Item Summary Analysis: 2016-17 ACCESS for ELLs 2.0

WIDA ACCESS for ELLs 2.0 Writing and Speaking Tasks

Overview

During the WIDA ACCESS for ELLs 2.0 testing season, DRC provides WIDA and CAL with a weekly Item Summary Report for the scoring of Speaking and Writing tasks for all grade clusters. The Item Summary Reports contain data on inter-rater reliability (the degree of agreement among raters) and score point distribution for all Speaking and Writing tasks across grades 1-12. These reports provide WIDA and CAL regular and ongoing insight into scoring trends for the Speaking and Writing tasks. In addition, we are able to monitor rater performance on all tasks and flag any tasks where reliability rates and score point distributions may indicate that specified minimum performance rates are not being met.

This year we reviewed the WIDA ACCESS for ELLs 2.0 Item Summary Reports for Speaking and Writing tasks generated by DRC in order to track inter-rater reliability for Speaking and Writing tasks. A high rate of inter-rater reliability indicates that DRC-trained raters score Speaking and Writing responses both accurately and consistently. Tasks were flagged for further review if their inter-rater reliability fell below 70% and/or if their score point distributions deviated from normal trends.

Overall, we observed that rater reliability improved markedly in the 2016-17 academic year (AY). Inter-rater reliability had not been problematic in AY 2015-16, with all tasks meeting or exceeding the minimum of 70%, but the improvements in reliability for both Speaking and Writing were very encouraging. Indeed, inter-rater reliability for Writing tasks was consistently above 90% in AY 2016-17. The high rates of rater reliability attained during AY 2016-17 support the claim that ACCESS for ELLs 2.0 Speaking and Writing domain scores provide valid and reliable information about the relevant language proficiency of students who take the test.

Use of the Data

Upon receiving the weekly Item Summary Report from DRC, WIDA conducted an analysis of the data. This analysis was then shared with members of the WIDA Assessment Team and collaborators at CAL and DRC. WIDA, CAL and DRC met weekly to review the data on the Speaking and Writing tasks and to discuss any questionable scoring trends.

On the next page of this end-of-year analysis, we have included sample Speaking and Writing raw data from the May 12, 2017 Item Summary Report.

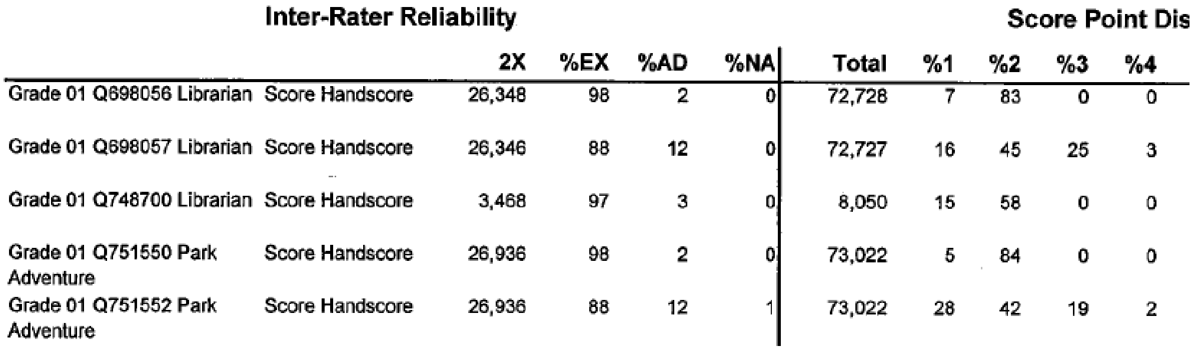
Speaking Task Review

Table 1 contains an explanation of the terms and symbols used in the Speaking data listed in Figure 1. Please note that the first line of the Grade 1 Speaking data only contains responses at score points 1 and 2 because these Tier A tasks only aim to elicit word-level responses. Response data to other tasks (targeting Tiers B and C) show all score points.

