ACCESS for ELLs® Series 302 Media-based Listening Field Test: Technical Brief

Prepared by:

CAL/WIDA Partnership Activities
Psychometrics/Research Team

Center for Applied Linguistics

July 13, 2015

2015 WIDA Consortium Members

Alabama
Alaska
Colorado
Delaware
District of Columbia
Georgia
Hawaii
Idaho
Illinois
Indiana
Kentucky
Maine
Maryland
Massachusetts
Michigan
Minnesota
Missouri
Montana
Nevada
New Hampshire
New Jersey
New Mexico
North Carolina
North Dakota
Northern Mariana Islands
Oklahoma
Pennsylvania
Rhode Island
South Dakota
Tennessee
Utah
Vermont
Virginia
Wisconsin
Wyoming

2015 Non-member States Formally Adopting the WIDA ELD Standards

Florida
South Carolina

WIDA™ World-Class Instructional Design and Assessment

WIDA advances academic language development and academic achievement for linguistically diverse students through high quality standards, assessments, research, and professional development for educators. The WIDA vision is to be the most trusted resource in the education of PreKindergarten through Grade 12 language learners.

Contents

1. Purpose of the ACCESS for ELLs® Series 302 Media-Based Listening Field Test

The purpose of this document is to summarize the results of a special study that links the ACCESS for ELLs® (hereafter ACCESS) Series 302 Media-Delivered (MD) Listening Test to the Series 301 Script-Based (SB) Listening Test. The technical information herein is intended for use by those with technical knowledge of test construction and measurement procedures, as stated in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014).

1.1 Description of the Study

Beginning with ACCESS Series 302 (operational year 2013-2014), the Listening test switched from a traditional test administrator-read script to a media-delivered format, either from CD or from streaming audio available online, for all grade level clusters except for Kindergarten. In response to the change in the test delivery mode, two main psychometric issues were addressed. First, given the numerous changes that accompanied the new media-delivery mode, it was necessary to determine whether the same construct is being measured in both the script and media methods of delivery. The second challenge was establishing a method to maintain the continuity of the reporting scale between the two delivery methods. The annual equating method employed for equating the SB Listening Tests between years is the common-item equating method, where refreshed items from the previous series are used as the anchors. Since the Series 302 Listening Test consists of all MD items, however, there are no common items between the Series 301 and 302 operational tests that can serve as anchors. Therefore, an alternative equating method was needed in order maintain the continuity of the Listening scale.

In order to maintain the Listening score scale, a MD Listening Field Test was conducted during the Series 301 administration in the 2012-2013 operational year. The main psychometric goals of the field test were to (a) collect empirical data to examine whether the MD and SB Listening Test measure the same construct via factor analysis and (b) use Rasch analysis to link student performances on the 302 MD Listening Field Test to the 301 SB Listening Test to derive item difficulty parameters on the ACCESS listening scale in preparation for the Series 302 administration in the 2013-2014 operational year. In addition to these main psychometric objectives, an additional goal of the field test was to analyze the test forms and individual items of the MD assessment for the purposes of final item selection for Series 302 Listening Test.

The MD Listening Field Test was designed with the aim of collecting a sample of student performances on both Series 302 MD Listening Field Test and 301 SB Listening Test items such that a common-person design can be used to establish the link between series. In a common-person design, the same students take two tests that consist of different test items but are designed to measure the same construct. Since the same people are being assessed, their performances should be similar on both test forms. If student performances differ, this indicates that one test was harder than the other and item parameters are adjusted so that scores on the new assessment (Series 302) are on the same scale as scores on the original assessment (Series 301).

Each Series 302 Listening Test was field tested as an entire test form. The field test form was administered to samples of students within two weeks after administration of the operational Series 301 test. Since the common-person linking design relies on the performance of a group of participants who are common to both assessments to establish the linking relationship between assessments, it is essential to ensure that students' performances across the two assessments are stable. After the factor analysis and prior to conducting the linking analysis, a Rasch-based outlier analysis was conducted to identify and remove students whose comparative performance on the operational and field test administrations was considered unstable.

1.2 Description of the MD Listening Test

As in previous ACCESS assessments, test forms are organized based on grade-level cluster (K, 1-2, 3-5, 6-8, 9-12) and English language proficiency (ELP) level. In general, Tier A folders were designed to assess the lowest proficiency level followed by Tier B and C folders[1]. Test folders are written to support WIDA's Standards of English Language Development (2012). These standards are Social and Instructional Language (SIL), Language of Language Arts (LoLA), Language of Mathematics (LoMA), Language of Science (LoSC), and Language of Social Studies (LoSS). Table 1 shows the distribution of test folders by tier and standard. This structure is the same across all grade-level clusters. In the Field Test, students were administered complete test forms, including two practice folders in addition to the folders listed in Table 1.

Table 1
Folder Structure of the ACCESS Series 302 Listening Field Test

Tier	# SIL Folders	# LoLA Folders	# LoMA Folders	# LoSC Folders	# LoSS Folders
A	2	1	1	1	1
B	1	2	2	1	1
C	1	2	2	1	1

[1] Items in each tier are based on the ELP levels identified in the WIDA Amplified English Language Development Standards (2012). Tier A folders have items at Proficiency Levels 1 (Entering), 2 (Emerging), and 3 (Developing); Tier B folders have items at Proficiency Levels 2 (Emerging), 3 (Developing), and 4 (Expanding); and Tier C folders have items at Proficiency Levels 3 (Developing), 4 (Expanding), and 5 (Bridging).

1.2.1 Changes in the MD Listening Test

The transition from script to media delivery resulted in several changes to the Listening Test. As previously noted, a CD or streaming audio from the MetriTech website is used to deliver the test in lieu of a test administrator reading the test items aloud. This enhances standardization in test administration, as all students hear the same voices presenting the information in exactly the same way. This consistent input removes potential issues like dialect differences, variability in teacher delivery, and lack of capacity to monitor every teacher to ensure that the script was read only once and at a proper speed with appropriate enunciation.

To take advantage of the new delivery mode, all new items were created for the Series 302 MD Listening Test. Given the added use of professional voice actors in recording the items, many of the new items introduce multiple speakers, including children. Consequently, the MD items represent a wider range of the types of language that students hear in school, allowing for the inclusion of student-student and student-teacher interactions rather than being restricted to the student-directed teacher speech of the traditional SB delivery.

1.2.2 Similarities between the SB and MD Listening Tests

In spite of the differences between the SB and MD versions of the Listening Test, a number of key characteristics of the two assessments remain unchanged. First, the underlying construct being measured has not changed (see Chapter 3 for more information on the factor analysis confirming this assumption). Additionally, as shown in Table 2, the SB and MD versions of the Listening Test are developed using the same test specifications, resulting in consistent content coverage across the two delivery methods.

Note that in the Listening test specifications, there is one folder for each of the content areas targeted at the specified tier levels, along with one (for Tier A) or two (for Tiers B and C) “extra” folders targeted at the next higher level; these folders are marked with a subscript 2. For Tier A, an SIL folder is used as the extra folder, because it represents the language most likely to be acquired first by learners. For Tiers B and C, the extra folders are in LoLA and LoMA, because these are the content areas that are emphasized in the No Child Left Behind legislation.

Table 2

Listening Test Specifications: Number of Items by Standard and Proficiency Level

Listening Tier A

		Proficiency Level			
Folder	Standard	1	2	3	4
1	SI$_1$	x	x	x	
2	LA	x	x	x	
3	MA	x	x	x	
4	SC	x	x	x	
5	SS	x	x	x	
6	SI$_2$		x	x	x
	Number of items	5	6	6	1

6 folders = 18 items

Listening Tier B

		Proficiency Level			
Folder	Standard	2	3	4	5
1	SI	x	x	x	
2	LA$_1$	x	x	x	
3	MA$_1$	x	x	x	
4	SC	x	x	x	
5	SS	x	x	x	
6	LA$_2$		x	x	x
7	MA$_2$		x	x	x
	Number of items	5	7	7	2

7 folders = 21 items

Listening Tier C

		Proficiency Level		
Folder	Standard	3	4	5
1	SI	x	x	x
2	LA$_1$	x	x	x
3	MA$_1$	x	x	x
4	SC	x	x	x
5	SS	x	x	x
6	LA$_2$	x	x	x
7	MA$_2$	x	x	x
	Number of items	7	7	7

7 folders = 21 items

1.3 Student Survey

Because a media format poses a new method of presenting ACCESS test items, additional data were collected using a student survey to gauge students' overall experiences with the MD version of the Listening Test. The survey was designed by the Test Development team at the Center for Applied Linguistics (CAL). Since the design used to link student performances on the ACCESS Series 302 MD Listening Test to the ACCESS Series 301 SB Listening Test is a common-person design, it is essential that the data collected from both assessments are true reflections of students' English language listening proficiency. If, for example, students failed to answer the questions correctly because they were unable to hear the recording, not because they did not understand the question, then the data would not accurately reflect the English language listening skills of the students. In addition, since the MD Listening Test was administered as a field test, students may have been less focused and motivated than they would be during the operational ACCESS test. Two questions on the student survey were designed to collect information on these potential issues from participating students. Question 1 of the survey asked the students whether they could hear the recording. Question 6 of the survey asked whether they tried their best. Responses to these questions from students in the 6-8 and 9-12 grade-level clusters were analyzed as part of the data cleaning process. Students from lower grade levels were excluded from the survey because there was concern that younger students would not understand the survey questions fully and therefore their answers might not accurately reflect their experiences and efforts. The full survey is included in the Appendix.

