Listening. Learning. Leading.®

Research Report

ETS RR–11-27

Score Comparability for Language Minority Students on the Content Assessments Used by Two States

John W. Young

Steven Holtzman

Jonathan Steinberg

May 2011

Technical Review Editor: Daniel Eignor

Technical Reviewers: Heather Buzick and Elizabeth Stone

Score Comparability for Language Minority Students on the Content Assessments Used by Two States

John W. Young, Steven Holtzman, and Jonathan Steinberg

ETS, Princeton, New Jersey

June 2011

Abstract

In this research investigation of score comparability for language minority students (English language learners [ELLs] and former English language learners), we examined 3 indicators of score comparability (reliability, internal test structure, and differential item functioning) for 4th and 8th grade students who took the NCLB-mandated content assessments in English-language arts and mathematics in 2 different U. S. states. Overall, for the 8 assessments we examined, a high degree of score comparability was found for ELLs and former ELLs, when compared with native English speakers. The results from this study showed that although the assessments from the 2 states differed somewhat with respect to the 3 indicators, a high degree of score comparability was found for both states’ content assessments.

Key words: score comparability, test validity, large-scale assessments, English language learners, language minority students, K–12 assessments

Table of Contents

List of Tables

Background

At present, English language learners (ELLs) are one of the fastest growing subpopulations of K–12 students in American classrooms (Kindler, 2002). During the 2006–07 school year, there were over 5 million ELLs from pre-kindergarten to grade 12, which represents one in nine students in American classrooms (National Clearinghouse for English Language Acquisition, 2008). Since the proportion of students who are ELLs is increasing, and because ELLs are an at-risk group academically, the academic achievement of ELLs is of vital national concern to educators. In addition, under the No Child Left Behind (NCLB) federal legislation, ELLs are one of ten subgroups identified for accountability purposes, so that any underperformance by ELLs has had consequences for their schools. For these reasons, it is important to ascertain whether the large-scale assessments used to determine students' proficiencies in academic subject areas (such as in English-language arts, history-social science, mathematics, science, and other subjects) are valid for all students, including those who speak a language other than English, referred to in this paper as language minority students. The group of language minority students includes students who are currently classified as ELLs as well as students who have been re-classified as former ELLs.

The purpose of this study was to investigate the degree of score comparability, for language minority students, of the content assessments that have been used in two different U. S. states for NCLB accountability purposes. State A is located in the Northeast and State B is located in the Midwest; both are currently among the most populous states in the country. The data for this study were from the content assessments in English-language arts and mathematics taken by 4th and 8th graders in the regular Spring 2008 administration for State A and the regular Fall 2006 administration for State B, so that a total of eight assessments were included. Students were classified by their schools into one of three categories of English proficiency: Native English speaker, ELL, or former ELL. These categories were used to group the students for the comparisons in this study.

Prior Research

In order to achieve educational equity among U.S. students, NCLB has mandated that all students are to be held accountable to the same academic standards. One of the implications of this mandate for ELLs is that they should be administered the same statewide academic achievement tests that are taken by other students. This poses unique challenges for assessing

ELLs because English is the language used for these tests, which generally places ELLs at a linguistic disadvantage. Research indicates that it can take up to seven years for ELLs to acquire the academic language that is needed to fully access academic knowledge from English-based sources (Hakuta, Goto, & Witt, 2000). The discrepancy between the level of English used in assessments and the English language skills of ELL examinees is the main source of concern in assessing the academic achievement of ELLs. The *Standards for Educational and Psychological Testing* caution that "…for all test takers, any test that employs language is, in part, a measure of their language skills. This is of particular concern for test takers whose first language is not the language of the test. Test use with individuals who have not sufficiently acquired the language of the test may introduce construct irrelevant components to the testing process. In such instances, test results may not reflect accurately the qualities and competencies intended to be measured" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 91). To address this issue, all states implement policies regarding the use of testing accommodations for ELLs taking achievement tests, but research is limited on the effectiveness and validity implications of the use of accommodations for ELLs (Young & King, 2008).

