



Research Report
ETS RR-11-38

**Examining the Factor Structure of
a State Standards-Based Science
Assessment for Students with
Learning Disabilities**

Jonathan Steinberg

Frederick Cline

Yasuyo Sawaki

September 2011

**Examining the Factor Structure of a State Standards-Based Science Assessment for
Students with Learning Disabilities**

Jonathan Steinberg, Frederick Cline, and Yasuyo Sawaki
ETS, Princeton, New Jersey

September 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

<http://www.ets.org/research/contact.html>

Technical Review Editor: Daniel Eignor

Technical Reviewers: Linda Cook and Donald Rock

Copyright © 2011 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, GRE, and LISTENING. LEARNING.
LEADING. are registered trademarks of Educational Testing
Service (ETS).

SAT is a registered trademark of the College Board



Abstract

This study examined the scores on a state standards-based Grade 5 Science assessment obtained by a group of students without learning disabilities who took the standard form of the test and by three groups of students with learning disabilities: one taking the standard form of the test without accommodations or modifications, a second taking the test with accommodations, and a third group taking the test with modifications. The groups received accommodations or modifications that were specified in their 504 or IEP plans. Unlimited time was granted to complete the test. A series of item-level, then parcel-level, exploratory and confirmatory factor analyses investigated whether or not the assessment demonstrated factorial invariance for the four groups of students studied. The results of this study help substantiate the validity of test scores for students with disabilities who take the test with the particular set of accommodations or modifications that were used in this study. In addition, the results lend support to modifying this state's policy in order to aggregate scores obtained by students without learning disabilities and by students with learning disabilities who have taken the assessment with accommodations or modifications required by their 504 plans or IEPs for AYP purposes.

Key words: science assessment, students with learning disabilities, factor analysis, factorial invariance, item parcels

Acknowledgments

We would like to thank many people for their contributions to this report. We recognize the contributions of our ETS colleagues: Linda Cook and John Young, who jointly directed the project under which this research was conducted, through a grant led by Cara Laitusis; John Cope and Elizabeth Stone for helping procure access to the data; Diana Munoz and Joe Sipper for their leadership roles in Assessment Development for the state under study; Yeonsuk Cho, Dan Eignor, and Guangming Ling for their insights and support throughout the project; our reviewers, Linda Cook, Don Rock, and Dan Eignor; and Ruth Greenwood for her efforts in copyediting this report. A previous version of this paper was presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, held March 24–28, 2008 in New York, NY.

Table of Contents

1. Background.....	1
2. Review of Relevant Research.....	4
2.1 Experimental Studies on Accommodations.....	5
2.2 Studies Focusing on Science.....	6
2.3 Studies of Internal Test Structure.....	7
3. Overview of the Study.....	11
4. Methods.....	12
4.1 Description of the Test.....	12
4.2 Description of the Samples.....	13
4.3 Description of Possible Factor Structure.....	14
5. Analyses.....	14
5.1 General Descriptive and Psychometric Statistics.....	14
5.2 Item-Level Factor Analyses.....	18
5.3 Parcel-Level Factor Analyses.....	22
6. Discussion and Conclusions.....	31
References.....	34
List of Appendices.....	42

List of Figures

Figure 1. Percent correct on the total test and by strand by group.....	15
Figure 2. Partial scree plot for groups 1, 2, and 3.....	20
Figure 3. Complete scree plot for all groups.....	23
Figure 4. Proposed Grade 5 Science three-factor parcel design.....	25

List of Tables

Table 1	List of Approved Accommodations and Modifications Used in Grade 5 Science	3
Table 2	Number of Items by Content Area and Grade Level for Grade 5 Science	12
Table 3	Raw Score Summary Statistics for Grade 5 Science Factor Analysis Samples	13
Table 4	Total Test and Strand Length-Adjusted Reliabilities by Group	15
Table 5	Correlations of Observed Strand Scores to Total Test Score	16
Table 6	Summary of Proposed Factor Analyses.....	17
Table 7	Original Grade 5 Science Parcel Design.....	22
Table 8	Summary of Parcel Factor Loadings on General Science Factor	24
Table 9	Summary of Individual-Group Parcel-Level Confirmatory Factor Analysis Results ...	26
Table 10	Latent Factor Intercorrelation Matrices From Three-Factor Individual-Group Confirmatory Factor Analysis Models.....	27
Table 11	Summary of Proposed Parcel-Level Multi-Group Confirmatory Factor Analyses	28
Table 12	Summary of Parcel-Level Multi-Group Confirmatory Factor Analysis Results	29
Table 13	Summary of Revised Parcel-Level Multi-Group Confirmatory Factor Analysis Results	30
Table 14	Summary of Standardized Factor Loadings for Final Factor Invariance Model	30

1. Background

The historical record regarding accommodations and modifications for testing examinees with disabilities dates back over seven decades. The implications of interpreting test scores that result from accommodated or modified administrations have been debated for just as long. Pitoniak and Royer (2001) provided a thorough context for understanding the evolution of the practice, its effects on score validity, and its place in mainstream society.

One of the first key distinctions mentioned in Pitoniak and Royer (2001) is between an accommodation and a modification in the context of testing. Hollenbeck, Tindal, and Almond (1998) distinguished the two terms as follows:

“Accommodations do not change the nature of the construct being tested, but differentially affect a student’s or group’s performance in comparison to a peer group...modifications result in a change in the test (how it is given, how it is completed, or what construct is being assessed)...because of the lack of interaction between “group” and “change in test,” the modification itself does not qualify as an accommodation. (pp. 175–176).

The definition of these two key terms led to a description by Pitoniak and Royer (2001) of the evolution of legal policy concerning those with disabilities. The highlights are discussed here.

As referenced by Rosenfeld, Tannenbaum, and Wesley (1995) in Pitoniak and Royer (2001), accommodations were first allowed on the SAT[®] in the 1930s. Fischer (1994) is quoted in Pitoniak and Royer (2001) discussing how the U.S. Civil Service Commission first thought about modifying assessments measuring job-related abilities for candidates with disabilities in 1946, with the validity investigations of such tests starting a decade later. From a legislative perspective, according to Pitoniak and Royer (2001), the Rehabilitation Act of 1973, the Americans with Disabilities Act (ADA) of 1990, and the Individuals with Disabilities Educational Acts (IDEAs) of 1991 and 1997 set the stage for current practices in this area.

Section 504 of the Rehabilitation Act of 1973 stipulated how federally-funded programs or activities must require the provision for accommodations to those with disabilities in major life activities (e.g., speaking, learning, and working) to ensure equal access and participation. The ADA extended the 1973 legislation to employment and educational opportunities. The IDEA requires Individual Educational Programs (IEPs) for students with disabilities so they receive

proper instruction and services (including testing) in the least restrictive environments that are possible. The 1997 revision to the IDEA mandated that students with disabilities be included in general assessments at the district and state levels when given appropriate accommodations to take these assessments.

The No Child Left Behind Act or NCLB (2001) reinforced the objectives of the 1997 IDEA amendments and focused on improving the education of students with disabilities in two ways: (a) requiring states and districts to report scores for subgroups of students, including students with disabilities; and (b) holding schools accountable for Adequate Yearly Progress (AYP) of these subgroups on a state's academic standards.

Elbaum, Arguelles, Campbell, and Saleh (2004) discussed how the rising level of participation by students with disabilities in statewide assessments has stimulated considerable research and discussion concerning how to appropriately assign testing accommodations, how accommodations impact performance for students with and without disabilities, and the validity of interpretations of that performance when students are granted particular accommodations. Of particular concern is whether or not the scores obtained on a test where accommodations and/or modifications are permitted have the same meaning as scores obtained on a standard administration of the test. A second very important question is whether accommodations on a standard assessment truly lead to more valid interpretations of scores for groups of students who receive the accommodations.

This study examined the scores on a state standards-based Grade 5 Science assessment obtained by students without learning disabilities who took the standard form of the test and by three groups of students with learning disabilities: (a) those taking the standard form of the test without accommodations or modifications, (b) those taking the test with accommodations as specified in their IEPs or 504 plans, and (c) those taking the test with state-approved modifications.

The overall goal of the study was to determine whether or not this assessment had the same internal structure for those students with varying learning disabilities compared to those without learning disabilities taking the test under standard conditions. This was investigated by factor analyses of item-level, and then parcel-level, scores for all groups studied, using those students without learning disabilities as the primary reference group. Specifically, this study addressed the following research questions:

1. Does the assessment measure the same construct(s) for examinees with learning disabilities who take the test under standard conditions as it does for the corresponding nondisabled population?
2. Does the assessment measure the same underlying construct(s) for examinees with learning disabilities who take the test with a test change (accommodations or modifications) as it does for the nondisabled population who take the test under standard conditions?

The accommodations and modifications that were evaluated in this study were a general set of accommodations and modifications specified in students' 504 plans or IEPs approved by this particular state for use in testing. The set of allowable accommodations and modifications on this Grade 5 Science test is described in Table 1.

Table 1

List of Approved Accommodations and Modifications Used in Grade 5 Science

Accommodations	Marked in test booklet	Had supervised breaks
	Dictated responses to scribe	Tested at most beneficial time of day
	Used non-interfering assistive device	Administered at home or in a hospital
	Used Braille test	Used a dictionary
	Used large-print test	Examiner presented with Manually Coded English (MCE) or American Sign Language (ASL)
	Tested over more than 1 day	Examiner read test questions aloud
Modifications	Used a calculator	
	Used an arithmetic table	
	Used math manipulatives	
	Used interfering assistive device	

It should be pointed out that the most frequently used supports for students on assessments are extra time and what is often referred to as an audio *accommodation*. For the

state under study, there were no time constraints for any students taking the test. When using an audio accommodation, a student either listened to a recording of the assessment questions and (sometimes) the assessment text, or the student had the questions/text read aloud to him or her by a human reader. Most states, including the state participating in this study, consider the audio accommodation for an English-language arts (ELA) assessment to actually be a modification. However, some states consider audio assistance to be an accommodation for science tests. The accommodation “Examiner read test questions aloud” displayed in Table 1 was used in this study. Therefore for this report, those students receiving the audio accommodation were included in the group with learning disabilities who took the test with accommodations required by their 504 plans or IEPs. Modifications are thought to change the construct(s) being measured by the assessment, and consequently this state, like many other states, does not currently aggregate scores obtained using modifications with other scores for NCLB purposes. This research will attempt to show whether the scores obtained by students using approved modifications are valid and have similar meanings compared to scores obtained by students taking the test with accommodations.

2. Review of Relevant Research

Some of the most common accommodations or modifications for students with reading-based learning disabilities were examined in the studies reviewed in this section. These accommodations or modifications were typically specified in 504 or IEP plans (including extra time and audio presentation, 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:

- the multiple types of accommodations or modifications that are employed.
- the variety and severity of disabilities of the examinees.
- controversy regarding how each accommodation or modification might or might not change the test’s construct(s).
- the inability to aggregate data across administrations because of database shortcomings (e.g., information about the type of accommodation or modification an individual receives 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 (Sireci, Li, & Scarpati, 2003).

Section 2.1 will focus on studies that explain differences in test performance between groups on a test, typically done under true experimental conditions¹. The focus has mostly been on reading and mathematics assessments, but some studies involving science assessments will be referenced in Section 2.2. Section 2.3 deals specifically with methods relating to examining the internal structure of a test.

