exploratory factor analysis that determines the number of possible factors found in the test. Such analyses are generally undertaken if there are no a priori hypotheses about the number of factors that exist in the data, or as a check that a factor analytic model is appropriate if sample sizes for some groups are small. It would be expected that a test would measure the same number of factors regardless of the group membership of the examinees; if not, then it is possible the test is measuring one set of constructs for one group and a different set of constructs for another. If the same number of factors is found across groups, the second step is confirmatory factor analysis which tests if the number of factors is the same and if the composition of the factors for each group are statistically similar; if not, the nature of the factors may not be comparable even though these may be similar in number. This may often be true when working with small populations, when differences due to chance are more likely.

By focusing on students who are deaf or hard of hearing, this study extends work previously carried out on the internal structure of the same English-Language Arts assessment for students with learning disabilities, students who are blind or visually impaired, and students

who are English-language learners (Cook, Eignor, Sawaki, Steinberg, and Cline, 2006a, 2006b; Pitoniak, Cook, Cline, and Cahalan-Laitusis, 2007; and Stone, Cook, Cahalan-Laitusis, and Cline, 2007).

Literature Review

Cawthon (2004) surveyed a variety of settings where deaf or hard of hearing students are educated and gathered information on the type of accommodations these settings offered to students on state-wide accountability assessments. She found that the most prevalent accommodations used in state-wide standardized assessments in math and reading with deaf or hard of hearing students were extended time, an interpreter for directions, and a separate room for test administration. She also found that particularly on math assessments, the signing of both the question and the possible responses was a prevalent accommodation. Most importantly for this discussion, Cawthon reported that there is very little information on the impact of testing accommodations for students who are deaf or hard of hearing, which she attributes to the small sample sizes that are available for this type of research.

Some recent studies of tests given to students who are deaf or hard of hearing have focused on expert reviews of test content and basic analyses of examinee results. A study carried out by Martin (2005) examined the appropriateness of the New York State (NYS) ELA Grade 8 Test for deaf students. Eight experts (three of whom were deaf) were used to rate the passages and items found on the test. For the passages, over 60% of the reviewers rated the overall difficulty of the eight passages as "hard" for students who are deaf. With regard to the test items, reviewers found that 18% of the MC items and 28% of the CR items failed to pass the item quality indicators. Specifically, literary device, author point of view and author technique were

the three skills identified by reviewers as not being taught to students prior to test administration. The test data of 44 deaf students who had taken the exam revealed that expert opinions on the test items were reflected in students' performances on the test. Only 6 of 25 MC items could be correctly answered by 50% of students, and these items belonged to passages that had been judged as comparatively "easier" by the experts. Similarly, items with the lowest percentage of correct responses belonged to passages ranked as "hardest" for deaf students.

A recent study by Lollis and LaSasso (2008) used a similar design as the one used by Martin to examine the appropriateness of the North Carolina state-mandated Reading Competency Test when administered to students who are deaf. Eight reviewers (four deaf or hard of hearing) examined the reading passages of the exam for difficulty, appropriateness, interest, and structure. Individual item analyses included whether performance on the items could be impacted by the student's hearing loss, for instance due to cultural bias or disability-related stereotyping, content that would be less familiar to students who are deaf, or item format. Half of the reviewers judged six of the ten passages as "hard". Similarly five items from the content passages and two items from the literary passages received negative comments from half the reviewers. The study concluded that a test for deaf students on which one could confidently base high-stakes decisions would require prior input on test construction from teachers of deaf or hard of hearing students.

While not specifically related to students who are deaf or hard of hearing, some studies have looked at the internal structure of tests when given to students with other disabilities under different types of accommodations. Cahalan-Laitusis, Cook, and Aicher (2004) examined DIF on third and seventh grade assessments of English-Language Arts by comparing students with learning disabilities that received a read-aloud accommodation or extra time to two separate

reference groups (students with and without disabilities who received no accommodations). The research results indicated that 7-12% of the test items functioned differently for students with learning disabilities that received read-aloud accommodations when compared to either of the reference groups, while extra time resulted in only 0-1% of the items showing DIF when the focal group received extra time and the reference groups did not. A similar study by Bolt (2004) compared small samples of students on three state assessments of reading or English-Language Arts. In all three states the read-aloud accommodation resulted in significantly more items with DIF than for other accommodations.

Tippetts and Michaels (1997) factor analyzed data from the Maryland School Performance Assessment Program (MSPAP) and found that scores from students with disabilities who received accommodations and students with disabilities who received no accommodations had comparable factor structures and concluded that this similarity of factor structures provided evidence of test fairness for the two populations taking the MSPAP.

Huynh and Barton (2006) used confirmatory factor analysis to examine the effect of accommodations on the performance of students who took the reading portion of the South Carolina High School Exit Examination (HSEE) in Grade 10. Three groups of students were studied. The first group was students with disabilities who were given the test with an oral administration, the second group of students was students with a disability who were given the regular form of the test, and the third group was students without a disability who took the regular form of the test. The authors concluded that the results of their study clearly indicated that a one-factor model could be used to describe the data for those students taking the accommodated form (the first group) and the regular form (the other two groups).