1.4 Administration the ACCESS for ELLs® Series 302 Listening Field Test

Field test recruiters aimed for a target sample size of 300 students per MD Listening Test form, as recommended by the WIDA Technical Advisory Committee. Students were recruited from all active states that have been members of the WIDA consortium for more than one year. Every effort was made to ensure that the sample was representative in terms of geographic region, gender, and ethnicity across the consortium. Students participating in the ACCESS for ELLs Series 302 Listening Field Test took the field test form within 2 weeks of the operational ACCESS for ELLs test and were assigned to the same test tier and grade-level cluster as on the ACCESS for ELLs test.

2. Description of Field Test Participants

2.1 Data Cleaning Procedure

A total of 5,048 students participated in the Series 302 MD Listening Field Test. Students were drawn from 177 schools and 108 school districts across 10 WIDA states. A series of data cleaning procedures were applied to the original data set provided by MetriTech to identify and remove data that may not be accurate or reliable. Table 3 shows that a total of 356 student records were removed after the initial data cleaning. These records were removed for various reasons ranging from having missing values on the operational test records to discrepancies in the test forms reportedly taken by the students. The final number of student records retained for analysis was 4,692.

Table 4 Table 4 shows the frequency distribution of the field test sample by state after the data cleaning.

Table 3
Summary of Initial Data Cleaning Results

Reason for Removing Students	Number of Students
All Missing Values on Operational Test	190
All Missing Values on Field Test	34
Missing Grade Information for Field Test	4
Took Incorrect Cluster on Field Test	1
Discrepant Field Test/Operational Test Tiers	125
Discrepant Field Test/Operational Test Forms	2
Total Removed	356

Note: The two students identified as taking discrepant test forms took different grade level cluster test form on the Field Test and Operational Test. These students, both in Grade 6, took the 35A test form on the Field Test, but took the 68A Test form on operational ACCESS for ELLs.

Table 4
The Frequency Distribution of Field Test Sample by State

State	Frequency	Percent
Alabama	330	7.0
Georgia	347	7.4
Illinois	673	14.3
Kentucky	195	4.2
Maine	15	0.3
Missouri	396	8.4
New Jersey	2314	49.3
New Mexico	217	4.7
Oklahoma	10	0.2
Wisconsin	195	4.2
Total	4692	100.0

2.2 Student Demographic Characteristics

The demographic characteristics of the Series 302 MD Listening Field Test sample were examined to determine the degree to which the field test sample is representative of the geographic regions, the grade-level and gender distributions, and the ethnic compositions of the WIDA consortium. Since WIDA states vary widely in the timing and length of testing and the field test sample consists of students from states with earlier testing windows, however, it is not appropriate to compare the demographic characteristics of the field test sample with those of all the states in the WIDA consortium. A better frame of reference to use for this comparison is the group of WIDA states that have Series 301 testing windows that are similar to the field test states. The following process was used to identify this comparison group.

First, the Series 301 testing windows of all WIDA states were compiled and analyzed. A review of this list revealed that the WIDA states differed in terms of the exact beginning and end date of testing. Using such a stringent standard, only states already included in field test sample were eligible for inclusion in the comparison group. Participant selection, therefore, needed to be expanded to in order to include states that are not already part of the field test states in the comparison group. To achieve this goal, the comparison group was initially set as all the states whose Series 301 testing windows are within one week of the Series 301 testing windows of the field test states.

Next, the Series 301 testing windows of the states participating in the field test were compiled and analyzed. Among them, Maine and Oklahoma had very different testing windows compared to the rest of the field test states. Maine’s testing window is very early while Oklahoma’s testing window is much later and wider than the other field test states. A preliminary analysis suggested that including these two states in the demographic analysis could significantly affect the composition of the comparison group and therefore should not be included in the demographic analysis. In addition, since these two states had a very small number of participating students in the field test sample (15 in Maine and 10 in Oklahoma), it was determined that excluding them from the field test sample for this analysis only should have a minor impact on the results while considerably simplifying the process of identifying an appropriate comparison group. Consequently, all demographic analyses described in this section are based on the field test sample of 4,667 student field test records from the remaining eight states; the 25 students from Maine and Oklahoma have been excluded from all analyses. Table 5 summarizes the Series 301 testing windows and the geographic regions of the states in the field test sample.

Table 5
Series 301 Testing Windows and the Geographic Regions of the Field Test Sample

	Series 301 Testing Windows		Geographic Region
States	Opens	Closes	
Wisconsin	12/03/12	02/08/13	Mid West
Kentucky	01/02/13	02/08/13	South
Missouri	01/07/13	03/01/13	Mid West
Illinois	01/14/13	02/15/13	Mid West
New Mexico	01/14/13	02/22/13	West
Georgia	01/22/13	03/05/13	South
New Jersey	03/18/13	04/30/13	Mid Atlantic
Alabama	03/25/13	05/03/13	South

Table 6 summarizes the Series 301 testing windows and the geographic regions of the comparison group. Since the eight field test sample states are also part of the comparison group, their testing windows are included in Table 6.

Table 6

Series 301 Testing Windows and the Geographic Regions of the Comparison Group States

	Series 301 Testing Windows		Geographic Region
States	Opens	Closes	
Wisconsin	12/03/12	02/08/13	Mid West
Kentucky	01/02/13	02/08/13	South
Missouri	01/07/13	03/01/13	Mid West
Illinois	01/14/13	02/15/13	Mid West
New Mexico	01/14/13	02/22/13	West
Georgia	01/22/13	03/05/13	South
New Jersey	03/18/13	04/30/13	Mid Atlantic
Alabama	03/25/13	05/03/13	South
Montana	12/05/12	01/30/13	West
Colorado	01/07/13	02/08/13	West
New Hampshire	01/07/13	03/01/13	North East
Rhode Island	01/14/13	02/15/13	North East
Massachusetts	01/10/13	02/13/13	North East
Maryland	01/14/13	02/22/13	Mid Atlantic
North Dakota	01/28/13	03/08/13	Midwest
Pennsylvania	01/28/13	03/08/13	Mid Atlantic
North Carolina	02/01/13	03/15/13	South
Mississippi	04/01/13	04/29/13	South

Note: Hawaii (HI) was not included in the reference group. Although HI had a similar testing window to Georgia (GA), HI has a very different ethnicity distribution compared to the rest of the operational states.

Once the comparison group was identified, the characteristics of the field test sample and the comparison group were compared in terms of their geographic locations, as well as their grade-level, gender (male versus female), and ethnic (Hispanics versus Non-Hispanics) compositions.

2.2.1 Changes in the MD Listening Test Geographic Region Comparison

Table 6 summarizes the Series 301 testing windows and the geographic regions of the comparison group. Since the eight field test sample states are also part of the comparison group, their testing windows are included in Table 6.

Table 6 shows, the field test sample covered fairly similar geographic regions as the states in the comparison group. All but one (North East) of the geographic regions included in the comparison group states are presented in the field test sample.

2.2.2 Demographic Comparisons

Demographic comparisons between the field test sample and the comparison group are shown in Table 7 through Table 9. The demographic analyses were conducted at the level of the grade-level cluster. Table 7 reports the numbers and the percentages of students at each grade level for the field test sample and the comparison group. The results show that the differences in the percentages of students by grade-level cluster between the field test sample and the comparison group were very small (at or below 3 percent) for all grade-level clusters, except for the 6-8 grade-level cluster. For this grade-level cluster, the 6th graders were underrepresented in the field test sample by seven percentage points while the7th graders were over represented in the field test sample by six percentage points.

Table 7

Grade Level Comparison between the Field Test Sample and Comparison Group

Cluster	Grade	Field Test Sample (N=4,667)		Comparison Group (N=758,873)	
		Frequency	Percent	Frequency	Percent
1-2	1	775	54%	132,762	52%
	2	648	46%	123,757	48%
	Total	**1,423**		**256,519**	
3-5	3	545	48%	104,724	45%
	4	321	28%	71,616	30%
	5	263	23%	58,753	25%
	Total	**1,129**		**235,093**	
6-8	6	326	**28%**	50,304	35%
	7	472	**40%**	48,006	34%
	8	381	32%	43,926	31%
	Total	**1,179**		**142,236**	
9-12	9	360	38%	49,282	39%
	10	268	29%	32,020	26%
	11	171	18%	24,512	20%
	12	137	15%	19,211	15%
	Total	**936**		**125,025**	

Table 8 reports the numbers and the percentages of males and females for the field test sample and the comparison group by grade-level cluster. The results showed that the percentages of males and females in the field test sample differed by two percentage points or less from those of the comparison group for all grade-level clusters.

Table 8
Gender Comparison between the Field Test Sample and Comparison Group

Cluster	Gender	Field Test Sample (N=4,667)		Comparison Group (N=758,873)	
		Frequency	Percent	Frequency	Percent
1-2	F	687	48%	121,578	47%
	M	719	51%	134,647	52%
	Missing	17	1%	294	0.1%
	Total	**1,423**		**256,519**	
3-5	F	524	46%	108,320	46%
	M	605	54%	126,462	54%
	Missing	-	-	311	0.1%
	Total	**1,129**		**235,093**	
6-8	F	512	43%	63,673	45%
	M	659	56%	78,329	55%
	Missing	8	1%	234	0.2%
	Total	**1,179**		**142,236**	
9-12	F	425	45%	56,616	45%
	M	511	55%	68,086	54%
	Missing	-	-	323	0.3%
	Total	**936**		**125,025**	

Table 9 shows the numbers and the percentages of and Hispanics and Non-Hispanics in the field test sample and the comparison group by grade-level cluster. For the 1-2 and 9-12 grade-level clusters, the percentages of Hispanics and Non-Hispanics in the field test sample are similar (i.e., 3 percentage points or less) to those in the comparison group. However, for the 3-5 and the 6-8 grade-level clusters, the differences in the percentages of Hispanics between the field test sample and the comparison group were larger at 14% and 11%, respectively. Overall, Hispanic students were overrepresented in the field test sample across grade-level clusters, particularly in the 3-5 and the 6-8 grade-level clusters.

