2.3 Maintaining the ACCESS for ELLs Speaking Score Scale

Prepared by:
Shu Jing Yen, Cary Lin, and Meg Montee
October, 2015

Issue

The purpose of this document is to provide an overview of the planned approach to maintain the score reporting scale from the ACCESS for ELLs (hereafter ACCESS) assessment of academic English language proficiency development expressed in the Speaking domain (hereafter "Speaking Test") for use on the ACCESS 2.0 Online (hereafter ACCESS 2.0) assessment.

Background

As presented at the 2014 TAC meeting, because of several changes between the ACCESS Speaking Test and the ACCESS 2.0 Speaking Test, the ACCESS 2.0 Speaking Test score scale must be reconstructed using data from the ACCESS 2.0 Speaking Test and the operational ACCESS Listening Test. It was determined that a standard setting and a cut score study would follow in summer 2015 so that new score interpretations would be applied to the ACCESS 2.0 operational administration. However, as discussed in the March 2015 TAC conference call, that plan has since changed; standard setting studies for ACCESS 2.0 will now be conducted in fall 2016. The revised standards and cut scores will not be used to report scores until the ACCESS 2.0 Series 401 administration. Therefore, the current ACCESS standards and cut scores will be used to report scores for the ACCESS 2.0 Series 400 administration. Such a delay was necessary in order to align ACCESS 2.0 performance standards to the new Common Core State Standards (See read-ahead 4.)

This delay creates an issue for the score interpretation of the Speaking domain for Series 400 since, unlike the other three domains, the ACCESS 2.0 Speaking score scale cannot be linked directly to the ACCESS Speaking score scale. More importantly, as discussed in greater detail below, the speaking score distribution for ACCESS and ACCESS 2.0 are likely to be different at the higher end of the distribution: Currently, about 40-50% of students are at or above PL 5 in Speaking; if the current ACCESS standards and cut scores are used to report scores for Series 400, it is expected that these percentages will drop and potentially resulting in a negative impact for the Annual Measurement Achievements Objects (AMAOs) for the 2014-2015 academic year. This read-ahead presents a plan to mitigate this potential negative impact to the AMAOs as well as to update the plan to reconstruct the ACCESS Speaking score scale.

ACCESS 2.0 Speaking Test

Several important enhancements have been made to the ACCESS 2.0 Speaking Test. These new features were designed with the aim of producing a more accurate

measurement of academic English language proficiency development as expressed in the Speaking domain. These features are:

- Enhanced standardization in the test administration procedure;
- Alignment to key academic language uses (Recount, Explain, and Argue), which are communicative functions that WIDA has identified as central to academic language;
- A model student who demonstrates the language expectation of speaking tasks and who serves to provide scaffolding for task input;
- An improved polytomous score scale that allows for better distinctions among performances that are aligned to the WIDA English Language Development (ELD) performance levels as described in the 2012 Amplification of the WIDA ELD Standards; and
- Enhanced standardization in the scoring procedure through centralized scoring, including the ability to measure and monitor the reliability of rater scoring for the summative test.

For each grade cluster, the ACCESS 2.0 Speaking Test is tiered as Pre-A, A, and B/C, as shown in Figure 1. Within a Tier, each box represents a folder that consists of one (Pre-A) or two tasks (A, B/C), with the number and color for each task indicating the targeted proficiency level (PL, as described in the WIDA English language development standards). For example, a single folder in Tier B/C has two speaking tasks, one of which targets PL 3 and the other which targets PL 5. Each Tier includes a total of three folders. For Tier A and Tier B/C, the three folders add up to a total of six tasks each. For Tier Pre-A, the three folders amount to three tasks. Note that each column is based on one or two of the WIDA Standards: SIL (Social and Instructional Language), LoLA/SS (Language of Language Arts and Language of Social Studies) and LoMA/SC (Language of Mathematics and Language of Science), and the number of unique tasks within each column is three because the Level 3 (P3) task is shared between Tier B/C and Tier A, and the Level 1 (P1) task is shared between Tier A and Tier Pre-A.

