ACCESS for ELLs
Speaking and Writing Scoring Quality Control Meeting

2017-18 Administration

Series 402

April 12, 2018

ACCESS for ELLs Quality Control- Speaking and Writing Scoring

April 12, 2018

Executive Summary

On April 12, 2018 representatives from four WIDA consortium states (CO, OK, VA, VT) met at the DRC scoring facility in Plymouth, MN to review processes that support the scoring of ACCESS for ELLs 2.0 Speaking and Writing domain tests. This quality control (QC) meeting was facilitated by staff from WIDA, DRC, and CAL.

The aims of this QC meeting were to review the processes that support the scoring of ACCESS for ELLs 2.0 Speaking and Writing responses. The main processes covered during this QC meeting were:

- Hiring of appropriate raters
- Rater training
- Scoring – the use of DRC's ScoreBoard platform
- Rater monitoring
- Analysis of inter-rater reliability data

These processes are essential for the reliable scoring of students' ACCESS for ELLs 2.0 Speaking and Writing responses. The participants were asked to evaluate a series of claims made about the scoring processes summarized above. The participants were required to evaluate each of the claims, based on the evidence presented by WIDA, DRC, and CAL using the four options shown below.

- *Sufficient relevant evidence was presented*
- *Some relevant evidence was presented*
- *Little relevant evidence was presented*
- *No relevant evidence was presented*

The evaluations completed by the SEA participants provided strong support for the claims WIDA, DRC, and CAL make for the robustness of the processes that support the scoring of the ACCESS for ELLs 2.0 Speaking and Writing domain tests. All claims made about the scoring processes were evaluated using the two highest of the four options (with a large majority being awarded the highest rating), indicating that sufficient relevant evidence was presented at the QC meeting for all claims made.

These results indicate that the ACCESS for ELLs 2.0 Speaking and Writing domain tests are scored by well-trained raters who use an appropriate scoring platform, and are effectively monitored throughout the scoring process. Most importantly, the reliability data show that scoring is done consistently and reliably, meaning that stakeholders should have confidence in the meaningfulness of the scores that are reported for the Speaking and Writing domain tests of ACCESS for ELLs 2.0.

Aims of QC Meeting

The aims of this QC meeting were to review the processes that support the scoring of ACCESS for ELLs 2.0 Speaking and Writing responses. The main processes covered during this QC meeting were:

- Hiring of appropriate raters
- Rater training
- Scoring – the use of DRC's ScoreBoard platform
- Rater monitoring
- Analysis of inter-rater reliability data

These processes are essential for the reliable scoring of students' ACCESS for ELLs Speaking and Writing responses. State education agency representatives (SEAs) were invited to review the processes DRC, WIDA, and CAL have implemented for the ACCESS for ELLs program and to evaluate the robustness of these processes. These participants were asked to review specific processes developed separately (but with some overlap) for the Speaking and Writing domain tests.

WIDA believes that if appropriate raters are hired and then well trained, the resulting pool of certified raters will, if the scoring processes are not prone to user or technical error, score students' responses consistently and reliably, with the assumption that scores are routinely monitored to guard against rater drift. Rater drift describes the rating behavior whereby a rater may systematically deviate (either more lenient or more severe) from the intended application of the scale. Participants were asked to evaluate the processes described above to provide a thorough overview of the scoring processes.

Methodology

WIDA adopted an Assessment Use Argument (AUA) approach to the evaluation of the scoring processes for this QC Meeting. Within an AUA framework, claims are made about an aspect of a test and these claims must be supported by the presentation of evidence that provides backing (support) for the claim. If the backing is seen as reasonable then the claim for the test is supported. WIDA aims to provide detailed and extensive support for the claims we make about the strong processes we have in place to support the scoring of the ACCESS for ELLs Speaking and Writing domain tests.

To provide an example of this AUA methodology, WIDA makes the following claims (Table 1) about the training of raters who score the ACCESS for ELLs Speaking domain test.

Raters are provided with training that ensures reliable scoring of ACCESS for ELLs Speaking responses.
The training period provided is of an appropriate duration.
Trainee raters are provided a sufficient quantity of Speaking samples.
Trainee raters are provided sufficient opportunity to practice scoring authentic samples.
Trainee raters are provided feedback when their scores differ from the pre-assigned score.
Certification criteria required to successfully complete training and score operational responses are set appropriately.
Staff conducting the rater training have the appropriate background and experience to lead the training.

Table 1. Claims for Rater Training - Speaking

During the Scoring QC Meeting, WIDA, DRC, and CAL staff presented the SEA participants with a range of evidence that supports the claims shown above. The participants engaged with the evidence presented and were encouraged to ask questions and request additional evidence if they felt that was necessary. The participants were then asked to evaluate each of the claims shown in Table 1 based on the evidence they had been provided. The participants were required to evaluate each of the claims, using the following four options.