Table 9

Ethnicity Comparison between the Field Test Sample and Comparison Group

Cluster	Ethnicity	Field Test Sample (N=4,667)		Comparison Group (N=758,873)	
		Frequency	Percent	Frequency	Percent
1-2	Hispanic	1,084	76%	189,124	74%
	Non-Hispanic	321	23%	65,147	25%
	Missing	18	1%	2,248	1%
	Total	**1,423**		**256,519**	
3-5	Hispanic	1,007	89%	175,185	75%
	Non-Hispanic	121	11%	57,604	25%
	Missing	1	0.1%	2,304	1%
	Total	**1,129**		**235,093**	
6-8	Hispanic	974	83%	102,726	72%
	Non-Hispanic	188	16%	37,830	27%
	Missing	17	1%	1,680	1%
	Total	**1,179**		**142,236**	
9-12	Hispanic	634	68%	81,159	65%
	Non-Hispanic	294	31%	42,187	34%
	Missing	8	1%	1,679	1%
	Total	**936**		**125,025**	

The results from the geographic analysis indicate that, overall, the field test sample is fairly representative of the geographic regions of the comparison group. The results from the grade-level and gender comparison suggest that the field test sample matched the grade-level and gender distribution of the comparison group very well. The results from the ethnicity comparison were mixed. For the lowest and highest grade-level cluster, the ethnicity compositions of the field test sample were fairly similar to those of the comparison group. However, for the two middle grade-level clusters, Hispanic students were overrepresented.

2.3 Student Survey Results

As described in Chapter 1.3, the responses to Questions 1 and 6 of the student survey were analyzed for students in the 6-8 and 9-12 grade-level clusters to determine whether construct irrelevant causes may have impacted student performance. Questions were included to determine whether students experienced technical issues or lacked motivation while taking the test. The complete results of the student survey, although not discussed here, were shared with the Test Development team and used to develop the operational Test Administration Manual. The full survey is included in the Appendix of this document.

Table 10 presents the numbers and percentages of students responding to each of the three response categories for Questions 1 and 6 on the student survey for the 6-8 and the 9-12 grade-level clusters. Students who did not respond to these survey questions were excluded from the summary. Since the main purpose of the investigation is to identify students who expressed *affirmatively* that they experienced difficulty with the media or were not motivated, only data from the students who responded “No” to both questions were examined in this brief. Although the numbers and the percentages of students who responded to “Sometimes” to both questions are informative, it was not the focus of the investigation.

Table 10 also shows that there were very small percentages (0-6%) of students who indicated that they definitely had trouble hearing the recording or were not motivated when taking the field test. A close examination of the survey data confirmed that these students did not come from the same school districts or testing cohorts. There are two implications associated with these findings. First, the difficulty these students experienced were isolated and random incidents that did not point to systematic administration issues. Second, the motivational issues these students reported appeared to be isolated and random cases that did not point to across-the-board motivational problems of a particular student group.

Table 10

Participants' Responses to Survey Questions 1 and 6

Grade-level cluster	Tier	Question	"Yes"	"Sometimes"	"No"	Total
6-8	A	1. Did you hear the recording?	226 (80%)	46 (16%)	10 (4%)	282
		6. Did you try your best on the test?	227 (81%)	36 (13%)	17 (6%)	280
	B	1. Did you hear the recording?	327 (87%)	44 (12%)	3 (1%)	374
		6. Did you try your best on the test?	345 (92%)	17 (5%)	12 (3%)	374
	C	1. Did you hear the recording?	436 (85%)	70 (14%)	7 (1%)	513
		6. Did you try your best on the test?	463 (90%)	45 (9%)	5 (1%)	513
9-12	A	1. Did you hear the recording?	204 (78%)	52 (20%)	5 (2%)	261
		6. Did you try your best on the test?	217 (83%)	30 (12%)	13 (5%)	260
	B	1. Did you hear the recording?	283 (82%)	56 (16%)	6 (2%)	345
		6. Did you try your best on the test?	299 (87%)	35 (10%)	10 (3%)	344
	C	1. Did you hear the recording?	259 (91%)	25 (9%)	1 (0%)	285
		6. Did you try your best on the test?	247 (87%)	33 (12%)	4 (1%)	284

3. Factor Analysis

3.1 Similarity of Constructs in the MD and SB Test

As discussed in Chapter 1, since the 302 MD Listening Test was developed using the same WIDA ELD standards and test blue prints as the 301 SB Listening Test, the constructs measured by the MD test are assumed to be similar to those measured by the SB test. A series of confirmatory factor analyses (CFA) were performed before the linking study was conducted to empirically evaluate the similarity of the constructs measured by the two Listening tests. The analyses were conducted for all grade-level clusters (i.e., 1-2, 3-5, 6-8, 9-12) and tiers (i.e., A, B, C). Conducting multiple replications provided an opportunity to examine consistency of the findings across grade-level clusters and tiers.

Table 11 presents descriptive statistics and reliability indices (i.e., Cronbach's Coefficient alpha) by test. The table shows that, overall, mean raw scores for the MD and the SB tests are similar across grade-level clusters and tiers; most of the differences in mean raw scores are less than 0.08 with the exception of 68C[2], which has a difference of 0.13. Looking across grade-level clusters and tiers, there were cases in which the mean raw scores for the MD test were slightly higher than the SB test and cases where the opposite was true (i.e., the SB raw scores were slightly higher than the MD raw scores), suggesting that there is no consistent bias against either format. In other words, although the MD Listening Test is a new format that is not familiar to the ACCESS test takers, students did not seem to be disadvantaged as a result.

The Cronbach's alpha values ranged from .44 (68C SB) to .77 (12C SB). With the exception of 68C, these values are similar in magnitude to those typically observed on the ACCESS Listening Test, which range from .60 to .80 (Center for Applied Linguistics, 2014). The relatively low Cronbach's alpha of .44 for the 68C SB test form is an anomaly, as the Cronbach's alpha values for 68C Listening are typically in the .60 range. This may be an indication that the students in 68C field test sample have different characteristics than the population. Table 11 shows that the Cronbach's alpha values are generally similar between the MD and SB tests for most grade-level clusters and tiers; the exceptions are 912B and 68C, where the difference in the Cronbach's alpha values between tests are greater than .10. For 912B, the SB test has a higher Cronbach's alpha than the MD test while the opposite is true for 68C.

[2] In this document, individual test forms are abbreviated using the grade-level cluster and tier. For example, the Tier A test form for the 6-8 grade-level cluster is abbreviated as "68A," the Tier C test form for the 9-12 grade-level cluster is abbreviated as "912C," and so forth. In cases where the administration mode is relevant, it is indicated as SB (script-based) or MD (media-delivered). For example, "35B SB" is the script-based test form administered to Tier B students in Grades 3-5.

Table 11

Descriptive Statistics and Cronbach's alpha by Test

	1-2		3-5		6-8		9-12	
	MD	SB	MD	SB	MD	SB	MD	SB
Tier A								
Number of Items	18	18	18	18	18	18	18	18
Number of Students	347	347	203	203	292	292	272	272
Mean Raw Score	14.94	13.90	11.97	11.68	10.30	11.04	10.92	12.37
SD of Raw Score	2.78	2.91	3.11	3.48	3.30	3.26	3.00	2.64
Mean P Value	0.83	0.77	0.67	0.65	0.57	0.61	0.61	0.69
Cronbach's alpha	0.74	0.70	0.68	0.75	0.70	0.71	0.64	0.55
Tier B								
Number of Items	21	21	21	21	21	21	21	21
Number of Students	743	743	515	515	385	385	366	366
Mean Raw Score	17.56	15.86	15.49	14.72	14.73	14.60	12.63	14.18
SD of Raw Score	2.45	2.98	2.71	3.23	3.19	2.94	3.11	3.57
Mean P Value	0.84	0.76	0.74	0.70	0.70	0.70	0.60	0.68
Cronbach's alpha	0.62	0.65	0.56	0.66	0.66	0.62	0.60	0.71
Tier C								
Number of Items	21	21	21	21	21	21	21	21
Number of Students	333	333	421	421	517	517	298	298
Mean Raw Score	16.5	14.88	13.32	13.66	15.7	12.94	12.89	12.48
SD of Raw Score	3.01	3.79	3.12	3.05	2.83	2.63	3.30	3.52
Mean P Value	0.79	0.71	0.63	0.65	0.75	0.62	0.61	0.59
Cronbach's alpha	0.69	0.77	0.59	0.57	0.58	0.44	0.63	0.67

3.2 Confirmatory Factor Analysis (CFA)

CFA seeks to confirm that a set of observed variables share common variance characteristics that define theoretical constructs, and it statistically tests the significance of hypothesized measurement models in terms of whether the models fit the collected data. Using CFA to confirm the unidimensionality of the construct is important in order to support Rasch scaling within test because individual items from both the MD and the SB tests were used to measure a theoretical construct called "Listening." The factor structure of the MD and the SB Listening Tests were established by applying a one-factor confirmatory model to each test individually. This baseline one-factor model postulates that there is a common Listening construct underlying both the MD test items and the SB test items[3].

[3] Unlike the CFA model, the Rasch model allows for persons and item parameters to be estimated independently of each other, and it does not model the item discrimination parameters (item loadings) as they are assumed to be equal across all items. Therefore, testing the fit of the baseline one-factor model to the data is not the same as testing the fit of the Rasch model.