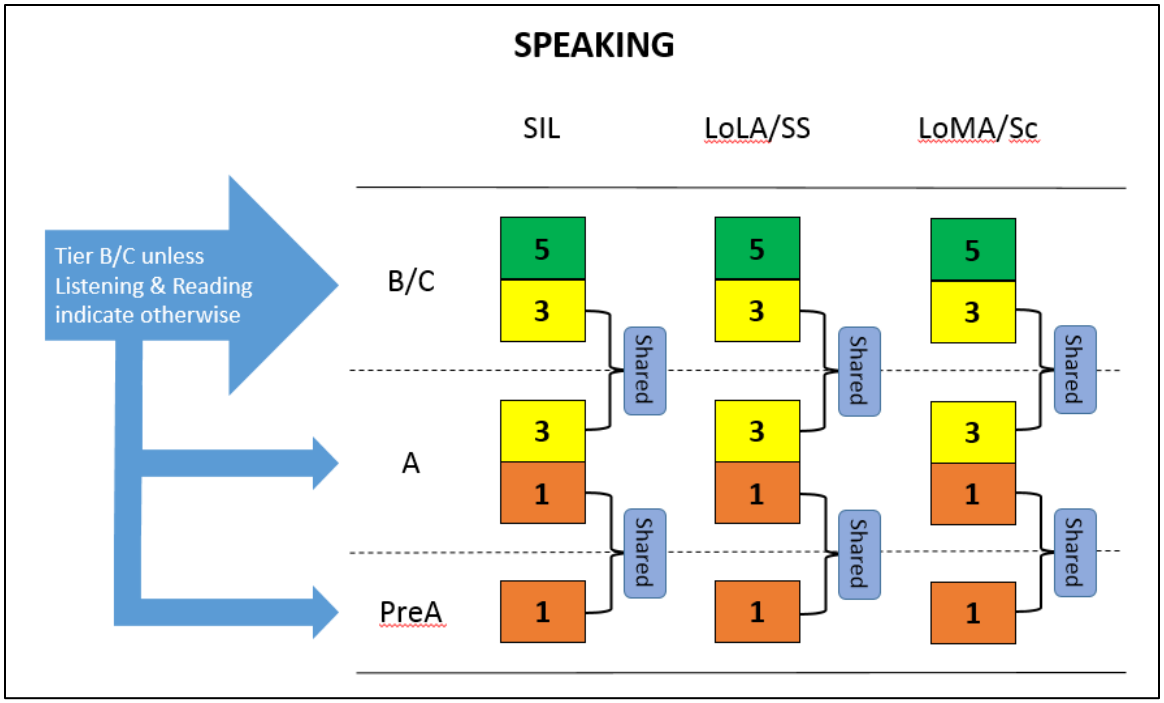

Figure 1. Task structure of the ACCESS 2.0 Speaking Test

In ACCESS 2.0, tier placement for the Speaking Test will be informed by student performance on the Listening and Reading Tests, as indicated in Figure 1. Most test-takers will be administered the Tier B/C test, though some students may be routed into Tier A if their performance on the Listening and Reading Tests warrants. Students whose scores on the Listening and Reading Tests indicate that they are at the earliest stages of English language development (e.g., students score at or below the chance level) will be routed into a Pre-A test form.

Figure 2 presents the task structure of the face-to-face Speaking Test that was used for ACCESS Series 100-303. By comparing Figures 1 and 2, one can see similarities and differences in task structures between the Speaking Test on ACCESS 2.0 and ACCESS. Similarities include the administration of tasks targeted to specific proficiency levels and the administration of three folders of tasks organized in the same way on both tests. However, there are several differences. The ACCESS Speaking Test was not tiered: one test form was developed for each grade-level cluster, for a total of 13 tasks as shown by the 13 rectangles in Figure 2. There were three tasks in Folder A and five tasks each in Folders B and C. The three tasks in Folder A targeted PLs 1 to 3; the five tasks in Folders B and C targeted PLs 1 to 5. The Speaking tasks were scored dichotomously; during testing, the administrator decided whether the student's performance either "Approaches" (= 0) or "Meets" (= 1) task-level expectations. The maximum raw score for Folder A was 3, and the maximum raw score for Folders B and C was 5. Therefore, raw scores for the full test ranged from 0-13. Based on these 13 raw score points, students were placed into one of six proficiency levels defined in the WIDA ELD Standards.