Research on the performance of ELLs on standardized tests has a relatively recent history, with studies dating back about two decades. Much of this research has been conducted by Abedi and his colleagues (Abedi, 2002, 2006; Abedi, Hofstetter, & Lord, 2004; Abedi & Lord, 2001; Abedi, Lord, & Hofstetter, 1998). Many of these studies have found significant achievement gaps between ELLs and native English speakers (or non-ELLs) (Abedi, 2002; Abedi & Lord, 2001; Abedi, Lord, & Hofstetter, 1998; Young et al., 2008). More specifically, the average test scores of ELLs are substantially lower across most, if not all, subjects and grade levels. Duran (2006) reported that while about 30% of non-ELLs performed at or above the Proficient level on the 2003 NAEP Mathematics and Reading tests, only about 10% of ELLs did so. Furthermore, the magnitude of the achievement difference between ELLs and non-ELLs is greatest for tests that require substantial verbal processing, such as English-language arts, and smallest for mathematics tests. Studies have also found that the internal structures of tests may differ for ELLs and non-ELLs, with the data having a level of dimensionality for ELLs that is different than that from data for native English speakers (Abedi, 2002; Young et al., 2008). In contrast, to date, fewer research studies have investigated the validity and fairness of content assessments for former ELLs (see Young et al., 2010, for an example).

Validity Issues in the Assessment of ELLs

The general topic of investigating differences in test validity between examinee groups is known as differential validity. Differential validity has been a major research topic in educational measurement since the 1960s, with demographic subgroups (i.e., subgroups formed based on sex or race/ethnicity) being the most commonly studied subgroups of examinees (see e.g., Linn, 1978; Young, 2001). Since ELLs and former ELLs represent subgroups of special interest, the primary validity concerns are with regard to whether the test scores of ELLs and former ELLs have the same meanings and interpretations as for other subgroups, such as native English speakers. Because all assessments measure language proficiency to some degree, whether it is part of the construct of interest or not, the main validity threat for ELLs (and to a lesser extent, former ELLs) when they take content tests is that their scores will measure English proficiency in addition to content knowledge. When performance is associated with factors other than the target construct, construct-irrelevant variance may be introduced into the test scores. For ELLs, the most likely source of construct-irrelevant variance is the use of language, syntax, or terms that are differentially difficult for them, since by definition, their English proficiency is still not fully developed.

A Framework for Test Validity Research for ELLs

Under the Title I regulations of NCLB, all students are required to be assessed in academic content areas, which currently include English-language arts, mathematics, and science. At present, all students are required to take assessments in English-language arts and mathematics annually in grades 3–8 and once in high school. Students are also required to take assessments in Science once each in grades 3–5, 6–9, and 10–11. To determine whether the assessments that have been used for NCLB accountability purposes are accurate measures of proficiency for all students, we employed the concept of test comparability. Willingham et al (1988) identified eight indicators of test comparability, which have been modified so as to be appropriate for language minority students (Young, 2009). The eight indicators are classified into two types of measures of test comparability: The first five indicators are measures of score comparability and the last three indicators are measures of task comparability. Score comparability indicates that the meanings and interpretations of test performance are the same across groups, while task comparability means that the cognitive demands of the test are the same across groups (Willingham, et al., 1988). The eight indicators of test comparability are:

1. Reliability: Equal precision of measurement across examinee groups;
2. Factor structure: Relationships among test items and components are similar across examinee groups;
3. Differential item functioning (DIF): No differential item difficulty due to group membership;
4. Predictive validity: No differential prediction due to group membership;
5. Educational decisions: No differential decision-making due to group membership;
6. Test content: Content and cognitive processes used are the same across examinee groups;
7. Testing accommodations: Accommodations are appropriate, perceived as such, and have minimal impact on scores for examinees who do not require them; and
8. Test timing: No differential speededness due to group membership.

To date, much of the research on the comparability of large-scale content assessments for ELLs and former ELLs has focused on Indicators 1, 2 and 3 (see e.g., Abedi, 2002; Abedi & Gandara, 2006; Abedi & Lord, 2001; Martiniello, 2008; Young et al., 2008; Young et al., 2010) as these are the comparability issues of greatest concern to psychometricians, researchers, and test developers. In addition, for Indicators 1, 2 and 3, all of the evidence required for judging test comparability is contained within the test results so that no reference to or use of an external criterion or additional sources of information is necessary. Thus, for this study, test comparability in the form of score comparability was assessed based solely on Indicators 1, 2, and 3.

