Using Factor Analysis to Investigate the Impact of Accommodations on the Scores of Students with Disabilities on a Reading Comprehension Assessment

Linda Cook
Daniel Eignor
Jonathan Steinberg
Yasuyo Sawaki
Frederick Cline

Educational Testing Service

Abstract

The purpose of this study was to investigate the impact of a read-aloud test change administered with the Gates-MacGinitie Reading Test (GMRT) on the underlying constructs measured by the Comprehension subtest. The study evaluated the factor structures for the Level 4 Comprehension subtest given to a sample of New Jersey fourth-grade students with and without reading-based learning disabilities. Both exploratory and confirmatory factor analyses were used to determine whether or not the GMRT Comprehension subtest measures the same underlying constructs when administered with and without a read-aloud test change. The results of the analyses indicated factorial invariance held when the Comprehension subtest was administered to groups of students without disabilities who took the test under standard conditions and with a read-aloud test change and for groups of students with reading-based learning disabilities who also took the test under standard conditions and with a read-aloud test change.

Introduction

The No Child Left Behind Act of 2001 mandates that states, districts, and schools be accountable for the academic achievement of all students, including students with disabilities. Although this requirement is extremely important for states and school districts to implement, it is also an extremely challenging one for a number of reasons, particularly when applied to students with disabilities. One reason this requirement is challenging for states and school districts is that the number of students with disabilities now currently educated in US public schools is not trivial. According to the United States Government Accountability Office, in the 2003-4 school year, more than 6 million students with disabilities—approximately 13 % of all students—attended US public schools (GAO-05-618, Special Education Assessments, 2005). A second complicating factor is that assessments for students with disabilities are also required under the Individuals with Disabilities Education Act (1997), and this act clearly stipulates that states must provide a means for participation (through accommodations and/or modifications) in statewide assessments for students with disabilities. This requirement often raises challenges for the interpretation of the scores from these assessments because of the possible lack of standardization of test administrations brought about by the use of these test accommodations/modifications and the potential for the test accommodations or modifications to impact the construct(s) that these tests measure.[1]

Cahalan-Laitusis and Cook (2008) discuss a recent review of state testing accommodations carried out by Clapper, Morse, Lazarus, Thompson, and Thurlow (2005) that found that the majority of states agree on their classifications of most changes in testing

[1] The term "accommodation" is typically reserved for test changes that a state believes do not change the underlying construct measured by the test. The term "modification" typically refers to test changes that a state believes may change the underlying construct measured by the test.

procedures or methods as either an accommodation or modification. Cahalan-Laitusis and Cook point out, however, that "States are not in agreement on whether to consider the audio presentation of test content (i.e., read-aloud) on reading assessments to be an accommodation or a modification. These differences are largely due to different specifications of reading in each state's reading standards. States that consider read-aloud a modification on tests of reading have either (a) determined that reading involves visual or tactile decoding of text or (b) argue that scores are not comparable because the test scores that are obtained with read-aloud represent a measure of listening comprehension rather than reading comprehension. On the other hand, states that allow read-aloud accommodations on tests of reading or English-language arts (ELA) have either (a) defined reading as comprehension of written material that is presented in a visual, tactile, or audio format or (b) allow only portions of the test to be read aloud (e.g., test questions but not passages)". (p. 15).

The purpose of the study described in this report was to investigate the impact of a read-aloud test change administered with the Gates-MacGinitie Reading Tests (GMRT) on the underlying construct or constructs measured by the tests. The study evaluated the factor structures for Level 4, Form S of the GMRT given to a sample of New Jersey fourth grade students with and without reading based-learning disabilities. The GMRT Level 4 Form S contains three subtests, Word Decoding, Word Knowledge and Vocabulary, and Comprehension; this study employed only the Comprehension subtest. The data for this study were collected as part of a larger study that examined differential boost and that involved both Forms S and T of the GMRT and students at both the fourth and eighth grade level. (See Cahalan-Laitusis, Cook, Cline, King, & Sabatini, 2008, for a report of the larger study.)

For this study, the tests were administered with and without a read-aloud test change that was delivered using a compact disc (CD) player with headphones. Exploratory and confirmatory factor analyses were used to evaluate whether or not the GMRT Comprehension subtest measures the same underlying construct or constructs when administered with and without a read-aloud test change. The specific questions examined in this study were:

1. Does the GMRT Comprehension subtest measure the same construct(s) for examinees without disabilities (NLD) who take the test under standard conditions as it does for NLD examinees who take the test with a read-aloud test change?
2. Does the GMRT Comprehension subtest measure the same construct(s) for examinees with reading-based learning disabilities (RLD) who take the test under standard conditions as it does for RLD examinees who take the test with a read-aloud test change?

Review of Relevant Research

Some of the most common accommodations for students with RLD were examined in the studies reviewed in this section. These accommodations were those typically specified in 504 or IEP plans, including extra time and audio presentation accommodations; e.g., having the test read aloud, administered via audio cassette or administered with a screen reader. It should be noted that research in this area is difficult to conduct due to (a) the multiple types of accommodations available, typically in combination, (b) the variety and levels of severity of disabilities, (c) controversy regarding how each accommodation might change the test's construct, and (d) inability to aggregate data across administrations because of database shortcomings (e.g., information about type of accommodation is typically not collected). Tindal and Fuchs (2000) completed an exhaustive review of research on testing accommodations for

students with disabilities and this review has been updated more recently (Pitoniak & Royer, 2001; Sireci, Li, & Scarpati, 2003).