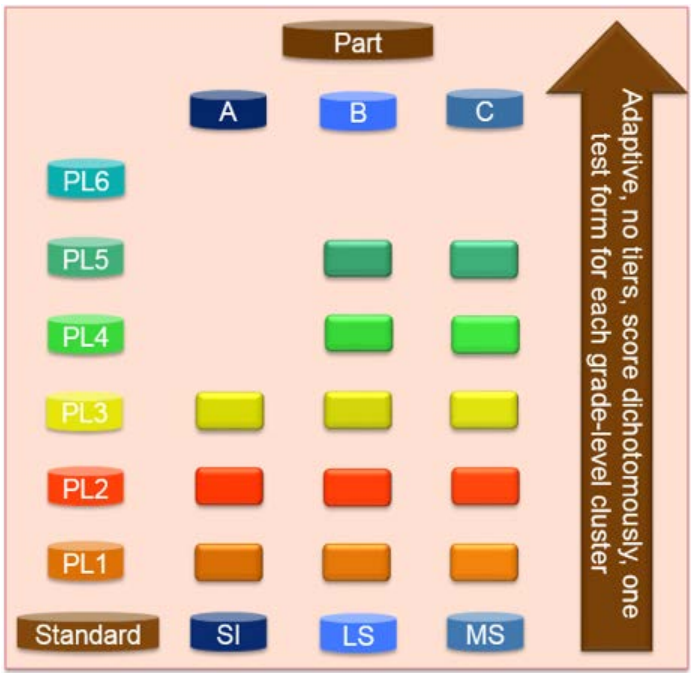

Figure 2. Task structure of the ACCESS Speaking Test

Unlike the dichotomous scoring on the ACCESS Speaking Test, each task on the ACCESS 2.0 Speaking Test is polytomously scored as shown in Table 1: No Response (0), Attempted (1), Adequate (2), Strong (3), and Exemplary (4). P1 tasks are scored on a scale of 0-2, and P3 and Level 5 (P5) tasks are scored on a scale of 0-4. According to the task structure by Tier shown in Figure 1, raw scores for Tier Pre-A and Tier A then range from 0-6 and 0-18, respectively. For Tier B/C, because students who are routed to this tier are expected to successfully complete any P1 task (i.e., say a minimum of two words in English) , raw scores are calculated as if the Tier B/C students were administered three P1 tasks and were awarded full marks on the P1 tasks. As a result, the raw scores for Tier B/C range from 6-30. In sum, raw scores for the full ACCESS 2.0 Speaking Test across tiers range from 0-30.

Table 1. ACCESS 2.0 Speaking Score scale

ACCESS for ELLs 2.0 Speaking Scoring Scale	
Score point	**Response characteristics**
Exemplary use of oral language to provide an elaborated response	• Language use comparable to or going beyond the model in sophistication • Clear, automatic, and fluent delivery • Precise and appropriate word choice
Strong use of oral language to provide a detailed response	• Language use approaching that of model in sophistication, though not as rich • Clear delivery • Appropriate word choice
Adequate use of oral language to provide a satisfactory response	• Language use not as sophisticated as that of model • Generally comprehensible use of oral language • Adequate word choice
Attempted use of oral language to provide a response in English	• Language use does not support an adequate response • Comprehensibility may be compromised • Word choice may not be fully adequate
No response (in English)	• Does not respond (in English)

It must be emphasized that, although the P1, P3, and P5 tasks on the ACCESS 2.0 Speaking Test are scored using the unified score scale in Table 1, an awarded score of "Adequate" for P1 tasks is not equivalent to the same score of "Adequate" for P3 tasks, because different level-specific tasks are designed to elicit responses with varying degrees of linguistic complexity. A conceptualization of how the three level-specific tasks on the unified score scale are related is shown in Figure 3. Clearly, an "Adequate" response for a P1 task may not reach the "Adequate" threshold for P3 and P5 tasks, while an "Adequate" response for a higher-level task will meet the language expectation of the "Adequate" threshold for a lower-level task.

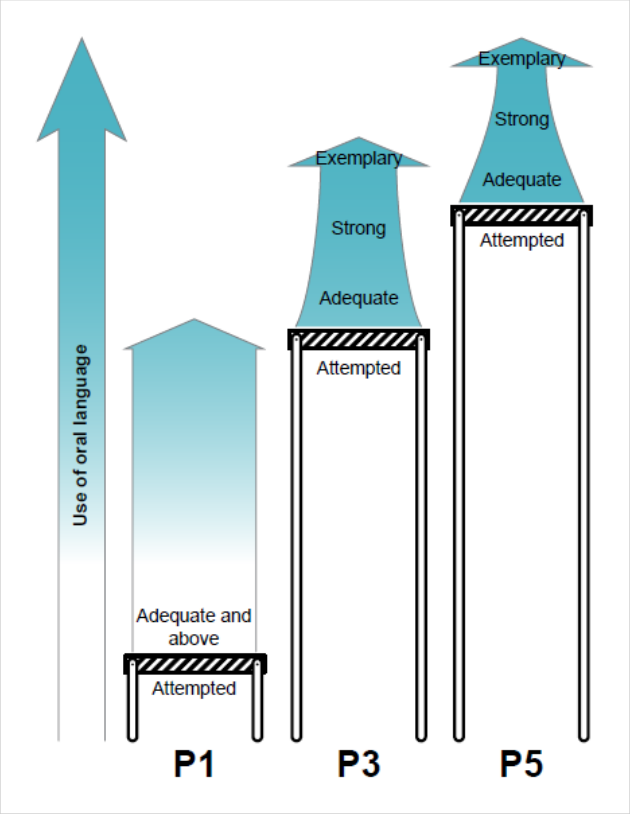

Figure 3. Score scale in relation to level-specific language expectation for the ACCESS 2.0 Speaking Test

Another difference in the task structures between the Speaking Test on ACCESS 2.0 and ACCESS is the proficiency levels at which the tasks are targeted. As mentioned above and as shown in Figure 2, for the ACCESS Speaking Test, the three tasks in Folder A target PLs 1 to 3 and the five tasks in Folders B and C target PLs 1 to 5. On the ACCESS 2.0 Speaking Test, as shown in Figure 1, tasks in Tier Pre-A target PL 1, tasks in Tier A target PLs 1 and 3, and tasks in Tier B/C target PLs 3 and 5.