It is important to understand that evaluating the validity of an assessment for different subgroups should be based on more than comparability with respect to these eight indicators alone. Although the outcomes of an assessment may be comparable on the eight indicators of test comparability for different subgroups, the assessment may not necessarily be valid across all groups with regard to content, construct, and score interpretation. Because validity is an argument based on logic, theory, and data, the strength of the argument depends on the depth of the theory, the rigor of the logic, the quality and amount of the data, and connections among all three of these argument components (Kane, 2006). Thus, test comparability, as assessed based on the eight indicators, provides an organizational framework for the test validity argument and indicates the logic, theory, and data needed to support this argument.

Research Methods

In order to support the claim that NCLB content assessments are valid and fair for ELLs and former ELLs, we must provide evidence that the scores from the English-language arts and mathematics assessments for these subgroups have the same meanings as for other students, such as native English speakers.

Research Questions

In this study, we were interested in answering the following three research questions:

- Are the assessments being studied equally reliable for native English speakers, ELLs, and former ELLs, as measured by internal consistency reliability estimates?
- Do the assessments being studied measure the same underlying constructs for native English speakers, ELLs, and former ELLs as assessed using confirmatory factor analysis?
- Are the items on these assessments fair for ELLs and former ELLs as evaluated using a DIF detection method?

Assessments

This study focused on the English-language arts (ELA) and mathematics assessments for two states at grades 4 and 8. The item types included in each of the eight assessments are shown in Table 1.

Table 1

Item Types for Each Assessment – State A

Assessment	Number of items	Maximum score	1/2-point multiple-choice items	Multiple-choice items	Short-response items	Constructed-response and essay items
Grade 4 English-language arts	16	43	0	11	0	5
Grade 4 Mathematics	37	43	8	24	0	5
Grade 8 English-language arts	44	78	0	36	0	8
Grade 8 Mathematics	44	52	0	32	8	4

Table 2

Item Types for Each Assessment – State B

Assessment	Number of items	Multiple-choice items	Constructed-response items
Grade 4 English-language arts	37	34	3
Grade 4 Mathematics	58	57	1
Grade 8 English-language arts	37	34	3
Grade 8 Mathematics	55	54	1

Samples

All students who took the regular administrations of the assessments being studied (in Spring 2008 for State A and Fall 2006 for State B) were included to create the samples used in this study. Including students with disabilities in the study's samples would have confounded English proficiency with disability status, so we chose to exclude those students from our analyses. This decision is also consistent with previous studies conducted on the performance of ELLs on large-scale assessments. For ELLs and former ELLs, all students who met the inclusion criteria were included in the study samples. For the native English speakers, we drew a 50% random sample for each assessment such that the score distributions were comparable to those from the total sample, in order to more appropriately compare results across groups. All of the analyses were conducted separately by grade and subject within each state.

The language classification status for each student that was used in this study was the one that was designated by his or her school and reported on the student's answer document. Although the categories for the language classification variable are consistent across states, the manner in which a student is assigned to a category can vary widely across districts and states. Among the several sources of information used to classify students by English proficiency is the score from the English-language proficiency (ELP) assessment that the state has selected for this purpose (Wolf, Farnsworth, & Herman, 2008). In this study, the two states used different ELP assessments, so it is possible that a student would have received a different language classification (e.g., ELL vs. former ELL) in the other state.

In both states, the samples for the native English speakers were larger for the 8th grade tests than for the 4th grade tests. Within the same grade level for each state, the number of students with the same language proficiency classification who took the English-language arts and mathematics tests differed because some students completed only one of the two tests. In State A, the test with the largest number of ELLs was the Grade 8 Mathematics test, while the Grade 8 English-language arts test had the smallest number of ELLs. There were roughly twice as many former ELLs who took the 4th grade tests as compared to the 8th grade tests. In State B, the test with the largest number of ELLs was the Grade 4 Mathematics test, while the Grade 8 English-language arts test had the smallest number of ELLs. There were about 45% more ELLs who took the 4th grade tests than the 8th grade tests, and there were roughly 60% more former ELLs who took the 4th grade tests as compared to the 8th grade tests.