Method

The data in this study come from 4th and 8th grade state standards-based tests in English-Language Arts that were administered in 2004. Both the 4th and 8th grade assessments consist of reading and writing portions designed to reflect the state content standards and consist of 75 multiple choice items (the Grade 4 exam also has a brief constructed response writing task, which is not included in the analyses). The reading portion has three strands/reporting clusters spread over 42 items: Word Analysis and Vocabulary, Literary Response and Analysis, and Reading Comprehension. The writing multiple choice portion has two strands/reporting clusters spread over 33 items: Written and Oral Language Conventions and Writing Strategies.

Table 1

Number of Items for the English-Language Arts Assessments: Grades 4 and 8

		Number of Items	
Test	Content	Grade 4	Grade 8
Reading	Word Analysis and Vocabulary	18	9
	Reading Comprehension	15	18
	Literary Response and Analysis	9	15
Writing	Writing Strategies	15	17
	Written and Oral Language Conventions	18	16

Both the DIF and the factor analyses for the Grade 4 and Grade 8 ELA tests included four groups of students, all of whom took the test under standard conditions.

- Students without disabilities and not classified as ELL (Group A).
- Students without disabilities and classified as ELL (Group B).
- Deaf or hard of hearing students not classified as ELL (Group C).
- Deaf or hard of hearing students classified as ELL (Group D).

The initial sizes of Groups A and B in each grade were quite large, so a random sample of approximately 30,000 students was selected for the DIF analysis. From those students a second sample of 500 students was selected for the factor analysis, primarily to facilitate the running of the various computer programs used in the analyses. This step was also done so that the results could be better compared to those from Groups C and D which were not large samples, and therefore left intact. Both the DIF and Factor Analysis samples showed similar means and distributions as their respective full populations. The mean score for the full Group A population in Grade 4 was 47.9 (N = 298,622); for Group B the mean was 34.6 (N = 133,915). In Grade 8, the mean score for Group A was 46.4 (N = 357,374) and the mean score for Group B was 32.1 (N = 81,023). Table 2 shows sample sizes and mean scores for the sampled groups.

Table 2

Raw scores on the English-Language Arts Assessments: Grades 4 and 8, by group

	DIF Sample			Factor Analysis Sample		
Grade 4	N	Mean	SD	N	Mean	SD
Group A: Non-disabled/Non-ELL	30225	47.9	14.1	500	47.2	13.7
Group B: Non-disabled/ELL	30134	34.7	12.0	500	34.9	12.6
Group C: Deaf or hard of hearing/Non-ELL	236	36.8	15.8	236	36.8	15.8
Group D: Deaf or hard of hearing/ELL	174	28.6	11.1	174	28.6	11.1

	DIF Sample			Factor Analysis Sample		
Grade 8	N	Mean	SD	N	Mean	SD
Group A: Non-disabled/Non-ELL	30069	46.4	12.1	500	46.4	12.1
Group B: Non-disabled/ELL	29865	32.1	9.5	500	32.2	9.2
Group C: Deaf or hard of hearing/Non-ELL	289	34.0	14.2	289	34.0	14.2
Group D: Deaf or hard of hearing/ELL	165	24.9	8.6	165	24.9	8.6

For the DIF analyses, the two main comparisons were students who are not disabled vs. students who are deaf or hard of hearing, based on ELL status: Group A (Reference) with Group C (Focal) and Group B (Reference) with Group D (Focal). Additional DIF analyses, which used groups with sample sizes too small for a comparable factor analysis, focused on those students

who were deaf or hard of hearing who took the tests with accommodations. For Grade 4, students who are deaf who took the test with accommodations (Group E) and students who are hard of hearing who took the test with accommodations (Group F) were compared to Group A. Those two groups were combined for Grade 8 due to small sample size, so Group G consists of Grade 8 students who are deaf or hard of hearing who took the test with accommodations. Information on the additional DIF analysis samples can be found in Table 3.

Table 3

Raw scores on the English-Language Arts Assessments: Grades 4 and 8, by Group, for additional samples used for DIF

Grade 4	N	Mean	SD
Group E: Deaf, tested with accommodations	104	23.3	11.0
Group F: Hard of Hearing, tested with accommodations	113	26.9	10.0
Grade 8	**N**	**Mean**	**SD**
Group G: Deaf or hard of hearing, tested with accommodations	130	24.7	8.5

The DIF analyses were conducted using the Mantel-Haenszel procedure (Holland & Thayer, 1988). The Mantel-Haenszel (MH) procedure uses contingency tables to compare the item performance between two groups of examinees (i.e., reference and focal groups) who were previously matched on ability estimates (e.g., observed total test score). Contingency tables are constructed for each item at each possible total test score, and the analysis of the contingency tables indicates whether or not there is a difference between groups in the odds of getting an item correct at a given score level. The Mantel-Haenszel Chi-square (MH $\_^2$) statistic provides the statistical significance test, while the Mantel-Haenszel delta scale (MH D-DIF) statistic provides the measure of the effect size. The higher the MH D-DIF statistic, the greater the difference between the matched groups, with a negative value indicating the test question is more difficult for the focal group and a positive value indicating it is more difficult for the reference group.

Based on the magnitude of the MH D-DIF statistic, items are categorized into three categories (*A*, *B*, *C*) - category *A* contains items with negligible DIF, category *B* contains items with slight to moderate values of DIF, and category *C* contains items with moderate to large values of DIF. Items categorized as *C* are flagged as being problematic and on operational tests would be reviewed for possible exclusion.