2.1 Experimental Studies on Accommodations

The studies reviewed for this report indicate that the most common accommodations and modifications for students with reading-based learning disabilities are extra time and audio presentations. Research on extra time indicates that students with those types of disabilities do differentially benefit when compared with students without such disabilities (i.e., a differential boost² is demonstrated when the two groups are compared and students with disabilities achieved greater 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 an audio presentation on tests of reading or English-language arts is less conclusive than the research on timing, but will be described here.

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 not from extended time or from the use of large-print text.

Three studies on the effects of audio presentation reviewed by Sireci, Li, and Scarpati (2003) indicated no gains for students with or without disabilities (Kosciolek & Ysseldyke, 2000; McKeivitt & Elliot, 2003) or similar gains for both groups (Meloy, DeVille, & Frisbie, 2002). Sample sizes may have contributed to the different findings among the studies that tested the interaction model for differential boost. The Fuchs, et al. (2000) study had the largest total sample size ($n = 365$) and did detect a differential boost, while the study with the next largest total sample size (Meloy et al., 2002; $n = 260$) found similar gains for students with and without disabilities. The last two studies that tested the interaction model had small samples (31 in the

Kosciolek and Ysseldyke study and 79 in the McKevitt and 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 in the studies.

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 [LD]) in Grades 6 through 10. The researchers administered alternate forms of an assessment constructed of 3rd- to 5th-grade level reading passages with accompanying comprehension questions. All students first took the assessment in the standard condition and then with instructions to read the passages aloud at their own pace. The researchers found that “as a group of students, test performance did not differ in the two conditions, and students with LD did not benefit more from the accommodation than students without LD.” The researchers noticed, however, that the scores of LD students were more variable in the accommodated condition than were the scores of students without disabilities. They emphasized that the findings of their study, “. . . underscore the need to go beyond the interpretation of group mean differences in determining the validity of testing accommodations.”

The data analyzed for this report, as is typical with studies carried out using data from large-scale state assessments, are considered to have come from a *nonexperimental* design. This means that students only took the exam one time under one set of conditions compared to two times in a typical experimental design study where the standard form and a parallel form of the test are used.

2.2 Studies Focusing on Science

Compared to reading and mathematics, fewer studies examine performance on science assessments when accommodations or modifications are provided. However, the number of existing studies is still quite large compared to other academic subjects. Sireci, Scarpati, and Li (2005) mentioned a few studies that look at accommodations for students with disabilities, on science exams. Meloy, DeVille, and Frisbie (2002) looked at the read-aloud accommodation and its effect on performance of middle school students on the Iowa Test of Basic Skills (ITBS) in four subjects, one of which was science. There were 260 participants across Grades 6 through 8, and the ratio of students without disabilities to students with reading-based learning disabilities was 3 to 1. Random assignment of all participants to the read-aloud or standard condition was

made and was consistently applied to all four tests. As expected, students without reading-based disabilities performed better than those with such disabilities under both conditions on all four tests, and scores were higher on the tests taken with the read-aloud test change compared to those on the standard forms. This is consistent with the view of the interaction hypothesis according to Fuchs, Fuchs, Eaton, Hamlett, and Karns (2000).

A study by Brown and Augustine (2001) is referenced in the Sireci et al. (2005) paper as well, regarding the use of screen-reading software in administering publicly available National Assessment of Educational Progress (NAEP) science items to 96 participants. Standard and computer-read forms of the test were given. There was no significant difference in performance between conditions, after controlling for the reading ability of the students.

A study by Koretz and Hamilton (2000) used data from the Kentucky Instructional Results Information System (KIRIS) assessment where multiple-choice items were added to the previously exclusive open-ended response exam for all academic subjects. Given a matrix sampling design of 12 of the 28 multiple-choice items, the authors focused their attention on the common set of 16 items. The state's science assessment was given in Grades 4, 7, and 11.

Across grades, students with disabilities scored 0.7–1.0 standard deviations (SDs) below those students without disabilities. Comparisons were also made within the disability group between those taking the test with and without accommodations. The accommodations most often used were oral presentation, paraphrasing, and dictation. Across grades, those taking the test with accommodations scored between 0.1 and 0.3 SDs below those students taking the test without accommodations on the multiple-choice items. Comparing mutually exclusive groups of accommodated students, those receiving an oral reading of the test scored almost 0.2 SDs higher compared to those receiving paraphrasing. Those receiving an oral reading and dictation or paraphrasing performed at least 0.1 SDs better than those receiving only an oral reading accommodation. The results from this study need to be interpreted in the context that students with disabilities found these multiple-choice items to be more difficult and that the items were less discriminating for students with disabilities than for students without disabilities.

2.3 Studies of Internal Test Structure

Studies using differential item functioning (DIF). In addition to the 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 that of a comparison group of students matched on total test score 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 matched on total test score (students with disabilities who received no accommodations and students without disabilities who received no accommodations). The results indicated that 7–12% of the test items functioned differently for the focal group (students with learning disabilities who received read-aloud accommodations) when compared to either of the reference groups. Extra time was also examined, but no more than 1 percent 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 for other accommodations. Both of these studies provide evidence that a read-aloud accommodation may change the construct being assessed.

Factor analysis studies using items or subscores. Only a small number of published studies have examined and compared the factor structures of assessments given to students without disabilities with those of assessments given to students with disabilities under accommodated and nonaccommodated conditions. Tippetts and Michaels (1997) analyzed data from the Maryland School Performance Assessment Program (MSPAP) and found that scores obtained by students with disabilities who received accommodations and scores obtained by 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 (2002) compared factor structures for students with disabilities taking the Iowa Tests of Basic Skills assessments with a read-aloud accommodation and for students without a disability taking the assessments without such an accommodation. These authors concluded that the read-aloud accommodation appeared to change the construct being measured for most accommodated students relative to students who were assessed under standard conditions.

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 Grade 10

South Carolina High School Exit Examination (HSEE). Three groups of students were studied: (a) students without disabilities who took the regular form of the test, (b) students with disabilities who were given the regular form of the test, and (c) students with disabilities who were given the test with a read-aloud accommodation. The purpose of their study was to assess the comparability of accommodated and non-accommodated scores. One of the specific issues they addressed was whether or not the accommodation changed the internal structure of the test.

Initially, these authors carried out a principal components analysis on the correlation matrix of the assessment's six subscores for each group of examinees. This indicated a single factor was adequate to summarize the data for each group. This was followed by a multigroup maximum likelihood confirmatory factor analysis of the six sub-scores to determine whether a one-factor model could best describe the data for all three groups considered together. The authors concluded 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). They concluded that the accommodations provided on the South Carolina High School Exit Examination did not substantially change the test's internal structure and preserved the major construct underlying the test.

Cook, Eignor, Sawaki, Steinberg, and Cline (2006) used both exploratory and confirmatory factor analyses to investigate whether the Grade 4 ELA test in a large state measured the same construct(s) for students without disabilities taking the test without accommodations compared to mutually exclusive groups of students with disabilities taking the test without and with accommodations. Accommodations used by the latter group were defined by the students' 504 plans or IEPs. Interfactor correlations of the five content strands measured by the test were all generally high, indicative of a possible reduction in the number of factors present in the data.

Exploratory analyses based on tetrachoric correlations from item-level data suggested a single factor provided a more parsimonious fit to the data for all groups under study, compared to separate reading and writing factors. As is often the case with large-scale assessments across subjects when item-level data is used, there was a presence of a very large general first factor for all groups, despite a low proportion of explained variance. This was supported through maximum likelihood confirmatory factor analyses among the individual groups and then in a

multigroup setting, as all model fit indices were within acceptable ranges for single-factor models and no appreciable improvement in model fit was found for two-factor models.

Factor analysis studies using item parcels. Rock, Bennett, Kaplan, and Jirele (1988) examined the factor structure of the GRE[®] and the SAT for groups of examinees with and without learning disabilities. These authors felt that if the factor structure was the same for each test across these groups, this would lend support to the notion that the test scores have the same meaning for students with and without learning disabilities. At the time of the analysis, the SAT was composed of two tests (verbal and quantitative) and the GRE had three tests (verbal, quantitative, and analytical). These authors fit a two-factor model to the SAT and a three-factor model to the GRE, using 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 learning disabilities who took a cassette-recorded version of the test, but the factors were less correlated for this group than for examinees without learning 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 GRE verbal, quantitative, and analytical factors were examined for test-takers without disabilities who did not receive accommodations on the assessment and for test-takers with visual impairments or physical disabilities who did receive accommodations. While the parcels created for the verbal and quantitative sections each formed a single factor for examinees with disabilities, the parcels created for the analytical section broke out into two factors (logical reasoning and analytical reasoning), rather than the one overall hypothesized factor.

Cook, Steinberg, Sawaki, and Eignor (2008) examined a large-scale mathematics assessment in the same state and with the same groups as those described and used in this paper. Item parcels, balanced by difficulty within the five reported content strands, were created to facilitate interpretation of the internal structure of the test. The same three-factor structure emerged for all groups: (a) number sense; (b) algebra and functions combined with statistics, data analysis, and probability; and (c) measurement and geometry. However, the groups with learning disabilities each had a different variance on the latent measurement and geometry factor compared to the students without learning disabilities.

3. Overview of the Study

The analyses carried out for this Grade 5 Science examination from the selected state were similar to those employed in the Cook et al. (2008) study that made use of mathematics data for the state in a previous year. Exploratory and confirmatory factor analyses (EFA and CFA, respectively) were first conducted at the item level, starting from matrices of tetrachoric correlations of dichotomously-scored item-level data. The focus of the analyses was to determine and compare the number of factors that accounted for the data among four different groups taking the Grade 5 Science test. The four groups were:

- students without learning disabilities who took the test without accommodations (Group 1),
- students with learning disabilities who took the test without accommodations (Group 2),
- students with learning disabilities who took the test with accommodations defined by a 504 plan or IEP (Group 3), and
- students with learning disabilities who took the test with modifications specified in a 504 plan or IEP, including calculators, arithmetic tables or formulas, or manipulatives (Group 4).

Item-level exploratory factor analyses, first without then with rotation of factors, were carried out to obtain a preliminary indication of the possible dimensionality of the data. This analysis was to be followed by a series of confirmatory factor analyses separately for each group to determine the number of factors for the Grade 5 Science assessment.

However, there were limitations to following this item-level approach for each of the four groups examined in the study. As has been mentioned previously, sample size issues often arise in studies involving groups of students with specific disabilities or who need specific accommodations to take tests. The current study was conducted after a change in this particular state's internal database classification system for accommodations, thereby making it more difficult to obtain sufficiently large samples to conduct confirmatory factor analyses. This had not been an issue for the Cook et al. (2006) study. It was for this reason that Group 4 could not be included in the item-level factor analyses. Also, for reasons to be mentioned later, there are psychometric concerns with performing item-level factor analyses.

Therefore, to complete the analysis of the internal structure for all groups under study, the items were grouped into parcels within content strands, balanced for difficulty, so that each parcel would have a minimum of four items (Cook, Dorans, & Eignor, 1988). Variance-covariance matrices of parcel scores were used as inputs in model testing for this phase of the analysis. Individual-group and multi-group confirmatory models were then tested under a modified strand structure so that appropriate measurement model identification could be achieved and the research questions could be properly answered. The ultimate objective was to demonstrate the similarity, or invariance, of the factor structure across groups.