The studies reviewed for this report indicate that the most common accommodations for students with reading-based learning disabilities are extra time and audio presentations. Most research on extra time indicates that students with disabilities do benefit differentially when compared with students without disabilities (i.e., a differential boost[2] is demonstrated when the two groups are compared and students with disabilities achieved larger gains than students without disabilities) and that extra time does not appear to alter the construct of most state achievement tests (Sireci, Li, & Scarpati, 2003). Research on the impact of audio presentation on tests of reading or English language arts is less conclusive than the research on timing. Five studies have examined the impact of audio presentation accommodations on tests of reading. One study by Fuchs, Fuchs, Eaton, Hamlett, Binkley, and Crouch (2000) researched the impact of commonly used testing accommodations on the performance of elementary school students with and without learning disabilities on a reading comprehension test. Results indicated that students with learning disabilities had a differential boost from the read-aloud accommodation, but surprisingly, not from extended time or from the provision of large print text. A study by Harker and Feldt (1993) indicated that high school students without disabilities performed better on English assessments when the test was read aloud. Unfortunately, the study did not include students with disabilities.

[2] Sireci et al., define the Interaction Hypothesis (also referred to as differential boost in the literature) as follows. The interaction hypothesis states that (a) when test accommodations are given to the students with disabilities who need them, their test scores will improve, relative to the scores they would attain when taking the test under standard conditions; and (b) students without disabilities will not exhibit higher scores when taking the test with those accommodations. If this hypothesis holds, the test change is considered an accommodation. If the hypothesis does not hold, the test change is considered a modification.

Three studies on the effects of audio presentation reviewed by Sireci, Li, and Scarpeti (2003) indicated no gains for students with or without disabilities (Kosciolek, & Ysseldyke, 2000; McKevitt & Elliot, in press) or similar gains for both groups (Meloy, Deville, & Frisbie, 2000). Sample sizes may have contributed to the different findings among the four studies just mentioned that tested the interaction model for differential boost. The Fuchs, et al., study had the largest sample size (n = 365) and did detect a differential boost, while the study with the next largest sample size (Meloy, et al., 2000; n = 260) found similar gains for students with and without disabilities. The last two studies that tested the interaction model had small samples (n = 31 in the Kosciolek & Ysseldke study and n = 79 in the McKevitt & Elliot study) and found no significant gains for students with or without disabilities. Other possible reasons for the inconsistent results are differences in the item types employed and in the grade levels of the students.

Elbaum, Arguelles, Campbell, and Saleh (2004) examined the effect of students themselves reading a test aloud as an accommodation. Their study included 456 students (283 with learning disabilities) in Grades 6-10. The researchers administered alternate forms of an assessment constructed of third to fifth grade level reading passages and accompanying comprehension questions. All students took the assessment first in the standard condition and second with instructions to read the passages aloud at their own pace. The researchers found that test performance did not differ in the two conditions, and students with learning disabilities (LD) did not benefit more from the accommodation than students without LD. The researchers noticed, however, that the scores of students with learning disabilities LD were more variable in the accommodated condition than were the scores of students without disabilities.

In addition to the experimentally designed studies reviewed above, two recent studies have used operational test data to examine differential item functioning (DIF) by comparing the performance of students who received read-aloud accommodations to those that did not receive accommodations on K-12 reading assessments. 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 to two separate reference groups (students with and without disabilities who received no accommodation). The research results indicated that 7-12% of the test items functioned differently for the focal group (students with learning disabilities that received read-aloud accommodations) when compared to either of the reference groups. Extra time also was examined, but less than 1% of the items had DIF when the focal group received extra time and the reference group did not. A similar study by Bolt (2004) compared smaller 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 other accommodations. Both of these studies provide evidence that a read-aloud accommodation may change the construct being assessed.

On the other hand, Pitoniak, Cook, Cline and Cahalan-Laitusis (2007) examined differential item functioning on large-scale state standards-based English-language Arts assessments at grades 4 and 8 for students without disabilities and students with learning disabilities who took the test with and without accommodations, including a read-aloud accommodation. Only one item at each grade was flagged as having moderate to large DIF, in each case favoring students without disabilities who did not receive an accommodation over students with disabilities who received the read-aloud accommodation. At both grades, additional items were flagged as having slight to moderate DIF, with both positive and negative

DIF being found. The results of this study are seen as supporting the validity of the accommodations given to students with disabilities.

Only a small number of studies that have examined and compared the factor structures of assessments given to students with disabilities under accommodated and non-accommodated conditions with scores obtained by students without disabilities are available in the literature. Consequently, the review presented here is not limited to just those studies that involved a read-aloud test change. Rock, Bennett, Kaplan, and Jirele (1988) examined the factor structure of the GRE and the SAT for groups of examinees with disabilities and students without disabilities. These authors felt that if the factor structure is the same for the tests across these populations, this finding would lend support to the notion that the test scores have the same meaning for students with and without these disabilities. These authors fit a two-factor model to the SAT and a three-factor model to the GRE, using analysis of item parcels and maximum-likelihood confirmatory factor analysis. They found that for the SAT, the two factors of verbal and quantitative ability fit the data reasonably well for examinees with disabilities who took a cassette recorded version (read-aloud test change) of the test, but that the factors were less correlated with each other for this group than for examinees without disabilities who did not receive this accommodation. Although the two-factor model fit overall, additional examination of the SAT verbal and quantitative factors showed evidence of differential meaning of scores for examinees with learning disabilities taking the cassette-recorded version of the SAT.

The results of the analysis of the GRE verbal, quantitative, and analytical factors were examined for examinees with no disabilities that did not receive accommodations on the assessment and for examinees with visual impairments or physical disabilities that did receive accommodations. The analyses revealed some questions regarding the proposed three-factor

structure for the scores of examinees with these disabilities. For examinees with disabilities the analytical factor appeared to actually be two factors, logical reasoning and analytical reasoning.

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. Meloy, Deville, and Frisbie (2000) factor analyzed data from the Iowa Tests of Basic Skills (ITBS). These researchers compared factor structures for students with disabilities taking the assessment with a read-aloud accommodation and students without a disability taking the assessment without such an accommodation. Meloy, et al., concluded that the read-aloud accommodations appeared to change the construct being measured for most accommodated students relative to the scores of students who were assessed under standard conditions.