- *Sufficient relevant evidence was presented*
- *Some relevant evidence was presented*
- *Little relevant evidence was presented*
- *No relevant evidence was presented*

Participants were instructed to choose only one of these four options for each claim that they were provided. The participants completed an evaluation checklist by hand during the Scoring QC Meeting. In addition to completing this evaluation checklist, the participants were also given multiple opportunities to record open ended comments about the processes they were evaluating. These detailed comments were intended to allow the participants to provide more specific and focused feedback on the scoring processes that the evaluation checklist may have failed to capture.

In the following Results section, we report the evaluation checklist data quantitatively. The four evaluation statements are transformed into numeric evaluations, as shown in the following table.

Evaluation Statement	**Numeric Value**
Sufficient relevant evidence was presented	4
Some relevant evidence was presented	3
Little relevant evidence was presented	2
No relevant evidence was presented	1

Table 2. Evaluation Statement Converted to Numeric Value

This transformation allows data to be presented in a way that makes for easy comparison between the different processes that were evaluated at the Scoring QC. The following Results section reports on both the quantitative data and all open-ended comments made by the participants.

Results

1. Hiring of Raters

The claims made for the processes supporting the hiring of raters to score the ACCESS for ELLs Speaking and Writing domain tests are shown in the following table.

Appropriate staff are hired to score ACCESS for ELLs Speaking and Writing responses.
The minimum educational requirements for raters are sufficient and appropriate for scoring ACCESS for ELLs.
Raters are familiar or familiarized with typical spoken or written language features of grades 1-12 ELLs.
Skilled raters are identified and retained from year to year, whenever possible.

Table 3. Claims Made for Hiring Raters

The participants evaluated these claims made on the evidence presented by WIDA, DRC, and CAL at the QC. The results of the participants' evaluation of these claims are shown in the following table.

Evaluation	Numeric value	Percentage
Sufficient relevant evidence was presented	4	87.5
Some relevant evidence was presented	3	12.5
Little relevant evidence was presented	2	0
No relevant evidence was presented	1	0

Table 4. Participant Evaluation of Claims Made for Hiring Raters

The following comment was added by one participant.

> *Would like to know the percent of those hired with degrees either in education, educating ELs, or something related.*

The quantitative responses and comment indicate that there were no major concerns expressed about the processes supporting the hiring of raters to score the ACCESS for ELLs 2.0 Speaking and Writing domain tests.

2. Training of Writing Raters

The claims made for the processes supporting the training of raters to score the ACCESS for ELLs 2.0 Writing domain test are shown in the following table.

Raters are provided with training that ensures reliable scoring of ACCESS for ELLs Writing responses.
The training period provided is of an appropriate duration.
Trainee raters are provided a sufficient quantity of Writing samples.
Trainee raters are provided sufficient opportunity to practice scoring authentic samples.
Trainee raters are provided feedback when their scores differ from the pre-assigned score.
Certification criteria required to successfully complete training and score operational responses are set appropriately.
Staff conducting the rater training have the appropriate background and experience to lead the training.

Table 5. Claims Made for Training Raters - Writing

The participants evaluated these claims made on the evidence presented by WIDA, DRC, and CAL at the QC. The results of the participants' evaluation of these claims are shown in the following table.

Evaluation	Numeric value	Percentage
Sufficient relevant evidence was presented	4	96.5
Some relevant evidence was presented	3	3.5
Little relevant evidence was presented	2	0
No relevant evidence was presented	1	0

Table 6. Participant Evaluation of Claims Made for Training Raters - Writing

No open-ended comments were made by participants in this section of the evaluation.

The quantitative responses and lack of negative comments indicate that there were no concerns expressed about the processes supporting the training of raters to score the ACCESS for ELLs 2.0 Writing domain test.

3. Scoring of Writing Responses

The claims made for the processes supporting the scoring of ACCESS for ELLs 2.0 Writing responses are shown in the following table.

ScoreBoard is an appropriate tool for scoring ELLs Writing responses.
Writing responses are clear enough for raters to read and score reliably.
ScoreBoard interface will allow Writing scores to be recorded without error.
Working environment in the scoring facility is appropriate for raters to score Writing responses productively.
Working environment in the scoring facility is appropriate for raters to score Writing responses without being distracted.
Scoring facility is secure and only accredited staff may enter the premises.

Table 7. Claims Made for Scoring Writing Responses

The participants evaluated these claims made on the evidence presented by WIDA, DRC, and CAL at the QC. The results of the participants' evaluation of these claims are shown in the following table.