4. Methods

4.1 Description of the Test

The Grade 5 Science assessment consisted of 60 multiple-choice items, covering three major content areas (Physical Science – 20 items, Life Science – 21 items, and Earth Science – 19 items) and was a two-year cumulative exam so that each content area had Grade 4 and Grade 5 components. Students received scores on six reported content strands and this design reflected the state content standards. Table 2 provides a summary of the content strands and the numbers of items making up each strand.

Table 2

Number of Items by Content Area and Grade Level for Grade 5 Science

Content area	Grade 4	Grade 5	Total items
Physical science	8	12	20
Life science	10	11	21
Earth science	7	12	19
Total items	25	35	60

The test placed more emphasis on material covered in Grade 5 (35 items) than Grade 4 (25 items). Most questions were stand-alone items without a stimulus, but some questions were grouped by a common stimulus such as a diagram.

4.2 Description of the Samples

This study evaluated the factor structure of the Grade 5 Science assessment when administered to four groups of students: students without disabilities who took the test under standard conditions (Group 1), students with learning disabilities who took the test under standard conditions (Group 2), students with learning disabilities who took the test with an accommodation defined in their 504 plan or IEP (Group 3), and students with learning disabilities who took the test with a modification specified in their 504 plans or IEPs (Group 4).

The size of Group 1 was initially well over 230,000 students, so in preparation for the item-level analyses, a random sample of 30,000 examinees was created, primarily to facilitate the running of the various computer programs used in the analyses. Group 2 was reduced to roughly 10% of its original size while Groups 3 and 4 were left intact. For the parcel-level analyses Groups 1, 2, and 3 were sampled down to 500 students, so that the results could be more comparable to those of Group 4. Sample sizes and raw score summary statistics for all groups for both the item-level and parcel-level analyses are shown in Table 3, with raw score summary statistics for the sampled-down groups (Groups 1 and 2) being comparable to their respective total samples.

Table 3

Raw Score Summary Statistics for Grade 5 Science Factor Analysis Samples

Group	Item-level analyses			Parcel-level analyses		
	<i>N</i>	Mean	SD	<i>N</i>	Mean	SD
1 – Non-LD	30,000	34.5	10.3	500	35.0	9.9
2 – LD, no accommodations	11,394	23.4	9.1	500	23.3	8.9
3 – LD, IEP/504 accommodations	3,231	23.4	8.5	500	23.0	8.0
4 – LD, IEP/504 modifications	295	22.8	9.1	295	22.8	9.1

It can be seen from a review of the information provided in Table 3 that, as expected, the performance of students without learning disabilities (Group 1) was approximately one SD above that of students with learning disabilities who took the test without accommodations (Group 2). Additionally, the performance of students with learning disabilities who took the test with accommodations (Group 3) and of students taking the test with modifications (Group 4) were not

substantially different from that of Group 2. More information on the construction of the samples used in the parcel-level analyses will be provided later. However, it is evident that the performance of the samples used in the parcel-level analyses were not significantly different from comparable samples used in the item-level analyses.

4.3 Description of Possible Factor Structure

An initial hypothesis about the underlying structure of the test was that it corresponded to the existing strand structure, that is, the test had six underlying factors, one for each strand. Such a configuration made sense given that separate scores were reported for each of the six strands. Given the description of the strands, these were likely to be correlated.

However, a closer look at the items, both within and across the strands, reinforced the observation that the strands were likely to only be conceptual in nature, and may not in any way have corresponded to the underlying empirical structure of the test. It might then be reasonable to hypothesize that the test had either three underlying factors, one for each content area (Physical Science, Life Science, and Earth Science), or two underlying factors, one for Grade 4 content and one for Grade 5 content, and that in each instance these factors would be correlated. Finally, since the test is by title a science assessment, it was reasonable to hypothesize that the test had only one underlying factor, accounting for data from each content area and each grade level. All items appeared to heavily depend on the examinee being able to read and understand a wide array of scientific concepts and terms, possibly suggesting one underlying factor. The exploratory and confirmatory factor analyses that were conducted were influenced by the various hypotheses about the number of underlying dimensions that explained the data, and whether these were similar in nature across groups.

5. Analyses

5.1 General Descriptive and Psychometric Statistics

Table 3 displayed the raw score summary statistics for the groups of interest on the entire test, but it was worth looking at performance on the individual strands to see if any patterns emerged to help form hypotheses for factor analyses to be conducted later. Figure 1 shows each group's performance measured by percent correct³ on the total test and each of the six strands.

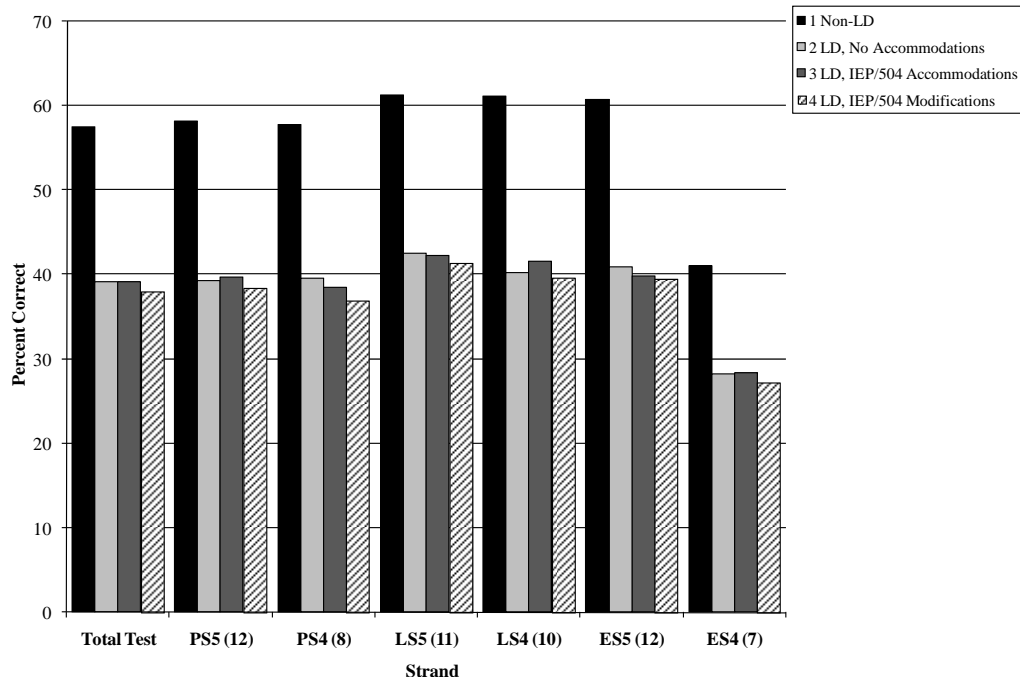


Figure 1. Percent correct on the total test and by strand⁴ by group.

We begin to see differing patterns of performance between Group 1 and Groups 2, 3, and 4. We also see that content area performance tended to be ordered by Life Science, then Physical Science and Earth Science. All groups were weakest on the Earth Science Grade 4 strand. Grade 5 strand performance tends to be approximately equal to that of Grade 4, with the exception of Earth Science. Table 4 displays the Cronbach’s Alpha reliabilities for the total test and at the strand level. Strand-level values are adjusted for strand length.⁵

Table 4

Total Test and Strand Length-Adjusted Reliabilities by Group

Group	Total test	PS5 (12)	PS4 (8)	LS5 (11)	LS4 (10)	ES5 (12)	ES4 (7)
1: Non-LD	0.89	0.75	0.81	0.77	0.80	0.77	0.84
2: LD, no accommodations	0.85	0.74	0.78	0.75	0.77	0.75	0.76
3: LD, IEP/504 accommodations	0.82	0.73	0.78	0.71	0.76	0.73	0.74
4: LD, IEP/504 modifications	0.85	0.74	0.76	0.75	0.77	0.75	0.78

The unadjusted strand reliabilities ranged from 0.31 to 0.63. Based on Spearman-Brown adjusted values, the reliabilities of the individual strands were closer to the reliabilities for the total test in each group. This allows for factor solutions generated from the item-level data positing that the number of factors is equal to the number of strands to be readily interpretable.

Finally, Table 5 displays the correlation of observed strand scores to total test score.

Table 5
Correlations of Observed Strand Scores to Total Test Score

Group	PS5 (12)	PS4 (8)	LS5 (11)	LS4 (10)	ES5 (12)	ES4 (7)
1: Non-LD	0.81	0.74	0.80	0.82	0.82	0.71
2: LD, no accommodations	0.79	0.69	0.76	0.77	0.80	0.58
3: LD, IEP/504 accommodations	0.75	0.67	0.73	0.76	0.77	0.56
4: LD, IEP/504 modifications	0.78	0.65	0.78	0.77	0.77	0.68

The range of correlations across groups and strands based on the item-level data ranges from 0.56 to 0.82, which is quite a large range. The observed intercorrelations of the strands range from 0.29 to 0.63 (raw) and 0.61 to 1.00 (adjusted for attenuation) across groups, which may raise issues about whether a measurement model could be represented by the individual strands based on item-level data. Appendix A displays the matrices of observed intercorrelations among the strands adjusted for attenuation. While observed descriptive statistics are helpful in interpreting how different groups of students did on the test, formal factor analyses must also be conducted before making a decision on the number of factors to use. Table 6 shows the factor analyses to be conducted and the sections of the paper where these steps are referenced.

Table 6***Summary of Proposed Factor Analyses***

Level	Analysis	Type	Objective	Number of factors expected	Groups	Section
Item	1	Eigenvalue estimation	Rough estimate of number of factors	NA	1,2,3	5.2
	2	Common exploratory factor analysis – no rotation	Rough estimate of number of factors	1	1,2,3	5.2
	2	Common exploratory factor analysis – rotations performed	Finer indication of the number of factors	3 or 6	1,2,3	5.2
	3	Confirmatory factor analysis	Final indication of the number of factors	1 or 3	1,2,3	5.2
Parcel	4	Eigenvalue estimation	Rough estimate of number of factors	NA	All	5.3
	5	Common exploratory factor analysis – no rotation	Rough estimate of number of factors	1	All	5.3
	5	Common exploratory factor analysis – rotations performed	Finer indication of the number of factors	3	All	5.3
	6	Confirmatory factor analysis – Part 1	Final indication of the number of factors	1 or 3	All	5.3
	7	Confirmatory factor analysis – Part 2	Tests to determine whether factor structure is invariant across groups	1	All	5.3

5.2 Item-Level Factor Analyses

As previously mentioned, the factor analyses performed in this study were first done at the item level. However, the analyses could not be done for all groups, due to sample size limitations in conjunction with test length.⁶ For that reason, only individual-group confirmatory analyses were conducted using item-level data where sample sizes were sufficient. Since matrices of tetrachoric correlations were used in these analyses, the following is a brief justification of their uses in these circumstances.