Cook, Eignor, Sawaki, Steinberg, and Cline (in press) carried out an item-level exploratory and an item-level confirmatory factor analysis of a large state standards-based English-language arts (ELA) assessment using data from students without disabilities and students with learning disabilities who took the test with and without accommodations. These researchers concluded that the ELA assessment was unidimensional; i.e., measured a single factor for all three groups investigated. It should be noted that the accommodations used in this study did not include a read-aloud test change.

Finally, 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 (read-aloud) administration, the second group of students was students with disabilities who were given the regular form of the test and the third group was students without disabilities who also took the regular form of the test. The purpose of their study was to assess the comparability of accommodated and non-accommodated scores. The specific issues they addressed were whether or not the accommodation changed the internal structure of the test and to what degree the accommodation impacted student performance on the test. Only the investigation of the structure of the test is relevant for this review.

In order to evaluate the structure of the test, the authors initially carried out a principal components analysis on the matrix of the correlations among the six subtests making up the test for each group of examinees and this indicated a single factor was adequate to summarize the data. They followed up the principal components analysis with a multi-group maximum likelihood confirmatory factor analysis of the subtest scores to determine whether a one-factor model could best describe the data for all three groups considered together. The authors concluded that the results of their study clearly indicted a one-factor model could be used to describe the data for the accommodated form given to students with disabilities and the regular form given to students with and without disabilities.

Overview of the Study

For this study, we carried out a series of exploratory and confirmatory factor analyses using variance/covariance matrices of scores on item parcels as input to the analyses. The focus of the analyses was to determine and compare the number of factors that account for the data for students with and without disabilities taking Form S, Level 4 of the GMRT. Data for the study came from fourth grade New Jersey public school students with and without reading-based

learning disabilities. Level 4 of the GMRT was administered under standard conditions and with a read-aloud test change. First, single-group exploratory analyses were carried out to define the underlying factor structure for the four groups of students used in the study:

- Students without disabilities (NLD) taking the test under standard conditions (Group 1)
- Students without disabilities (NLD) taking the test with a read-aloud test change (Group 2)
- Students with a reading-based learning disability (RLD) taking the test under standard conditions (Group 3)
- Students with a reading-based learning disability (RLD) taking the test with a read-aloud test change (Group 4)

Next, single-group confirmatory analyses were carried out on the same four groups of students to confirm the factor structures uncovered by the exploratory analyses. Finally, we carried out two multi-group confirmatory analyses; one analysis tested the hypothesis of factorial invariance for the test given to students without disabilities (NLD) under the standard and read-aloud conditions (Groups 1 and 2), and the second analysis tested the factorial invariance for the groups of students with reading- based learning disabilities (RLD) who took the test under standard and read-aloud conditions (Groups 3 and 4). If the hypothesis of invariant factor structures was accepted, we would be able to infer that the test measures the same underlying construct(s) for students without disabilities who take the test with and without a read-aloud test change and the same construct(s) for students with disabilities who take the test with and without a read-aloud test change.

Method

Description of Tests

Form S Level 4 of the GMRT Fourth Edition (Reading Comprehension subtest only) was used for this study. The Comprehension subtest measures a student's ability to read and understand different types of prose. The test contains 11 passages of various lengths and about various subjects, all selected from published books or periodicals. A total of 48 multiple-choice questions examine the student's understanding of the passages. Some of the questions require constructing an understanding based on information that is explicitly stated in the passage; others require constructing an understanding based on information that is only implicit in the passage. The time for the test is set for 35 minutes; however, for this study, the test was administered with extra time (time and a half) for both the standard and read-aloud conditions. Also, students were allowed to mark their answers in the test book for both the standard and read-aloud testing conditions. The audio (read-aloud) presentation was delivered using a compact disc (CD) player with headphones. The passage and each test question with answer choices were recorded on separate tracks and students were allowed to replay the tracks. Passages were read at rates of 150-160 words per minute. Students had access to the test form in paper format as well as being able to listen to it.

Description of Samples

As previously mentioned, data used for this study was collected as part of a larger study carried out by Cahalan-Laitusis, Cook, Cline, King, and Sabatini (2008). For the larger study, all public schools in New Jersey with fourth and eighth grades were contacted and invited to participate. A total of 84 schools accepted the invitation. The full sample for the study included 1,181 fourth grade students (527 with reading-based learning disabilities and 654 with no

disabilities) and 847 eighth grade students (376 students with reading-based learning disabilities and 471 students without disabilities). Raw-score summary statistics for the Grade 4 Form S sample used for this study are shown in Table 1. It should be pointed out that within each group (RLD or NLD) students were randomly assigned to either the standard or read-aloud condition. The reader is referred to Cahalan-Laitusis, et al., for additional information on sampling and on the summary statistics and demographics of the fourth and eighth grade samples taking Forms S and T that were used for the larger study.

Table 1

Summary Statistics for Grade 4 Form S Factor Analysis Samples

Group	N	Mean	SD
NLD Standard (Group 1)	326	30.08	9.68
NLD Read-aloud (Group 2)	328	32.44	8.81
RLD Standard (Group 3)	258	19.18	9.05
RLD Read-aloud (Group 4)	269	24.36	8.81

It is clear, from an examination of the data shown in Table 1, that students without disabilities had higher mean scores than students with reading-based learning disabilities under both the standard and read-aloud conditions. It is also clear that the read-aloud test change was more beneficial (resulted in larger score gains) for the RLD students than for the NLD students.

Hypothesized Factor Structure

Standardized reading tests such as the GMRT are designed to assess proficiency in reading comprehension and thus to rank order individuals on a one-dimensional scale. According to Ozuru, Rowe, O'Reilly and McNamara (in press), the GMRT does not have a

theoretical basis and is not designed to measure specific diagnostic components. In an analysis that these researchers completed of Levels 7/9 and 10/12 of the GMRT, they concluded that, "…the GMRT is suited for assessing a broad range of abilities involved in reading comprehension from a variety of text materials in a broad stroke." (p. 27). Given this assessment and the manner in which the test is designed, it seems reasonable to hypothesize that the 48-item Comprehension subtest is a measure of a single dimension that could be labeled reading comprehension, although it was thought to be prudent to investigate the possibility of additional factors in the exploratory analyses.