Statistical Analyses

Summary statistics on the number correct scores and internal consistency reliability values (based on Cronbach's alpha) were computed for each of the assessments for native English speakers, ELLs, and former ELLs. Confirmatory factor analyses (CFA) were conducted for these student groups using item-level data. Mantel-Haenszel DIF analyses (Holland & Thayer, 1988) were conducted for the following three comparisons: (a) native English speakers (reference group) versus ELLs (focal group), (b) native English speakers (reference group) versus former ELLs (focal group), and (c) ELLs (reference group) versus former ELLs (focal group). The Mantel-Haenszel DIF statistic compares the odds of a correct response for two subgroups after controlling for differences in overall ability. If an item functions similarly for the reference and focal groups, their respective odds of a correct response should be equal after controlling for differences in ability (Schnipke, Roussos, & Pashley, 2000). For the factor analyses and the DIF analyses, only the multiple-choice items were included in the analyses.

Results

Summary Statistics

Summary statistics on the number correct scores and internal consistency reliability values were computed for each of the assessments for native English speakers, ELLs, and former ELLs and are shown in Tables 3 and 4. For State A, the highest mean scores on each of the assessments were earned by the native English speakers, while ELLs had the lowest average scores. The former ELLs scored, on average, about halfway between the means of the other two groups, with the means for the former ELLs being somewhat closer to those of the native English speakers on the 4th grade tests compared to the 8th grade tests. In State B, the highest mean scores on each of the assessments were earned by the former ELLs, native English speakers had slightly lower mean scores than the former ELLs, and ELLs had the lowest average scores.

Table 3

Summary Statistics for Grade 4 and Grade 8 Assessments by Group – State A

State A tests	Group	*N*	Mean	SD	Internal consistency reliability
Grade 4 ELA (16 items)	Native English speakers	48,070	22.5	4.8	0.78
	English language learners	2,530	17.9	5.7	0.80
	Former ELLs	1,793	20.7	4.5	0.75
Grade 4 Mathematics (37 items)	Native English speakers	48,380	27.9	8.8	0.87
	English language learners	2,542	21.4	9.0	0.87
	Former ELLs	1,794	25.3	8.4	0.85
Grade 8 ELA (44 items)	Native English speakers	50,569	50.0	9.7	0.90
	English language learners	2,488	35.6	10.6	0.87
	Former ELLs	956	43.5	9.6	0.88
Grade 8 Mathematics (44 items)	Native English speakers	50,445	33.5	11.3	0.92
	English language learners	2,840	21.5	11.2	0.91
	Former ELLs	954	27.6	10.9	0.91

Note. ELA = English-language arts, ELLs = English-language learners.

Table 4

Summary Statistics for Grade 4 and Grade 8 Assessments by Group – State B

State B tests	Group	*N*	Mean	SD	Internal consistency reliability
Grade 4 ELA (34 items)	Native English speakers	49,663	23.7	5.9	0.84
	English language learners	4,087	19.0	6.0	0.81
	Former ELLs	939	24.3	5.2	0.79
Grade 4 Mathematics (57 items)	Native English speakers	49,753	40.1	8.9	0.89
	English language learners	4,288	34.7	9.8	0.90
	Former ELLs	941	41.6	8.3	0.88
Grade 8 ELA (34 items)	Native English speakers	57,073	25.1	6.3	0.87
	English language learners	2,803	18.9	6.5	0.84
	Former ELLs	577	25.3	6.0	0.86
Grade 8 Mathematics (54 items)	Native English speakers	57,576	25.9	8.2	0.84
	English language learners	2,961	20.7	6.8	0.77
	Former ELLs	577	26.4	8.8	0.86

Note. ELA = English-language arts, ELLs = English-language learners.

ELLs had the largest score variability (as measured by the group standard deviation) of the three student groups in both states for both English-language arts tests and the Grade 4 Mathematics test. These patterns of score variability differed somewhat from those reported in previous studies (e.g., Young et al., 2008; Young et al., 2010). In these other studies, lower score variability was found for the ELL and former ELL groups. This is because many ELLs are concentrated at the lower end of the score distribution on content assessments, which can lead to restriction of range problems, while the former ELLs are only reclassified after achieving a certain level of proficiency on their state's content assessments.

With regard to internal consistency reliability, for all of the assessments in both states, the reliability values were similar across all of the language proficiency groups, although the values for the Grade 4 English-language arts tests were generally lower than for the other tests. In addition, on the Grade 8 Mathematics test in State B, the reliability estimate for the ELLs (0.77) was lower than for the native English speakers (0.84) or the former ELLs (0.86).