Exploratory linear factor analysis of test-level data from a single population can be conducted utilizing a covariance matrix. However, when the data to be come from dichotomously-scored items, as used here, a sample covariance matrix, may in some instances, when used in confirmatory fashion, lead to incorrect inferences about the underlying structure of the data (Hoyle & Panter, 1995). In such a situation, a correlation matrix should be used. The next question is whether to use phi correlations (product-moment correlations at the item level) or tetrachoric correlations.

Problems inherent in the factor analysis of item-level phi coefficients are well documented (Carroll, 1945; Mislevy, 1986; Cook, Dorans, & Eignor, 1988). Much of the early discussion focused on the fact that a factor analysis of phi coefficients typically resulted in factor solutions containing artifactual or *difficulty* factors. McDonald and Ahlawat (1974) stated that the artifactual factors are not due to the way the items are scored (i.e., dichotomously), but rather result from the fact that a nonlinear model is needed to characterize the regression of the scores on the underlying factors, instead of the assumed linear model. However, because of the assumptions underlying the creation of tetrachoric correlations, artifactual difficulty factors should not result (Christofferson, 1975; Muthen, 1978, 1989), or rather that a linear factor model should be appropriate. With phi correlations, the assumption of a linear model is inappropriate.

The use of tetrachoric correlations is not free from other problems. Carroll (1945) documented the problems involved with a linear factor analysis of tetrachoric correlation coefficients based on binary-scored multiple-choice items, where guessing is possible, and has provided formulas to correct the tetrachoric correlations for the effects of guessing. Mislevy (1986) discussed that even if tetrachorics have been corrected for guessing, the sample

tetrachoric correlation matrix may not necessarily be positive definite, in which case a factor solution may not be obtainable.

In lieu of using a linear factor analysis of tetrachorics resulting from dichotomously-scored data in exploratory analyses, a number of researchers have developed nonlinear factor analytic procedures to be used with such data (Bock, Gibbons, & Muraki, 1988; Fraser & McDonald, 1988). Such procedures operate directly on the item scores rather than on the correlation matrix, and are similar to multidimensional item response theory techniques. Waller (1991) pointed out that in an extensive simulation study done by Knol and Burger (1988), multiple linear and nonlinear factor analytic procedures for dealing with binary-scored item data were employed, and these authors concluded that “a common [iterated] factor analysis of the matrix of tetrachoric correlations yields the best estimates” (p. 199). Since the Knol and Burger study was a simulation study, “best” could be defined as the method that generated the smallest mean-squared error of parameter recovery. Based on these results, Waller went on to apply the linear factor analysis of tetrachoric correlations to data from the Minnesota Multiphasic Personality Inventory (MMPI) in looking at its underlying structure.

The item-level exploratory factor analyses in this study were therefore conducted employing a linear factor analytic model with a matrix of tetrachoric correlations. The program PRELIS (Joreskog & Sorbom, 2005b) produced the tetrachoric correlation matrices. There is no specific correction for guessing in producing these estimates, but the aim is to look at interfactor correlations and factor loadings, and guessing mostly affects factor intercepts, which were not examined.

It is important to discuss concerns raised by Mulaik (1972) about the use of correlation matrices. Comparing factors derived from correlation matrices instead of variance-covariance matrices computed for samples from different experimental populations, “violates the principle that the analysis must be in the same metric for the factor-pattern-matrix coefficients to be comparable across populations. Using correlation coefficients in each analysis forces the variables to have unit variances in each population, thereby creating a different matrix for each population. Before factors obtained from correlation matrices are compared, the factor-pattern matrices should be modified to express a common metric for the observed variables” (p. 356). Mulaik’s concerns were addressed by directly comparing factor structures across groups, not based on the results from item-level confirmatory factor analyses, but rather based on multi-

group confirmatory factor analyses that utilized variance-covariance matrices of item parcel scores.

Item-level exploratory factor analyses. The matrices of tetrachoric correlations 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 more than one factor). Scree plots (Child, 1970) of the first 12 eigenvalues computed from the matrix of tetrachoric correlations between items are displayed in Figure 2 for Groups 1, 2, and 3 since Group 4 was omitted from the item-level analyses. The scree plot was generated from preliminary eigenvalues generated from the tetrachoric correlation matrix with prior communality estimates on the diagonals.⁷

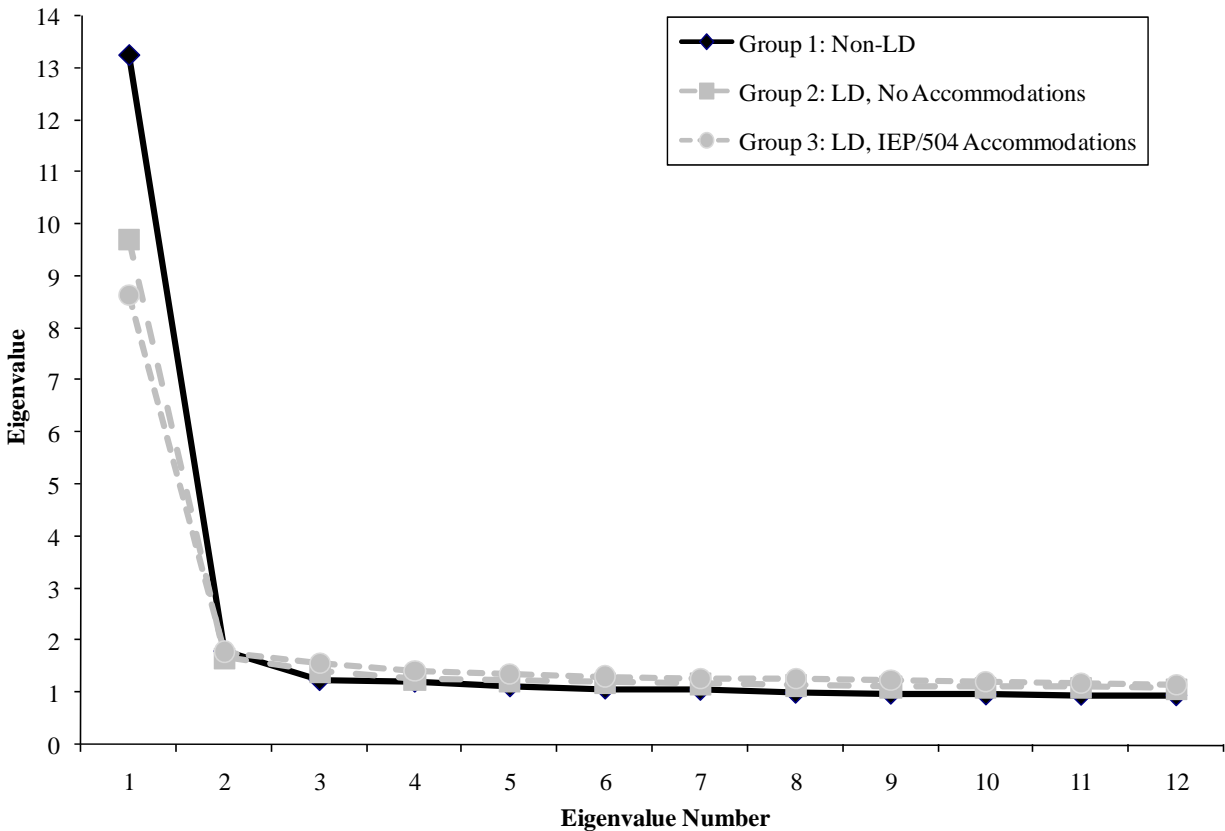


Figure 2. Partial scree plot for groups 1, 2, and 3.

The first eigenvalue in each group is very large compared to subsequent eigenvalues. The first eigenvalue explained 22% of the variance in the data for Group 1, 16% of the variance for

Group 2, and 14% of the variance for Group 3. More noise in the data was likely present for the groups with learning disabilities. The low proportion of explained variance was expected from the item-level data.

It is clear from the scree plot that beyond the first dimension, there was little additional explanatory power to be gained from further dimensionalizing the data. Attempts to extract more dimensions using promax rotation so that the factors could be correlated (Hendrickson & White, 1964) sometimes led to Heywood cases for some groups, which was evidence of possibly overfactoring the data. Also, chi-square tests hypothesizing whether the number of factors extracted was sufficient yielded significant *p*-values even for a six-factor structure (i.e., indicating that more than six factors are needed), was dubious in this context. Therefore, more weight was placed on the results from the confirmatory factor analyses for the following reasons: (a) items may be assigned to specific factors, which is not the case in exploratory models; (b) 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 (c) the fit indices obtained from confirmatory models can be more readily interpreted .

Item-level confirmatory factor analyses. Item-level confirmatory factor analyses were conducted using LISREL 8.72 (Joreskog & Sorbom, 2005a) using maximum likelihood estimation, the tetrachoric correlation matrix, and the asymptotic covariance matrix for each group separately to determine the proper number of factors to fit the data. The results showed that a single factor resulted in the most parsimonious fit to the data compared to a model based on content areas (three factors) or a model based on all the reporting strands (six factors). This finding was based on examination of fit statistics as suggested by Hoyle and Panter (1995), such as the *root mean square error of approximation* (RMSEA)⁸ which was less than 0.02 for all groups and all models. Values of the *comparative fit index* (CFI)⁹ and *goodness-of-fit index* (GFI)¹⁰ were above 0.90 for all groups and models, yet slightly below preferred thresholds.

The finding of adequate fit with a one-factor model was supported by high inter-factor correlations of 0.90 or greater when multiple factors were extracted. As suggested by Bagozzi and Yi (1988), the confidence intervals for the correlations based on ± 2 standard errors included 1.00, among factors in the three and six-factor designs for Groups 1, 2, and 3. It was for this reason that item parcels were subsequently created so that multi-group confirmatory analyses could be properly carried out for all four groups in the study.

5.3 Parcel-Level Factor Analyses

Rock, Bennett, and Kaplan (1985), in a factor analysis of the SAT comparing students without disabilities to several different groups of students with disabilities, recommended the use of item parcels over individual items in factor analyses based on a few guiding principles: (a) reliability is naturally higher when more items are put together, (b) non-linear relationships exist between items that are dichotomously scored which under certain circumstances can create more factors than are really present, (c) statistical power increases in hypothesis testing with parcels when the parcel distributions are multivariate normal, which is more likely to be the case than with item-level data.¹¹

In this Grade 5 Science assessment, the 60 items were divided into 11 parcels with approximately the same level of difficulty within each of the six individual reporting strands, based on item statistics from Group 1.¹² The Earth Science Grade 4 strand which had seven items could not be split into parcels since the minimum number of items going into a parcel was set at four, and was therefore left as its own parcel. Group sample sizes were further reduced to 500 to make interpretation of findings more similar across groups. Distributional properties of test performance were preserved when this sampling was conducted, as previously shown in Table 3. The overall sample size of Group 4 was 295, so the entire group was used for this set of analyses. Table 7 displays the parcel design for the Grade 5 Science assessment.

Table 7

Original Grade 5 Science Parcel Design

Strand	Number of items in strand	Parcel number	Number of items in parcel	Average parcel difficulty
Physical Science Grade 5	12	1	6	0.57
		2	6	0.57
Physical Science Grade 4	8	3	4	0.57
		4	4	0.58
Life Science Grade 5	11	5	6	0.62
		6	5	0.61
Life Science Grade 4	10	7	5	0.61
		8	5	0.61
Earth Science Grade 5	12	9	6	0.61
		10	6	0.61
Earth Science Grade 4	7	11	7	0.41

While the range of percent correct values ranged from 0.27 to 0.89 at the individual item level, the average difficulties of the parcels ranged from 0.41 to 0.62, yet were balanced within strands as is evident from Table 7.¹³

It should be noted that only one parceling design was used, and a number of alternative methods could also have been tried.