Analyses

Input Matrices for Analyses. Variance-covariance matrices of item-parcel scores were used as input to all of the analyses carried out for this study. (See Rock, Bennett, & Kaplan, 1985; Cook, Dorans, & Eignor, 1988; and Steinberg, Cline & Sawaki, 2008, for a discussion of the use of item-parcel scores in factor analytical studies.) The 48-item Level 4 version of the GMRT was separated into 8 parcels of test items with 6 items in each parcel. The item parcels were balanced as much as possible to have similar levels of average difficulty. Table 2, shown below, provides information on the range of intercorrelations of the item parcel scores and the lower-bound reliabilities (Kuder-Richardson formula 20) of the scores on the individual parcels for each of the groups used in the study.

Table 2

Summary Data for Item Parcels (eight parcels each containing six items)

Group	Range of Parcel Intercorrelations	Parcel Reliabilities							
		1	2	3	4	5	6	7	8
NLD Standard (Group 1)	.479-.675	.548	.419	.539	.537	.606	.640	.555	.483
NLD Read-aloud (Group 2)	.427-.640	.500	.464	.543	.520	.583	.519	.529	.419
RLD Standard (Group 3)	.396-.587	.469	.383	.513	.501	.511	.520	.468	.407
RLD Read-aloud (Group 4)	.344-.610	.540	.331	.468	.379	.526	.430	.431	.351

Exploratory Factor Analyses (EFA). Exploratory factor analyses for each of the four groups were carried out using the computer program SAS. The inputs for each of the four analyses were the variance-covariance matrices of item parcel scores for the relevant group. Scree plots based on the 8 eigenvalues for each of the four single-group analyses were studied, along with the percentage of total test variance accounted for by the largest eigenvalue. SAS was used to explore the possibility of a single factor or two or three correlated factors accounting for the data for each of the four groups. Maximum likelihood procedures were employed and all solutions were rotated obliquely using promax rotation (Hendrickson & White, 1964). A factor loading of .30 or above was used as an arbitrary value to designate a salient factor loading when interpreting the results of the exploratory analyses.

Single-group Confirmatory Analyses (CFA). The single-group confirmatory analyses were carried out using EQS (Bentler & Wu, 2006). As was the case for the exploratory analyses, the inputs for the confirmatory analyses were variance-covariance matrices of item parcel scores for each of the four groups. Based on the results of the single-group exploratory analyses, only

the fit of a single-factor model was tested for each of the groups. The following goodness-of-fit criteria, largely based on Hoyle and Panter's (1995) suggestions, were used to test the overall fit of the models in both the single-group and multi-group confirmatory analyses;

- *Goodness of Fit Index (GFI):* An absolute model fit index, which is analogous to a model R^2 in multiple regression analysis. A GFI of .90 or above indicates an adequate model fit.
- *Non-Normed Fit Index (NNFI):* An incremental fit index, NNFI is an extension of the Tucker-Lewis Index (TLI) (Tucker & Lewis, 1973). An NNFI assesses whether a particular confirmatory factor analysis model is an improvement over a model that specifies no latent factors, taking into account the model complexity (Raykov & Marcoulides, 2000). An NNFI of .90 or above indicates an adequate model fit.
- *Comparative Fit Index (CFI):* 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.

Besides the indices above, two more criteria was taken into account:

- *Normal Theory Chi-Squar*e: A fit index that addresses the degree to which the variances and covariances implied by the specified model match the observed variances and covariances. The chi-square is expected to roughly equal its degrees of freedom. A ratio greater than 2 or 3 suggests important lack of fit.
- *Root Mean Square Errors of Approximation (RMSEA):* A RMSEA 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 a value of .08 or below as an indication of adequate fit (Browne & Cudeck, 1993).

Multi-group Confirmatory Analyses (CFA)

The multi-group confirmatory analyses were also carried out using EQS with variance-covariance matrices of item parcel scores used as inputs to the analysis. Informed by the results of the single-group confirmatory analyses, only the invariance of a single-factor structure was tested. The following set of nested models with increasing restrictions was tested: equality of factor loadings; equality of factor loadings and factor variances; equality of factor loadings, factor variances, and residuals. The goodness-of-fit indices described above that were used to evaluate the results of the single-group confirmatory analyses were also used to evaluate the overall fit of the nested models. Chi-square difference tests were also conducted to evaluate the fit of the nested models tested for the multi-group confirmatory analyses. Table 3, shown below, summarizes the analyses carried out for this study.

Table 3

Summary of Factor Analyses Carried Out for the Study

Analysis Number	Type	Questions to be Answered	Number of Hypothesized Factors	Level of Analysis
1	EFA	Number of Factors	--	Single Group
2	Single-Group CFA	Confirm Single Factor	1	Single Group
3a	Multi-Group CFA	Base-line Model	1	Two Groups
3b	Multi-Group CFA	Equality of Factor Loadings	1	Two Groups
3c	Multi-Group CFA	Equality of Factor Loadings and Variances	1	Two Groups
3d	Multi-Group CFA	Equality of Factor Loadings, Variances and Residuals	1	Two Groups

Results

Single-Group Exploratory Analyses (EFA)

As mentioned earlier in this report, items for the 48-item Level 4 GMRT Comprehension subtest were separated into 8 parcels balanced in both numbers of items and level of difficulty. The exploratory analyses were carried out separately for each of the four groups included in this study; Group 1, students without disabilities (NLD) who took the test under standard conditions; Group 2, students without disabilities (NLD) who took the test with a read-aloud test change; Group 3, students with reading-based learning disabilities (RLD) who took the test under standard conditions; and Group 4, students with reading-based learning disabilities (RLD) who took the test with a read-aloud test change. Figure 1, shown below, contains the scree plots for the 8 eigenvalues for each of the four groups.