Evaluation	Numeric value	Percentage
Sufficient relevant evidence was presented	4	100
Some relevant evidence was presented	3	0
Little relevant evidence was presented	2	0
No relevant evidence was presented	1	0

Table 8. Participant Evaluation of Claims Made for Scoring - Writing

No open-ended comments were made by participants in this section of the evaluation.

The quantitative responses indicate that there were no concerns expressed about the processes supporting the scoring of ACCESS for ELLs 2.0 Writing responses.

4. Monitoring of Writing Raters

The claims made for the processes supporting the monitoring of raters who score the ACCESS for ELLs 2.0 Writing responses are shown in the following table.

Monitoring of raters is done sufficiently so as to ensure consistently reliable scoring.
Monitoring of raters is done with and without notice to the raters.
After monitoring, appropriate feedback is provided to raters to help improve their scoring.
Minimally acceptable rating performance is clearly documented.
Raters who fail to meet minimally acceptable performance criteria are retrained or removed from scoring.
Staff leading the monitoring have appropriate background and experience to evaluate the raters.

Table 9. Claims Made for Monitoring Raters - Writing

The participants evaluated these claims made on the evidence presented by WIDA, DRC, and CAL at the QC. The results of the participants' evaluation of these claims are shown in the following table.

Evaluation	Numeric value	Percentage
Sufficient relevant evidence was presented	4	100
Some relevant evidence was presented	3	0
Little relevant evidence was presented	2	0
No relevant evidence was presented	1	0

Table 10. Participant Evaluation of Claims Made for Monitoring Raters - Writing

The following comment was added by one participant.

> *I'm sure staff leading the monitoring (i.e. team leaders) are appropriate, but we didn't (or I don't remember) discuss the credentialing of the Team Leaders. Is it just "time of service"?*

The quantitative responses and comment indicate that there were no concerns expressed about monitoring of raters who score the ACCESS for ELLs 2.0 Writing responses.

5. Training of Speaking Raters

The claims made for the processes supporting the training of raters who score ACCESS for ELLs 2.0 Speaking responses are shown in the following table.

Raters are provided with training that ensures reliable scoring of ACCESS for ELLs Speaking responses.
The training period provided is of an appropriate duration.
Trainee raters are provided a sufficient quantity of Speaking samples.
Trainee raters are provided sufficient opportunity to practice scoring authentic samples.
Trainee raters are provided feedback when their scores differ from the pre-assigned score.
Certification criteria required to successfully complete training and score operational responses are set appropriately.
Staff conducting the rater training have the appropriate background and experience to lead the training.

Table 11. Claims Made for Training Raters - Speaking

The participants evaluated these claims made on the evidence presented by WIDA, DRC, and CAL at the QC. The results of the participants' evaluation of these claims are shown in the following table.

Evaluation	Numeric value	Percentage
Sufficient relevant evidence was presented	4	96.5
Some relevant evidence was presented	3	3.5
Little relevant evidence was presented	2	0
No relevant evidence was presented	1	0

Table 12. Participant Evaluation of Claims Made for Training Raters - Speaking

No open-ended comments were made by participants in this section of the evaluation.

The quantitative responses and lack of negative comments indicate that there were no concerns expressed about the processes supporting the training of raters who score ACCESS for ELLs 2.0 Speaking responses.

6. Scoring of Speaking Responses

The claims made for the processes supporting the scoring of ACCESS for ELLs 2.0 Speaking responses are shown in the following table.

ScoreBoard is an appropriate tool for scoring ELLs Speaking responses.
Speaking responses are clear enough for raters to hear and score reliably.
ScoreBoard interface will allow Speaking scores to be recorded without error.

Table 13. Claims Made for Scoring Speaking Responses

The participants evaluated these claims made on the evidence presented by WIDA, DRC, and CAL at the QC. The results of the participants' evaluation of these claims are shown in the following table.

Evaluation	Numeric value	Percentage
Sufficient relevant evidence was presented	4	100
Some relevant evidence was presented	3	0
Little relevant evidence was presented	2	0
No relevant evidence was presented	1	0

Table 14. Participant Evaluation of Claims Made for Scoring Speaking Responses

The following comment was added by one participant.

Sufficient because there is an evaluator supervisor that can listen as well.

The quantitative responses and lack of negative comments indicate that there were no concerns expressed about the processes supporting the scoring of ACCESS for ELLs 2.0 Speaking responses.

7. Monitoring of Speaking Raters

The claims made for the processes supporting the monitoring of raters who score the ACCESS for ELLs 2.0 Speaking responses are shown in the following table.

Monitoring of raters is done sufficiently so as to ensure consistently reliable scoring.
Monitoring of raters is done with and without notice to the raters.
After monitoring, appropriate feedback is provided to raters to help improve their scoring.
Minimally acceptable rating performance is clearly documented.
Raters who fail to meet minimally acceptable performance criteria are retrained or removed from scoring.
Staff leading the monitoring have appropriate background and experience to evaluate the raters.