Factor Analysis Results

Since all of the tests were hypothesized to be essentially unidimensional in their underlying factor structure for all of the language proficiency groups, one-factor CFAs were conducted using item-level data for the three language proficiency groups for each of the tests by state, grade, and subject area separately. First, single-group CFAs were conducted for each assessment for each of the student groups. These analyses were conducted using LISREL 8.72 (Joreskog & Sorbom, 2005) employing maximum likelihood estimation and the asymptotic covariance matrix for each group to determine the proper number of factors to fit the data. The results for all of the assessments from both states showed that a single factor fit the data relatively well based on the fit statistics employed and when compared to models with more factors extracted (those models and comparisons are not reported here). This was based on the examination of fit statistics suggested by Hoyle and Panter (1995) such as the root mean square error of approximation (RMSEA),[1] which was less than 0.04 for all groups and all models and the values of the comparative fit index (CFI),[2] which were above 0.90 for all groups and all models.

The results from the single-group CFA models indicated that a single factor could best explain the structure of each of the four assessments for both states. This became the basis for proceeding to the next step in the analysis where a multi-group one-factor confirmatory model was tested using LISREL. The goal was to assess factorial similarity, or invariance for each test, across the groups under study. The four steps undertaken to complete this process are shown in Table 5. At each step in the process, the model fit indices are checked for reasonableness before proceeding on to the next step. If there is any model misfit, equality constraints may be relaxed, if necessary. Testing the most restrictive model would show true invariance (Byrne, 1998), but this test can be too stringent. The establishment of the baseline model described in Step 1 involves taking all the individual-group models described earlier and combining them into a multi-group model (with three groups) with the degrees of freedom from the individual models being additive. For seven of the eight assessments we investigated, the baseline multi-group models showed good fit. Fit statistics for each of the multi-group models is shown in Tables 6 through 12. For the Grade 4 English-language arts assessment for State B, we were not able to obtain CFA results for the multi-group model as the correlation matrices were nonpositive definite.

Table 5

Summary of Multi-Group Confirmatory Factor Analyses

Model	Objective	Constraints imposed
1 (Least restrictive)	Establish a baseline multi-group model	None
2	Test whether factor loadings are invariant across groups	Factor loadings equal across groups
3	Test whether factor loadings and factor errors of measurement are invariant across groups	Factor loadings and errors of measurement equal across groups
4 (Most restrictive)	Test whether factor loadings, errors of measurement, and factor variances, are invariant across groups	Factor loadings, errors of measurement, and variances equal across groups

Table 6

Summary of Multi-Group Confirmatory Analysis Results for Grade 4 English-Language Arts Assessment in State A

Model	Model DF	Satorra-Bentler chi-square	RMSEA	SRMR	AIC	CFI	GFI
1	312	5906.633	0.032	0.047	22129.127	0.993	0.932
2	342	6379.433	0.032	0.082	22497.457	0.992	0.926
3	374	6054.955	0.030	0.084	23096.046	0.993	0.912

Note. RMSEA = root mean square error of approximation, SRMR = standardized root mean residual, AIC = Akaike information criterion, CFI = comparative fit index, GFI = goodness of fit index.

Table 7

Summary of Multi-Group Confirmatory Analysis Results for Grade 8 English-Language Arts Assessment in State A

Model	Model DF	Satorra-Bentler chi-square	RMSEA	SRMR	AIC	CFI	GFI
1	2706	22941.994	0.020	0.054	153079.028	0.996	0.806
2	2792	24138.992	0.021	0.095	156471.672	0.996	0.795
3	2880	23556.829	0.020	0.102	162440.248	0.996	0.779

Note. RMSEA = root mean square error of approximation, SRMR = standardized root mean residual, AIC = Akaike information criterion, CFI = comparative fit index, GFI = goodness of fit index.

Table 8

Summary of Multi-Group Confirmatory Analysis Results for Grade 4 Mathematics Assessment in State A

Model	Model DF	Satorra-Bentler Chi-Square	RMSEA	SRMR	AIC	CFI	GFI
1	1887	22087.955	0.025	0.050	97498.666	0.994	0.842
2	1959	22813.923	0.025	0.068	97881.881	0.994	0.839
3	2033	21916.304	0.024	0.070	98975.772	0.994	0.836

Note. RMSEA = root mean square error of approximation, SRMR = standardized root mean residual, AIC = Akaike information criterion, CFI = comparative fit index, GFI = goodness of fit index.