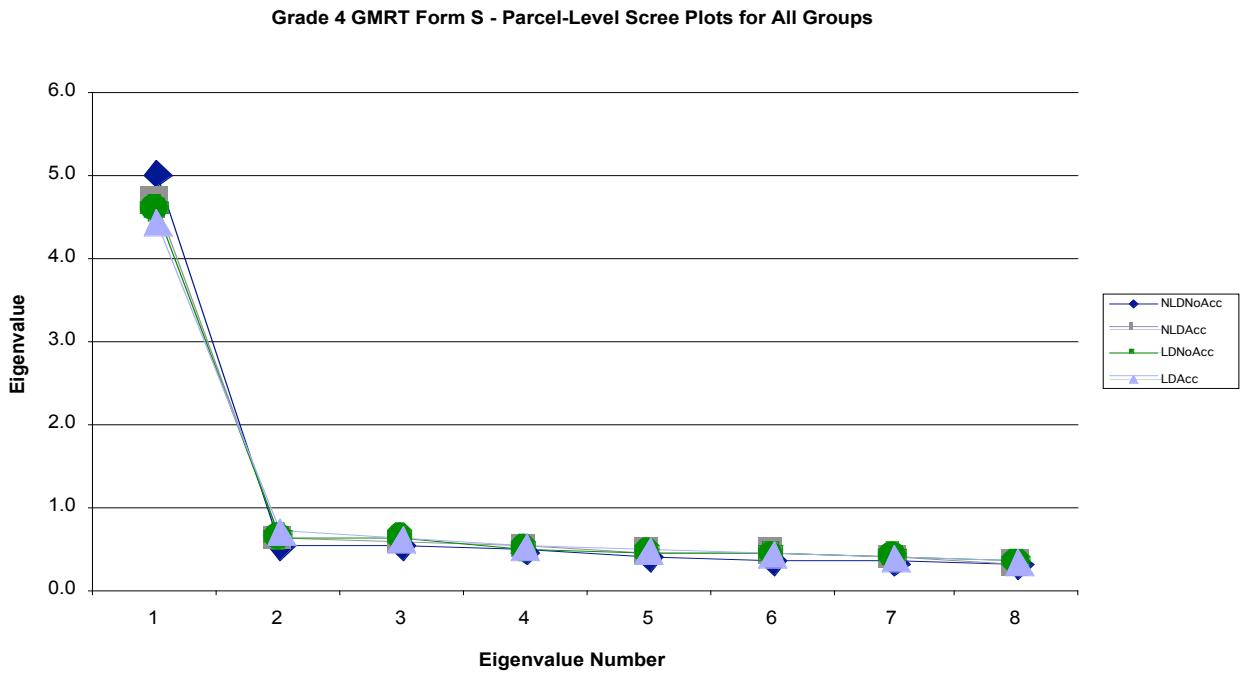

Figure 1. Scree Plots for the 8 Eigenvalues Obtained from the Exploratory Factor Analyses of the Level Four Comprehension Subtest

An examination of the information in the plots, indicates that, for all four groups, the test was measuring a single factor. The percentage of total variance accounted for by the largest eigenvalue for each of the four groups was: 58% for Group 1; 53% for Group 2; 51% for Group 3; and, 49% for Group 4.

For each of the four groups, from three correlated factors down to one factor were extracted and rotated obliquely using the promax method. The factor solutions were examined for each group using .30 as a cutoff value for a salient factor loading. The results of the analyses were very similar for all four groups. The two-and three-factor solutions extracted factors that were highly correlated. The correlations for the two-factor solution were; .75, .62, .58, and .71 for Groups 1-4, respectively. Also, the chi-square tests, indicating that the number of factors were sufficient, were non-significant for the single-factor solutions for all groups with the exception of Group 4. Based on these results, it would appear that the single-group exploratory analyses were indicating a single factor for each of the four groups studied and that using a hypothesis of one factor for the single-group CFAs was a reasonable next step.

Single-Group Confirmatory Factor Analysis (CFA)

Figure 2 contains a diagram of the confirmatory factor analysis model that was tested for both the single-group and the two multi-group analyses. It can be seen from this figure that the model hypothesized that the Level 4 GMRT Comprehension subtest measured a single factor labeled reading.