Table 1: Explanation of Speaking Sample Data

Data	Explanation
Q698057 Librarian	Internal item number and Folder theme
2X - 26,346	Number of double scored responses
%EX - 88	Inter-rater reliability rate of exact agreement
%AD - 12	Inter-rater reliability rate of agreement of adjacent scores
%NA	Inter-rater reliability rate of agreement of non-adjacent scores
TOTAL - 72,727	Total number of responses scored
%1 – %4	Percentage of responses at each score point (SP1: Attempted, SP2: Adequate, SP3: Strong, and SP4: Exemplary)

Figure 1: Speaking Grade 1



		Inter-Rater Reliability				Score Point Dis				
		2X	%EX	%AD	%NA	Total	%1	%2	%3	%4
Grade 01 Q698056 Librarian	Score Handscore	26,348	98	2	0	72,728	7	83	0	0
Grade 01 Q698057 Librarian	Score Handscore	26,346	88	12	0	72,727	16	45	25	3
Grade 01 Q748700 Librarian	Score Handscore	3,468	97	3	0	8,050	15	58	0	0
Grade 01 Q751550 Park Adventure	Score Handscore	26,936	98	2	0	73,022	5	84	0	0
Grade 01 Q751552 Park Adventure	Score Handscore	26,936	88	12	1	73,022	28	42	19	2

For Speaking tasks, we monitored not only the inter-rater reliability data but also the distribution of score points (SP) which are SP1: Attempted, SP2: Adequate, SP3: Strong and SP4: Exemplary. We flagged score points if SP1 was greater than 25% or if SP4 was greater than 10% for Speaking tasks in order to monitor the concentration of student responses at both ends of the WIDA Speaking Scoring Scale (see appendix). Tasks that were flagged because of their score distributions are not necessarily problematic but would be reviewed by WIDA, CAL, and DRC to confirm that the tasks are eliciting spoken language as intended.

Writing Task Review

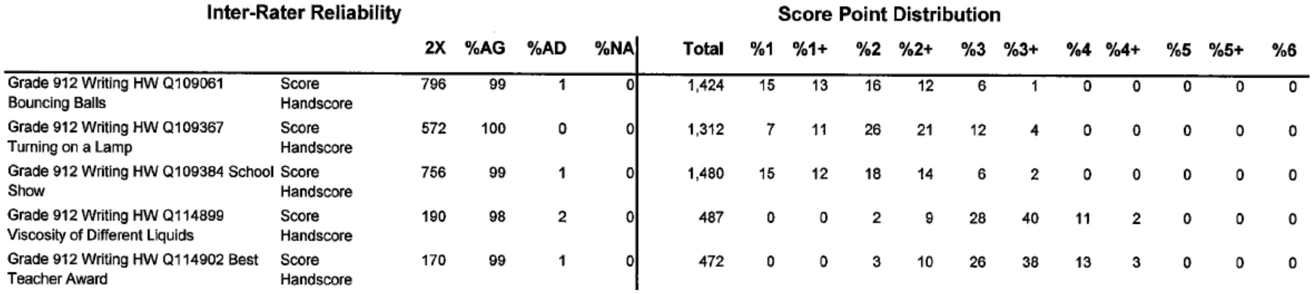
Table 2 contains an explanation of the terms and symbols used in the Writing data listed in Figure 2. Please note that the first line of the Grade 1 Writing data only contains responses at score points 1 and 2 because these Tier A (P1) tasks only aim to elicit word-level responses.

Table 2: Explanation of Writing Sample Data

Data	Explanation
HWQ109061	Internal item number, Folder theme, and response mode (handwritten)
2X - 796	Number of double scored responses
%AG - 99	Inter-rater reliability rate of agreement
%AD - 1	Inter-rater reliability rate of agreement of adjacent scores
%NA - 0	Inter-rater reliability rate of agreement of non-adjacent scores
TOTAL – 1,424	Total number of responses scored
%1 – %6	Percentage of responses at each score point

Please note that when interpreting the score point distribution data in Figure 2 that Tier A and Tier B/C tasks typically demonstrate different distributions. While tasks at both Tiers may be scored up to the maximum raw score of 6, it is rare to observe raw scores above 3+ for responses to Tier A tasks.

