2.1 ACCESS Online Score Scale Maintenance: Listening

(Series 400 to Series 303)

Prepared by:
Shu Jing Yen, James Marcus, and Melissa Amos
October 2015

Issue

The purpose of this document is to provide an update of the approach to link the Grades 1-12 ACCESS for ELLs® 2.0 Series 400 Online (hereafter ACCESS 2.0) assessment of academic English language proficiency development expressed in the Listening domain (hereafter "Listening Test") to the ACCESS for ELLs® (hereafter ACCESS) Listening Test in order to maintain the ACCESS Listening reporting scale.

Background

The ACCESS Listening Test is a paper-based (PB) media-delivered test; items are delivered via CD or streaming audio from the MetriTech website, and students record their responses in a paper test booklet. ACCESS 2.0 Online differs from ACCESS in the following ways:

- Beginning with the operational 2015-2016 ACCESS 2.0 Series 400 Online Test, all Listening Test stimuli and audio are presented on the computer. Instead of marking their answer choices on paper, students select their answers on the computer.

- The clustering of the grade levels in the elementary school grades has changed, as shown in Table 1.

Crucially, the construct being measured has not changed; ACCESS 2.0 uses the same test blueprint and covers the same content as ACCESS. Thus, even though the Listening Test items for ACCESS 2.0 Series 400 Online are newly developed, their specifications are founded on those used for the items in ACCESS Listening Test.

Table 1 summarizes the differences between the ACCESS 2.0 Series 400 Online Listening Test and the ACCESS Listening Test with regard to grade-level clusters, items, stimulus presentation and response mode.

Table 1
Differences between Listening Tests for ACCESS 2.0 Series 400 Online and ACCESS Series 303

	ACCESS 2.0 Series 400 Online	ACCESS Series 303
Grade-level clusters	1 2–3 4–5 6–8 9–12	1–2 3–5 6–8 9–12
Items	New Development	Same as 302
Stimulus Presentation	Computer	Paper & Media
Response Mode	Computer	Paper

Method

The Listening Field Test was designed with the aim of collecting a sample of student performances on both the ACCESS 2.0 Series 400 Online field test and the operational ACCESS Test such that a common-person design could be used to link the two tests. The ACCESS 2.0 Series 400 Online field test folders were administered to samples of students within two weeks of taking the operational ACCESS Listening Test Series 303. The target sample size was 350 students per field test folder. The ACCESS Listening Test's operational item parameters are used as anchors in a concurrent calibration with the ACCESS 2.0 Series 400 Online field test items to establish the link between the two tests. Figure 4, located in the Appendix at the end of this document, illustrates the three step plan of linking via a common-person design: (a) Step I, an outlier analysis; (b) Step II, concurrent calibration; and (c) Step III, a verification study. As of October 2015, steps I and II have been completed.

Step I: Outlier Analysis

To safeguard the accuracy of the common-person linking, an outlier analysis was conducted to identify and remove those students whose comparative performance on the operational and field test administrations was considered to be too inconsistent given the short time interval between administrations. Although the two assessments are designed to measure the same construct, random factors like student fatigue or motivation can affect how a student performs on a given assessment, leading to spuriously large differences in a student's ability estimates. This instability negatively impacts the linking results since the random errors created by these outlier test-takers would mistakenly be treated as real differences between the assessments.

The scatter plot procedure in the Rasch-based software Winsteps was used to identify participating common students who exhibited statistically significant differences in performance on the ACCESS 2.0 Series 400 Online Listening field test and the ACCESS Listening Test (as evidenced by their person measures). Graphically, the scatter plot procedure plots the joint distribution of common students' Rasch-analysis-derived equivalent statistics, such as

person measures, on the two tests. It also constructs a 95% confidence interval of the joint distribution using the standard errors of the equivalent statistics. In addition to the graph, Winsteps also outputs t-statistics testing the null hypothesis that the differences between the pairs of equivalent statistics can be attributed to measurement error; these values serve to numerically identify aberrant cases. The advantage of this procedure, in contrast to other statistical procedures such as linear regression, is that it takes into account the standard errors of the parameter estimates.