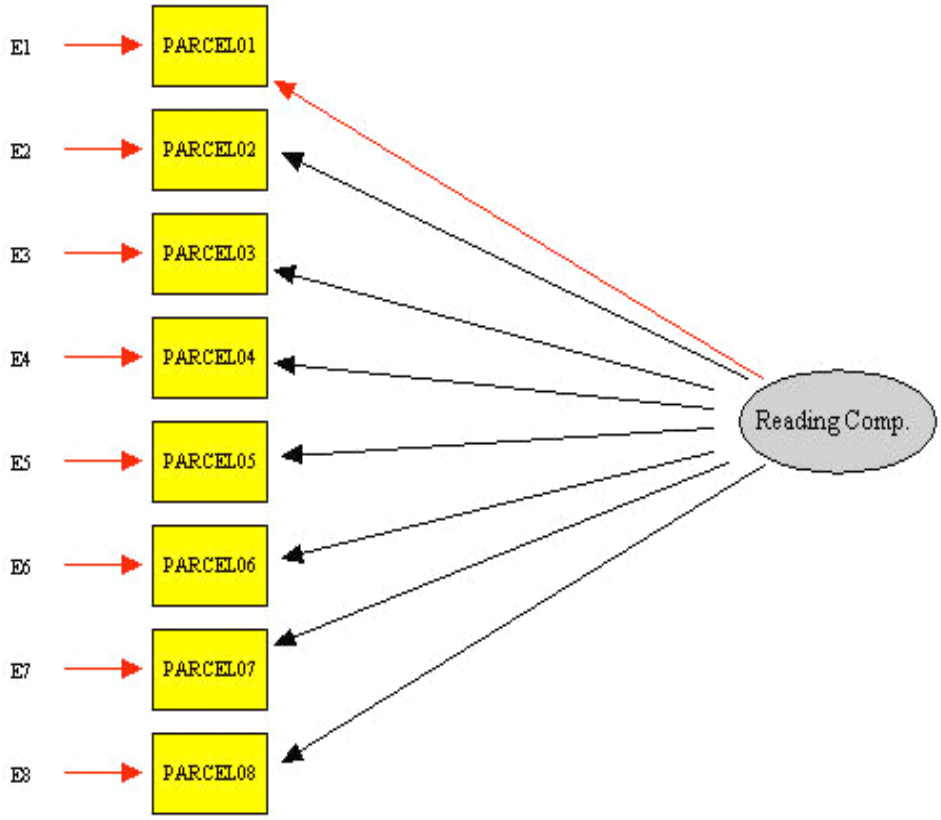

Figure 2. Confirmatory Factor Analysis Model used for Both Single-and Multi-group Analyses

Given that the exploratory analyses indicated that a single-factor solution would be reasonable for all four groups, we decided to verify this hypothesis by separately testing the fit of a one- factor model to each of the four groups used in the study (single-group confirmatory analyses) prior to testing the fit of a single-factor solution simultaneously to the two groups in each pair (multi-group confirmatory analyses).

The results of these analyses showed that a one-factor model fit the data well for all four groups. Table 4, shown below, contains the fit statistics for the one-factor, single-group confirmatory analyses carried out for each of the four groups. Mardia's coefficient (Mardia, 1970), computed for each group, (-.688, 1/275, -.830, 0.434; for groups 1-4, respectively)

indicated that the distributions of item-parcel scores are multivariate normal, supporting the use of the normal theory chi-square statistic to evaluate the fit of the models.

Table 4

Summary of Fit Statistics for Single-Group Confirmatory Analyses (one factor)

Group	DF	Normal Theory Chisq	GFI	NNFI	CFI	RMSEA
1	20	14.590	0.989	1.005	1.000	0.000
2	20	25.634	0.981	0.994	0.995	0.029
3	20	25.215	0.976	0.992	0.994	0.032
4	20	33.110	0.970	0.980	0.985	0.049

It can be seen, from examination of the fit statistics presented in Table 4, that all of the statistics evaluating the fit of a one-factor model to the data from each of the four groups met the criteria for model fit. Consequently, it was decided to proceed with the multi-group analyses, testing the invariance of a single-factor solution for each of the two pairs of groups simultaneously.

Multi-Group Confirmatory Factor Analysis (CFA)

The purpose of the multi-group analyses was to test for factorial invariance across Groups 1 and 2 and then, in a separate analysis, to test for factorial invariance across Groups 3 and 4. If factorial invariance could be demonstrated, it could be concluded that the read-aloud test accommodation/modification did not change the underlying construct measured by the GMRT Comprehension subtest for students without disabilities as well as for students with reading based-learning disabilities.

Factorial invariance was tested in four steps, carried out separately for each of the two pairs of groups (Group 1 vs. Group 2 and Group 3 vs. Group 4). The four steps taken were:

- Establish a base line multi-group model that establishes the fit of a single factor simultaneously to the two-groups used for each multi-group analysis
- Test invariance of factor loadings across two groups
- Test invariance of factor loadings and factor variances across two groups
- Test invariance of factor loadings, factor variances, and residuals across two groups

The results of the multi-group confirmatory analysis for Groups 1 and 2 are shown in Table 5 and those for Groups 3 and 4 are shown in Table 6.

Table 5

Summary of Multi-Group CFA for NLD Standard (Group 1) vs. NLD Read-aloud (Group2)

Model	DF[3]	Normal Theory Chisq	Chisq Difference	DF	P-value	RMSEA	GFI	CFI
Baseline	40	40.224				.004	.986	1.000
Invariance of factor loadings	47	50.078	9.854	7	0.197	.014	.981	.999
Invariance of factor loadings and factor variances	48	53.087	12.863	8	0.117	.018	.980	.998
Invariance of factor loadings, factor variances and residuals	56	63.644	23.420	16	0.103	.020	.976	.997

[3] The loading of the first parcel on the factor was arbitrarily set to 1 to set the scale for the factor and for model identification purposes.

Table 6

Summary of Multi-Group CFA for RLD Standard (Group 3) vs. RLD Read-aloud (Group4)

Model	DF[1]	Normal Theory Chisq	Chisq Difference	DF	P-value	RMSEA	GFI	CFI
Baseline	40	58.325				.042	.973	.990
Invariance of factor loadings	47	66.916	8.591	7	0.283	.040	.970	.989
Invariance of factor loadings and factor variances	48	67.044	8.719	8	0.367	.039	.970	.989
Invariance of factor loadings, factor variances and residuals	56	72.604	14.279	16	0.578	.034	.967	.991

[1] The loading of the first parcel on the factor was arbitrarily set to 1 to set the scale for the factor and for model identification p urposes.

It is very clear, from examination of the fit indices shown in Tables 5 and 6, that the hypothesis of factorial invariance across Groups 1 and 2 and across Groups 3 and 4 can clearly be accepted. In Table 5 all values of RMSEA are well below .05 and all values of both GFI and CFI are close to 1. The chi-square difference tests are all non-significant. The chi-square difference test that compares the baseline model and the model constraining factor loadings, factor variances, and residuals to be equal across the two groups has a value of 23.44, df =16 and p>.05. Similarly the information presented in Table 6 for the RLD comparisons shows the values of RMSEA all well below .05 and the values of the CFI and GFI all very high. The chi-square difference between the baseline model and the model constraining the factor loadings, factor variances and residuals to be equal across the two RLD groups equals 14.279, df=16, p>.05. The results of these analyses indicate that the GMRT Comprehension subtest measures the same underlying single factor regardless of whether it is administered under standard conditions or with a read-aloud test change. Furthermore, it is clear that this hypothesis is

equally true for students without disabilities and for students with reading-based learning disabilities.