Once the baseline one-factor model was established within test, two factor analytical models were applied to the combined the MD and SB test data to evaluate whether the test items measure a single construct. The first model was a one-factor confirmatory model. Since the MD and SB tests were developed using the same WIDA standards and blue prints, the initial hypothesis was that the one-factor model would fit the combined MD and SB test data well. However, given that the MD and SB tests consist of totally different test items, there might be some variance not accounted for by the simple one-factor model. Furthermore, given the changes that accompanied the new media-delivery mode, it is possible that the renovated MD Listening Test might in fact capture something slightly different from the SB version. In factor analytical terms, there might be some variation between the SB and the MD test items that cannot be totally explained by the common factor; as a result, a more complex model might better account for the observed variance and covariance structure. While many different CFA models might be postulated to try to fully explain the variance covariance structure of the combined MD and SB data on the basis of different hypothesized relationships between the test items and the proposed construct, the main purpose of this report was to examine whether the construct changes when the delivery method is modified from SB to MD (i.e., to identify a potential form effect).

To examine the possibility of a form effect, a two correlated-factor model was applied to the combined MD and SB test data. This model stipulates that the MD test items measure something distinct from the SB test items. Specifically, each test item is hypothesized to measure either the SB or MD factor; these two factors are correlated, and the measurement error variances are not related. The assumption is that although the SB and MD factors are related to a more general latent factor conceptualized as “Listening,” the individual SB and MD factors capture some type of form effect. Because of the shared association with the general Listening factor, a high correlation between the MD and SB-specific factors was expected.

All analyses were conducted in Mplus, Version 7 (Muthen & Muthen, 2013) using the Robust Weighted Least Square (WLSMV) Estimator to take into account that the item responses are dichotomous (Muthen & Muthen, 2013).

3.3 CFA within Test

CFA was first used to test the one-factor measurement model of the items within each test: one for the MD test and one for the SB test. The chi-square statistic is commonly used to assess how well the model reproduced the covariance matrix. However, because this statistic is sensitive to sample size and may not be a practical test of model fit (Byrne, 2012; Hu & Bentler, 1999), two other goodness of fit statistics that are less sensitive to sample size are used to assess model fit: the comparative fit index (CFI) and the root mean square error approximation (RMSEA). CFI values near 1.0 are optimal, with values greater than .90 indicating acceptable model fit (Byrne, 2012). RMSEA values of 0.0 indicate the best fit between the population covariance matrix and the covariance matrix implied by the model and estimated with the sample data. Generally speaking, values less than .08 are considered reasonable, with values less than .05 indicating a closer approximate fit (Kline, 2005).

All models converged successfully on the first attempt except for the 12B MD test form. The error message provided by Mplus suggested that Item 2 may have high collinearity with other items on the test; as a result, the residual covariance matrix was not positive definite. A close examination of Item 2 showed that this item was extremely easy (only four examinees got this item wrong). As a result, this item had perfect correlations with two other very easy items on the test. In order to attain convergence (Schumacker & Lomax, 2004), Item 2 was removed from all analyses for the 12B MD test form. The within-test model fit statistics are summarized in Table 12. Using the previously discussed criterion of a CFI greater than .90 and RMSEA less than .05, the one-factor model fit the data well for all MD test forms except for 35B, 912B, and 912C (in **bold**) and for all SB test forms except for 68B, 912A, and 912C (in ***italic bold***.) RMSEA values were less than .05 for all test forms except for the 912A SB test. The one-factor model did not fit the data for either the SB or MD format. The 912A SB test form had the lowest CFI value (.72) across all test forms.

The analyses that showed less than ideal fit were closely examined using the modification indices and residual statistics reported in Mplus (Byrne, 2012). The modification indices provide diagnostic information in terms of the amount of reduction in chi-square statistics expected if a particular parameter is allowed to be freely estimated (Muthen & Muthen, 2009). In Mplus, the default minimum modification index value necessary for reporting the modification index is 10.000.

For the MD test, the residual covariance between Item 13 and Item 16 of the 35B test form was flagged as having a high modification index (10.924). When the error variances (or measurement errors) of these two items were allowed to be estimated, then the CFI value increased from .86 to .89. Similarly for the 912C test form, the residual covariance between Item 2 and Item 8 were flagged as having a high modification index (10.072). When the error variances of these two items were allowed to correlate, the CFI value improved from .88 to .91. In both cases, items flagged by the modification indices measure the same proficiency level, which may explain the high residual covariance between those items. For the 912B test form, Mplus modification indices did not flag any parameter as problematic. A close examination of the estimated parameters, however, indicated that Item 3 had a standardized factor loading of -0.03, suggesting that this item has very little communality with the underlying factor. Items with little communality with the common factor are not a valid measure of the underlying construct. When this item parameter was constrained to be zero, then the CFI values improved from .89 to .90.

For the SB test, Item 1 and Item 8 of the 68B test form and Item 10 and Item 7 of the 912C test form had a high modification index (28.466 and 17.421, respectively) for their residual covariance. In both cases, items flagged by the modification indices measure the same proficiency level, which may explain the high residual covariance between items. No parameter was flagged by the modification indices in the 912A test forms; however, Item 8 had a negative covariance with the rest of the test items in 912A, suggesting that it may measure some unique factor that is not accounted for by the simple one-factor model.

Comparing the model fit between the MD and SB tests by grade-level cluster/tier, the one-factor model fits the data well for both tests for 7 out of the 12 grade-level cluster/tier forms. Four of the remaining test forms (35B, 912B, 68B, 912A) showed mixed results and one (912C) showed that the one-factor model did not fit the data for either the SB or the MD test. The results from the within-test CFA analyses suggest that the one-factor model fits the data adequately for more than half of the test forms. For the test forms that showed less than ideal CFI and RMSEA values, the misfit is related to minor fit issues associated with selected items.

Table 12

Summary of Model Fit of the One-Factor Measurement Model within Test Form

Test Form	Delivery Method	*df*	CFI	RMSEA	Chi-square	p-value
12A	SB	135	.97	.02	156.35	.10
	MD	135	.98	.02	151.45	.16
12B	SB	189	.99	.01	202.14	.24
	MD	170	.94	.02	202.80	.04
12C	SB	189	.93	.03	262.43	.00
	MD	189	.95	.02	213.04	.11
35A	SB	135	.97	.03	154.63	.12
	MD	135	.95	.03	156.77	.10
35B	SB	189	.97	.02	213.93	.02
	MD	**189**	**.86**	**.02**	**237.83**	**.01**
35C	SB	189	.93	.02	210.66	.13
	MD	189	.97	.01	200.49	.27
68A	SB	135	.97	.01	167.69	.03
	MD	135	1.00	.01	137.56	.42
68B	***SB***	***189***	***.89***	***.03***	***244.45***	***.00***
	MD	189	.93	.02	228.36	.03
68C	SB	189	.98	.01	193.16	.40
	MD	189	.98	.01	198.77	.30
912A	***SB***	***135***	***.72***	***.05***	***210.32***	***.00***
	MD	135	.91	.03	167.00	.03
912B	SB	189	.97	.02	215.82	.09
	MD	**189**	**.89**	**.03**	**236.50**	**.01**
912C	***SB***	***189***	***.81***	***.04***	***268.05***	***.00***
	MD	**189**	**.88**	**.03**	**237.17**	**.01**

3.4 CFA between Tests

Two confirmatory factor analytic models were applied to the data to examine whether the MD and the SB items measure the same construct. Model I is a one-factor model where items from both the MD and the SB test load on the same latent variable. If this model fits the data well, then it would support the contention that the MD and the SB items measure a single common construct. Model I was statistically compared to a more complex Model II that specifies a potential form effect in the model. Model II is a two correlated-factor model where items from the SB test load on one factor, items from the MD test load on a different factor, and the two factors are correlated. The measurement error variances are not related. If Model II fits better than Model I, then it will suggest that although the SB and MD factors are related to a more general latent factor, the individual SB and MD factors capture some type of form effect. To determine whether an argument can be made that the there is no form effect, the fit of the two models was compared using the DIFFTEST option. The DIFFTEST null hypothesis asserts that the restricted model (MODEL I) does not worsen the model fit. In other words, if the null hypothesis is not rejected, the restrictions in the one-factor model cannot be rejected; thus, there is no statistically significant form effect. If the null hypothesis is rejected, then the less restricted model (MODEL II) is a better fit and there is a statistically significant form effect. Two indicators are used as criteria to determine whether the null hypothesis cannot be rejected (i.e., no form effect is found): a) if there is a non-significant change in the chi-square; and b) a change in the CFI of less than .01 (Cheung & Rensvold, 2002). These two criteria are commonly used in research examining construct invariance of parallel instruments.

A summary of the model fit statistics as well as the model comparison results are reported in Table 13. When examining the absolute fit of the models, the criteria of a CFI greater than .90 and an RMSEA smaller than .05 were used. Table 13 shows that the one-factor model fit the data well for all test forms except for 12B, 35B, 912A, and 912C and the two correlated-factor model fit the data well for all test forms except for 12B, 35B, and 912C. Since neither of the models fit 35B and 912C, strictly speaking, both are excluded from the discussion of relative model fit. These two grade-level clusters are included in the discussion below, however, for the purpose of trying to detect patterns of model comparison results across grade-level clusters and tiers. In terms of the relative model fit between the one-factor and the two correlated-factor model, 7 out of the 12 comparisons showed no change in the CFI and the RMSEA values. The remaining five test form comparisons resulted in four (12A, 12B, 35C, and 68C) that indicated a minimal change (.01-.03) in the CFI values and one (912A) that exhibited a significant change (.09) in the CFI values. All five of these comparisons also had significant chi-square difference test statistics, and the correlation between the MD factor and the SB factor ranged from .70 (12B) to .85 (912A) for these five comparisons.