Figure 2: Writing Grades 9-12



Inter-Rater Reliability						Score Point Distribution															
		2X	%AG	%AD	%NA	Total	%1	%1+	%2	%2+	%3	%3+	%4	%4+	%5	%5+	%6
Grade 912 Writing HW Q109061 Bouncing Balls	Score Handscore	796	99	1	0	1,424	15	13	16	12	6	1	0	0	0	0	0
Grade 912 Writing HW Q109367 Turning on a Lamp	Score Handscore	572	100	0	0	1,312	7	11	26	21	12	4	0	0	0	0	0
Grade 912 Writing HW Q109384 School Show	Score Handscore	756	99	1	0	1,480	15	12	18	14	6	2	0	0	0	0	0
Grade 912 Writing HW Q114899 Viscosity of Different Liquids	Score Handscore	190	98	2	0	487	0	0	2	9	28	40	11	2	0	0	0
Grade 912 Writing HW Q114902 Best Teacher Award	Score Handscore	170	99	1	0	472	0	0	3	10	26	38	13	3	0	0	0

Comparisons

Writing Tasks: Inter-rater Reliability

For Writing, inter-rater reliability was at or above 93% for all tasks across all grade clusters, significantly exceeding the minimum requirement of 70% inter-rater agreement. These data represent an improvement when compared to last year's end-of-year analysis. For the AY 2015-16 testing season, inter-rater reliability for Writing was at or above 75%. In addition, there were no issues with the Score Point Distribution of any Writing tasks.

Please see the appendix for a copy of the WIDA Writing Scoring Scale. These data illustrate the high rates of rater agreement achieved when scoring the Writing domain of ACCESS for ELLs 2.0 and should provide confidence in the reliability of the Writing domain scores.

Speaking Tasks: Inter-rater Reliability

There were no Speaking tasks with inter-rater reliability below 70% and we categorized only seven of the 75 Speaking tasks as having borderline reliability data (below 75%) across all grade clusters. The rate of agreement for these seven tasks was between 70%-74%. These data represent an improvement when compared to last year's end-of-year analysis. For AY 2015-16, we ended the testing season with three flagged Speaking Tasks (based on score distributions) and eleven borderline tasks (please see Table 3 for a comparison).

It should be noted that the very high rates of inter-rater reliability across all tasks indicate that tasks with lower rates of reliability (around or below 70%) may well be attributed to the task characteristics rather than of the raters or the training materials. Speaking tasks that are flagged for either inter-rater reliability and/or score distributions are investigated by WIDA, CAL, and DRC to confirm whether there are issues with the task that may contribute to the reliability data. Tasks that are reviewed and identified as problematic from both a content and scoring perspective are targeted for refreshment during the following operational testing season.

Table 3: Inter-rater Reliability for Speaking Tasks

Tasks Identified	N	Tier		PL	
Flagged (AY 2016-17)	N = 0	Tier A = 0	Tier B/C = 0	P3 = 0	P5 = 0
Flagged (AY 2015-16)	N = 3	Tier A = 0	Tier B/C = 3	P3 = 1	P5 = 2
Borderline (AY 2016-17)	N = 7	Tier A = 0	Tier B/C = 7	P3 = 3	P5 = 4
Borderline (AY 2015-16)	N = 11	Tier A = 2	Tier B/C = 9	P3 = 7	P5 = 4

The data presented in Table 3 show that scoring reliability of the ACCESS for ELLs 2.0 Speaking domain improved in AY 2016-17, compared to AY 2015-16. All Speaking tasks attained inter-rater reliability at or above 70%.

Speaking Tasks: Score Point Distribution

We also tracked the score point (SP) distribution data for all Speaking tasks. As a part of this weekly review, we searched for concentrations of scores on the WIDA Speaking Scoring Scale that exceeded 25% for SP1 (Attempted) and 10% for SP4 (Exemplary). This year we flagged twelve Speaking tasks that met the established criteria. As we observed with inter-rater reliability, this data on score point distribution is an improvement on the final data for last year's testing season (please see Table 4 below for details).