There is also no clear rule-of-thumb for minimum sample size to carry out DIF analyses, though various guidelines do exist. These guidelines vary based on the circumstances, as smaller sample sizes (minimum of 100 in the focal group) considered acceptable at the test assembly phase and larger sample sizes (minimum of 500 in the focal group) are required if changing reported test scores (Zieky, 1993). As with most statistical tests, a larger sample size would make the results more stable, and smaller samples may result in items flagged for DIF falsely as well as items not flagged for DIF that should have been. However, the ultimate goal here is to identify the number of items that may favor a focal or reference group as a means to investigate the overall internal structure of the test and not to identify the nature of specific items, In this research-oriented context, the authors felt justified in applying this DIF method to groups with at least 100 subjects, especially given that a large sample size for these populations was not available.

For the factor analysis phase, the individual item data was combined into four or five item-long "parcels". There are a few benefits of using item parcels over individual items in factor analyses, including: (a) allowing the analysis to be run on a smaller sample by reducing the number of variables being analyzed while increasing statistical power in hypothesis testing, (b) increased reliability and variance of the data when items are analyzed together, and (c) non-

linear relationships exist between items that are dichotomously scored which can create more factors than are really present.

Factor analyses are generally feasible if the sample size is greater than approximately one-half the square of the number of variables analyzed. This allows for proper formation of the matrix generally used in confirmatory factor analysis (see Joreskog and Sorbom (2005) for a justification). Given that the test consisted of 75 items (variables), the minimum sample size necessary for an item level analysis was 2775, which was not possible with Groups C and D.

The item parcels were created by selecting items within each of the five content strands, while balancing the same level of difficulty within each of the five individual strands, based on item statistics from Group A. The minimum number of items going into a parcel was set at four. The result was a total of 16 parcels for the 4^{th} grade data and 17 parcels for the 8^{th} grade data. Tables 4 and 5 display the parcel designs for the Grade 4 and Grade 8 ELA assessments. The range of average difficulties of the parcels ranged from 0.55 to 0.71 in Grade 4 and 0.56 to 0.66 in Grade 8, but these are balanced within strands.

Table 4

Parcel Design for Grade 4 ELA Assessment

Section	Strand	Items in Strand	Parcel	Items in Parcel	Average Parcel Difficulty (P+)
Reading	Word Analysis	18	1	5	0.71
			2	4	0.70
			3	5	0.70
			4	4	0.71
	Reading Comprehension	15	5	5	0.62
			6	5	0.61
			7	5	0.61
	Literary Response	9	8	5	0.55
			9	4	0.55
Writing	Written & Oral English Conventions	18	10	5	0.64
			11	5	0.62
			12	4	0.65
			13	4	0.65
	Writing Strategies	15	14	5	0.59
			15	5	0.60
			16	5	0.59

Table 5

Parcel Design for Grade 8 ELA Assessment

Section	Strand	Items in Strand	Parcel	Items in Parcel	Average Parcel Difficulty (P+)
Reading	Word Analysis	9	1	5	0.65
			2	4	0.65
	Reading Comprehension	18	3	5	0.61
			4	4	0.61
			5	5	0.61
			6	4	0.60
	Literary Response	15	7	5	0.59
			8	5	0.61
			9	5	0.61
Writing	Written & Oral English Conventions	16	10	4	0.56
			11	4	0.59
			12	4	0.60
			13	4	0.58
	Writing Strategies	17	14	4	0.65
			15	4	0.66
			16	4	0.66
			17	5	0.66

The reason for parceling within content strand was due to considerations of the possible underlying structure of the test. There were three reasonable options for the number of possible factors: 1) the underlying factor structure corresponds to the existing strand structure, resulting in five underlying factors; 2) the underlying factor structure corresponds to the two test sections, one for Reading and one for Writing; or 3) as an overall ELA assessment, it is reasonable to hypothesize that the test has only one underlying factor, simultaneously accounting for data from each content area. This initial thinking of the possible existence of five, two, or one underlying dimensions that explain the data, and whether the same number of factors exists across all groups, influenced the exploratory and confirmatory factor analyses that were conducted.

Results

Differential Item Functioning (DIF) Analyses

The main DIF analyses on Groups A through D showed only one item with C-level DIF in Grade 4 and one item with C-level DIF in Grade 8. However, additional DIF analyses beyond the main groups indicated minimal to no C-level DIF items for both grades as well.

For Grade 4, when comparing Group A (Non-disabled, non-ELL) to Group C (Deaf or Hard of Hearing, non-ELL), one item showed C-level DIF, favoring Group A. This indicates that for one question out of the 75 administered, the non-ELL deaf or hard of hearing students were less likely than expected to answer that item correctly than their non-ELL, non-disabled counterparts when controlling for ability. No items showed C-level DIF for the comparison of Group B (Non-disabled, ELL) to Group D (Deaf or Hard of Hearing, ELL).

The second set of DIF analyses looked at the Grade 4 students who were either deaf or hard of hearing and took the test with accommodations. For this DIF comparison, while the students who were classified as deaf or hard of hearing were analyzed separately (Groups E and F). Two items showed C-level DIF for Group E (Deaf and took the test with accommodations) when compared to Group A (Non-disabled, non-ELL), both favoring the non-disabled students. No items showed C-level DIF for Group F (Hard of hearing and took the test with accommodations) when compared to Group A (Non-disabled, non-ELL). One final comparison done for Grade 4 compared Group C with Group D, and also resulted in no C-level DIF items.