Parcel-level exploratory factor analyses. As with the item-level analyses, a scree plot was generated to get a rough approximation of an appropriate number of factors to fit to the parcel-level data. Figure 3 shows the scree plot for the groups.

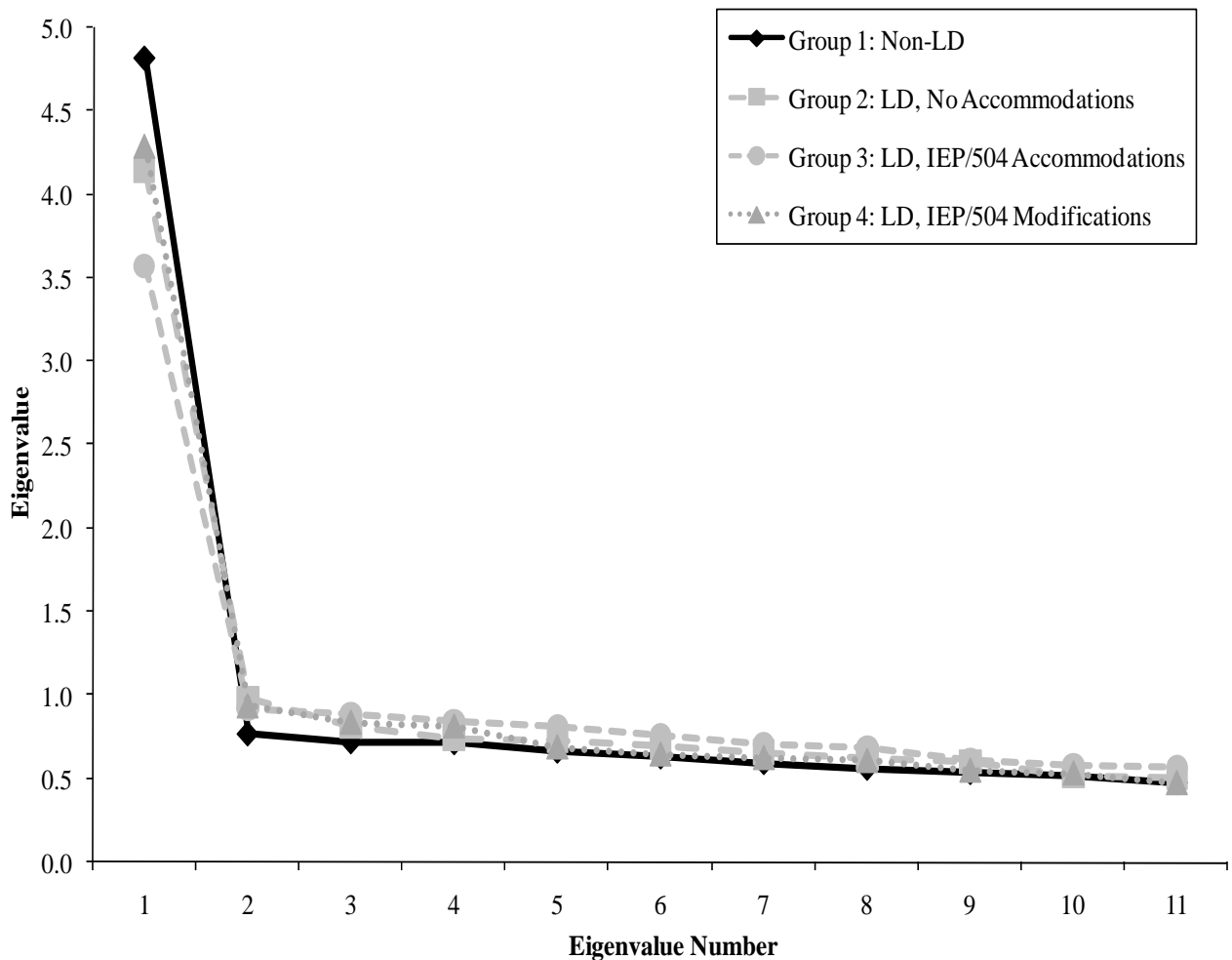


Figure 3. Complete scree plot for all groups.

Given the first eigenvalue was again large compared to subsequent eigenvalues, it made sense that a single factor may fit the data. The proportion of explained variance accounted for by the first eigenvalue ranged from 32 to 44 percent across the four groups.

Exploratory factor analyses were conducted in SAS using each group’s variance-covariance matrix as an input and using maximum likelihood extraction with no rotation at first (for one factor) and later, promax rotation (for more than one factor).¹⁴ All group sample sizes were sufficient given 11 parcels¹⁵. For each group, *p*-values testing for whether one factor was sufficient to fit the data were all above 0.05, indicative of adequate fit, and all parcels had loadings of at least 0.30 on the general factor. (See Table 9 for a summary of these results.) Individual factor loadings by parcel can be found in Appendix C.

Table 8
Summary of Parcel Factor Loadings on General Science Factor

Group	Model DF	Model χ^2	<i>p</i> -value	Range of loadings
1: Non-LD	44	35.340	0.821	0.509-0.672
2: LD, no accommodations	44	54.377	0.136	0.440-0.669
3: LD, IEP/504 accommodations	44	42.391	0.541	0.369-0.610
4: LD, IEP/504 modifications	44	46.474	0.371	0.414-0.643

When rotations were applied in extracting multiple factors, the *p*-values further increased, but there was no clear interpretability of additional factors. Therefore, only a single factor was considered. However, as mentioned previously, confirmatory models provided more concrete evidence to support this hypothesis.

Parcel-level individual-group confirmatory factor analyses. Even though evidence showed that a single factor may be sufficient to fit the data for all groups, attempts were made to confirm more than a single factor for this test. The reason was to identify any disparate patterns in fit indices, factor loadings, or interfactor correlations across groups that might suggest a different course of action in further analyses. Proposed additional models were a grade-level design (two factors) and a content area design (three factors). The six-factor model originally proposed in the item-level analyses with one strand assigned to each factor was not considered

for the parcel-level analyses because based on the scree plot, little additional explanatory power would be gained from such a design. The two-factor model was ruled out after consultation with the developers of the test, so the only designs under study were a single-factor and a three-factor design.

The conditions for proper model identification (Kline, 1998) were met under the proposed three-factor model design as displayed in Figure 4.

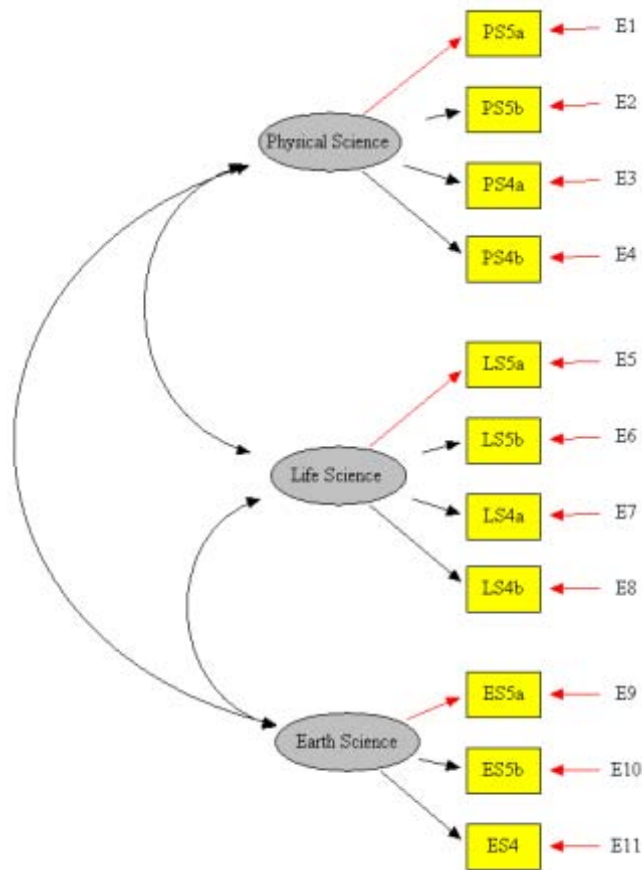


Figure 4. Proposed Grade 5 Science three-factor parcel design.

The factor loading for the first parcel (or in the three-factor case, the first parcel of each factor) was fixed to one to aid in model identification.¹⁶

The testing of multi-group confirmatory factor analysis models in an extensive and exhaustive fashion to obtain estimates of multivariate normality and robust parameter estimates

was done using EQS 6.1 (Bentler & Wu, 2006) using maximum likelihood estimation for free parameters. The results from the single-factor and three-factor individual-group models are displayed in Table 10.

Table 9

Summary of Individual-Group Parcel-Level Confirmatory Factor Analysis Results

Group	N	Model DF	ML χ^2	RMSEA	CFI	GFI	Mardia normalized estimate
Group 1: 1 factor	500	44	35.709	0.000	1.000	0.987	-2.260
Group 1: 3 factors	500	41	30.607	0.000	1.000	0.989	-2.260
Group 2: 1 factor	500	44	54.946	0.022	0.991	0.980	-2.327
Group 2: 3 factors	500	41	49.480	0.020	0.993	0.982	-2.327
Group 3: 1 factor	500	44	42.834	0.000	1.000	0.985	-3.882
Group 3: 3 factors	500	41	34.743	0.000	1.000	0.988	-3.882
Group 4: 1 factor	295	44	47.305	0.016	0.996	0.971	-1.612
Group 4: 3 factors	295	41	34.963	0.000	1.000	0.979	-1.612
χ^2 difference tests			Δ DF	$\Delta \chi^2$	p-value		
Group 1: 3 factors – 1 factor			3	5.102	0.164		
Group 2: 3 factors – 1 factor			3	5.466	0.141		
Group 3: 3 factors – 1 factor			3	8.091	0.044		
Group 4: 3 factors – 1 factor			3	12.342	0.006		

EQS outputs the Mardia (1970) coefficient of multivariate kurtosis, which is important in determining whether maximum likelihood estimates for factor loadings, standard errors, variances, and covariances (where applicable) are sufficient for interpretation under the assumption of multivariate normality. The authors felt that the normalized values of the Mardia coefficient were low enough across groups not to reject the hypothesis that the data came from a multivariate normal distribution.

The results from the individual confirmatory analyses did somewhat agree with the results from the exploratory analyses on the parcels as the model chi-square was lowest for Group 1 and highest for Group 2. The RMSEA values for both the one-factor and three-factor

solutions were well below the 0.05 cutoff, and the CFI and GFI values were close to 1.00, all indicative of a good fit to the model for both designs across all four groups.

There is no clear consensus as to whether fit indices or changes in model chi-square values should be used as a guide to judge model adequacy, but both approaches should be acknowledged. To test whether the three-factor model sufficiently improved the fit, the changes in model chi-square values were 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 Table 10. The changes in model chi-square values were significant at the 0.05 level of significance for Groups 3 and 4, which would suggest that a three-factor model might fit the data better for these groups.