Table 4: Score Point Distribution for Speaking Tasks

Tasks Identified	N	SP		Tier			PL		
Flagged (AY 2016-17)	N = 12	SP1 = 12	SP4 = 0	Pre-A = 0	A = 9	B/C = 3	P1 = 0	P3 = 9	P5 = 3
Flagged (AY 2015-16)	N = 17	SP1 = 13	SP4 = 4	Pre-A = 0	A = 11	B/C = 6	P1 = 0	P3 = 14	P5 = 3

WIDA and CAL use these data to review Speaking tasks that demonstrated unusual score point distributions. Such tasks may not be ideal for continued operational use and an analysis of these tasks, along with qualitative feedback from DRC raters are used to determine whether any of the Speaking tasks are problematic and therefore prioritized for revision and/or replacement.

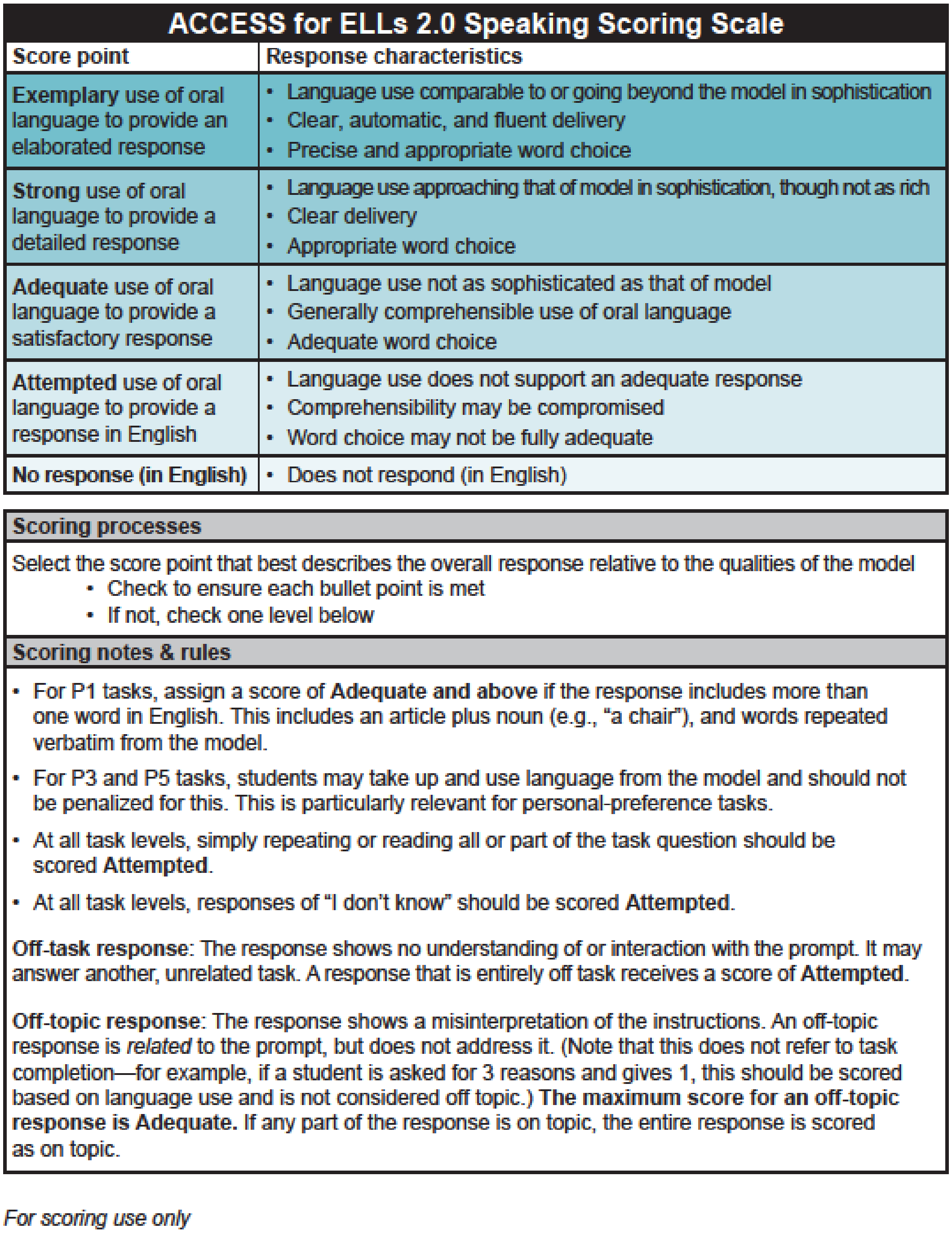
Conclusion

The Item Summary Reports for Speaking and Writing tasks highlighted the scoring trends during the AY 2016-17 testing season for WIDA ACCESS for ELLs 2.0. These reports demonstrate the reliability of the scoring conducted by raters at DRC. The very strong reliability data also support the claim that WIDA ACCESS for ELLs 2.0 is a valid and reliable measure of English language proficiency for emergent bilinguals.