For the five test forms where the correlated-factor model fit the data better than the one-factor model, the modification indices for the one-factor model were closely examined to identify potential causes of misfit. Two (12A and 912A) of the test forms did not have modification indices exceeding the Mplus criterion of 10 or higher, two (68B and 68C) had significant modification indices associated with one item pair's residual covariance, and one (12B) had significant modification indices for six item pairs' residual covariance. For 12A and 912A, where Mplus did not report modification indices higher than 10 for the one-factor model, the parameter estimates were examined to try to find potential sources of misfit. These examinations did not find any clear source of misfit for either test form.

For the three comparisons with significant modification indices reported by Mplus, two different misfit patterns were observed. The first type of misfit is the result of high residual covariance between item pairs, as observed in the 68B test form where a large modification index (19.5) was reported for the residual covariance between SB Item 1 and SB Item 8. The same item pair was also flagged for high residual covariance in the within-test analysis. A similar issue was found in the 12B test form where one of the six pairs of residuals flagged by the modification indices pointed to a high residual covariance between Items 12 and 16 of the MD test. A second type of misfit is due to high correlations between the MD and the SB test items. For the 68C test form, Item 14 of the MD test and Item 16 of the SB test were flagged as having a high modification index (11.103) on their item residual covariance. Similarly, for the 12B test form, five out of the six pairs of residuals flagged were related to the high item residual covariance between the MD and the SB items. This is an indication that for these test forms, a more complicated model that directly takes into account the correlated measurement errors between items may be needed in order to fully explain the observed covariance structure.

The results from the between-test CFA analyses suggest that the MD and the SB items measure a single common construct for 7 of the 12 test forms. The remaining five show a statistically significant form effect, shown in bold in Table 13. The presence of a form effect is not unexpected as the MD and SB tests consist of entirely different test items and the changes in test delivery mode from script to media delivery may also contribute to the observed form effect. Although a form effect was detected for these five test forms, the correlation between factors suggests that the MD test and the SB test have a high amount of shared variance and that they are related to a more general ability factor. Essentially, the factor correlation implies an indirect relationship between test items with factor loadings constrained to zero on a factor. Given the high correlations found in all grade-level cluster and tier combinations, the two correlated-factor model implies that all test items are correlated with both the MD and SB factors since the factors are correlated. Based on these findings, it can be concluded that although there is a form effect present for some of the test forms, the MD and the SB items appear to measure similar constructs.

Table 13

Summary of Model Comparisons

Test Form	Model	*df*	CFI	RMSEA	Chi-square Difference Test	*p*-Value	Correlation between Factors
12A	One Factor	594	.96	.02	20.756	.00	.80
	Two Corr Factor	**593**	**.98**	**.01**			
12B	One Factor	779	.89	.02	27.77	.00	.70
	Two Corr Factor	**778**	**.92**	**.02**			
12C	One Factor	819	.92	.02	13.919	.00	.85
	Two Corr Factor	818	.92	.02			
35A	One Factor	594	.94	.03	3.025	.10	.95
	Two Corr Factor	593	.94	.03			
35B	One Factor	819	.90	.02	1.050	.30	.96
	Two Corr Factor	818	.90	.02			
35C	One Factor	819	.95	.02	19.066	.00	.78
	Two Corr Factor	**818**	**.97**	**.01**			
68A	One Factor	594	.96	.01	6.915	.02	.97
	Two Corr Factor	593	.96	.02			
68B	One Factor	819	.94	.02	9.374	.00	.87
	Two Corr Factor	818	.94	.02			
68C	One Factor	819	.93	.01	6.484	.01	.84
	Two Corr Factor	**818**	**.94**	**.01**			
912A	One Factor	594	.82	.03	6.748	.01	.85
	Two Corr Factor	**593**	**.91**	**.03**			
912B	One Factor	819	.94	.02	8.530	.00	.89
	Two Corr Factor	818	.94	.02			
912C	One Factor	819	.90	.02	2.841	.10	.93
	Two Corr Factor	818	.90	.02			

4. Linking Analysis

Student performances on the 302 MD Listening Field Test were linked to the 301 SB Listening Test in order to derive item difficulty parameters used to score all students operationally on the 302 MD Listening Test during the 2013-2014 school year. Since the common-person linking design relies on the performance of a group of participants who are common to both assessments to establish the linking relationship between assessments, it is essential to ensure that students' performances across the two assessments are stable. Prior to conducting the linking analysis, an outlier analysis was conducted to identify and remove students whose comparative performance on the operational and field test administrations was considered unstable.

4.1 Rasch-Based Outlier Analysis

A major assumption underlying the common-person linking design is that the sample of students used are relatively stable in terms of how they behaved in the operational test and field test administrations. Given the short time interval between the operational test and the field test administrations, some consistency or stability in how participating students performed on the two assessments would be expected; if this assumption is violated, then the linking relationship derived using the data set will not be very accurate. Random factors like student fatigue or motivation can affect how a student performs on an assessment. Therefore, it is not improbable that some students may perform in an unexpected way (i.e., may exhibit incongruous performances) on two assessments that were designed to measure the same construct (Linacre, 2012). These students are considered outliers since their test performances are unstable and spurious. Retaining outliers in the linking study could negatively impact the linking results since the random errors created by these outliers would be mistakenly treated as real differences between the SB and MD assessments.

Although the assumption of common-person stability is hard to examine directly, this assumption can be evaluated through statistical procedures using empirical data. The Scatter Plot procedure in Winsteps (Linacre, 2012) was used to identify participating students that exhibited statistically significant differences in their person measures on the 302 MD Listening Field Test and the 301 SB Listening Test. Graphically, the Scatter Plot procedure produces plots of equivalent statistics, such as person measures, from two separate Rasch analyses (Linacre, 2012). It constructs a 95% confidence interval around the pairs of equivalent statistics using the standard errors of the equivalent statistics (Linacre, 2012). T-statistics that test the null hypothesis that the differences between the pairs of equivalent statistics are attributable to measurement error were output and used to numerically identify aberrant cases. The advantage of this procedure, in contrast to other statistical procedures such as linear regression, is that this procedure takes into account the standard errors of the parameter estimates. Procedurally, an independent Rasch calibration was first conducted on the 302 MD Listening Field Test data and the 301 SB Listening Test data, respectively, by test form. Then, the person measures estimated from the two Rasch analyses were entered into the Winsteps Scatter Plot procedure to identify students with t-statistics greater than +/-2. These students were classified as outliers and were removed from the linking analysis.

Figure 1 provides a graphical illustration of how the person calibration from the 302 MD Listening Field Test was compared with the person calibration from the 301 SB Listening operational test for test form 68B. The x-axis represents the logit values of the person measures from the 301 SB Listening Test and the y-axis represents the logit values of the person measures from the 302 MD Listening Field Test. The points representing the logit values of the students from both tests are plotted by their labels. The curved lines are the approximate 95% two-sided confidence bands. The confidence bands were computed and smoothed across all points. They are not straight because the standard errors of the person measures differ across persons. The dotted line in the plot is the empirical equivalence or best-fitting line. Points that are farther away from the best-fitting line indicate students whose scores on the two assessments are more divergent. The points that fall below the lower right curve line of the scatter plot are students whose performances on the 301 SB Listening Test are statistically significantly higher than their performances on the 302 MD Listening Field Test. Conversely, the points fall above the upper left curve line of Figure 1 are students whose performances on the 301 SB Listening Test are statistically significantly lower than their performances on the 302 MD Listening Field Test.

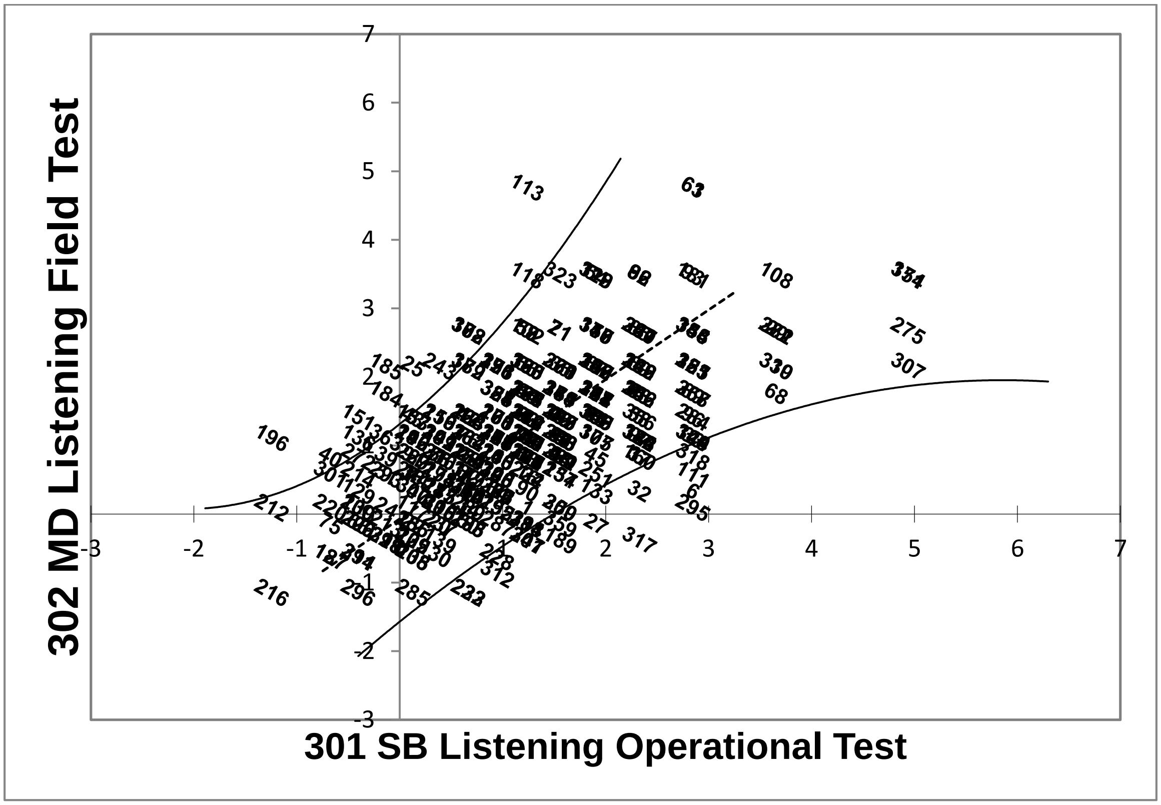

Figure 1: Comparison of scores on the 301 SB and 302 MD Listening Tests by student