The results were similar for Grade 8. When comparing Group A (Non-disabled, non-ELL) to Group C (Deaf or Hard of Hearing, non-ELL), one item showed C-level DIF, favoring Group A. No items showed C-level DIF for the comparison of Group B (Non-disabled, ELL) to Group D (Deaf or Hard of Hearing, ELL).

As with Grade 4, the second set of DIF analyses for Grade 8 looked at the students who were deaf or hard of hearing and took the test with accommodations. Sample sizes required combining the deaf or hard of hearing populations, regardless of ELL status, who received accommodations into one focal group. One item showed C-level DIF for Group G (students who were deaf or hard of hearing and took the test with accommodations) when compared to Group A (Non-disabled, Non-ELL), both favoring the non-disabled students. One final comparison done for Grade 8 included comparing Group C with Group D, which as in Grade 4 resulted in no C-level DIF items.

Factor Analyses

The factor analyses, using the same four main groups for both grades as used in the DIF analyses, were conducted in stages, starting with the exploratory phase and ending with the confirmatory phase. As previously mentioned, item parcels were created utilizing the individual items. The reason the individual items were not considered for analysis was due to sample size limitations in conjunction with test length. This also avoided concerns that tend to arise with item-level data.

The variance-covariance matrices of parcel scores were entered into SAS (2003) to perform exploratory factor analyses using maximum likelihood extraction with no rotation at first (for one factor) and later promax rotation (for two correlated factors). Given the inherent correlation of the sub-scores on this type of test, it is desirable to have the latent factors correlate as well, which is why a promax rotation was used compared to a varimax rotation where latent factors are orthogonal and not correlated. Scree plots (Child, 1970) of the eigenvalues computed from the variance-covariance matrices among the parcels are displayed in Figures 1 and 2.