Since chi-square difference tests are heavily dependent on sample size, the changes in fit indices ultimately took precedence over the changes in model chi-square values as the criterion for model adequacy. In comparing the three-factor design to the one-factor design for each group, changes in the CFI were less than 0.01, representing model equivalence (Cheung & Rensvold, 2002). A key additional piece of information, the latent factor intercorrelations (Table 11), led to the final choice of a single-factor model for all groups.

Table 10

Latent Factor Intercorrelation Matrices From Three-Factor Individual-Group Confirmatory Factor Analysis Models

Group	Correlations		
	Physical Science with Life Science	Physical Science with Earth Science	Life Science with Earth Science
1: Non-LD	0.941	1.000	0.970
2: LD, no accommodations	0.947	0.906	0.960
3: LD, IEP/504 accommodations	0.925	0.892	0.869
4: LD, IEP/504 modifications	0.876	0.864	0.901

Since the latent correlation between Physical Science and Earth Science was 1.0, the results from the three-factor model were treated somewhat cautiously, and thus it was determined that a single factor could best explain the data for all groups. This was based on the Bagozzi and Yi (1988) criterion for sufficiently high correlations mentioned earlier.

Parcel-level multi-group confirmatory factor analyses. The results from the individual-group confirmatory models indicating that a single factor could best explain the Grade 5 Science assessment data became the basis for testing a one-factor multi-group confirmatory model. The goal was to show factorial invariance across the groups under study. Brown (2006) discussed different approaches to enter constraints. Table 12 displays the four proposed steps to complete this process according to his recommendations.

Table 11

Summary of Proposed Parcel-Level Multi-Group Confirmatory Factor Analyses

Model	Objective	Constraints imposed
1 (Least restrictive)	Establish the same number of factors across groups	None
2	Test whether factor loadings are the same across groups	Factor loadings equal across groups
3	Test whether factor standard errors (SEs) are the same across groups	Factor loadings and standard errors (SEs) equal across groups
4 (Most restrictive)	Test whether factor variances are the same across groups	Factor loadings, standard errors (SEs), and variances equal across groups

At each step in the process, model fit indices produced by EQS were checked for reasonableness before proceeding to the next step. If there was any model misfit, equality constraints could be relaxed, if necessary. Testing the most restrictive model suggested above would show true invariance as suggested by Byrne (1998), but she cautioned that this test can be too stringent.

The establishment of the baseline model described in Step 1 combined the individual-group models described in the previous section and stacked them together as one model with the degrees of freedom from the individual models being additive. More degrees of freedom were added as constraints are imposed. Table 13 summarizes the results from these multi-group confirmatory models.

Table 12***Summary of Parcel-Level Multi-Group Confirmatory Factor Analysis Results***

Model	Constraints	DF	ML χ^2	RMSEA	CFI	GFI	Δ DF	$\Delta\chi^2$	<i>p</i> -value
1	None	176	180.795	0.008	0.999	0.982	-	-	-
2	Loadings	206	238.158	0.019	0.993	0.976	30	57.363	0.002
3	Loadings + SEs	239	290.491	0.022	0.989	0.971	33	52.333	0.018
4	Loadings + SEs +variances	242	310.399	0.025	0.985	0.969	3	19.908	< 0.001

As constraints were imposed in Steps 2 through 4, the model fit indices decreased as one might expect. However, the RMSEA, CFI, and GFI were all within normally accepted boundaries for adequate model fit at all steps. The change in CFI between Models 1 and 2 was less than 0.01, indicating partial factorial invariance, but reached the 0.01 threshold between Models 2 and 3, compared to Model 1. The interpretation was that while the factor loadings could be considered to be equivalent across groups, the standard errors of these loadings may differ across groups.

Given the misfit between Models 2 and 3, attempts were made to locate the source of the misfit. This was done for Model 3 using the Lagrange Multiplier test for releasing constraints provided by EQS. The cumulative multivariate statistics were examined to determine which constraints had univariate chi-square increments with significant associated probabilities ($p < 0.05$). There were seven constraints whose univariate chi-square values were large enough to result in significant changes in the cumulative chi-square. Three of these were identified as attempts to constrain the factor standard errors to be equal between Group 1 and the other three groups for Parcel 11, corresponding to the Earth Science Grade 4 strand. Given the source of the misfit could be clearly identified, Model 3 was rerun with the three factor standard error constraints relaxed, and the fit statistics were reexamined.¹⁷ Table 14 displays the results from the modified model where factor loadings were constrained to be equal and the standard errors for all parcels except for Earth Science Grade 4 were constrained to be equal.

Table 13***Summary of Revised Parcel-Level Multi-Group Confirmatory Factor Analysis Results***

Model	Constraints	DF	ML χ^2	RMSEA	CFI	GFI	Δ DF	$\Delta\chi^2$	<i>p</i> -value
1	None	176	180.795	0.008	0.999	0.982	-	-	-
2	Loadings	206	238.158	0.019	0.993	0.976	30	57.363	0.002
3A	Loadings + SEs except Parcel 11	236	268.662	0.018	0.993	0.973	30	30.504	0.440

The relaxing of the one factor standard error equality constraint for Parcel 11 achieved strong factorial invariance as the change in CFI was not significant compared to Model 1. The analysis continued from this point; however when the factor variances were constrained to be equal across the groups, model misfit was detected again, as the latent factor variances between Group 1 (Non-LD) and Group 3 (LD, IEP/504 accommodations) appeared to be significantly different. The results from Model 3A confirmed this finding, as the latent factor variance of Group 1 was 0.859 with a standard error of 0.085, and the latent factor variance of Group 3 was 0.539 with a standard error of 0.056. The corresponding latent factor variance of Group 2 was 0.704 with a standard error of 0.071, and the latent factor variance of Group 4 was 0.727 with a standard error of 0.086. Given that Group 3 was the only LD group to show this difference compared to the non-LD group, the authors chose to end the invariance analyses at this point. Table 15 displays the standardized factor loadings of the revised factor standard error equality constraint invariance model (Model 3A).

Table 14***Summary of Standardized Factor Loadings for Final Factor Invariance Model***

#	Strand	Groups			
		1	2	3	4
1	PS5a ¹⁸	0.636	0.598	0.547	0.604
2	PS5b	0.623	0.585	0.534	0.591
3	PS4a	0.482	0.445	0.399	0.451
4	PS4b	0.567	0.529	0.479	0.535
5	LS5a	0.664	0.626	0.575	0.632
6	LS5b	0.508	0.471	0.424	0.477
7	LS4a	0.628	0.589	0.538	0.595
8	LS4b	0.665	0.627	0.576	0.633
9	ES5a	0.664	0.626	0.575	0.632
10	ES5b	0.608	0.570	0.519	0.576
11	ES4	0.523	0.542	0.491	0.583

Most parcels across groups have standardized factor loadings above a substantive level of 0.50 suggested by Bagozzi and Yi (1988). Other parcels are considered to have moderate factor loadings. A full display of the information in Table 15, including the final unstandardized factor loadings, standard errors, and explained variance for Model 3A can be found in Appendix D. The proportion of explained variance, expressed as an average of the individual parcel R^2 values, was 36% for Group 1, 32% for Group 2, 27% for Group 3, and 33% for Group 4.

Therefore, based on this particular parcel construction of the Grade 5 Science test data, the test can be considered unidimensional for all groups. Factor loading equality invariance was achieved for all groups and invariance of the factor standard errors was achieved when the constraints on the standard errors for the one parcel related to Earth Science Grade 4 were relaxed. More importantly, a similar factor structure exists for students without learning disabilities compared to students with learning disabilities taking this test without or with accommodations and/or modifications, which the research was intended to explore. The results provide an indication that the accommodations and modifications do not change the overall science construct being measured. This demonstrates that the scores from this assessment can be compared because the same number of factors was found to exist across groups. The results are similar to those found in several studies (Rock et al., 1988; Meloy et al., 2002; Cook et al., 2008). However, properties of this latent construct may differ slightly across groups based on some of the model results.

6. Discussion and Conclusions

The purpose of this study was to examine the scores on a state standards-based science assessment obtained by a group of students without learning disabilities who took the standard form of the test and by three groups of students with learning disabilities: one taking the test without accommodations or modifications, a second taking the test with accommodations, and a third group taking the test with modifications. The investigation focused on whether or not the science assessment demonstrated factorial invariance for the four groups of students studied. A series of exploratory and confirmatory factor analyses were conducted first at the item level and then at the parcel level. Full analyses could not be completed at the item level since the sample size for students with disabilities taking the test with modifications was insufficient to carry out confirmatory analyses, and was therefore removed from all item-level analyses. However, there was evidence that a single factor might exist for the remaining three groups.

Item parcels created within content strands were balanced for difficulty, and exploratory and confirmatory factor analysis models made use of parcel-level scores. Exploratory analyses again suggested the presence of a single factor that fit the data. Individual-group confirmatory models tested whether a single factor truly fit the data well or whether a three-factor model based on content areas across grades would fit better. There was no significant improvement in model fit based on a three-factor model for students without learning disabilities taking the test under standard conditions (Group 1) and students with learning disabilities taking the test under standard conditions (Group 2), compared to a single-factor model. For students with learning disabilities taking the test with accommodations (Group 3) or modifications (Group 4), examination of the matrices of estimated latent factor intercorrelations provided more convincing evidence beyond the fit indices that a single factor existed for these groups as well.

Finally, a single-factor multi-group confirmatory model tested whether a single factor would best fit the data across groups, including the gradual inclusion of model constraints (factor loadings, then factor standard errors, then factor variances). When the factor loading equality constraints were imposed, invariance was detected across all groups, that is, the changes in the CFI for the more restrictive model were below the 0.01 threshold compared to the baseline model. All but one latent factor standard error was invariant across all groups, that for the Earth Science Grade 4 strand. When this equality constraint was relaxed, the factor variances for non-LD students (Group 1) and LD students taking the test with accommodations specified in their IEPs or 504 plans (Group 3) were found to be different. In summary, the results of the analyses indicated that the Grade 5 Science assessment was unidimensional; that is, the test measured a single factor for all four groups investigated, but this factor slightly differed in composition across all the groups studied.

The results were consistent with Rock, Bennett, Kaplan, and Jirele (1988) and Meloy, DeVille, and Frisbie (2002) who found that when factor structures were compared for students with disabilities who received a read-aloud accommodation to students without disabilities, there was reason to suspect a difference in factor structures between these two groups. Additionally, Cook et al. (2008) found that when factor structures were compared for students without learning disabilities to students with learning disabilities taking a mathematics assessment with and without accommodations, of the three factors hypothesized to fit the data, the disability groups together behaved differently on the dimension characterizing measurement and geometry in

terms of factor standard errors, variances, and covariances with other factors compared to students without disabilities taking the test under standard conditions. In that instance, similar multi-group models were constructed as for the science assessment described in this paper, but equality constraints were relaxed for the students without disabilities in some instances.

The results of this study are important for a number of reasons. One very important reason is that the presence of a single factor across groups lends support to aggregating together scores that are obtained by students without learning disabilities and that were obtained by students with learning disabilities who have taken the assessment with accommodations or modifications required by their 504 plans or IEPs for AYP purposes. In addition, the results provide evidence that the tests have some degree of validity for students with learning disabilities who take the test with and without the particular set of accommodations or modifications that were employed in this study. It is also very important to point out that while invariance across the four groups was limited to the factor loadings and the factor standard errors after a minor modification of the multi-group confirmatory model, there was still evidence that the test measured the same overall construct for all groups, and therefore the test scores may have similar meanings for all of the groups participating in this study.