Table 14 presents the results of the outlier analysis for all test forms. The second column of Table 14 presents the initial number of students taking each test form before the outliers were removed. The next two columns indicate the number of students flagged as outliers. Students with t-statistics greater than 2 are considered to have performed unusually well on the 301 SB Listening Test while students with t-statistics less than -2 are considered to have performed unusually well on the 302 MD Listening Field Test. The last column presents the final number of students analyzed from each test form after the outliers were removed.

Table 14
Results of the Outlier Analysis for All Grade-level Clusters and Tiers

Test Form	Initial Number of Students	Number of Students Performed Unusually Well on the 301 SB Test	Number of Students Performed Unusually Well on the 302 MD FT	Final Number of Students
12A	347	5	23	319
12B	743	3	142	598
12C	333	6	37	290
35A	203	5	8	190
35B	515	2	26	487
35C	421	23	7	391
68A	292	19	4	269
68B	385	16	11	358
68C	517	5	49	463
912A	272	17	2	253
912B	366	25	6	335
912C	298	9	8	281

4.2 Concurrent Calibration

A concurrent calibration procedure was used to estimate the item difficulty parameters of the 302 MD Listening Field Test while anchoring on the item difficulty parameters of the 301 SB Listening Test using Winsteps (Linacre, 2012). Through this concurrent calibration procedure, the item difficulty parameters of the 301 MD Listening Field Test were placed on the same scale as the 301 SB Listening Test. Once the 302 MD Listening Field Test items parameters were placed on the same scale as the 301 SB Listening Test, the quality of the linking was evaluated. Under the common-person linking design, if the linking was successful, the mean Rasch measures across students and the logit score distributions from the two tests should be similar. In addition, students' Rasch measures between the two tests should correlate reasonably well.

Table 15 presents the means and standard deviations of the students' Rasch measures for the 301 SB Listening Test and the 302 MD Listening Field Test. The Standardized Mean Difference (SMD), which is an effect size measure of group mean differences, between the two tests is presented in the last column. Overall, the mean Rasch person measures of the two tests are very close, suggesting that the linking was successful in putting the item parameters from the two tests on the same scale. The SMDs ranged from 0.01 to 0.08. Based on the guidelines suggested by Cohen (1988), an effect size of 0.0-0.2 is a small effect, indicating that the differences in group means between the two tests are small.

Table 15

Descriptive Statistics for the Rasch Analyses

		301 SB Operational			302 MD Field Test			
Test Form	Number of Items	Number of Students	Mean Rasch Measures	SD of Rasch Measures	Number of Students	Mean Rasch Measures	SD of Rasch Measures	Standardized Mean Difference
12A	18	319	0.347	1.250	319	0.390	1.334	.02
12B	21	598	1.101	1.101	598	1.161	1.208	.04
12C	21	290	0.693	1.194	290	0.776	1.225	.05
35A	18	190	0.469	1.817	190	0.403	1.499	.03
35B	21	487	1.353	1.024	487	1.311	0.936	.03
35C	21	391	1.452	0.821	391	1.437	0.831	.01
68A	18	269	0.211	1.237	269	0.276	1.418	.03
68B	21	358	1.472	0.985	358	1.458	0.985	.01
68C	21	463	2.256	0.680	463	2.325	0.912	.06
912A	18	253	0.373	0.970	253	0.385	1.004	.01
912B	21	335	1.944	1.020	335	1.843	0.826	.08
912C	21	281	2.264	0.945	281	2.182	0.885	.06

Figure 2 to Figure 13 present the cumulative frequency distributions of the 302 MD Listening Field Test and the 301 SB Listening Test by test form. In these plots, the x-axis represents the logit value of the students' Rasch measures and the y-axis presents the cumulative percentage of students with that particular Rasch value. The solid line represents the cumulative frequency distribution of the 301 SB Listening Test while the dashed line represents those of the 302 MD Listening Field Test. Figure 2 to Figure 13 show that, overall, the cumulative frequency distributions of the two tests aligned very well for most of the test forms. Only four test forms (12A, 68A, 68C, and 912C) exhibited slightly imperfect alignment of the two logit score distributions. Given the differences in test delivery mode between tests, the differences in test items between tests, and the time gap between the two test administrations, some discrepancies in the logit score distributions between tests are not unexpected. Overall, the amount of misalignment in the logit score distributions between tests is very small and tends to occur only in selected regions on the logit scale.

The correlation, the corrected or dis-attenuated correlation between the person measures of the two tests, and the correlation between the MD and the SB factor from the between-test CFA analysis are presented in Table 16. When two sets of person measures are correlated, measurement error lowers the correlation coefficient below the level it would have reached had the measures been perfect. Person measures are estimated with errors; the shorter the test is, the larger the measurement errors in the estimates (Wang, 2004.) Measurement error can be removed from a correlation coefficient to estimate the correlation coefficient dis-attenuated of measurement error (Spearman, 1904; Munchinsky, 1996) using the Spearman correction (Charles, 2005). If the dis-attenuated coefficient is near one based on the analysis conducted on the sample, it can be concluded that the two tests are measuring the same trait (Joreskog, 1971). The Spearman correction uses the population reliabilities of the instruments to make corrections to the correlation, while the factor-analytic approach involves estimating the correlation between latent variables associated with the instruments being correlated. Since the dis-attenuated formula assumes that measurement errors are constant across person measures while in the Rasch Model context, measurement errors are different depending on the person measure and applying the Spearman correction to the correlation between person measures may not be appropriate (Wang, 2004.) It has been suggested that the factor-analytic approach takes the measurement error of the instruments directly into account in the model and therefore is dis-attenuated of measurement error (Charles, 2005). The factor-analytic approach to estimating the dis-attenuated correlation has not been seen in applied research due to the complexities involved in applying the model, and this discussion is beyond the scope of this report. However, the between-test CFA analysis presented earlier in this report provides estimates of the correlations between the MD and SB factors; they can be used as an alternative but indirect measure of the relationship between the person measures of the two tests that are dis-attenuated of measurement error.

The Spearman correction requires estimates of the population reliabilities for both the 302 MD Listening Field Test and the 301 SB Listening Test. The Cronbach's alphas for the 301 Listening Test based on all WIDA ACCESS test takers (Center for Applied Linguistics, 2014) were used in the computation of the Spearman correction since they are more reliable estimates of the population reliabilities than those computed using the MD Listening Field Test sample data. One controversy related to the use of the Spearman correction, however, is that the correction could result in a correlation greater than one. For this reason, Spearman correction values that are greater than 1 are usually reported as 1. The Spearman correction is known to be inflated if the assumptions are not met, population reliabilities are underestimated, the measurement error is not normally distributed, and the sample size is small (Munchinsky, 1996; Charles, 2005). It is well known that the population reliabilities are underestimated and that Cronbach's alpha tends to underestimate the reliability of heterogeneous tests. Therefore, it is expected that some of the Spearman correction values in this analysis may be greater than one.

Table 16 shows the correlation and the Spearman correction between the person measures before and after outliers are removed, as well as the correlation between the MD and the SB factors from the between-test CFA analysis. The correlation between person measures of the two tests before outliers are removed is reported only as a reference, as the main focus this analysis is the correlation between the person measures of the two tests after outliers are removed. The correction coefficients ranged from .54 (68C) to .76 (35A), and all correlation coefficients are statistically significant. The Spearman correction values ranged from .74 (12A) to 1 (12C, 35B, 68B, and 912B). There is no test of statistical significance for the Spearman correction, and it cannot be compared with uncorrected correlation coefficient (Munchinsky, 1996.) However, the values of the Spearman correction suggest that several low correlations between the person measures of the two tests (12A, 35C, and 68C, for example) can be attributed to measurement error. Lastly, in general, the Spearman correction values are higher than the correlations between the MD and SB factor, suggesting that Spearman correction may have been inflated due to the reasons described earlier. The correlations between the MD and the SB factor ranged from .70 (12B) to .97 (68A) and most of the correlations are in the high 80s and 90s, suggesting that overall the (dis-attenuated) correlations between the two tests are pretty high.