The data indicate the scoring of Speaking and Writing tasks is moving in the right direction and that scoring in 2016-17 improved, demonstrating higher rates of scoring reliability when compared to 2015-16. For inter-rater reliability, we were pleased to observe no flagged Speaking tasks and a very high rate of reliability for all Writing tasks. The data on score point distribution for Speaking and Writing tasks also instill confidence in the scoring procedures. WIDA will continue to work with CAL and DRC to ensure that scoring processes are of the highest quality.

Appendix

WIDA Speaking Scoring 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)

Scoring processes

Select the score point that best describes the overall response relative to the qualities of the model

- Check to ensure each bullet point is met
- If not, check one level below

Scoring notes & rules

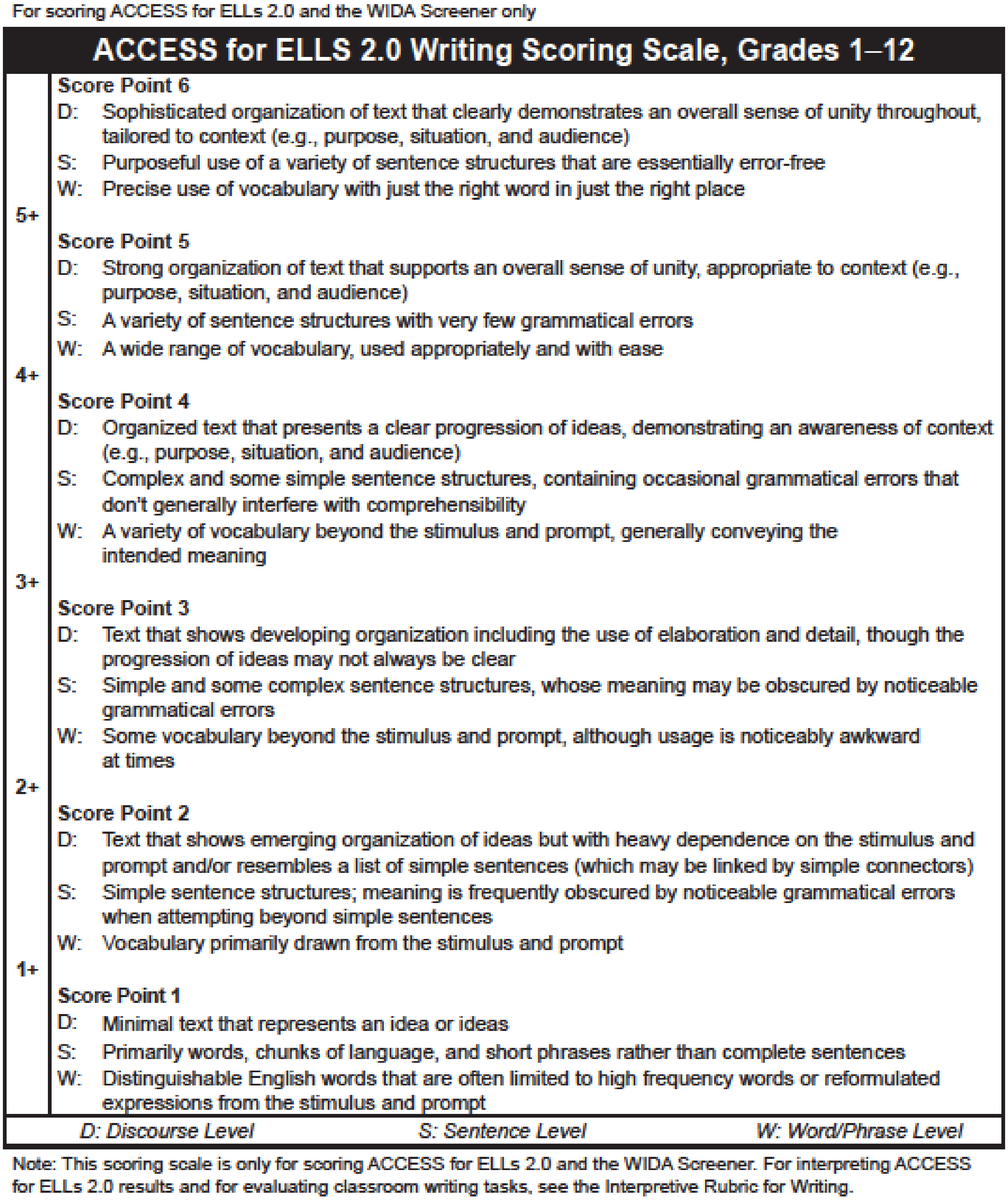
- For P1 tasks, assign a score of **Adequate and above** if the response includes more than one word in English. This includes an article plus noun (e.g., "a chair"), and words repeated verbatim from the model.
- For P3 and P5 tasks, students may take up and use language from the model and should not be penalized for this. This is particularly relevant for personal-preference tasks.
- At all task levels, simply repeating or reading all or part of the task question should be scored **Attempted**.
- At all task levels, responses of "I don't know" should be scored **Attempted**.