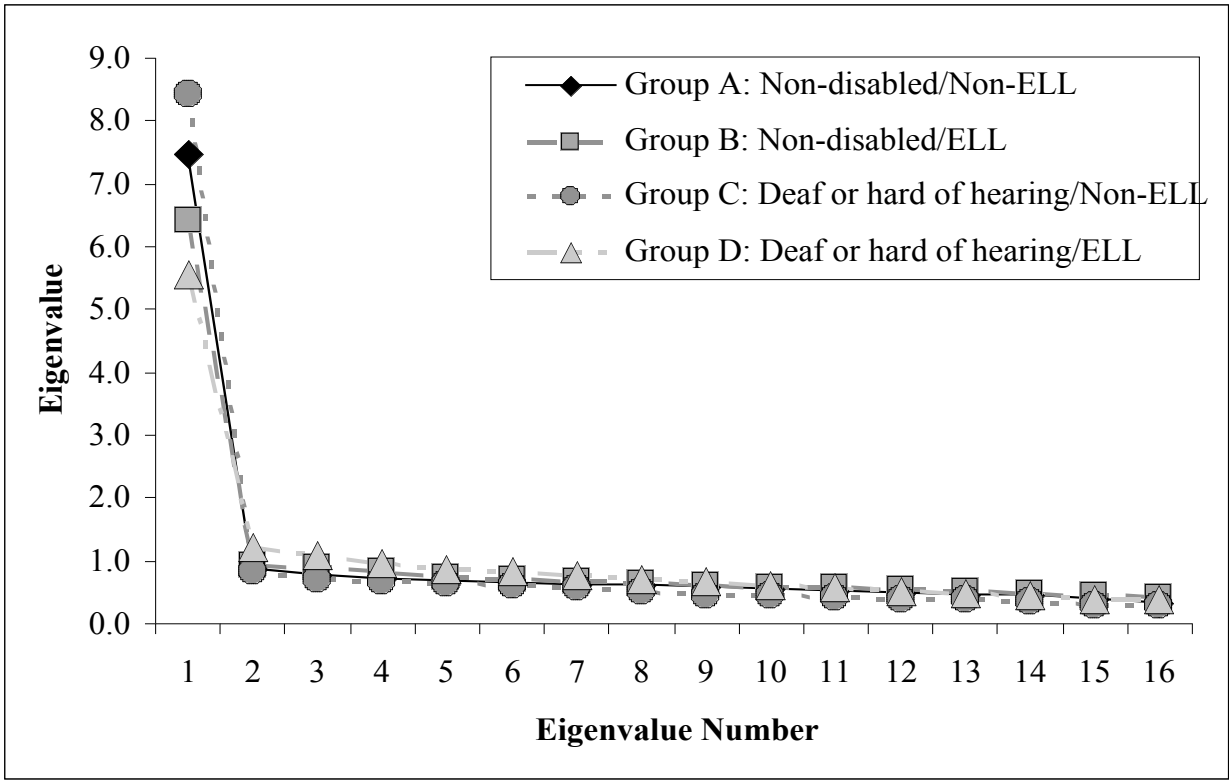

Figure 1. Scree Plot of Eigenvalues from Grade 4 ELA Assessment

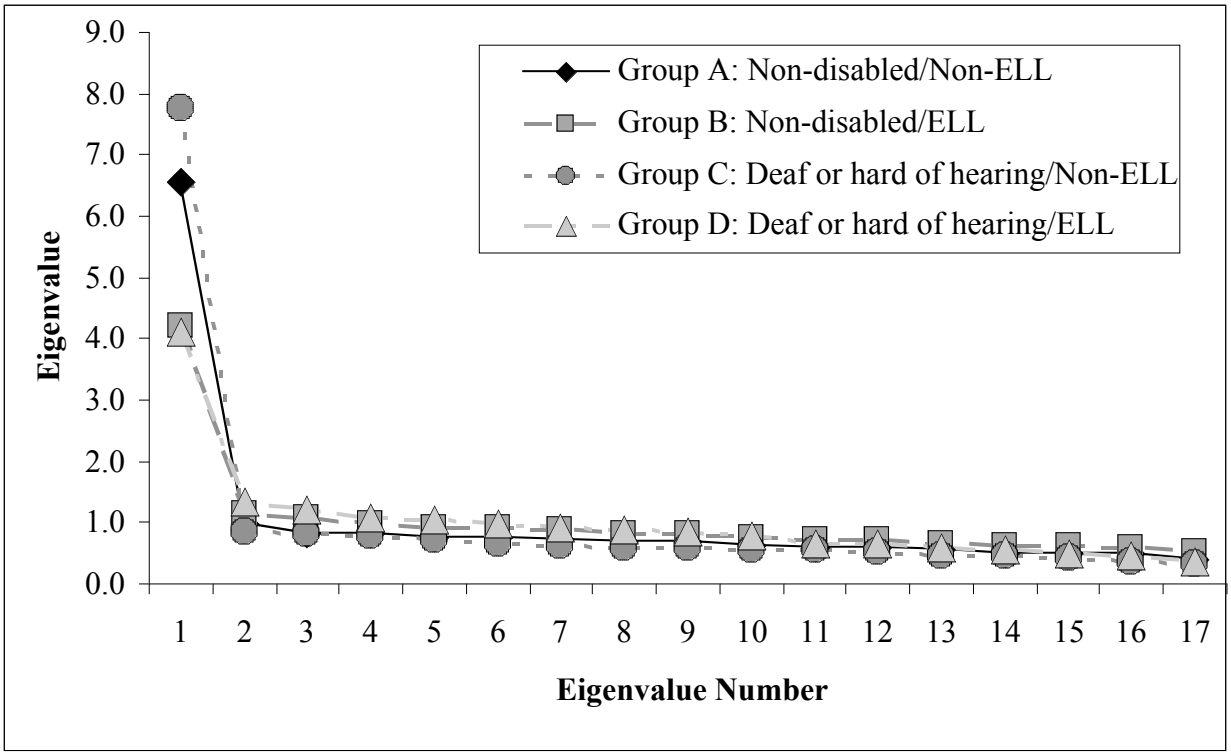

Figure 2. Scree Plot of Eigenvalues from Grade 8 ELA Assessment

The first eigenvalue in each group for both grades is very large compared to subsequent eigenvalues. In each case, a single factor seemed to fit the data best, even though some p-values testing for whether one factor was sufficient to fit the data were below 0.05, indicative of inadequate fit, and there were instances of parcels with loadings below 0.30 on the general factor for Group D in Grade 8. There were some indications that separate Reading and Writing factors might exist for some groups, but attempts to try and extract more dimensions (namely five) using promax rotation so that the factors could be correlated (Hendrickson and White, 1964), sometimes led to Heywood cases[1] for some groups in higher numbers of dimensions, which is evidence of possibly over-factoring the data. Also, as evidenced in the scree plots, additional factors do not suggest a large additional amount of explained variance.

Therefore, more weight is placed on the results from the confirmatory, rather than the exploratory factor analyses for the following reasons: (1) the fit indices obtained from confirmatory models can be more readily interpreted, (2) the results from hypothesis tests on the sufficiency on the number of factors correct for non-normality in the data, and therefore can also be more easily interpreted, and (3) items may be assigned to specific factors, which is not the case in exploratory models.

Attempts were made to confirm more than a single factor for this test, despite apparent evidence that a single factor may be sufficient to fit the data for all groups. The reason for this was to identify any disparate patterns in fit indices, factor loadings, or inter-factor correlations across groups that might suggest a different course of action in further analyses. The other proposed additional model was a content area-based design (two factors – reading and writing). The five-factor model originally proposed with one strand assigned to each factor was not

[1] When the percent of variance in a given variable explained by all the factors exceeds 100.

considered because as was evident from the scree plot, there was little additional explanatory power to be gained from such a design.

Confirmatory factor analyses were conducted in EQS (Bentler & Wu, 2006) using the variance-covariance matrices of item parcel scores as inputs under maximum likelihood estimation for free parameters. The results from the single factor and two-factor individual group models are displayed in Tables 6 and 7.

Table 6

Summary of Individual Group Parcel-Level Confirmatory Results for Grade 4

Group	N	Model DF	ML Chi-Square	RMSEA	CFI	GFI	Mardia Normalized Estimate
Group A - 1 Factor	500	104	159.702	0.033	0.983	0.962	4.691
Group A - 2 Factors	500	103	132.640	0.024	0.991	0.969	4.691
Group B - 1 Factor	500	104	169.384	0.035	0.974	0.958	-2.570
Group B - 2 Factors	500	103	145.067	0.029	0.983	0.965	-2.570
Group C - 1 Factor	236	104	125.740	0.030	0.989	0.937	-1.201
Group C - 2 Factors	236	103	103.230	0.003	1.000	0.949	-1.201
Group D - 1 Factor	174	104	134.822	0.041	0.954	0.913	-1.505
Group D - 2 Factors	174	103	129.865	0.039	0.960	0.917	-1.505