Additionally, the ACCESS 2.0 Speaking Test and the ACCESS Speaking Test differ in the conceptual relationship between score scale points and the WIDA proficiency levels. On the ACCESS Speaking Test, because each level-specific task is scored dichotomously ("Approaches" vs. "Meets"), a level-specific score scale point corresponds directly to a

WIDA proficiency level. For example, if a student meets the language expectation of a P5 task, the student will be awarded a score of 1, which is conceptually linked to the attainment of WIDA PL 5. This one-to-one correspondence between score scale points and the WIDA proficiency levels does not apply to the ACCESS 2.0 Speaking Test because of its polytomous score scale. Table 2 demonstrates the conceptual relationship between score scale points and the WIDA proficiency levels for each level-specific task on the ACCESS 2.0 Speaking Test. For a P5 task, each score scale point, starting from 1, is conceptually linked to at least two WIDA proficiency levels. For example, a score of 4 (Exemplary) on a P5 task corresponds to WIDA PLs 5 and 6. Moving from the one-to-one correspondence between the score scale and WIDA proficiency levels to the current one-to-more correspondence can better address the fact that student performances will vary in quality even within a level-specific task. As such, the polytomous score scale allows raters to better distinguish performances relative to the language expectation for each task.

Table 2.
A conceptual relationship among task level, score scale, and WIDA Proficiency Levels for the ACCESS 2.0 Speaking Test

Task level	**Score scale**	**Psychometric Score Point**	**Conceptual Interpretative "Crosswalk" to WIDA Proficiency Levels***
P5	Exemplary	4	PL 5/PL 6
	Strong	3	PL 4/PL 5
	Adequate	2	PL 3/PL 4
	Attempted	1	PL 1/PL 2/PL 3
	No Response	0	No Evidence
P3	Exemplary Strong	3/4	PL 3/PL 4
	Adequate	2	Pl 2/PL 3
	Attempted	1	PL 1/PL 2
	No Response	0	No Evidence
P1	Adequate and above	2	PL 1/PL 2/PL 3
	Attempted	1	PL 1
	No Response	0	No Evidence

*PL 1: Entering; PL 2: Emerging; PL 3: Developing; PL 4: Expanding; PL 5: Bridging; and PL 6: Reaching

Tasks on the ACCESS Speaking Test are administered face-to-face and scored live by local administrators, while tasks on the online format ACCESS 2.0 Speaking Test are presented via computer and the student responses are recorded and sent to a central repository, where they are rated by trained raters.

Being administered via computer, the tasks on the ACCESS 2.0 Speaking Test have a very different format from those on ACCESS Speaking Test. The presentation of the Speaking Test is meant to simulate communication between the examinee, a Virtual Test Administrator (VTA), and a model student. Task input includes audio, text, and graphics related to the theme of the folder. The VTA delivers the test instructions and the tasks, to which the model student and the test taker both respond. Everything said by the VTA is presented both as audio and text. Before asking the test taker a question, the VTA asks the model student a question that is similar or analogous to the question the test taker will be asked. After the model student responds, a question intended for the test taker is delivered, again, both aurally and as text. When ready to answer, the test taker responds to the task question by a clicking a record button on the test interface and speaking into the microphone on his or her headset. Test takers have a maximum of 20-50 seconds to record their answers, depending on the grade-level cluster and proficiency level of the task. Figure 4 models the components of the Speaking Test as they are laid out on the computer screen.

