Maintaining Rater Reliability in Scoring ACCESS for ELLs 2.0 Paper Speaking Test

WIDA Resources for Educators

Students who take the ACCESS for ELLs 2.0 Paper Speaking test have their spoken responses scored by the Test Administrator who administered the Speaking test. Another term for this Test Administrator is *rater*. Raters must be trained and certified, so we can be confident that they interpret students' spoken language consistently and fairly, and that the scores are reported according to the WIDA English language proficiency standards. WIDA provides several different types of resources to support raters' training and reliability.

The purposes of this document are to:

- Explain the resources that WIDA provides that enable raters to score the Speaking test reliably
- Explain the importance of raters being formally certified via the WIDA Secure Portal
- Give an overview of what rater reliability means
- Describe how rater reliability may be monitored and why this is an important process

It is particularly important for Test Administrators to be aware of the requirements for maintaining accuracy, impartiality, and reliability if they administer the paper-based Speaking test to ensure the fairness of the test scores for students.

Importance of Training and Monitoring

The ACCESS for ELLs 2.0 Speaking Test consists of a number of performance tasks that require students to engage and respond verbally with as much academic language as they can produce. The students' spoken responses must then be evaluated and assigned a score, based on the ACCESS for ELLs 2.0 Speaking Scoring Scale.

Students who take ACCESS for ELLs 2.0 Online have their spoken responses digitally recorded and then **scored centrally** by DRC's trained raters. Students who take ACCESS for ELLs 2.0 Paper have their spoken responses **scored in real time by the Test Administrator who administered the Speaking test**. In both cases, it is important that the individual who scores the spoken responses is trained and certified.

Standardized rater training should yield scores that are impartial, consistent, and do not reflect the personal experiences of the individual doing the scoring. Consistent scoring should allow educators to make defensible decisions about a student's ability to produce spoken English.

In addition, student scores may be used for cross-district, cross-school, and cross-classroom comparisons. Scores must be comparable across a range of different contexts, so that high-stakes decisions that are informed by test scores may be made with confidence. Valid and fair test scores are essential for accountability systems, and rater training and rater monitoring are key elements that contribute to the fairness of the scores we award students and use for accountability.

What is Rater Reliability?

Rater reliability is a technical term that refers to the consistency of scores awarded to a student by multiple raters. Once the same group of raters evaluates multiple students, rater reliability can be calculated. We can then compare how consistent those raters are in evaluating the same spoken responses. If the raters are very consistent in how they evaluate the students' spoken language, their rate of reliability will be high. However, if many different scores are awarded to the same students, then the rate of reliability will be low, which is a concern, because it indicates a high risk of measurement error, which can result in students receiving scores that are not fair or appropriate. Reliability can be measured easily and reported as a percentage from 1 to 100.

On the ACCESS for ELLs 2.0 Speaking Test, raters may award one of five different score points.

- Exemplary
- Strong
- Adequate
- Attempted
- No Response

In order to calculate the reliability of the raters, these score points can be converted into numbers, as shown in the table below.

Score Point	Numeric Score
Exemplary	4
Strong	3
Adequate	2
Attempted	1
No Response	0

The next table (Example 1) shows two sets of example scores, using the numeric scores to represent score points on the ACCESS for ELLs 2.0 Speaking Scoring Scale.

Example 1

	Rater 1 scores	Rater 2 scores
Student 1	4	3
Student 2	3	3
Student 3	1	2
Student 4	0	0
Student 5	2	3
Student 6	2	2
Student 7	2	1
Student 8	1	3

In the example above, the two raters awarded the same scores in only three cases. Therefore, there is a high rate of inconsistency between the two raters' scores. We can calculate their interrater reliability as 3/8, or 37.5% agreement. This is a low rate of reliability. Typically, WIDA considers a minimally acceptable rate of reliability to be 70%. Look at Example 2 in the following table.

Example 2

	Rater 1 scores	Rater 2 scores
Student 1	4	4
Student 2	3	2
Student 3	1	1
Student 4	0	0
Student 5	2	3
Student 6	2	2
Student 7	2	2
Student 8	1	1

In Example 2, 6/8 scores are in exact agreement, giving an interrater reliability rate of 75%. This represents an acceptable reliability rate among the raters.

It is important to have **high rates of reliability** between raters, so we can be sure that students are being scored accurately and fairly. If some raters are lenient and some raters are strict, or others simply be inconsistent, then student scores will not truly reflect the students' abilities.

It may also be useful to consider the number of *adjacent* scores awarded by raters. *Adjacent* means that the scores differ by one score point. For example, if the first rater awards a score of **Strong**, but the second rater awards a score of **Adequate**, then these scores are adjacent. Ideally, we would want the total of scores that are in *exact* agreement and that are *adjacent* to equal 100%. Scores that are non-adjacent (i.e., scores that differ by more than one score point) are an indication that raters have a problem with consistency and require additional training.

Rater Training

WIDA provides a series of training modules in the Secure Portal on the WIDA website. ACCESS for ELLs Speaking Test raters should complete three core modules:

1. Overview and Test Structure
2. Speaking Assessment Scoring Practice
3. Speaking Assessment Recommended Practice

WIDA strongly recommends that all new raters complete all three of these modules. These modules provide a comprehensive introduction to the ACCESS for ELLs 2.0 Speaking test and the opportunity to learn how to score students' spoken English reliably using the ACCESS for ELLs 2.0 Speaking Scoring Scale.