Chi-Sq Difference Tests	_ DF	_ Chi-Sq	P-Value
Group A: 2 Factors - 1 Factor	1	27.062	0.000
Group B: 2 Factors - 1 Factor	1	24.317	0.000
Group C: 2 Factors - 1 Factor	1	22.510	0.000
Group D: 2 Factors - 1 Factor	1	4.957	0.026

Table 7

Summary of Individual Group Parcel-Level Confirmatory Results for Grade 8

Group	N	Model DF	ML Chi-Square	RMSEA	CFI	GFI	Mardia Normalized Estimate
Group A - 1 Factor	500	119	154.693	0.025	0.986	0.964	0.280
Group A - 2 Factors	500	118	120.587	0.007	0.999	0.973	0.280
Group B - 1 Factor	500	119	174.597	0.031	0.951	0.960	-3.385
Group B - 2 Factors	500	118	151.710	0.024	0.970	0.966	-3.385
Group C - 1 Factor	289	119	121.166	0.008	0.999	0.954	-3.356
Group C - 2 Factors	289	118	115.023	0.000	1.000	0.957	-3.356
Group D - 1 Factor	165	119	143.121	0.035	0.934	0.914	-0.829
Group D - 2 Factors	165	118	142.652	0.036	0.933	0.914	-0.829

Chi-Sq Difference Tests	_ DF	_ Chi-Sq	P-Value
Group A: 2 Factors - 1 Factor	1	34.106	0.000
Group B: 2 Factors - 1 Factor	1	22.887	0.000
Group C: 2 Factors - 1 Factor	1	6.143	0.013
Group D: 2 Factors - 1 Factor	1	0.469	0.493

The results showed that a single factor may fit the data quite well compared to a model based on content areas (two factors). This was based on examination of fit statistics suggested by Hoyle & Panter (1995) such as the Root Mean Square Error of Approximation (RMSEA)[2] which was less than 0.05 for all groups and all models and values of the Comparative Fit Index (CFI)[3] and Goodness-of-Fit Index (GFI)[4] being above 0.90 for all groups and models. The Mardia

[2] Evaluates the extent to which the model approximates the data, taking into account the model complexity. A RMSEA of .05 or below is considered to be an indication of close fit and .08 or below for adequate fit as proposed by Browne & Cudeck (1993).

[3] An incremental fit index, which assesses overall improvement of a proposed model over an independence model where the observed variables are uncorrelated. A CFI of .90 or above indicates an adequate model fit

[4] An absolute model fit index, which is analogous to a model R2 in multiple regression analysis. A GFI of .90 or above indicates an adequate model fit

coefficient expresses the degree to which the parcel scores fit a multivariate normal distribution, and the values displayed in Tables 6 and 7 are satisfactory.

To test whether the two-factor model sufficiently improves the fit, the changes in chi-squares are compared to the critical values of the chi-square distribution given the changes in degrees of freedom. These changes are shown in the lower part of Tables 6 and 7. The changes in chi-squares were significant at the 0.05 level of significance for all groups, except Group D for Grade 8, which would suggest that a two-factor model might fit the data better for this group in both grades. However, one key piece of information was also used to help aid in reaching a final decision of a single-factor model for both of these groups. That was the factor inter-correlations, which are displayed in Table 8.

Table 8

Inter-Correlation Matrix from Two-Factor Individual Group Confirmatory Models

Group	Grade 4	Grade 8
Group A: Non-disabled/Non-ELL	0.939	0.913
Group B: Non-disabled/ELL	0.921	0.854
Group C: Deaf or hard of hearing/Non-ELL	0.937	0.965
Group D: Deaf or hard of hearing/ELL	0.927	0.963

The results were such that the authors felt that all inter-correlations were above, or close enough to the 0.90 rule-of-thumb suggested by Bagozzi & Yi (1988) to allow that a single factor could best explain the data for all groups, including Group D.

The results from the individual group confirmatory models indicating that a single factor could best explain the Grade 4 and Grade 8 ELA assessment data became the basis for proceeding to the final step in the analysis, where a one-factor multi-group confirmatory model was tested in each grade. Given the apparent comparable behavior of Groups A and C (non-

ELL, Set 1) and Groups B and D (ELL, Set 2) in each grade, demonstrated in test performance and the proportion of explained variance in the parcel data, two sets of multi-group confirmatory models were executed in each grade, based on ELL status. The goal in each analysis is to show factorial invariance across the groups under study. There were four steps undertaken to complete this process, which are described in Table 9.

Table 9

Summary of Parcel-Level Multi-Group Confirmatory Factor Analyses

Model	Objective	Constraints Imposed
1 (Least Restrictive)	Establish a baseline multi-group model	None
2	Test whether factor loadings are invariant across groups	Factor Loadings Equal Across Groups
3	Test whether factor loadings and factor variances are invariant across groups	Factor Loadings and Variances Equal Across Groups
4 (Most Restrictive)	Test whether factor loadings, variances, and residuals are invariant across groups	Factor Loadings, Variances, and Residuals Equal Across Groups

At each step in the process, model fit indices as produced by EQS are checked for reasonableness before proceeding to the next step. If there is any model misfit, equality constraints may be relaxed, if necessary. Testing the most restrictive model suggested above which would show true invariance, as suggested by Byrne (1998), is not always conducted, but is encouraged if it is felt the model will fit the data well.