Table 16

Correlations before and after Removal of Outliers

	Before Outliers Removed		After Outliers Removed		
Grade	Un-attenuated Correlation	Spearman Correction	Un-attenuated Correlation	Spearman Correction	Correlation between MD and SB Factor
12A	0.54	0.69	0.58	0.74	0.80
12B	0.42	0.60	0.64	0.91	0.70
12C	0.56	0.90	0.73	1.00	0.85
35A	0.67	0.85	0.76	0.96	0.95
35B	0.56	0.86	0.67	1.00	0.96
35C	0.48	0.83	0.57	0.98	0.78
68A	0.64	0.80	0.71	0.90	0.97
68B	0.57	0.85	0.69	1.00	0.87
68C	0.42	0.64	0.54	0.82	0.84
912A	0.58	0.78	0.66	0.89	0.85
912B	0.58	0.96	0.68	1.00	0.89
912C	0.57	0.79	0.67	0.92	0.93

Figure 4 to Figure 15 presents the bivariate plots for each test form between students' Rasch measures on the 302 MD Listening Field Test and the 301 SB Listening Test, as well as the best fitting line predicting the 302 MD Listening Field Test logit score from the 301 SB Listening Test logit score. The bivariate plots showed that for most grade clusters, there are statistical outliers that could affect the estimation of the best fitting lines and the correlations between the person measures of the two tests. A closer examination suggests the presence of one or more pairs of observations with large standardized residuals (larger than 3) for most grade clusters. The bivariate plots also revealed that for most grade clusters, some observations are more reliable than others; in other words, the variance is not constant across the entire range of logit scores. Specifically, the variance seems to increase when logit score increases, implying that the assumption of homoscedasticity may have been violated to some degree and the correlation coefficient may not provide the most accurate summary of the relationship between the person measures of the two tests. Furthermore, the bivariate plots are pretty dispersed at the high end of the logit score distribution, in general. In other words, the range of observed 302 MD Listening Field Test logit scores for a given 301 SB Listening Test logit score is wider at the high end of the logit score distribution. This suggests that the conditional distribution of the MD 302 Listening Field Test logit score has a larger standard deviation at the high end of the logit score distribution. This variation is partially due to measurement error as well as variation between students, and it is a function of the tests. The presence of statistical outliers and the violation of the assumption of homoscedasticity suggest that the bivariate plots and the correlation coefficients presented in Table 16 should be interpreted with caution.

In summary, based on (a) the comparison of the students’ mean Rasch measures; (b) the score distributions on the 302 MD Listening Field Test and the 301 SB Listening Test; and (c) the correlations between students’ Rasch measures on the 302 MD Listening Field Test and the 301 SB Listening Test, the linking analysis appears to have successfully placed the 302 MD Listening Field Test items on the Operational ACCESS Listening Test score scale.