Table 15. Claims Made for Monitoring Raters - Speaking

The participants evaluated these claims made on the evidence presented by WIDA, DRC, and CAL at the QC. The results of the participants' evaluation of these claims are shown in the following table.

Evaluation	Numeric value	Percentage
Sufficient relevant evidence was presented	4	92
Some relevant evidence was presented	3	8
Little relevant evidence was presented	2	0
No relevant evidence was presented	1	0

Table 16. Participant Evaluation of Claims Made for Monitoring Raters - Speaking

The following comment was added by one participant.

Not sure what the minimum cut off is? When do you let someone go?

The quantitative responses and lack of negative comments indicate that there were no concerns expressed about the monitoring of raters who score the ACCESS for ELLs 2.0 Speaking responses.

8. Rater Reliability

The claims made for the reliability of raters who score the ACCESS for ELLs 2.0 Speaking and Writing domain tests are shown in the following table.

Rater reliability data are clearly documented.
Rater reliability data are recorded with appropriate regularity.
Minimally acceptable reliability criteria are established appropriately.
All Speaking and Writing reliability data meet or exceed the minimally acceptable reliability criteria.
The rater reliability data indicate that ACCESS for ELLs Speaking and Writing responses are scored reliably.

Table 17. Claims Made for the Reliability of Raters

The participants evaluated these claims made on the evidence presented by WIDA, DRC, and CAL at the QC. The results of the participants' evaluation of these claims are shown in the following table.

Evaluation	Numeric value	Percentage
Sufficient relevant evidence was presented	4	100
Some relevant evidence was presented	3	0
Little relevant evidence was presented	2	0
No relevant evidence was presented	1	0

Table 18. Participant Evaluation of Claims Made for the Reliability of Raters

No open-ended comments were made by participants in this section of the evaluation.

The quantitative responses and lack of negative comments indicate that there were no concerns expressed about the reliability of raters who score the ACCESS for ELLs 2.0 Speaking and Writing domain tests.

9. Additional Notes & Comments

The following comments was left by the one participant at the end of the evaluation checklist.

> *What I've seen today seems very much in keeping with industry standards and best practices. I really enjoyed being part of this process and appreciate being afforded the opportunity. It will allow me to better communicate with my LEAs about how students are scored.*

Conclusion

The evaluations completed by the SEA participants provide strong support for the claims WIDA, DRC, and CAL make for the robustness of the processes that support the scoring of the ACCESS for ELLs Speaking and Writing domain tests. The quantitative findings and comments provided by the participants indicate that the evidence provided at the QC Meeting did indeed support the claims made by WIDA, DRC, and CAL. This is a reassuring conclusion and shows that the ACCESS for ELLs Speaking and Writing domain tests are scored by well-trained raters who use an appropriate scoring platform to record scores, and are effectively monitored throughout the scoring process. Most importantly, the reliability data demonstrate that scoring is performed consistently and reliably, indicating that stakeholders may have confidence in the meaningfulness of the scores that are reported for the Speaking and Writing domain tests of ACCESS for ELLs.

WIDA, DRC, and CAL would like to thank the SEA participants of this QC meeting. Their participation and questions helped make the meeting a success. WIDA greatly appreciates their commitment to the educators, students, and families who are served by the WIDA Consortium and the ACCESS for ELLs assessment.

Appendix

WIDA ACCESS for ELLs® 2.0 Speaking and Writing Scoring QC Meeting
April 12, 2018
Data Recognition Corporation – Plymouth, MN
AGENDA

Thursday, April 12

9:00-10:15 **Welcome, Introductions, and Overview**

1. Welcome, Introductions, and Agenda
2. Overview of how WIDA, DRC, and CAL collaborate on the design and preparation of scoring materials
3. Overview of how DRC scores Writing and Speaking responses
4. Rater recruitment

10:30-12:30 **Scoring of the Writing and Speaking tests**

5. Overview of Writing and Speaking scoring
6. Processes supporting the training of DRC raters who score the ACCESS for ELLs Writing and Speaking tests
7. The use of ScoreBoard in the scoring of ACCESS for ELLs Writing and Speaking responses
8. Processes supporting the monitoring of raters who score the ACCESS for ELLs 2.0 Writing and Speaking tests

12:30-1:30 **Lunch**

1:30-3:15 Sample training exercise and tour

9. Introduction to Writing rater training activity
10. Practice scoring authentic Writing samples based on Writing Scoring Scale and benchmark responses
11. Tour of scoring facility

3:30-4:00 **Wrap up and Q&A**

12. Review of inter-rater reliability data from the operational assessment; Q&A
13. Complete and return evaluations and wrap up