The establishment of the baseline model described in Step 1 involves simply taking the individual-group models in each set, as described in the previous section and stacking them together as one large model with the degrees of freedom from the individual models being

additive. More degrees of freedom are added as constraints are imposed. Tables 10 through 13 summarize the results from these multi-group confirmatory models.

Table 10

Summary of Parcel-Level Multi-Group Confirmatory Analysis Results – Grade 4, Non-ELL, Students without disabilities vs. Deaf or hard of hearing students

Model	Model DF	ML Chi-Square	RMSEA	CFI	GFI	Constraints
1	208	285.442	0.032	0.985	0.954	
2	223	305.719	0.032	0.984	0.951	Factor Loadings
3	224	313.191	0.033	0.983	0.950	Factor Loadings, Variances
4	240	330.785	0.032	0.983	0.947	Factor Loadings, Variances, Residuals

Model Difference	_ DF	_ Chi-Sq	P-Value
2 vs. 1	15	20.277	0.162
3 vs. 2	1	7.472	0.006
4 vs. 3	16	17.594	0.348

Table 11

Summary of Parcel-Level Multi-Group Confirmatory Analysis Results – Grade 4, ELL, Students without disabilities vs. Deaf or hard of hearing students

Model	Model DF	ML Chi-Square	RMSEA	CFI	GFI	Constraints
1	208	304.206	0.037	0.970	0.947	
2	223	344.736	0.040	0.962	0.940	Factor Loadings
3	224	349.259	0.041	0.961	0.940	Factor Loadings, Variances
4	240	364.491	0.039	0.961	0.937	Factor Loadings, Variances, Residuals

Model Difference	_ DF	_ Chi-Sq	P-Value
2 vs. 1	15	40.530	0.000
3 vs. 2	1	4.523	0.033
4 vs. 3	16	15.232	0.508

Table 12

Summary of Parcel-Level Multi-Group Confirmatory Analysis Results – Grade 8, Non-ELL, Students without disabilities vs. Deaf or hard of hearing students

Model	Model DF	ML Chi-Square	RMSEA	CFI	GFI	Constraints
1	238	275.860	0.020	0.992	0.960	
2	254	304.577	0.022	0.989	0.956	Factor Loadings
3	255	313.689	0.024	0.987	0.955	Factor Loadings, Variances
4	272	332.387	0.024	0.987	0.952	Factor Loadings, Variances, Residuals

Model Difference	_ DF	_ Chi-Sq	P-Value
2 vs. 1	16	28.717	0.026
3 vs. 2	1	9.112	0.003
4 vs. 3	17	18.698	0.346

Table 13

Summary of Parcel-Level Multi-Group Confirmatory Analysis Results – Grade 8, ELL, Students without disabilities vs. Deaf or hard of hearing students

Model	Model DF	ML Chi-Square	RMSEA	CFI	GFI	Constraints
1	238	317.719	0.032	0.947	0.949	
2	254	363.293	0.036	0.927	0.941	Factor Loadings
3	255	363.977	0.036	0.927	0.941	Factor Loadings, Variances
4	272	382.490	0.035	0.927	0.938	Factor Loadings, Variances, Residuals

Model Difference	_ DF	_ Chi-Sq	P-Value
2 vs. 1	16	45.574	0.000
3 vs. 2	1	0.684	0.408
4 vs. 3	17	18.513	0.357

As constraints are imposed in Steps 2 through 4, the model fit indices go down as one might expect. The chi-square difference tests reveal interesting patterns in how these models fit as the constraints are imposed. As the constraint on factor loadings was imposed, all models showed significant changes in model fit at the 0.05 level, i.e. the model fit is not as good when

the constraint is applied, except for the non-ELL sample in Grade 4. An examination of the factor loadings prior to the equality constraint being imposed may explain part of the reason why only one of four group comparisons demonstrated invariance of factor loadings. The range of differences between Groups A and C for Grade 4 was (-.106, .038). The range of differences between Groups B and D for Grade 4 was (-.128, .239), which was only slightly wider, yet was wide enough to lower the p-value enough to show a lack of invariance at this step. The differences for the Grade 8 comparisons were (-.139, .027) for Groups A and C and (-.204, .228) for Groups B and D, which are clearly too great to indicate that the factor loadings could be similar between groups in each set. However, the RMSEA, CFI, and GFI were all within normally accepted boundaries for adequate model fit at both steps and the changes in these fit indices between models were not practically significant. Nevertheless, it does raise questions about whether the single factor was statistically similar across all comparisons.

Discussion

Both the DIF and factor analysis results of this study indicate that both the Grade 4 and Grade 8 tests appear to function about the same for students who are deaf or hard of hearing as they do for students without disabilities. In terms of DIF, there appears to be little to no item-level performance differences on these exams with regards to students who are deaf or hard of hearing compared to non-disabled students. With only one C-level DIF item in either grade, and only for the non-ELL students, it is hard to argue that either test is inappropriate for the groups investigated in this study. For the factor analyses, a single factor was identified for all groups on each test, however, not all goodness-of-fit indices indicated factorial invariance across all of these groups. While this indicates that the internal structure of the test is not exactly the same for all the groups, there is also little evidence to suggest that the tests are truly behaving differently

for the four groups. However, the fact that this single factor may not be identical across all groups indicates that the test may be measuring a slightly different construct for the four groups of interest.