In addition to the modules described above, WIDA also releases supplemental training materials each year to refamiliarize experienced raters with the Speaking Scoring Scale and introduce new Speaking tasks and sample responses for the coming year. These materials, called Supplemental Training for the Speaking Assessment, reflect the Speaking tasks that will appear on the test in the current year. WIDA recommends that all raters (new and experienced) engage with these supplementary materials at the start of each scoring season. Reading and reviewing these materials will help raters maintain their reliability from year to year and contribute to the fairness of test scores awarded to all students.

Rater Certification

After completing the training modules described in the section above, new raters should take the relevant certification quiz. WIDA provides two quizzes: one for raters who will evaluate students

in Grades 1-5, and another for raters who will evaluate students in Grades 6-12. Raters should take the appropriate quiz.

The purpose of the quiz is to ensure that raters have internalized the Speaking Scoring Scale and can apply it consistently. Only raters who pass the quiz(zes) should administer and score the ACCESS for ELLs 2.0 Paper Speaking test.

Recertification for Experienced Raters

WIDA recommends that raters who have scored the ACCESS for ELLs 2.0 Speaking tests for several years also recertify before each testing season. This will familiarize them with the supplementary training materials for raters that reflect the new Speaking tasks that will appear in the coming test administration. Minimally, WIDA recommends that all raters review Module 2: Speaking Assessment Scoring Practice from the Core Speaking Assessment Training, then review the Speaking Tasks and Supplementary Training Materials for the coming year. Encouraging all raters to recertify annually by passing the quiz(zes) maximizes the likelihood that students will be scored consistently and fairly.

Rater Monitoring

After raters have been trained and certified, they are considered "qualified" and ready to score the ACCESS for ELLs 2.0 Speaking test. However, one risk to rater reliability may arise when educators evaluate students whom they teach. It can be challenging for educators who are familiar with a student's spoken language, as observed in the classroom, to set that knowledge aside during an administration of ACCESS for ELLs 2.0. If students do not perform as expected during the Speaking test, it is possible that educators who have taught the student may award a different score from a rater who has no other knowledge of that student. To avoid this situation, WIDA recommends that whenever possible, educators not evaluate their own students' spoken language proficiency on ACCESS for ELLs 2.0.

In addition, even when raters are well trained and certified, there is a possibility that their scores will drift over time. That is, as the rater scores students' spoken language, the rater may be influenced by the students whom they evaluate. If a rater evaluates a large number of high-proficiency students, for example, he or she may unconsciously raise their expectations and award lower scores than they should to students at intermediate levels of proficiency. The opposite may be true for raters who evaluate many low-proficiency students. This type of rater drift is a common occurrence and may be avoided via rater monitoring to check that raters are scoring in a manner that is consistent with the Speaking Scoring Scale. Typically, raters may be monitored from two perspectives, rater reliability, and score point distributions.

Rater Reliability

The reliability of raters can be monitored if two or more raters have evaluated the same students. In this situation, a simple Excel file (see sample provided by WIDA and associated instructions) can be used to calculate the rate of agreement between raters. Agreement rates below 70% indicate that raters are not assigning scores to students consistently. A recommended minimum target for agreement rates is 70%.

WIDA recommends that rater reliability be checked regularly, using a second rater. If possible, rater reliability should be checked daily during the first few days of operational scoring. After that, rater reliability should be reviewed at least once a week. WIDA recommends that at least 10% of Speaking tests are monitored.

Rater reliability can be checked using a second rater in one of two ways: 1) either by having a second rater sit in on the administration and score live, or 2) by recording the administration and having the rater listen to the recording sometime after the fact. WIDA appreciates the operational challenges of implementing a monitoring program, but such programs are effective ways to assuring the reliability of scores awarded by different raters.

If a second rater sits in on an administration, note that the second rater should be seated so as not to distract the student. In addition, the test administrator and the second rater should be positioned so that they cannot see each other's ratings.

If the administration is recorded, the recording device should be placed discreetly so that it does not distract the student. The second rater should listen to each speaking sample just one time.

Regardless of the method used, the agreement rate should be calculated as soon as possible, and feedback provided in a timely manner. If possible, raters should discuss ratings that are not in agreement. If raters can come to consensus on what score should have been awarded based on the Speaking Scoring Scale, then future ratings are likely to be more consistent.

Score Point Distributions

Score point distributions refer to the number of times a rater awards a certain score, and can provide further insight into rater reliability. The ACCESS for ELLs 2.0 Speaking Scoring Scale contains five main score points:

- Exemplary
- Strong
- Adequate
- Attempted
- No response

Recording the percentages of frequency for each score point awarded provides insight into the scoring trends of each rater. If one rater (Rater 1) awards 50% of all students Exemplary scores, yet another rater (Rater 2) awards only 30% of the same students Exemplary scores, this may provide an indication that Rater 1 scores more leniently than Rater 2.

While score point distribution trends over time provide a valuable insight into raters' scoring tendencies, interrater reliability data is the strongest indicator of consistency among different raters.

Checklist for Rater Training, Monitoring, and Recertification

- ✓ New raters complete all Speaking Assessment Training
- ✓ New raters take and pass the appropriate certification quizzes
- ✓ All raters recertify at the start of each testing season (review new materials, retake quiz)
- ✓ Only certified raters administer and score the ACCESS for ELLs 2.0 Speaking test
- ✓ Raters do not evaluate their own students, if at all possible
- ✓ Rater reliability and/or score point distributions are monitored regularly

Thank you for your attention to this information emphasizing the importance of rating ELLs accurately and fairly on the ACCESS Speaking Test.