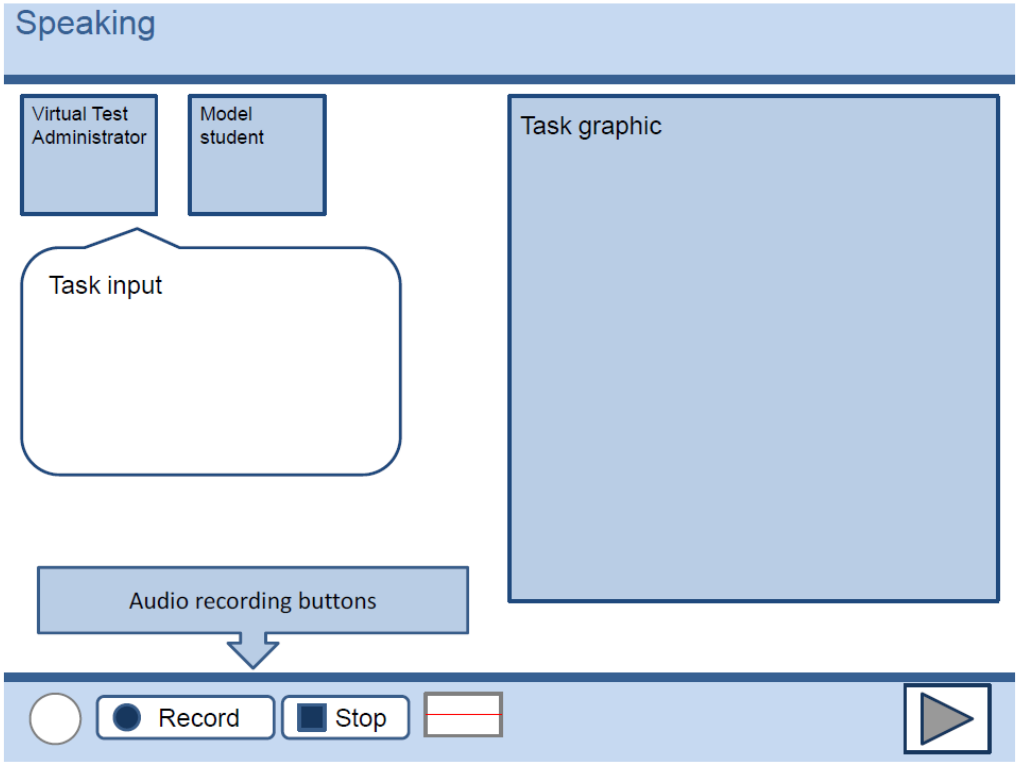

Figure 4. Components of the ACCESS 2.0 Speaking Test

As demonstrated, while the task and test design, as well as the scoring procedures (including the rubric) have changed from the ACCESS Speaking Test to the ACCESS 2.0 Speaking Test, there are still important similarities between the tests. These similarities help ensure that the construct measured by the two tests is comparable:

- They are both performance-based.
- They are both designed to allow test takers to show what they can do with their academic language proficiency in Speaking across the five WIDA standards.
- They both cover the range of proficiency levels defined by the WIDA ELD Standards (though ACCESS 2.0 is designed to more clearly assess the higher levels).
- They are both based on the Model Performance Indicators in or derived from the WIDA Standards.
- Many of the tasks on the ACCESS 2.0 Speaking Test were derived from operational ACCESS Speaking folders, adapted to meet the new task and test specifications.

Scaling Issues

Although the construct measured by the ACCESS 2.0 Speaking Test and the ACCESS Speaking Test is intended to be the same, albeit more refined in ACCESS 2.0, especially at the higher levels of proficiency, there are three unique technical issues related to the Speaking domain that make it complicated to maintain the ACCESS Speaking score scale. First, ACCESS 2.0 Speaking Test tasks will be scored using a four-point rubric as opposed to being dichotomously-scored as "Approaches" or "Meets," which requires a different method to calibrate the ACCESS 2.0 Speaking Test.

Second, the current procedure used to score the ACCESS Speaking Test does not appear to discriminate well among test takers at the high end of the student ability distribution. To illustrate this issue, Figure 5 shows a typical raw score distribution for the ACCESS Speaking Test, taken from the 2012-2013 Grades 3-5 ACCESS Speaking Test administration. This raw score distribution is negatively skewed, with more than 35% of students receiving a perfect score. One potential cause of this skewed distribution may be that the current scoring procedure requires test administrators to both administer the test tasks and to make the "Approaches" vs. "Meets" scoring distinction as soon as they hear the answer. Because time for rater training must be capped at 90 minutes, raters may not always be aware of what the task-level expectations are at the different proficiency levels. In addition, they may score more for the content of a student's response rather than for the type of language the student uses. In any case, this combined administration/scoring procedure may result in a tendency for administrators to give students the benefit of the doubt and award a "Meets" rating to any response that comes close to answering the question, whether or not expectations for the sophistication of language was met.

In addition, because the ACCESS Speaking Test is scored dichotomously as "approaching" or "meeting" task level expectations, there is no real way to demonstrate

performance at PL 6, since tasks at the PL 5 level are the highest available. This situation arose because the 2004 WIDA ELD Standards fully defined only five proficiency levels for Speaking at the time. (Note: It was decided in the first operational year of ACCESS that a "perfect score" on Speaking would automatically be interpreted as a PL 6, since a 6 was attainable in the other domains.)

The ACCESS 2.0 Speaking Test has been designed to more clearly distinguish between PLs 4, 5, and 6. First, ACCESS 2.0 Speaking Test tasks now feature the responses of a model student that illustrate for the test taker the sophistication of language at the word, sentence, and discourse levels that characterize performances at the highest level on the most challenging tasks. Second, the improved ACCESS 2.0 Speaking Test scoring rubric is more directly aligned to the six proficiency levels defined in the 2012 WIDA ELD Standards. Third, student responses are recorded and scored centrally by trained raters with stringent quality control. Because of these refinements, it is likely that there will be more variability in student scores on the ACCESS 2.0 Speaking Test compared to the ACCESS Speaking Test, particularly at the high end of the student ability distribution.