Procedurally, an independent Rasch calibration was first conducted on the ACCESS 2.0 Series 400 Online Listening field test data, and then on the ACCESS Listening Test data, by grade-level cluster and tier. 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 removed from the linking analysis.

Figure 1 presents a real example of the scatter plot procedure. The points representing the person-measure logit values of the students from both tests are plotted by their labels. The curved lines are the approximate 95% two-sided confidence bands 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 represent those students whose performances on the ACCESS Listening Test was statistically significantly higher than their performances on the ACCESS 2.0 Series 400 Online Listening field test; conversely, the points that fall above the upper left curve line of Figure 1 denote those students whose performance on the ACCESS Listening Test was statistically significantly lower than their performance on the ACCESS 2.0 Series 400 Online Listening field test.

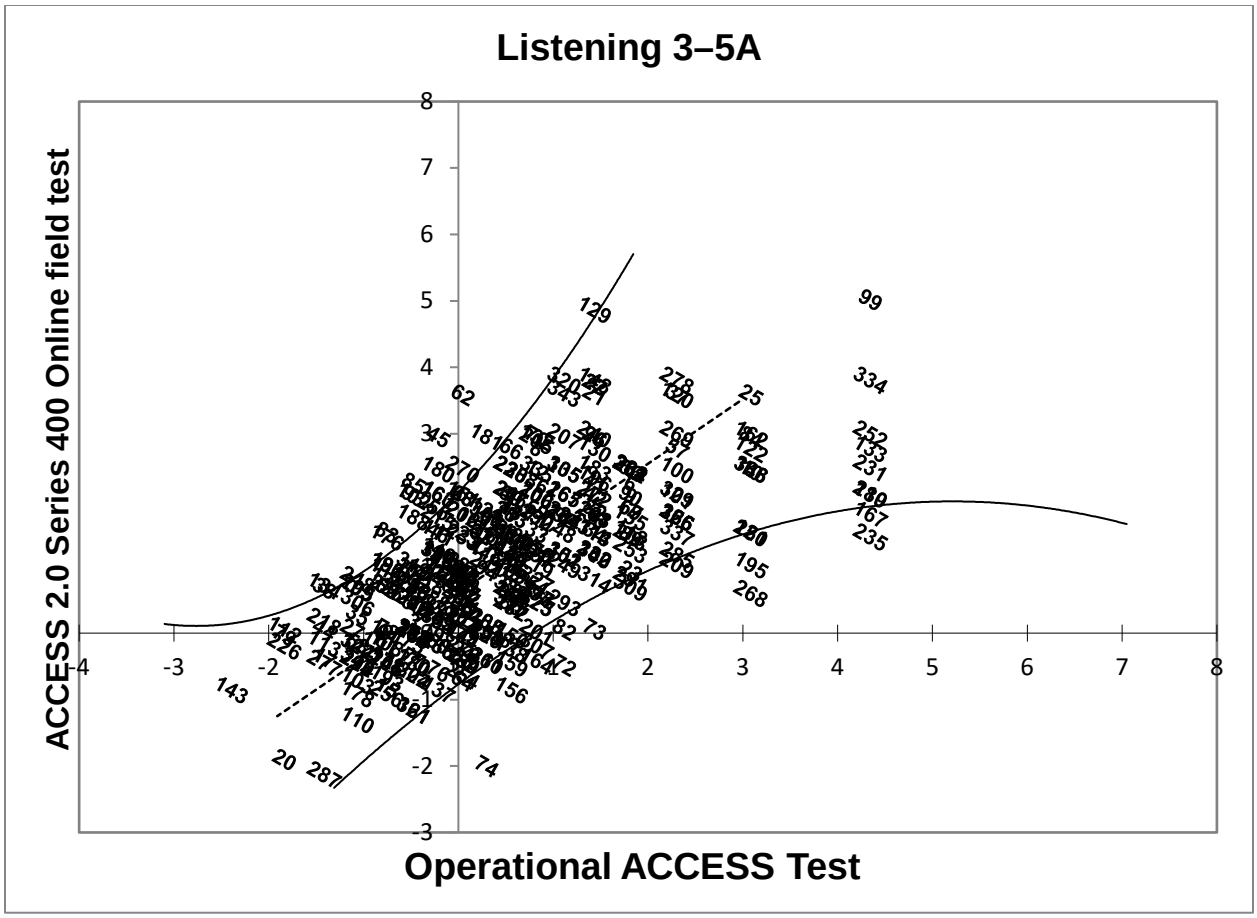

Figure 1. Example (Listening 3–5A) comparison of common students' person-measures (logit scale) on operational ACCESS Listening Test vs. ACCESS 2.0 Series 400 Online Listening field test