Table 9

Summary of Multi-Group Confirmatory Analysis Results for Grade 8 Mathematics Assessment in State A

Model	Model DF	Satorra-Bentler Chi-Square	RMSEA	SRMR	AIC	CFI	GFI
1	2706	25356.635	0.022	0.047	98432.820	0.996	0.821
2	2792	28845.402	0.023	0.079	108787.798	0.996	0.816
3	2880	27345.941	0.022	0.084	111311.264	0.996	0.803

Note. RMSEA = root mean square error of approximation, SRMR = standardized root mean residual, AIC = Akaike information criterion, CFI = comparative fit index, GFI = goodness of fit index.

Table 10

Summary of Multi-Group Confirmatory Analysis Results for Grade 8 English-Language Arts Assessment in State B

Model	Model DF	Satorra-Bentler Chi-Square	RMSEA	SRMR	AIC	CFI	GFI
1	1581	15928.391	0.021	0.069	72265.838	0.996	0.742
2	1649	16490.465	0.021	0.080	72497.268	0.996	0.739
3	1717	15791.816	0.020	0.082	74235.505	0.996	0.737

Note. RMSEA = root mean square error of approximation, SRMR = standardized root mean residual, AIC = Akaike information criterion, CFI = comparative fit index, GFI = goodness of fit index.

Table 11

Summary of Multi-Group Confirmatory Analysis Results for Grade 4 Mathematics Assessment in State B

Model	Model DF	Satorra-Bentler chi-square	RMSEA	SRMR	AIC	CFI	GFI
1	4617	77976.911	0.030	0.066	371069.987	0.988	0.694
2	4727	79572.217	0.029	0.075	372090.550	0.987	0.692
3	4837	72953.837	0.028	0.084	375078.956	0.989	0.676

Note. RMSEA = root mean square error of approximation, SRMR = standardized root mean residual, AIC = Akaike information criterion, CFI = comparative fit index, GFI = goodness of fit index.

Table 12

Summary of Multi-Group Confirmatory Analysis Results for Grade 8 Mathematics Assessment in State B

Model	Model DF	Satorra-Bentler Chi-Square	RMSEA	SRMR	AIC	CFI	GFI
1	4131	49968.337	0.023	0.072	171868.376	0.984	0.688
2	4239	51673.507	0.023	0.097	173938.997	0.983	0.671
3	4347	51750.947	0.023	0.092	176583.274	0.983	0.669

Note. RMSEA = root mean square error of approximation, SRMR = standardized root mean residual, AIC = Akaike information criterion, CFI = comparative fit index, GFI = goodness of fit index.

For the other seven assessments that we were able to obtain multi-group CFA results, we consistently found the same degree of measurement invariance. When used with large sample sizes, likelihood ratio tests, including the chi-square difference test between models, have been criticized for leading to incorrect conclusions of heterogeneity in parameters even though the practical differences are negligible (Vandenberg & Lance, 2000). Thus, the change in CFI has advantages over other indicators of model fit, since it is robust when large samples are involved and has been shown to be unaffected by model complexity (Cheung & Rensvold, 2002). Using the invariance criterion of a change in the CFI of less than 0.01, we found invariance of the factor loadings and of the factor errors of measurement for all seven assessments (Model 3 in Table 5). We did not test for invariance of the factor variances, since, in our analyses, the variances were fixed to be one. In summary, these results provide compelling evidence that similar factor structures exist for students in the different language proficiency groups.

DIF Results

Since the factor analysis results provided evidence that one-factor models fit well with data from each of the assessments (i.e., all of the assessments were unidimensional) for all groups, the use of the total test score as the matching criterion variable for the DIF analyses was supported. For the English-language arts assessments in State A, the matching criterion used was the total score on the reading items as the two open-ended writing items on these assessments were excluded. Mantel-Haenszel DIF analyses (Holland & Thayer, 1988) were conducted for the following three comparisons for each assessment: (a) native English speakers versus ELLs; (b) native English speakers versus former ELLs; and (c) ELLs versus former ELLs. The number of items that exhibited C-level (large) DIF on each assessment is shown in Table 13.