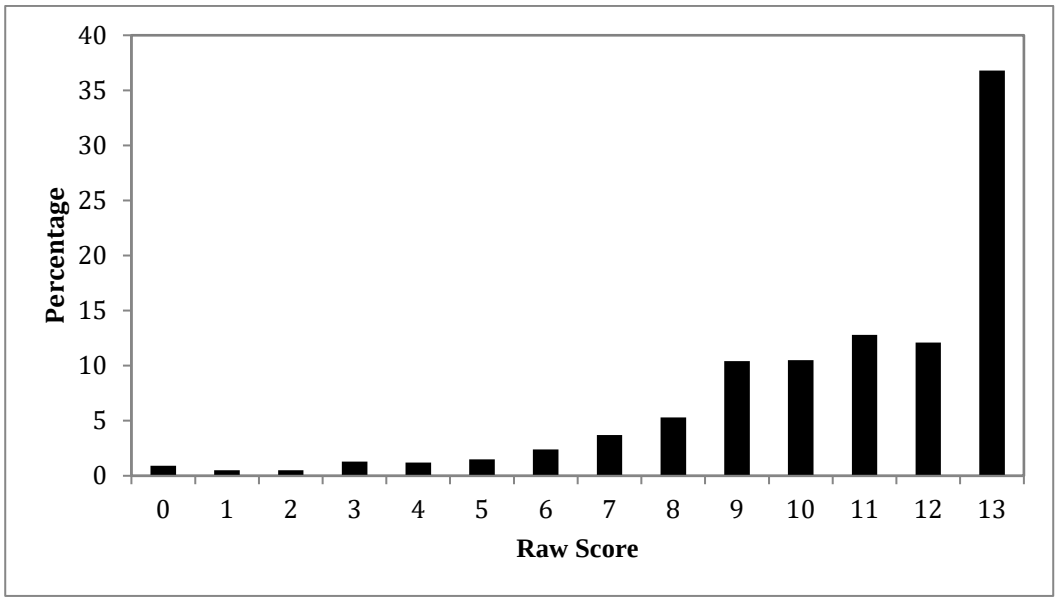

Figure 5. Grade 3-5 ACCESS Speaking Test raw score distribution from the 2013-2013 administration

Lastly, one aspect of the current ACCESS Speaking Test must be mentioned. The current Speaking tasks do not have pre-calibrated task difficulty parameters established on the ACCESS Speaking score scale; therefore, there is no easy way to use an Item Response Theory-based method to directly link performances on ACCESS 2.0 Speaking tasks to ACCESS Speaking tasks, even if the scoring method remains unchanged. (As some TAC members may recall, starting with ACCESS Series 200 (2008-2009), instead of conducting annual equating, pre-determined raw score to scale score tables were used in order to return to the original scaling used for ACCESS Series 100. This decision was based on the

belief that the original scaling best reflected the cuts that were established in the original standard setting study for the Speaking Test.)

The initial plan toward this end, presented at the 2014 TAC meeting, was to use the ACCESS 2.0 Speaking field test data to reconstruct the ACCESS Speaking score scale prior to the operational launch of ACCESS 2.0. In light of the delay of the ACCESS 2.0 standard setting studies to summer 2016, it is now feasible to reconstruct the scale using the operational data as opposed to field test data. For Series 400, the current ACCESS score scale and performance standards will be used to report scores. The complication is that the scores from the Series 303 and 400 administrations need to be linked in order to apply the current ACCESS score scale and performance standards to the new ACCCESS 2.0 Speaking Test. A special linking study, proposed by WIDA and CAL, will be conducted to link the Speaking raw scores of these two administrations in order to maintain the current Speaking score distribution during the transitional year. The basic design of the study is presented below.

Special Linking Study for the Transition Year

The linking study will implement an equipercentile procedure similar to that which is routinely implemented when a new state joins the consortium and a bridge study is conducted to establish the relationship between their old English Language Proficiency test and ACCESS. Procedurally, an equipercentile linking will be conducted using the Speaking raw score distributions obtained from the early return data of ACCESS 2.0 Series 400 administration and those from the population data of the ACCESS Series 303 administration. The early return data consists of all student data available to CAL by February 2016. Using the Early Return Data will be necessary since the linking study must be completed prior to score reporting for states with earlier testing windows.

The linking analysis will be conducted on the Speaking raw scores from the two administrations by grade-level cluster. Since the goal of the equipercentile procedure is to preserve the distribution of the ACCESS Series 303 Speaking Test, the proportion of student at each observable scale score and WIDA ELD proficiency level will be constrained to be more or less the same between series. Such an approach should provide some stability for the ACCESS 2.0 Series 400 Speaking score.