The use of item parcels facilitated the level of hypothesis testing carried out in this paper, since groups with small sample sizes can be more readily included in factor analysis studies and more explained variance can be obtained from parcel-level data compared to the use of individual items. A note of caution is that while the parcel design put into place here produced the desired result, it is unclear whether other parceling designs would yield similar outcomes, and is worthy of further study. The unidimensionality of this assessment also demonstrated how the total test score can be useful in relating the factor structure to overall performance and validity (e.g. for use in studies involving differential item functioning). However, as noted for the Cook et al. (2008) study on the mathematics test which originated from the same state, the same grade, and the same testing year as the Grade 5 Science test analyzed here at the parcel level, this may not be the case for all subjects, grades, and groups of students and must be considered for other analyses or studies.

References

- Americans with Disabilities Act of 1990, 42 U.S.C. § 12101 *et seq.* (1990).
- 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, CA: Multivariate Software.
- Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement*, 12, 261–280.
- 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.
- Brown, P. J. & Augustine, A. (2001, April). Screen reading software as an assessment accommodation: Implications for instruction and student performance. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.
- 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.
- Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.
- Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Erlbaum.
- 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.
- Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, 10, 1–19.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233–255.
- Child, D. (1970). *The essentials of factor analysis*. New York, NY: Holt, Rinehart, and Winston.
- Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40, 15–32.

- 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. (2006, 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., Steinberg, J., Sawaki, Y., & Eignor, D. R. (2008, March). Using factor analysis to compare the internal structure of a state standards-based math assessment. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.
- 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.
- Fabriger, L. R., Wegener, D. T. MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299.
- Fischer, R. J. (1994). The Americans with Disabilities Act: Implications for measurement. *Educational Measurement: Issues and Practice, 13*, 17–26, 37.
- Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267–269.
- 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.
- Fuchs, L. S., Fuchs, D., Eaton, S., Hamlett, C. L., & Karns, K. (2000). Supplementing teachers' judgments of mathematics test accommodations with objective data sources [Electronic version]. *School Psychology Review, 29*, 67–81.
- 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.
- Hollenbeck, K., Tindal, G., & Almond, P. (1998). Teachers' knowledge of accommodations as a validity issue in high-stakes testing. *Journal of Special Education, 32*, 175–183.

- Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In Hoyle, R. H. (Ed.), *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage Publications.
- Hu, L., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling*, 6, 1–55.
- 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.
- Huynh, H., Meyer, J. P., & Gallant, D. J. (2004). Comparability of student performance between regular and oral administrations for a high-stakes mathematics test. *Applied Measurement in Education*, 17, 38–57.
- Individuals with Disabilities Educational Act of 1997, 20 U.S.C. 1412(a) (17) (A). (1997).
- Joreskog, K. G., & Sorbom, D. (2005a). *LISREL 8 user's reference guide*. Chicago, IL: Scientific Software International.
- Joreskog, K. G., & Sorbom, D. (2005b). *PRELIS 2 user's reference guide*. Chicago, IL: Scientific Software International.
- Kline, R. B. (1998). *Principles and practices of structural equation modeling*. New York, NY: Guilford Press.
- Knol, D. L., & Berger, M. P. F. (1988). *Empirical comparison between factor analysis and item response models* (Research Report 88-11). Twente, the Netherlands: University of Twente, Department of Education.
- Koretz, D., & Hamilton, L. (2000). Assessment of students with disabilities in Kentucky: inclusion, student performance, and validity. *Educational Evaluation and Policy Analysis*, 22(3), 255–272.
- 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 from <http://education.umn.edu/NCEO/OnlinePubs/Technical28.htm>.
- Little, T. D., Cunningham, W. A., Shahar, G., & Widman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, 9(2), 151–173.

- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3): 519–530.
- McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, 27, 82–99.
- McKevitt, B. C., & Elliot, S. N. (2003). The effects and consequences of using testing accommodations on a standardized reading test. *School Psychology Review*, 32, 583–600.
- Meloy, L. L., DeVille, C., & Frisbie, D. (2002, 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, LA.
- Miles, J. & Shevlin, M. (1998). Effects of sample size, model specification and factor loadings on the GFI in confirmatory factor analysis. *Personality and Individual Differences*, 25, 85–90.
- Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, 11, 3–31.
- Muliak, S. A. (1972). *The foundations of factor analysis*. New York, NY: McGraw-Hill.
- Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43, 551–560.
- Muthen, B. (1989). Dichotomous factor analysis of symptom data. *Sociological Methods and Research*, 18, 19–65.
- No Child Left Behind Act of 2001, 20 U.S.C. 6301 *et seq.* (2001) (PL 107-110).
- Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education*, 7, 93–120.
- Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: a review of psychometric, legal, and social policy issues. *Review of Educational Research*, 71(1), 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.

- Rock, D. A., Bennett, R. E., & Kaplan, B. (1985). *The internal construct validity of the SAT across handicapped and nonhandicapped populations* (Research Report No. RR-85-50), Princeton, NJ: ETS.
- Rosenfeld, M., Tannenbaum, R. J., & Wesley, S. (1995). Policy issues with psychometric applications. In J.C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 59–87). Lincoln, NE: Buros Institute of Mental Measurements.
- SAS Institute. (2003). *SAS/STAT user's guide, Version 6*. Cary, NC: Author.
- Section 504 of the Rehabilitation Act of 1973, 29 U.S.C. § 701 *et seq.* (1973).
- 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: School of Education, University of Massachusetts.
- Sireci, S. G., Scarpati, S., & Li, S. (2005). Test accommodations for students with disabilities: an analysis of the interaction hypothesis. *Review of Educational Research*, 75(4), 457–490.
- Tindal, G., Hollenbeck, K., Health, W. & Almond, P. (1997). *The effect of using computers as an accommodation in a statewide writing test*. Eugene: University of Oregon.
- 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, IL.
- Waller, N. J. (1999). Searching for structure in the MMPI. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement, what every psychologist and educator should know*. Hillsdale, NJ: Erlbaum.

Notes

- ¹ Under permitted conditions, disabled and non-disabled groups would each take the test with and without accommodations, creating a set of four scores. This is not the same set of scores from the groups used in the analyses reported in this paper, because the non-disabled group did not take the test with accommodations.
- ² Sireci, et al. (2003). defined the *Interaction Hypothesis* (also referred to as *differential boost*, by Phillips (1994)) as, “The interaction hypothesis states that (a) when test accommodations are given to the students with disabilities (SWD) 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. Fuchs, Fuchs, Eaton, Hamlett, and Karns (2000) allowed for students without disabilities to also improve performance when taking the test with those accommodations.
- ³ Percent correct is defined here as the number of correct responses for the strand divided by the number of items in the strand.
- ⁴ The following strand abbreviations are used throughout the paper: PS5: Physical Science Grade 5; PS4: Physical Science Grade 4; LS5: Life Science Grade 5; LS4: Life Science Grade 4; ES5: Earth Science Grade 5; ES4: Earth Science Grade 4. The numbers of items in each strand are in parentheses.
- ⁵ The adjusted Cronbach’s Alpha was computed as $(60/SL)(\alpha)/(1 + (59/SL)(\alpha))$ where SL is strand length and α is Cronbach’s Alpha. The number 60 represents the number of items on the test.
- ⁶ Results from factor analyses are generally robust if the sample size exceeds the quantity of $.5(k)(k-1)$ where k is the number of test questions. This allows for proper accuracy in calculating the asymptotic covariance matrix obtained in confirmatory factor analysis (see Joreskog & Sorbom [2005b] for a justification) according to the weighted least squares (WLS) method, but could be applied in theory to the maximum likelihood method used here.

Given the test consisted of 60 items, the suggested minimum sample size was 1770, which was not possible with Group 4.

- ⁷ A more common practice is to use the correlation matrix with diagonal entries of 1 (Fabriger, Wegener, MacCallum & Strahan, 1999).
- ⁸ Evaluates the extent to which the model approximates the data, taking into account the model complexity. A RMSEA of approximately 0.06 or below is considered to be an indication of close fit (Hu & Bentler, 1999) and 0.08 or below indicates adequate fit as proposed by Browne and Cudeck (1993).
- ⁹ 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 0.95 or above indicates an adequate model fit (Byrne, 2006).
- ¹⁰ An absolute model fit index, which is analogous to a model R^2 in multiple regression analysis. A GFI of 0.95 or above indicates an adequate model fit (Miles & Shevlin, 1998).
- ¹¹ A generic description of the pros and cons of item parceling can be found in Little, Cunningham, Shahar and Widman (2002)
- ¹² Creating parcels of equal difficulty within content strands was employed by Cook et al. (1988) in an attempt to ensure that difficulty did not in any way affect the factor loadings that were obtained. The same practice was employed in this study.
- ¹³ In theory, a single parcel representing Earth Science Grade 4 would not be desirable. However, as shown later, a six-factor model using item parcel data was not retained, so this was not deemed to be a significant issue by the authors since this parcel could either load on a latent earth science factor for both grades or a single latent science factor.
- ¹⁴ Please refer to Appendix B for the variance-covariance matrix for each group.
- ¹⁵ Please refer to Footnote 7 for the formula to calculate the suggested minimum sample size.
- ¹⁶ This was an arbitrary decision, but was acceptable because whichever parcels were fixed would not change the results, since the results would be mathematically equivalent as suggested by Kline (1998) and Raykov and Marcoulides (2000).

¹⁷The strategy employed here was also used by Cook et al. (2008) in analyzing a mathematics assessment from the same state used in this analysis, whereby factor equality invariance held, and when further testing for invariance resulted in model misfit, specific equality constraints were identified and relaxed, and the analysis continued.