Step II. Concurrent Calibration

A concurrent calibration across grade-level clusters and tiers was conducted using Winsteps to estimate the item difficulty parameters of the ACCESS 2.0 Series 400 Online Listening field test items using the item difficulty parameters of the operational ACCESS Listening items as anchors. The presence of common field test folders both between tiers within a grade-level cluster and between grade-level clusters within a tier facilitated placing the item difficulty parameters of the ACCESS 2.0 Online Listening field test on the same vertical scale as the ACCESS Listening Test.

Once the ACCESS 2.0 Series 400 Online Listening field test item parameters are placed on the same scale as the ACCESS Listening Test, the quality of the linking can be evaluated. Under the common-person linking design, if the linking is successful, both the mean Rasch measures across students and the logit score distributions from the two tests should be similar.

Students' mean Rasch person measures, estimated based on the ACCESS 2.0 Series 400 Online Listening field test and the ACCESS Listening Test, are presented in Table 2 by Grade/Cluster. For Grades 1–3, the results are presented by grade level as opposed to grade-level cluster, as students in these grades are grouped in different grade-level clusters on the operational test and field test. The mean Rasch person measures of the two tests are close for three out of the six comparisons, namely Grade 3, Grades 6–8, and Grades 9–12. For the other three comparisons, namely Grade 1, Grade 2, and Grades 4–5, the mean Rasch person measures from ACCESS 2.0 are all lower than those from ACCESS. Overall, however, the linking was successful in putting the item parameters from the two tests on the same scale.

Table 2

Mean person measures (+SD) of operational ACCESS vs. ACCESS 2.0 Series 400 Online field test from Rasch analysis of common students, by Grade/Cluster

		ACCESS 303 Operational		ACCESS 2.0 Series 400 Online FT	
Grade/Cluster	Number of Students	Mean	SD	Mean	SD
1	2033	0.38	1.30	0.26	1.14
2	1951	1.26	1.27	1.04	0.97
3	1416	1.35	1.01	1.33	1.07
4–5	1522	1.70	1.14	1.55	1.07
6–8	1482	1.68	1.51	1.67	1.14
12	1367	1.56	1.38	1.62	1.06

As an example to illustrate another way the linking results are evaluated, Figure 2 presents the cumulative frequency distributions for the ACCESS 2.0 Series 400 Online Listening field test and the operational ACCESS Listening Test for Grade 3 students. In this plot, the x-axis represents the logit value of the students' Rasch measures and the y-axis represents the cumulative percentage of students with the particular Rasch measure on the x-axis. The red line represents the cumulative frequency distribution of the operational ACCESS Listening Test while the blue line represents those of the ACCESS 2.0 Series 400 Online Listening field test. Overall, the Grade 3 cumulative frequency distributions of the two tests aligned very well. Given the differences in the mode of test delivery, 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 to be expected. The amount of misalignment in the logit score distributions between tests is very small and occurs only in select regions on the logit scale.