Off-task response: The response shows no understanding of or interaction with the prompt. It may answer another, unrelated task. A response that is entirely off task receives a score of **Attempted**.

Off-topic response: The response shows a misinterpretation of the instructions. An off-topic response is *related* to the prompt, but does not address it. (Note that this does not refer to task completion—for example, if a student is asked for 3 reasons and gives 1, this should be scored based on language use and is not considered off topic.) **The maximum score for an off-topic response is Adequate.** If any part of the response is on topic, the entire response is scored as on topic.

For scoring use only

WIDA Writing Scoring Scale



For scoring ACCESS for ELLs 2.0 and the WIDA Screener only

ACCESS for ELLS 2.0 Writing Scoring Scale, Grades 1–12

Score Point 6

D: Sophisticated organization of text that clearly demonstrates an overall sense of unity throughout, tailored to context (e.g., purpose, situation, and audience)

S: Purposeful use of a variety of sentence structures that are essentially error-free

W: Precise use of vocabulary with just the right word in just the right place

5+

Score Point 5

D: Strong organization of text that supports an overall sense of unity, appropriate to context (e.g., purpose, situation, and audience)

S: A variety of sentence structures with very few grammatical errors

W: A wide range of vocabulary, used appropriately and with ease

4+

Score Point 4

D: Organized text that presents a clear progression of ideas, demonstrating an awareness of context (e.g., purpose, situation, and audience)

S: Complex and some simple sentence structures, containing occasional grammatical errors that don't generally interfere with comprehensibility

W: A variety of vocabulary beyond the stimulus and prompt, generally conveying the intended meaning

3+

Score Point 3

D: Text that shows developing organization including the use of elaboration and detail, though the progression of ideas may not always be clear

S: Simple and some complex sentence structures, whose meaning may be obscured by noticeable grammatical errors

W: Some vocabulary beyond the stimulus and prompt, although usage is noticeably awkward at times

2+

Score Point 2

D: Text that shows emerging organization of ideas but with heavy dependence on the stimulus and prompt and/or resembles a list of simple sentences (which may be linked by simple connectors)

S: Simple sentence structures; meaning is frequently obscured by noticeable grammatical errors when attempting beyond simple sentences

W: Vocabulary primarily drawn from the stimulus and prompt

1+

Score Point 1

D: Minimal text that represents an idea or ideas

S: Primarily words, chunks of language, and short phrases rather than complete sentences

W: Distinguishable English words that are often limited to high frequency words or reformulated expressions from the stimulus and prompt

D: Discourse Level *S: Sentence Level* *W: Word/Phrase Level*

Note: This scoring scale is only for scoring ACCESS for ELLs 2.0 and the WIDA Screener. For interpreting ACCESS for ELLs 2.0 results and for evaluating classroom writing tasks, see the Interpretive Rubric for Writing.