List of Appendices

A - Observed Reporting Strand Intercorrelation Matrices by Group Adjusted for Attenuation ..	43
B - Parcel-Covariance Matrices	45
C - Summary of Parcel-Level Exploratory Factor Analyses by Group	47
D - Summary of Factor Loadings and Residuals From Final Parcel-Level Multi-Group Confirmatory Factor Analyses	48

Appendix A

Observed Reporting Strand Intercorrelation Matrices by Group Adjusted for Attenuation

Table A1

Reporting Strand Intercorrelation Matrix for Grade 5 Science – Group 1

Strand	1	2	3	4	5
1. Physical Science Grade 5	1.000				
2. Physical Science Grade 4	0.784	1.000			
3. Life Science Grade 5	0.934	0.840	1.000		
4. Life Science Grade 4	0.952	0.862	1.000	1.000	
5. Earth Science Grade 5	0.656	0.606	0.735	0.808	1.000
6. Earth Science Grade 4	0.853	0.810	0.920	1.000	0.749

Table A2

Reporting Strand Intercorrelation Matrix for Grade 5 Science – Group 2

Strand	1	2	3	4	5
1. Physical Science Grade 5	1.000				
2. Physical Science Grade 4	0.964	1.000			
3. Life Science Grade 5	0.955	0.882	1.000		
4. Life Science Grade 4	0.993	0.922	1.000	1.000	
5. Earth Science Grade 5	1.000	0.902	0.936	0.968	1.000
6. Earth Science Grade 4	0.825	0.765	0.771	0.871	0.863

Table A3***Reporting Strand Inter-Correlation Matrix for Grade 5 Science – Group 3***

Strand	1	2	3	4	5
1. Physical Science Grade 5	1.000				
2. Physical Science Grade 4	0.964	1.000			
3. Life Science Grade 5	0.955	0.882	1.000		
4. Life Science Grade 4	0.993	0.922	1.000	1.000	
5. Earth Science Grade 5	1.000	0.902	0.936	0.968	1.000
6. Earth Science Grade 4	0.805	0.792	0.772	0.652	0.915

Table A4***Reporting Strand Inter-Correlation Matrix for Grade 5 Science – Group 4***

Strand	1	2	3	4	5
1. Physical Science Grade 5	1.000				
2. Physical Science Grade 4	0.988	1.000			
3. Life Science Grade 5	0.907	0.904	1.000		
4. Life Science Grade 4	0.915	0.816	1.000	1.000	
5. Earth Science Grade 5	0.941	0.756	0.923	0.867	1.000
6. Earth Science Grade 4	1.000	0.872	0.856	0.878	0.986

Appendix B
Parcel-Covariance Matrices

Table B1

Parcel Variance-Covariance Matrix - Group 1

#	1	2	3	4	5	6	7	8	9	10	11
1	1.859										
2	0.841	2.141									
3	0.522	0.551	1.240								
4	0.680	0.640	0.463	1.256							
5	0.848	0.948	0.586	0.660	2.191						
6	0.605	0.593	0.421	0.434	0.642	1.597					
7	0.623	0.749	0.419	0.525	0.745	0.473	1.632				
8	0.679	0.781	0.515	0.567	0.923	0.526	0.735	1.613			
9	0.829	0.852	0.493	0.696	0.830	0.589	0.715	0.687	2.031		
10	0.746	0.817	0.488	0.607	0.829	0.546	0.616	0.752	0.821	1.891	
11	1.010	1.056	0.610	0.854	1.221	0.719	0.901	0.915	1.033	0.872	3.206

Table B2

Parcel Variance-Covariance Matrix - Group 2

#	1	2	3	4	5	6	7	8	9	10	11
1	2.149										
2	0.787	2.026									
3	0.516	0.366	1.193								
4	0.429	0.464	0.383	1.123							
5	0.845	0.724	0.456	0.515	1.907						
6	0.494	0.467	0.173	0.213	0.499	1.674					
7	0.627	0.663	0.328	0.425	0.681	0.455	1.607				
8	0.630	0.618	0.395	0.412	0.742	0.475	0.560	1.578			
9	0.890	0.813	0.423	0.515	0.860	0.598	0.863	0.687	2.245		
10	0.689	0.615	0.379	0.453	0.673	0.486	0.626	0.688	0.919	2.035	
11	0.604	0.457	0.216	0.359	0.482	0.410	0.556	0.450	0.632	0.472	1.911

Table B3***Parcel Variance-Covariance Matrix - Group 3***

#	1	2	3	4	5	6	7	8	9	10	11
1	1.902										
2	0.688	1.983									
3	0.325	0.261	1.127								
4	0.270	0.348	0.216	1.116							
5	0.503	0.446	0.317	0.372	1.752						
6	0.435	0.424	0.148	0.273	0.348	1.445					
7	0.542	0.461	0.299	0.308	0.520	0.342	1.509				
8	0.574	0.491	0.289	0.346	0.564	0.410	0.626	1.568			
9	0.602	0.604	0.321	0.352	0.625	0.421	0.555	0.588	1.960		
10	0.498	0.538	0.226	0.379	0.379	0.337	0.468	0.507	0.699	1.850	
11	0.381	0.461	0.212	0.224	0.469	0.181	0.404	0.435	0.492	0.416	1.864

Table B4***Parcel Variance-Covariance Matrix- Group 4***

#	1	2	3	4	5	6	7	8	9	10	11
1	1.964										
2	0.752	1.944									
3	0.479	0.373	1.142								
4	0.516	0.502	0.263	1.167							
5	0.723	0.743	0.470	0.406	2.094						
6	0.622	0.519	0.311	0.366	0.625	1.665					
7	0.622	0.503	0.182	0.344	0.705	0.618	1.696				
8	0.650	0.641	0.408	0.415	0.779	0.558	0.629	1.592			
9	0.635	0.703	0.356	0.356	0.808	0.516	0.589	0.666	1.878		
10	0.657	0.690	0.336	0.415	0.838	0.581	0.628	0.594	0.876	2.206	
11	0.723	0.873	0.314	0.518	0.673	0.582	0.697	0.792	0.824	0.802	2.055

Appendix C

Summary of Parcel-Level Exploratory Factor Analyses by Group

#	Strand	Items	Group 1	Group 2	Group 3	Group 4
1	PS5a	6	0.645	0.605	0.548	0.598
2	PS5b	6	0.643	0.573	0.518	0.600
3	PS4a	4	0.527	0.440	0.369	0.414
4	PS4b	4	0.639	0.515	0.435	0.482
5	LS5a	6	0.672	0.647	0.536	0.624
6	LS5b	5	0.509	0.435	0.424	0.532
7	LS4a	5	0.601	0.616	0.580	0.558
8	LS4b	5	0.663	0.605	0.610	0.643
9	ES5a	6	0.633	0.669	0.593	0.614
10	ES5b	6	0.616	0.575	0.509	0.575
11	ES4	7	0.634	0.445	0.414	0.639

Note. The strand abbreviations are PS5 = Physical Science Grade 5, PS4 = Physical Science Grade 4, LS5 = Life Science Grade 5, LS4 = Life Science Grade 4, ES5 = Earth Science Grade 5, ES4 = Earth Science Grade 4.

Appendix D

Summary of Factor Loadings and Residuals From Final Parcel-Level Multi-Group Confirmatory Factor Analyses

Table D1

Summary of Factor Loadings, Standard Errors and Explained Variance – Group 1

#	Strand	Unstd. loading	Unstd. loading SE	Std. loading	Unstd. residual	Unstd. residual SE	Std. residual	R ²
1	PS5a	1.000	NA	0.636	1.264	0.047	0.772	0.405
2	PS5b	0.994	0.050	0.623	1.336	0.049	0.782	0.388
3	PS4a	0.576	0.036	0.482	0.945	0.033	0.876	0.232
4	PS4b	0.680	0.037	0.567	0.839	0.030	0.824	0.321
5	LS5a	1.048	0.050	0.664	1.199	0.046	0.748	0.440
6	LS5b	0.707	0.042	0.508	1.234	0.044	0.861	0.258
7	LS4a	0.889	0.045	0.628	1.044	0.039	0.779	0.394
8	LS4b	0.941	0.045	0.665	0.962	0.037	0.747	0.442
9	ES5a	1.067	0.051	0.664	1.243	0.047	0.748	0.440
10	ES5b	0.953	0.049	0.608	1.331	0.049	0.794	0.370
11	ES4	0.947	0.052	0.523	2.049	0.137	0.852	0.273

Note. The strand abbreviations are PS5 = Physical Science Grade 5, PS4 = Physical Science Grade 4, LS5 = Life Science Grade 5, LS4 = Life Science Grade 4, ES5 = Earth Science Grade 5, ES4 = Earth Science Grade 4. Unstd. = Unstandardized, Std. = Standardized.

Table D2

Summary of Factor Loadings, Standard Errors and Explained Variance – Group 2

#	Strand	Unstd. loading	Unstd. loading SE	Std. loading	Unstd. residual	Unstd. residual SE	Std. residual	R ²
1	PS5a	1.000	NA	0.598	1.264	0.047	0.801	0.358
2	PS5b	0.994	0.050	0.585	1.336	0.049	0.811	0.342
3	PS4a	0.576	0.036	0.445	0.945	0.033	0.895	0.198
4	PS4b	0.680	0.037	0.529	0.839	0.030	0.849	0.279
5	LS5a	1.048	0.050	0.626	1.199	0.046	0.780	0.392
6	LS5b	0.707	0.042	0.471	1.234	0.044	0.882	0.222
7	LS4a	0.889	0.045	0.589	1.044	0.039	0.808	0.347
8	LS4b	0.941	0.045	0.627	0.962	0.037	0.779	0.393
9	ES5a	1.067	0.051	0.626	1.243	0.047	0.780	0.392
10	ES5b	0.953	0.049	0.570	1.331	0.049	0.822	0.325
11	ES4	0.947	0.052	0.542	1.519	0.103	0.840	0.294

Note. Unstd. = unstandardized; Std. = standardized. See note below Table D1 for strand abbreviations.

Table D3***Summary of Factor Loadings, Standard Errors and Explained Variance – Group 3***

#	Strand	Unstd. loading	Unstd. loading SE	Std. loading	Unstd. residual	Unstd. residual SE	Std. residual	R ²
1	PS5a	1.000	NA	0.547	1.264	0.047	0.837	0.299
2	PS5b	0.994	0.050	0.534	1.336	0.049	0.846	0.285
3	PS4a	0.576	0.036	0.399	0.945	0.033	0.917	0.159
4	PS4b	0.680	0.037	0.479	0.839	0.030	0.878	0.229
5	LS5a	1.048	0.050	0.575	1.199	0.046	0.818	0.331
6	LS5b	0.707	0.042	0.424	1.234	0.044	0.906	0.179
7	LS4a	0.889	0.045	0.538	1.044	0.039	0.843	0.290
8	LS4b	0.941	0.045	0.576	0.962	0.037	0.817	0.332
9	ES5a	1.067	0.051	0.575	1.243	0.047	0.818	0.331
10	ES5b	0.953	0.049	0.519	1.331	0.049	0.855	0.269
11	ES4	0.947	0.052	0.491	1.525	0.103	0.871	0.241

Note. Unstd. = unstandardized; Std. = standardized. See note below Table D1 for strand abbreviations.

Table D4***Summary of Factor Loadings, Standard Errors and Explained Variance – Group 4***

#	Strand	Unstd. loading	Unstd. loading SE	Std. loading	Unstd. residual	Unstd. residual SE	Std. residual	R ²
1	PS5a	1.000	NA	0.604	1.264	0.047	0.797	0.365
2	PS5b	0.994	0.050	0.591	1.336	0.049	0.807	0.349
3	PS4a	0.576	0.036	0.451	0.945	0.033	0.892	0.203
4	PS4b	0.680	0.037	0.535	0.839	0.030	0.845	0.286
5	LS5a	1.048	0.050	0.632	1.199	0.046	0.775	0.400
6	LS5b	0.707	0.042	0.477	1.234	0.044	0.879	0.228
7	LS4a	0.889	0.045	0.595	1.044	0.039	0.803	0.355
8	LS4b	0.941	0.045	0.633	0.962	0.037	0.774	0.401
9	ES5a	1.067	0.051	0.632	1.243	0.047	0.775	0.400
10	ES5b	0.953	0.049	0.576	1.331	0.049	0.818	0.332
11	ES4	0.947	0.052	0.583	1.264	0.113	0.812	0.340

Note. Unstd. = unstandardized; Std. = standardized. See note below Table D1 for strand abbreviations.