However, the generally low scores on the ELA tests by deaf or hard of hearing students in both grades is an issue separate from the internal structure of the tests. The students who are deaf or hard of hearing did not perform as well on the exams as did the students without disabilities and who were not classified as ELL (See Table 2). For any exam, when one group is achieving scores approximately 1.5 standard deviations below the main population, the appropriateness of that test for that group should be examined as well as the classroom instruction that the students receive.

It is interesting, however, that the non-ELL deaf or hard of hearing students performed slightly better than the ELL students with no disabilities. One could argue that English is not the first language of students who are deaf, as their most commonly used form of communication, American Sign Language, uses a different grammatical structure than oral and written English. That all deaf students (and potentially some hard of hearing students) are effectively English language learners is a separate debate, but nevertheless the similarity in performance with the ELL population is worthy of future study. For tests in subjects such as math and science, the option for the equivalent of a read-aloud accommodation exists in the form of signing the questions as does the option of providing a glossary using signs. In those instances, the potential accommodations follow accepted practices already in place for students with learning disabilities or for ELL students. However, read-aloud accommodations for deaf or hard of hearing students on an English-Language Arts exam are problematic as many states currently do not consider such accommodations valid for reading tests. A study focused on signing questions to deaf or

hard of hearing students, following the methodology used by Cook, et al. (2009) would reveal whether signing is equivalent to read-aloud in relation to the internal structure of an ELA test.

References

American Educational Research Association, American Psychological Association. & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bagozzi, R. P., & Yi. Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16,* 74-94.

Bentler, P., & Wu, E. (2006). EQS 6.1 for Windows. Los Angeles: Multivariate Software, Inc.

Bolt, S. E. (2004, April). *Using DIF analyses to examine several commonly-held beliefs about testing accommodations for students with disabilities*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage Publications.

Cahalan-Laitusis, C., Cook, L. L., & Aicher, C. (2004, April). *Examining test items for students with disabilities by testing accommodation.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Cawthon, S. W. (2004). *National Survey of Accommodations and Alternate Assessments for Students Who Are Deaf or Hard of Hearing in the United States.* Online Research Lab, Walden University.

Cook, L. L., Eignor, D. R., Sawaki, Y., Steinberg, J., & Cline, F. (2006a, April). *Using factor analysis to investigate the impact of accommodations on the scores of students with*

disabilities on English-language arts assessments. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Cook, L. L., Eignor, D. R., Sawaki, Y., Steinberg, J., & .Cline, F. (2006b, April). *Investigating the Dimensionality of an English-Language Arts Assessment Administered to English-Language Learners With and Without Accommodations*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Cook, L. L., Eignor, D. R., Steinberg, J., Sawaki, Y., & .Cline, F. (2009). Using factor analysis to investigate the impact of accommodations on the scores of students with disabilities on a reading comprehension assessment. *Journal of Applied Testing Technology, XX,* pp-pp.

Elbaum, B., Arguelles, M. E., Campbell, Y., & Saleh, M. B. (2004). Effects of a student-read-aloud accommodation on the performance of students with and without learning disabilities on a test of reading comprehension. *Exceptionality, 12*(2), 71-87.

Hendrickson, A. E., & White, P. O. (1964). PROMAX: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology, 17,* 65-70.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Hoyle, R.H., & Panter, A.T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage Publications.

Huynh, H., & Barton, K. (2006). Performance of students with disabilities under regular and oral administrations of a high-stakes reading examination. *Applied Measurement in Education, 19*(1), 21-39.

Individuals with Disabilities Educational Act of 1997, 20 U.S.C. 1412(a) (17) (A). (1997).

Joreskog, K. G., & Sorbom, D. (2005). *LISREL 8 user's reference guide*. Chicago: Scientific Software International.

Lollis, J., & LaSasso, C. (2008). The appropriateness of the NC state-mandated reading competency test for deaf students as a criterion for high school graduation. *Journal of Deaf Studies and Deaf Education* Advance Access published on June 5, 2008, DOI 10.1093/deafed/enn017

Martin, P. (2005). An examination of the appropriateness of the New York state English Language Arts grade 8 test for deaf students. *Unpublished doctoral dissertation*, Gallaudet University, Washington D.C., March 2005.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education.

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

Pitoniak, M. J., Cook, L., Cline, F., & Cahalan-Laitusis, C. (2006, April). *Using differential item functioning to investigate the impact of accommodations on the scores of students with disabilities on English-language arts assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

SAS Institute. (2003). *SAS/STAT user's guide, Version 6*. Cary, NC: author.

Stone, E. A., Cook, L. L., Cline, F., & Cahalan-Laitusis, C. (2007, April). *Using differential item functioning to investigate the impact of testing accommodations on an English language arts assessment for students who are blind and visually impaired*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Thurlow, M., & Wiener, D. (2000). *Non-approved accommodations: Recommendations for use and reporting* (Policy Direction N0. 11). Minneapolis, MN: University of Minnesota, National Center for Educational Outcomes.

Tippets, E., & Michaels, H. (1997, March). *Factor structure invariance of accommodated and non-accommodated performance assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Zieky, M.. (1993). Practical Questions in the Use of DIF Statistics in Test Development. In P. Holland & H. Weiner (Eds.), Differential Item Functioning (pp. 337–347). Hillsdale, NJ: Erlbaum.