Maintaining the ACCESS Speaking Score Scale Using Series 400 Operational Data

The basic methodology that will be used to reconstruct the ACCESS Speaking score scale was presented during the 2014 TAC meeting. In this update, the ACCESS 2.0 Series 400 Speaking operational data will be used in the analysis as opposed to the field test data. The reconstructed Speaking score scale will be used in the ACCESS 2.0 standard setting studies (read-ahead 4) to set new performance standards and cut scores for Series 401.

The proposal to reconstruct the ACCESS Speaking Test score scale consists of three steps. An illustration of this proposal is presented in the Appendix of this document.

Step I. Reconstruct the Speaking Score Scale Using ACCESS 2.0 Series 400 Speaking and Listening Test Operational Data

A. Calibrate the ACCESS 2.0 Series Speaking Test tasks

A Rasch Rating Scale model will be used to put the task difficulty of the ACCESS 2.0 Speaking Test tasks and the ability of the students onto one common scale. However, as suggested by the TAC before, the Partial Credit model will also be investigated to determine if it provides a better fit to the data than the Rating Scale model. Since a generic scoring rubric is used to score ACCESS Speaking responses, a single rating scale will be modeled across all tasks and grade-level clusters, similar to how the ACCESS Writing Test was calibrated (Kenyon, 2006). The rating scale characteristics and the fit of the rating scale to the data will be examined to ensure that it is appropriate to model a single rating scale across tasks and grade-level clusters.

B. Concurrent calibration of ACCESS 2.0 Series 400 Speaking and Listening Data

The initial ACCESS Speaking vertical scale was built by linking the ACCESS Series 100 Speaking data to the ACCESS Series 100 Listening data, which was already on a vertical scale. (For more information see Kenyon [2006] and Kenyon et al. [2011].) A similar approach will be taken to reconstruct the vertical scale for the ACCESS 2.0 Series 400 Speaking Test. Student performance (i.e., ratings) on ACCESS 2.0 Series 400 Speaking tasks will be combined with performance (i.e., correct/incorrect responses) on the ACCESS 2.0 Series 400 Listening items. A concurrent calibration will be conducted to estimate the ACCESS 2.0 Series 400 Speaking task difficulty and scale step parameters while fixing the ACCESS 2.0 Listening item parameters to their pre-calibrated values. This joint calibration will put the task difficulty and scale step parameters of the ACCESS 2.0 Series 400 Speaking tasks on the same vertically equated scale as the ACCESS Listening Test items; that is, the joint calibration will vertically align all ACCESS 2.0 Online Speaking tasks and the common scale steps across tasks onto one common logit scale across all grade-level clusters and tiers.

C. Final Scale Adjustment

At the completion of Step I.B, the reconstructed ACCESS 2.0 Series 400 Speaking Test will have the same scale center, but may or may not have the same scale range as the current ACCESS Speaking Test. In order to maintain the characteristics of the current ACCESS Speaking scale, the range of the reconstructed ACCESS 2.0 Series 400 Speaking scale will be examined and adjustments will be made to either restrict or extend its range as needed.

Step II. Evaluate the Reconstructed Scale

Once the ACCESS 2.0 standard setting studies have been completed in summer 2016, student performances on the ACCESS 2.0 Series 400 Speaking Test, as defined by the old

ACCESS Speaking scale and the new ACCESS 2.0 Series 400 Speaking scale can be compared as a means of (a) evaluating the characteristics of the reconstructed scale and (b) projecting the potential impact of transitioning to the new ACCESS 2.0 Speaking scale in Series 401.

At the student level, the proficiency level classification outcomes of the same students can be compared on the two scales. At the group level, the percentage of students classified at various proficiency levels can be compared between the two scales, and similar distributions at the lower proficiency levels are anticipated. However, because the ACCESS 2.0 Speaking Test should better discriminate at the higher proficiency levels, we hope that a perfect score (which is interpreted as PL 6) is no longer seen as the modal score. While the modal score may still be at the high end for the Speaking domain, a better differentiation of performance among PLs 4, 5, and 6 is anticipated.

Step III. Pre-equating, Verification Study, and Final Adjustments to Score Tables

Similar to the Listening and Reading domains, the ACCESS 2.0 Speaking domain is moving to a pre-equated model starting with the Series 400 administration (see read-ahead 6.) Each student participating in the Series 400 administration will take one Series 401 Speaking folder after completing the Series 400 folders. The task difficulty parameters for the Series 401 tasks will be estimated while task difficulty and scale step parameters for the Series 400 tasks will be anchored to the pre-calibrated values, and displacement and quality control statistics will be evaluated to determine whether the anchor parameters require re-estimating. Through this concurrent calibration process, Series 400 and 401 task parameters will be placed on the same scale.

A verification study will be conducted using ACCESS 2.0 Series 401 Speaking Test operational data during the early testing window. The task difficulty parameters for the tasks and scale step parameters will be anchored to their pre-calibrated values, and displacement and quality control statistics will be evaluated to determine whether the task parameters require re-estimating. The final task difficulty parameters derived in this step will be used to score students for the ACCESS 2.0 Series 401 administration.