Table 13

Number of Items Exhibiting Significant DIF by Assessment for State A

Assessment	Number of items for DIF analyses	Native English speakers vs. ELLs	Native English speakers vs. former ELLs	ELLs vs. former ELLs
Grade 4 ELA	14	1	0	0
Grade 4 Mathematics	37	0	0	0
Grade 8 ELA	42	4	3	1
Grade 8 Mathematics	44	3	1	0

Note. ELA = English-language arts, ELLs = English-language learners.

For the assessments for State A, in the Grade 4 English-language arts assessment, the only item that exhibited C-level DIF was an item which asked for the definition of the word *predators* (the DIF for this item was in favor of the native English speakers). For the Grade 8 English-language arts assessment, one item exhibited DIF in favor of the native English speakers when compared with the ELLs, while three items exhibited DIF in favor of ELLs over the native English speakers. The item favoring the native English speakers was a question regarding the definition of a relatively uncommon word (*veer*) as used in a reading passage. One of the items that favored ELLs was a question regarding the definition of a word (*veneration*) as used in a reading passage. Veneration has a Spanish cognate (*veneracion*), which may have been familiar to many of the ELLs. In reviewing the other two items favoring ELLs, it was not evident as to why these items were relatively easier (in terms of DIF) for the ELLs.

For the Grade 8 English-language arts assessment, two items exhibited DIF in favor of the native English speakers when compared with the former ELLs, while another item exhibited DIF in favor of the former ELLs over the native English speakers. One of the items favoring the native English speakers was a question regarding the use of a word phrase (“inspired loopholes”), which may have been overly difficult for the former ELLs to interpret (but interestingly, this item did not exhibit DIF against the ELLs). Lastly, the other item that exhibited DIF in favor of native English speakers over ELLs also favored the former ELLs over the ELLs.

For the Grade 8 Mathematics assessment, one item exhibited DIF in favor of the native English speakers in comparisons with the ELLs and former ELLs, and two other items exhibited DIF in favor of the native English speakers in comparison with the ELLs. The item which exhibited DIF in favor of the native English speakers over both ELLs and former ELLs was a question that used a picture of a spinner to assess knowledge of probability. It is possible that ELLs and former ELLs may be less familiar with the use of a spinner or that they found the wording of the question to be difficult to understand. Martiniello (2008) has reported that a similar item, which included a spinner to assess a probability concept, from the Grade 4 Mathematics test of a different Northeastern state exhibited DIF against ELLs. One of the items which favored the native English speakers over ELLs was a three-part constructed response item which assessed number patterns. The stem of this question was extremely long, with a total of 140 words in the entire question. ELLs may have found the amount of information presented in the question overwhelming. The other item which favored the native English speakers over ELLs was one that asked for an interpretation of several graphs, and used the phrase, “vertical and horizontal line symmetry.” Although symmetry is a mathematical concept with which most 8th graders should be familiar, ELLs may have not fully understood its usage in this item.

For the assessments from State B, the only item that exhibited C-level DIF was an item on the Grade 8 English-language arts assessment. This item exhibited DIF in favor of the native English speakers when compared with the ELLs and also in favor of the native English speakers when compared with the former ELLs. This particular item was one that required students to use literacy strategies to determine meaning in context and an inspection did not produce any apparent reason that this item should exhibit DIF against these two groups.

Discussion

In this study, for the eight assessments we examined, varying degrees of score comparability were found for ELLs and former ELLs, when compared with native English speakers, based on the results for the three indicators examined (reliability, internal test structure, and differential item functioning). For the State A assessments, the internal consistency reliability values of the assessments were comparable across all of the language proficiency groups, although the reliability values for the Grade 4 English-language arts test were lower than for the other assessments from this state. This was likely due to the fact that this test was the shortest of the ones we studied with a total of only 16 test items. For State B, the reliabilities for the assessments were generally comparable across all three language proficiency groups, with the only exception being that the reliability estimate for the ELLs on the Grade 8 Mathematics test (0.77) was lower than for the native English speakers (.84) or the former ELLs (.86).

In terms of internal test factor structure, a higher degree of factorial invariance across language proficiency groups was found for the assessments from State A as invariance of the factor loadings and invariance of the factor errors of measurement was found for all four of this state's assessments. In contrast, for State B, invariance of the factor loadings and invariance of the factor errors of measurement was found for the three assessments we were able to analyze.