Further information can be gained about the analysis from an examination of the information provided in Tables 7 and 8. Table 7 provides the unstandardized and standardized factor loadings and residuals for the NLD groups taking the Level 4 GMRT reading subtest under standard conditions and with a read-aloud accommodation/modification. Table 8 provides the same information, but for the two RLD groups.

Table 7

Summary of Unstandardized and Standardized Factor Loadings and Residuals for Multi-group Analysis for Students without Disabilities (NLD)

Parcel Number	Unstandardized Factor Loadings[1][2][3]	Unstandardized Residuals[4]	Standardized Factor Loadings	Standardized Residuals
1	1.000 (-)	0.868 (0.055)	0.757	0.653
2	0.882 (0.050)	1.036 (0.063)	0.683	0.730
3	0.972 (0.053)	1.076 (0.066)	0.712	0.703
4	1.008 (0.053)	1.006 (0.063)	0.736	0.677
5	1.183 (0.055)	0.809 (0.056)	0.818	0.575
6	1.131 (0.055)	0.941 (0.062)	0.783	0.622
7	1.095 (0.055)	0.977 (0.063)	0.768	0.641
8	0.889 (0.051)	1.108 (0.067)	0.674	0.739

[1] Standard errors are given in parenthesis following the loading.
[2] All factor loadings are significant at $p < .05$
[3] The loading of the first parcel on the factor was arbitrarily set to 1 to set the scale for the factor and for model identification purposes.
[4] Standard errors are given in parenthesis following the residual.

Table 8

Summary of Unstandardized and Standardized Factor Loadings and Residuals for Multi-group Analysis for Students with Reading-based Learning Disabilities

Parcel Number	Unstandardized Factor Loadings [123]	Unstandardized Residuals [4]	Standardized Factor Loadings	Standardized Residuals
1	1.000 (-)	1.055 (0.075)	0.729	0.684
2	0.818 (0.058)	1.141 (0.077)	0.643	0.766
3	0.950 (0.063)	1.239 (0.085)	0.683	0.730
4	0.957 (0.062)	1.110 (0.078)	0.705	0.709
5	1.096 (0.064)	0.947 (0.072)	0.777	0.630
6	0.997 (0.062)	1.057 (0.076)	0.728	0.685
7	1.022 (0.062)	0.974 (0.071)	0.750	0.661
8	0.852 (0.060)	1.194 (0.081)	0.650	0.760

[1] Standard errors are given in parenthesis f ollowing the loading.
[2] All factor loadings are significant at p< .05
[3] The loading of the first parcel on the factor was arbitrarily set to 1 to set the scale for the factor and for model identification purposes.
[4] Standard errors are given in parenthesis following the residual.

It can be seen, from an examination of the information provided in Tables 7 and 8, that all estimates are reasonable, the factor loadings are all statistically significant and that the standard errors are in good order.

Discussion

The purpose of this study was to investigate the impact of a read-aloud test change administered with the Gates-MacGinitie Reading Test (GMRT) on the underlying constructs measured by the Comprehension subtest. The study evaluated the factor structures for the Level 4 Comprehension subtest given to a sample of New Jersey fourth-grade students with and without reading-based learning disabilities. Both exploratory and confirmatory factor analyses were used to determine whether or not the GMRT Comprehension subtest measures the same

underlying constructs when administered with and without a read-aloud test change. The specific questions examined in this study were:

1. Does the GMRT Comprehension subtest measure the same construct(s) for examinees without disabilities (NLD) who take the test under standard conditions as it does for NLD examinees who take the test with a read-aloud test change?
2. Does the GMRT Comprehension subtest measure the same construct(s) for examinees with reading-based learning disabilities (RLD) who take the test under standard conditions as it does for RLD examinees who take the test with a read-aloud test change?

The results of the analyses indicated factorial invariance held when the Comprehension subtest was administered to groups of students without disabilities who took the test under standard conditions and with a read-aloud test change and for groups of students with reading-based learning disabilities who also took the test under standard conditions and with a read-aloud test change. These results can be compared to the results of some of the studies cited earlier in this paper.

It will be recalled that a study by Fuchs, Fuchs, Eaton, Hamlett, Binkley and Crouch (2000) that was carried out using a reading comprehension test given to elementary school students found that a read-aloud accommodation provided a differential boost for students with learning disabilities, thus providing support for this type of test change as an accommodation rather than a modification. However, three studies on the effects of a read-aloud presentation reviewed by Sireci, Li, and Scarpati (2003) found no gains for students with or without disabilities as a result of a read-aloud test change contradicting the Fuchs, et al., findings.

A few other studies that employed factor analysis to examine the internal structure of tests given with and without a read-aloud test change were mentioned earlier. One study, carried out by Huynh and Barton (2006) used confirmatory factor analysis to evaluate the effects of a read-aloud test change for students with and without disabilities on the South Carolina High School Exit Examination. Similar to the results of the study reported in this paper, these researchers concluded that a one-factor model could be used to describe the data for the accommodated and regular form. On the other hand, Meloy, Deville, and Frisbie (2000) factor analyzed data from the Iowa Test of Basic Skills given to students with disabilities with and without a read-aloud test change, and concluded that the read-aloud condition changed the construct being measured by the assessment.