Questions for the TAC

- Question 1: Does the planned approach to maintain the ACCESS Speaking scores for the transition year appear sound?

- Question 2: Does the planned approach to reconstruct the ACCESS Speaking Test score scale appear sound?

- Question 3: Does the TAC have suggestions about additional ways to evaluate the psychometric quality of the reconstructed ACCESS Speaking Test score scale?

APPENDIX

Proposed Method to Reconstruct the ACCESS Speaking Test Score Scale

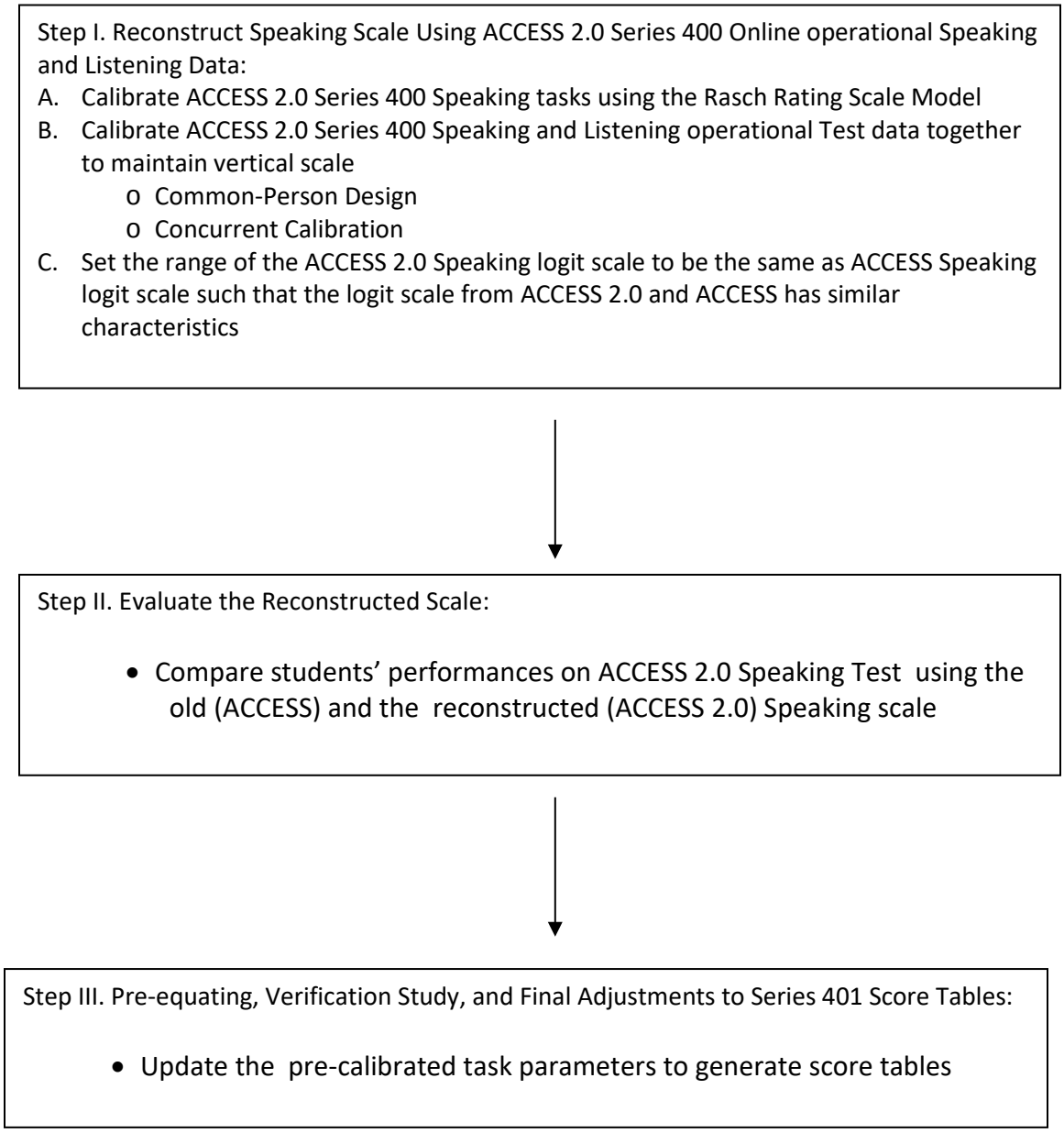

Figure 6. Proposed Method to Reconstruct the ACCESS Speaking Test Score Scale

References

Kenyon, D.M., MacGregor, D., Li, D., and Cooke, H. G. (2011). Issues in vertical scaling of a K-12 English language proficiency test. Language Testing, 28(3), 383-400.

Kenyon, D.M. (2006). *Development and Field Test of ACCESS for ELLs®* (WIDA Consortium Technical Report No. 1).