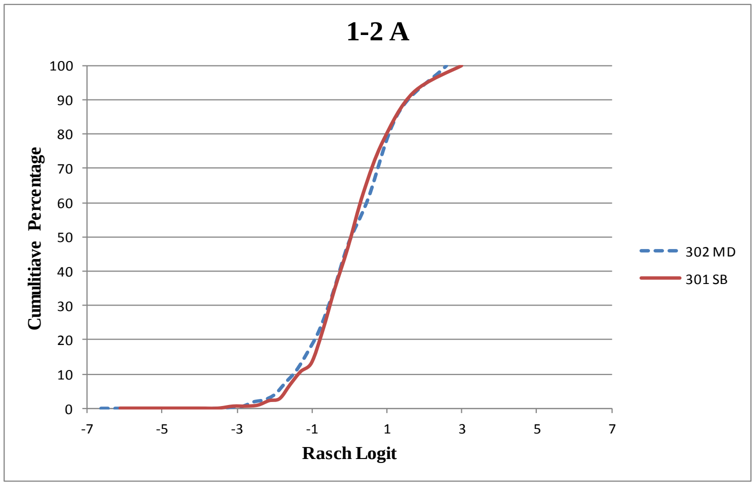

Figure 2: Cumulative frequency distributions for Test Form 12A

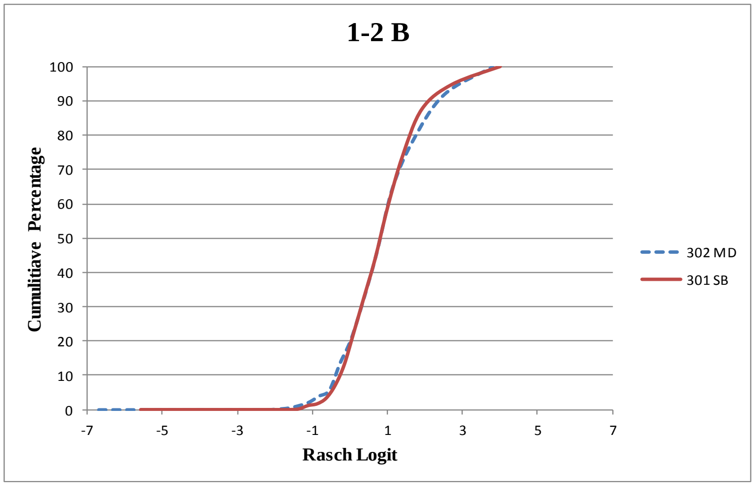

Figure 3: Cumulative frequency distributions for Test Form 12B

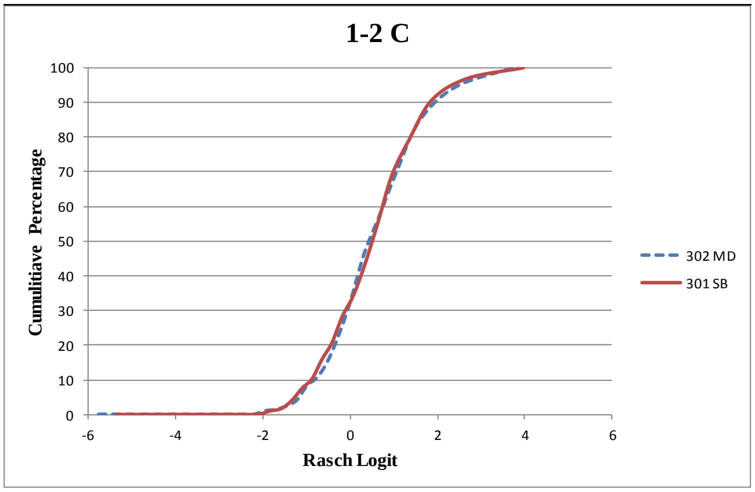

Figure 4: Cumulative frequency distributions for Test Form 12C

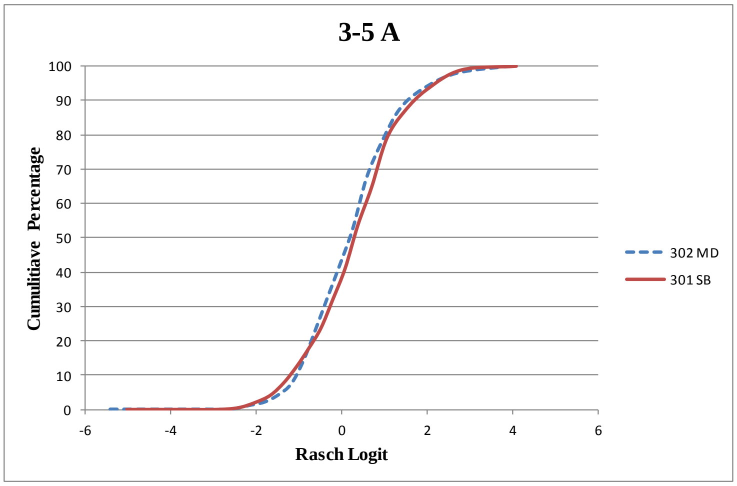

Figure 5: Cumulative frequency distributions for Test Form 35A

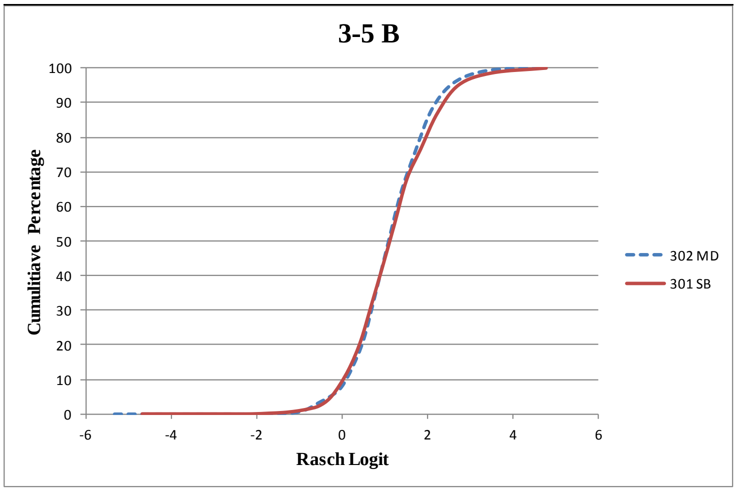

Figure 6: Cumulative frequency distributions for Test Form 35B

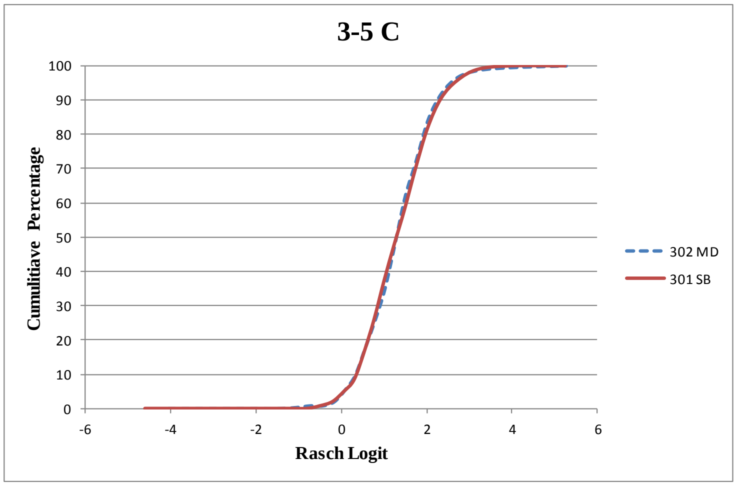

Figure 7: Cumulative frequency distributions for Test Form 35C

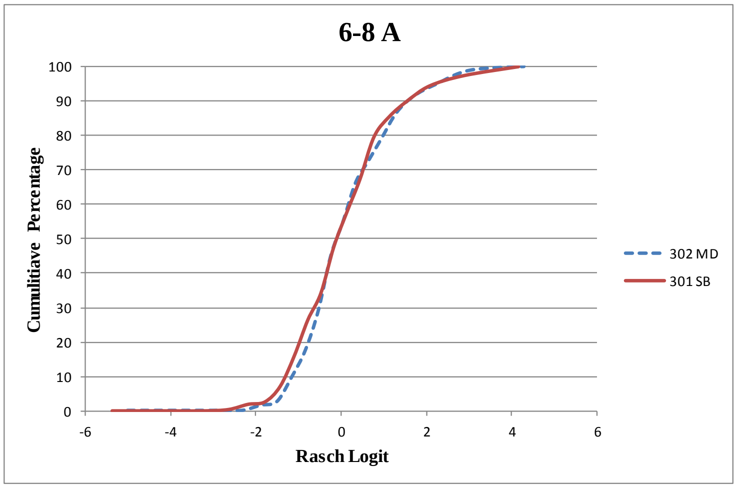

Figure 8: Cumulative frequency distributions for Test Form 68A

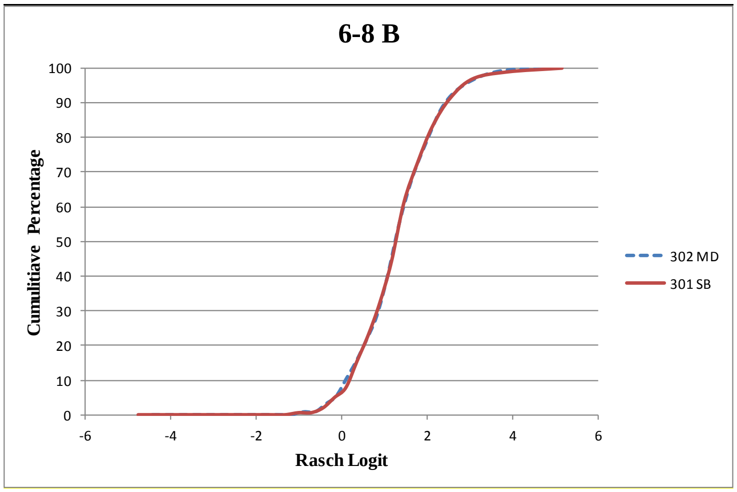

Figure 9: Cumulative frequency distributions for Test Form 68B

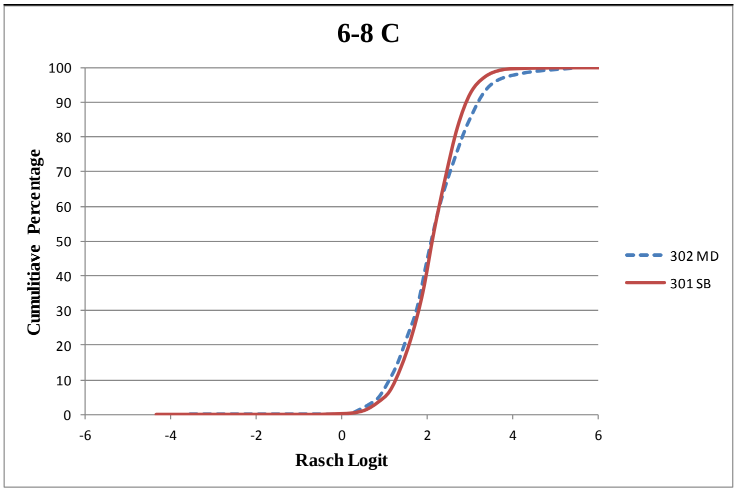

Figure 10: Cumulative frequency distributions for Test Form 68C

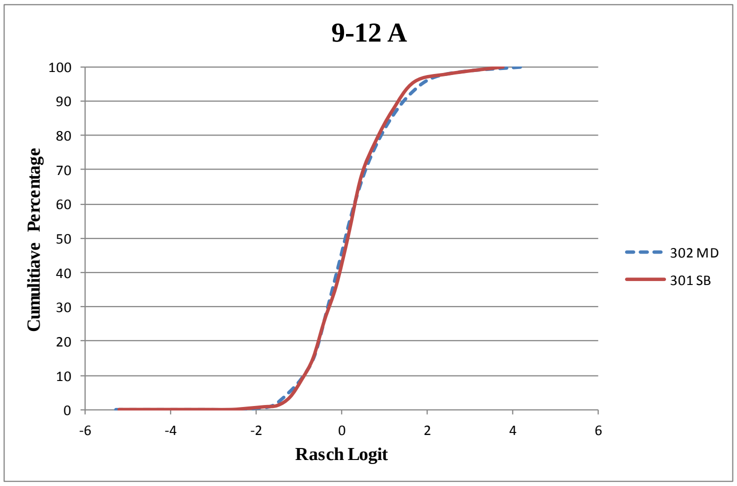

Figure 11: Cumulative frequency distributions for Test Form 912A

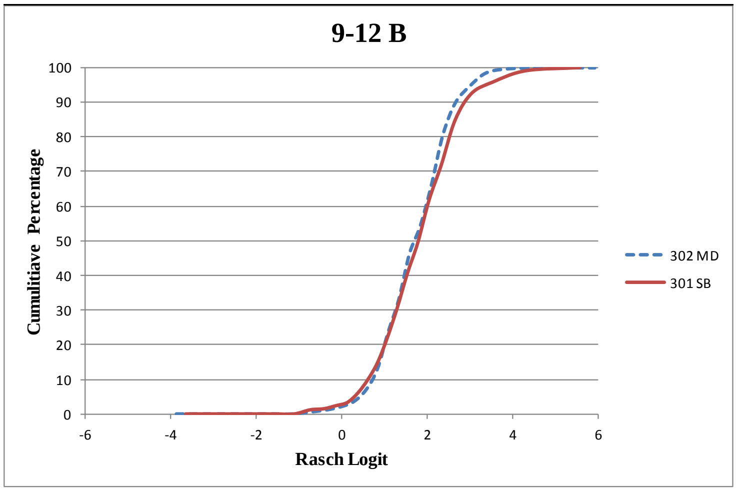

Figure 12: Cumulative frequency distributions for Test Form 912B

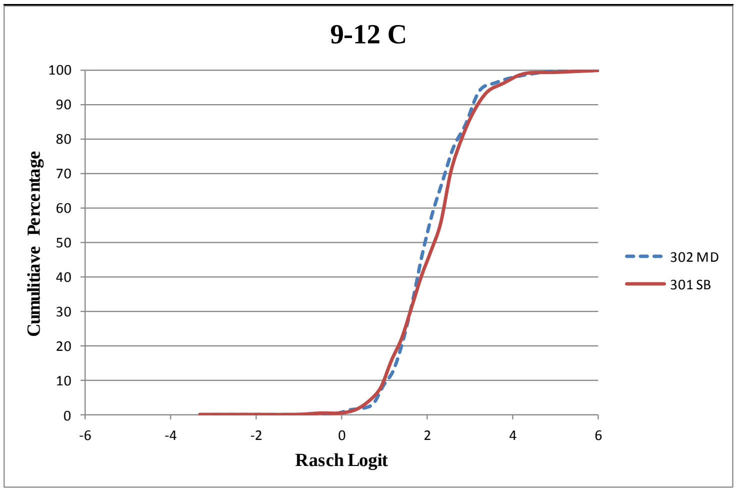

Figure 13: Cumulative frequency distributions for Test Form 912C

5. Appendix

WIDA ACCESS for ELLs® Listening Questionnaire

Instructions: Please answer the following questions about the test that you just took.

	Yes	Sometimes	No
1. Did you hear the recording?	□	□	□
2. Did you understand the instructions?	□	□	□
3. Did you follow along with the recording?	□	□	□
4. Did you have enough time to answer each question?	□	□	□
5. Did the people in the recording talk too fast?	□	□	□
6. Did you try your best on the test?	□	□	□
7. Did you like taking the test with a recording?	□	□	□

6. References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge.

Center for Applied Linguistics. (2014). Annual Technical Report for ACCESS for ELLs® English Language Proficiency Test, Series 203, 2011-2012 Administration (WIDA Consortium Annual Technical Report No. 8)

Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. Psychological Methods. 10, 2, 206-226.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, *9*(2), 233-255.

Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences (second ed.). Lawrence Erlbaum Associates.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Joreskog, K.G. (1971). Statistical analysis of sets of congeneric tests, Psychometrika, 36, 109-133.

Linacre, J. M. (2012). Winsteps® (Version 3.75. 0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2012.

Muchinsky P.M. (1996). The correction for attenuation. Educational and Psychological Measurement 56, 1, 63-75.

Muthén, L. K., & Muthén, B. O. (2013). Mplus 7.11. Los Angeles, CA: Muthén & Muthén.

Spearman C. (1904). The proof and measurement of association between two things. American Journal of Psychology, 15, 72-101.

Wang, W.C (2004). Direct Estimation of Correlation as a Measure of Association Strength Using Multidimensional Item Response Models. Educational and Psychological Measurement, 64, 6, 937-955