The DIF analyses showed that, for State A, there were many more test items exhibiting DIF on the 8th grade assessments than on the 4th grade assessments. For the 4th grade assessments, there was only one comparison for which C-level DIF was found, while for the 8th grade assessments, C-level DIF was found for a total of 12 comparisons (some items exhibited DIF in more than one comparison). In contrast, for State B, the DIF analyses showed that only one of the 179 multiple-choice items we analyzed exhibited C-level DIF. Since less than 1% of the multiple-choice items exhibited significant DIF, it is reasonable to conclude that with regard to score comparability on this indicator, the State B assessments showed little evidence of differential functioning in the test items for either group of language minority students.

Overall, for the eight assessments we investigated, we found a high degree of score comparability for language minority students based on the three indicators of score comparability we employed: reliability, internal test structure, and differential item functioning. However, there were notable between-state differences: For State A, the internal test structures of this state's assessments had a high degree of similarity across the language proficiency groups (since all of the assessments were invariant with respect to factor loadings and factor errors of measurement), but many more of the test items exhibited DIF against one or more groups. In contrast, the internal test structures of the State B assessments also showed a high degree of similarity since invariance of the factor loadings and factor errors of measurement was found for three of the assessments. With respect to differential item functioning, only one of the test items from the State B assessments exhibited significant DIF.

This investigation showed that, for the assessments we studied, there is a high degree of score comparability for language minority students on the content assessments used by these two states. The results showed that although the assessments from the two states differed somewhat with respect to the indicators, a high degree of score comparability was found for both states' content assessments. In conclusion, we found a number of interesting and informative results through our investigation of these content assessments, and we hope that this study has provided useful information to those who strive the meet the educational needs of all students.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment, 8,* 231–257.

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Mahwah, NJ: Lawrence Erlbaum Associates.

Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice, 25*(4), 36–46.

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74,* 1–28.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14,* 219–234.

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Technical Report No. 478). Los Angeles, CA: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

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

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

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

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.

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

Child, D. (1970). *The essentials of factor analysis.* New York, NY: Holt, Rinehart, and Winston.

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., Eignor, D., 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 American Educational Research Association, San Francisco, CA.

Duran, R. (2006, January). *State implementation of NCLB policies and interpretation of the NAEP performance of English language learners.* Paper commissioned by the NAEP Validity Studies (NVS) Panel. Palo Alto, CA: American Institutes for Research.

Hakuta, K., Goto, Y., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* (Policy Report No. 2000-1). Santa Barbara, CA: The University of California Linguistic Minority Research Institute.

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

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage Publications.

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

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kindler, A. L. (2002). *Survey of the states’ limited English proficient students and available educational programs and services: 2000-2001 summary report.* Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.

Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology, 63*, 507–512.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78,* 333–368.

National Clearinghouse for English Language Acquisition. (2008). *The growing numbers of limited English proficient students: 1995-96–2005-06.* Retrieved from: http://www.ncela.gwu.edu/files/uploads/4/GrowingLEP_0506.pdf

Schnipke, D. L., Roussos, L. A., & Pashley, P. J. (2000). *A comparison of Mantel-Haenszel differential item functioning parameters (*Law School Admission Council Research Report No. 98-03). Newtown, PA: Law School Admission Council.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3,* 4–69.

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (1988). *Testing handicapped people*. Needham, MA: Allyn & Bacon.

Wolf, M. K., Farnsworth, T., & Herman, J. (2008). Validity issues in assessing English language learners' language proficiency. *Educational Assessment, 13*(3), 80–107.

Young, J. W. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis* (College Board Research Report No. 2001-6). New York, NY: The College Board.

Young, J. W. (2009). A framework for test validity research on content assessments taken by English language learners. *Educational Assessment*, *14*, 122–138.

Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment, 13*(3), 170–192.

Young, J. W., & King, T. C. (2008). *Testing accommodations for English language learners: A review of state and district policies* (College Board Research Report No. 2008-6 and ETS Research Report No. RR-08-48). New York, NY: College Entrance Examination Board.

Young, J. W., Steinberg, J., Cline, F., Stone, E., Martiniello, M., Ling, G., & Cho, Y. (2010). Validity and fairness of standards-based assessments for initially fluent students and former English language learners. *Educational Assessment*, *15,* 87–106.

Notes

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

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