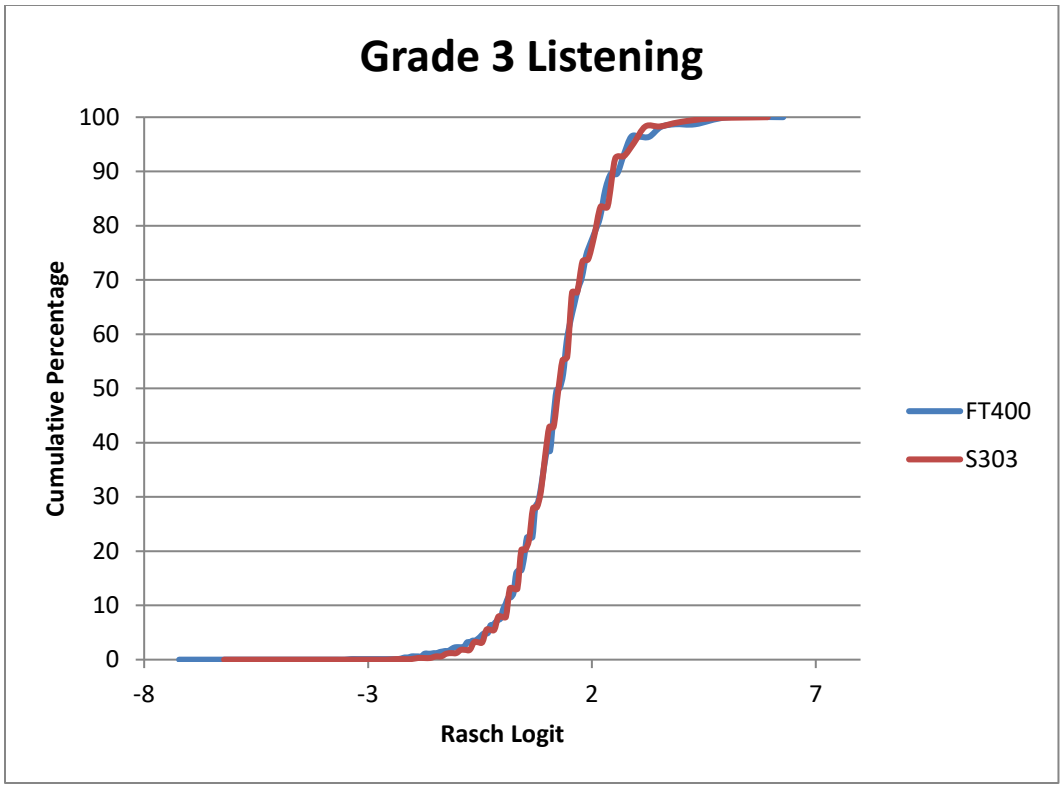

Figure 2. Cumulative score distribution for the ACCESS 2.0 Series 400 Online Listening field test (FT400) and the operational ACCESS Listening Test (S303)--Grade 3 Field Test Sample

Step III. Verification Study

After the field test linking analysis is completed, the item difficulty parameters derived from the field test administration will be used to prepare for the ACCESS 2.0 Series 400 Online Listening Test operational administration. For score reporting in the first operational year of ACCESS 2.0 Online (2015–16), these item difficulty parameters will be refined through a verification study using operational ACCESS 2.0 Series 400 Online Listening data collected during the early testing window; this is analogous to prior practices for equating the ACCESS operational administration with data collected early in the operational window. The ACCESS 2.0 Series 400 Online Listening Test item difficulty parameters will be initially anchored to the values derived from the ACCESS 2.0 Series 400 Online field test administration analyses, and displacement statistics will be evaluated to determine whether the parameters will need to be re-estimated based on the equating sample data. The final item difficulty parameters derived in this step will be used to score students for the ACCESS 2.0 Series 400 Online operational administration.

The field test design and proposed linking method used to maintain the ACCESS Listening Test score scale have been used successfully to link students' performances on the ACCESS Series 302 Media-Delivered (MD) Listening Test to the ACCESS Series 301 Script-Based (SB) Listening

Test. This change in delivery method represented a major transition in the Listening Test. In Series 301 and prior, the listening passages were read aloud by the test administrator from a script. Beginning in Series 302, the listening passages are pre-recorded. This transition may well have been a far greater change in the Listening Test than the shift from paper/media delivery to online delivery. Details of the ACCESS MD-based Listening Field Test study is available in the *ACCESS for ELL® Series 302 Media-Based Listening Field Test Technical Brief*.

Question for TAC

- Question: What additional analyses might the TAC recommend to ensure that the linking is conducted appropriately and that the Listening reporting scale is maintained?