Studies of Differential Item Functioning (DIF) sometimes have as their purpose, demonstrating that test items have the same meaning for different groups of students. DIF has been used by several researchers to examine the impact of accommodations and modifications on test scores, including read-aloud test changes. Bolt (2004) compared samples of students on three state assessments of reading or English-language arts. In all three studies by Bolt, the read-aloud test change resulted in considerable levels of DIF, indicating a change in construct. On the other hand, Pitoniak, Cook, Cline, and Cahalan-Laitusis (in press) examined DIF on a state assessment of English-language arts given to 4^{th} and 8^{th} grade students taking the test with and without accommodations, including a read-aloud test change. These researchers found only one item in each grade flagged as having moderate to large DIF.

Given the disparity of findings that exist for the few studies discussed above, and also the findings of the current study, it is difficult to generalize from the results of the current study to draw firm conclusions regarding the impact of a read-aloud test change on the underlying

constructs measured by a reading test. The reasons for the disparate findings of the studies discussed here, including the results of the current study, are most certainly complex; but are surely related to the fact that the reading tests studied differed in fundamental ways as did the populations of students used for the studies. Although the results of the analyses we carried out for GMRT Level 4 Form T given to 4th grade students and Level 7/9 Forms T and S given to 8th grade students were not reported here, the findings from these analyses were very similar to those presented in this paper. Consequently, we believe that it would be safe to conclude, from the research we have carried out on the GMRT, that there is some evidence that the Comprehension subtests for both Level 4 given to fourth grade students and Level 7/9 given to 8th grade students are most likely measuring the same underlying constructs regardless of whether they are given with or without a read-aloud test change.

However, given the research that has been carried out to date that provides contradictory evidence to this finding, we would hesitate to generalize our findings beyond this test (GMRT Comprehension subtest) and these samples. As more research on the question of the impact of a read-aloud test change on reading test scores is carried out, a clearer picture will most certainly emerge. Until that point in time, it would behoove anyone who is concerned about this question to view it as one with an empirical basis, and consequently, to address this question by carrying out the necessary analyses using data from the tests and student populations of interest.

References

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

Bolt, S. E. (2004). *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 Bollen, K. A., & Long, J. S. (Eds.), *Testing structural equation models* (pp. 136-62). Newbury Park, CA: Sage.

Cahalan-Laitusis, C., & Cook, L. (2008). Reading Aloud as an Accommodation for a Test of Reading Comprehension. Retrieved February 7, 2009, from: http://www.ets.org/Media/Research/pdf/SPOTLIGHT1.pdf.

Cahalan Laitusis, C. Cook, L., Cline, F., King, T., & Sabatini, J. (2008). *Examining the impact of audio presentation on tests of reading comprehension* (ETS RR-08-23). Princeton, NJ: Educational Testing Service.

Cahalan-Laitusis, C., Cook, L. L., & Aicher, C. (2004). *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

Clapper, A. T., Morse, A.B., Lazarus, S. S., Thompson, S. J., & Thurlow, M. (2005). *2003 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved March 29, 2006, from: http://education.umn.edu/NCEO/OnlinePubs/Synthesis56.html.

Cook, L. L., Dorans, N. J., & Eignor, D. R. (1988). An assessment of the dimensionality of three SAT – Verbal test editions. *Journal of Educational Statistics, 13,* 19-43.

Cook, L. L., Eignor, D. R., Sawaki, Y., Steinberg, J., & Cline, F. (in press). *Using factor analysis to investigate the impact of accommodations on the scores of students with disabilities on English-language arts assessments* (ETS RR- XX-XX*).* Princeton, NJ: Educational Testing Service.

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

Fuchs, L. S., Fuchs, D., Eaton, S., Hamlett, C. L., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children, 67*(1), 67-81.

Harker, J. K., & Feldt, L. S. (1993). A comparison of achievement test performance of non-disabled students under silent reading and reading plus listening modes of administration. *Applied Measurement in Education, 6*, 307-320.

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.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In Hoyle, R.H. (Ed.), *Structural equational 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 Education Act of 1997, 20 U.S.C 1412 (a) (17) (A). (1997).

Kosciolek, S., & Ysseldyke, J. E. (2000). *Effects of a reading accommodation on the validity of a reading test* (Technical Report 28). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved May 2005, from: http://education.umn.edu/NCEO/OnlinePubs/Technical28.htm.

Mardia, K. V.(1970). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhva, B36*, 115-128.

McKevitt, B. C., & Elliot, S. N. (in press). The effects and consequences of using testing accommodations on a standardized reading test. *School Psychology Review.*

Meloy, L. L., DeVille, C., & Frisbie, D. (2000, April). *The effect of a reading accommodation on standardized test scores of learning disabled and non learning disabled students*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

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

No Child Left Behind Act, Report to the Ranking Minority Member, Committee on Health, Education, Labor, and Pensions, U.S. Senate (2005). GAO-05-618, Washington, DC.

Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D.S. (in press). Where's the difficulty in standardized reading tests: The passage or the question? *Behavior Research Methods.*

Pitoniak, M., Cook, L., Cline, F., & Cahalan-Laitusis, C. (in press) *Using differential item functioning to investigate the impact of accommodations on the scores of students with disabilities on English-language arts assessments* (ETS RR-XX-XX*)*. Princeton, NJ: Educational Testing Service.

Pitoniak, M., & Royer, M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71,* 53–104.

Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling.* Mahwah, NJ: Erlbaum.

Rock, D. A., Bennett, R. E., Kaplan, B., & Jirele, T. (1988). Factor structure of the Graduate Record Examinations General Test in handicapped and nonhandicapped groups. *Journal of Applied Psychology, 73*, 383-392.

Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodations on test performance: A review of the literature* (Center for Educational Assessment Research, Report No. 485). Amherst, MA: School of Education, University of Massachusetts Amherst.

Steinberg, J., Cline, F., & Sawaki, Y. (2008, March). *Examining the internal validity of a state standards-based assessment of science*. Paper presented at the annual meeting of the American Research Association, New York.

Tindal, G., & Fuchs, L., (2000). A summary of research on test changes: An empirical basis for defining accommodations (ERIC Report ED442245). Washington, DC.

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.

Tucker, L. R. & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.