APPENDIX

Common-Person Design

	ACCESS 2.0 Series 400 Online Listening Test	ACCESS Listening Test
Field test sample (350 per test form)	X	X

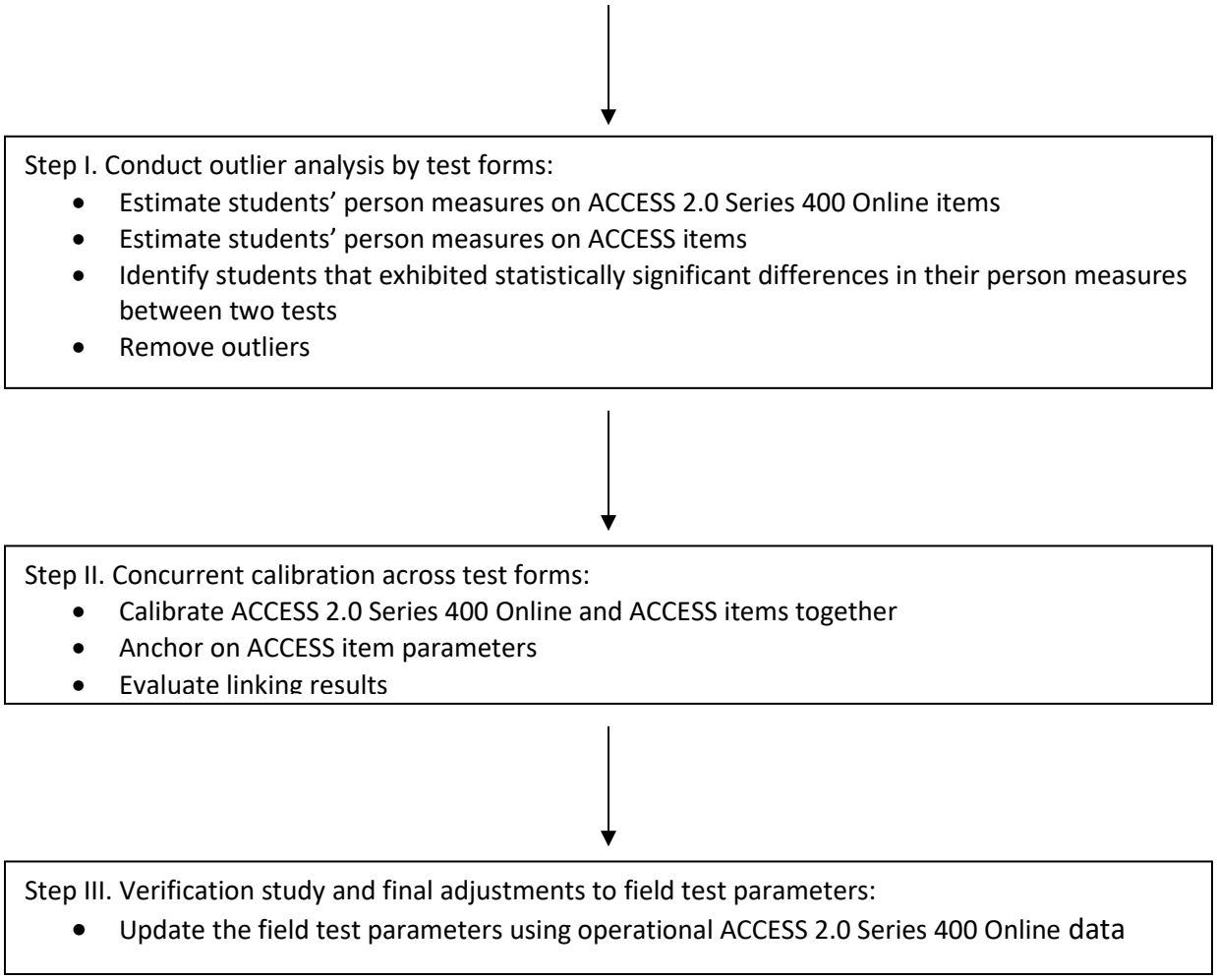

Figure 4. Proposed method to maintain the ACCESS Listening Test score scale