



Series 400 ACCESS Paper and Online Comparability Report

Kyongwon Bishop, Ph.D.

April 2017

Series 400 ACCESS Paper and Online Comparability Report

Introduction

In 2015-16, WIDA launched ACCESS 2.0, a multi-stage online test, along with the existing paper-based test, in the 400 series administration. The Listening and Reading domains in the paper-based test were from the refreshed ACCESS Series 302 form, and the newly developed online Listening and Reading tests in the adaptive computerized form. The new Speaking domain test was built sharing the same items between the paper-based and online tests, but varied slightly to accommodate each mode of administration. The paper-based and online versions of the Writing test also shared the same items. The online Writing test was made available for Grades 4-12 along with the paper-based form, and for Grades 1-3, only paper-based Writing forms were available. Note that students taking the online writing test can respond in handwritten format if needed. In grades 4, 5, states must choose whether writing is offered in keyboarded or handwritten response modes. In all other grades, the online writing test defaults to keyboarded responses. In this report, the terms handwritten/keyboarded and paper-based/online are used interchangeably. But it should be understood that analyses reported here refer only to online, keyboarded and paper-based, handwritten response comparisons. The measurement of the online test was based on the paper scale, so the scores between the paper-based and online tests are intended to be interpreted in a similar manner.

It is worth noting that despite the psychometric linking procedure between the paper-based and online tests there were differences between the paper-based and online tests because of the test designs. The online Listening and Reading domain tests adapt to students' abilities and allow them to test in tier-based folders unique to their proficiency level. This was done to provide for more reliable and accurate measurement, which is a goal for adaptive assessments. Tier placement in the online Speaking and Writing tests were determined by a logistic regression probability formula. Upon completion of the Listening and Reading tests, students were directed to specific tiers based on their likelihood of scoring at that tier. The adaptive online navigation (Listening and Reading) and placement rules (Speaking and Writing) enabled students to demonstrate their ability beyond the pre-assigned tier placement and capping structure of the paper-based form. The online Speaking and Writing tests were centrally scored by trained raters from the Data Recognition Corporation (DRC), consistent with industry standard centralized scoring inter-rater reliability checks and procedures. The paper-based Speaking test was administered and scored by local educators. No information is available to ascertain the inter-rater reliability of locally scored Speaking tests.

Based on the differences and similarities between the paper-based and online tests mentioned above, this report presents a comparability analysis between the paper-based and online tests for the ACCESS Series 400 administration.

Analyses

The comparability analysis consists of two parts. The first part compared the overall performance and distribution of the paper-based and online tests in the tested population. The second part of the analyses examined the similarities and differences between paper-based and online assessment performances when each domain score was matched to another domain score that closely correlated. The matching analysis was done to minimize possible bias that might confound analysis results. The matched domain scores were expected to provide similar patterns and yield a more homogeneous error variance. In both analyses, the online data contained only uninterrupted scores¹.

Table1

Correlations Between ACCESS Test Models (both online and paper)

Domains	Listening Scale score	Reading Scale score	Speaking Scale score	Writing Scale score
Listening Scale score	1	0.77	0.64	0.74
Reading Scale score	0.77	1	0.54	0.87
Speaking Scale score	0.64	0.54	1	0.55
Writing Scale score	0.74	0.87	0.55	1

Note. Correlations in separate mode found similar results.

Based on these correlations, the following matching routines are applied:

- The Listening domain scores are matched on Reading domain scores;
- The Reading domain scores are matched on Writing domain scores;
- The Speaking domain scores are matched on Listening domain scores, and
- The Writing domain scores are matched on Reading domain scores.

Part 1: Population performance analysis procedure:

- Step 1.** Identify the population counts by mode of administration (paper-based and online) per grade.
- Step 2.** Identify the number of uninterrupted test sessions in the online mode of administration.
- Step 3.** Calculate population distribution by mode of administration in each grade.

¹ During the Series 400 test administration, a significant proportion of students taking the online assessment had their test sessions interrupted. This analysis removed student records that were identified as *interrupted*.

Part 2: Matched sample performance analysis procedure:

- Step 1.** Identify cases in each domain that were to be matched based on the correlated domain in the same mode of administration (e.g., Speaking is matched with Listening scores where both scores were from the paper-based mode or both from the online mode).
- Step 2.** Identify scores from both the paper-based and online modes that matched (e.g., select cases where the Listening scores in the paper-based mode matched the Listening scores in the online mode).
- Step 3.** Randomly select matched scores from both modes.
- Step 4.** Repeated Step 3 20 times to collect different random samples. Calculated the mean scores from Step 4 of the domain of interest for each mode per grade.

Data and Results

The series 400 ACCESS administration consists of both paper-based and online test takers. Table 2 presents the counts used in analyses (less interruption) by mode.

Table 2

Count of Series 400 ACCESS Test Administration by Mode

N count	Listening		Reading		Speaking		Writing	
	Paper	Online*	Paper	Online*	Paper	Online*	KB*	HW
1	87092	128741	87092	132065	87093	141063	NA	257275
2	84179	123241	84179	118430	84179	154053	NA	258459
3	60949	120282	60949	113110	60949	152069	NA	232525
4	45277	77867	45277	65343	45277	92193	16184	78351
5	34526	59734	34526	49998	34526	70444	12298	60129
6	26067	46599	26067	44076	26067	66468	16679	38415
7	23993	46975	23993	43123	23993	66194	17427	35342
8	23687	48542	23687	42277	23687	66894	18419	36616
9	26982	61097	26982	50139	26982	71861	29317	40141
10	20631	38901	20631	30644	20631	45847	19431	28110
11	16130	27946	16130	20198	16130	31912	13459	20986
12	11564	20684	11564	16088	11564	23495	10849	15813
Total	461,077	800,609	461,077	725,491	461,078	982,493	154,063	1,102,162

Note. *Indicates online population without interrupted cases.

The ratio of paper-based to online test takers per domain is shown in Table 3. After the interrupted cases were removed from the online count, the ratio between the paper-based and online tests were about 60/40 in the Listening and Reading domains. The Speaking domain

online takers' ratio approached 70%. The Writing KB responses increased as grades went up from 17% to 41%. Still handwritten (HW) responses were higher than KB responses in all grades.

Table 3

Ratio Of Online and Paper Test Takers (Uninterrupted Online Testings Only)

Grade	Percent Listening		Percent Reading		Percent Speaking		Percent Writing	
	Paper	Online	Paper	Online	Paper	Online	KB	HW
1	40	60	40	60	38	62	NA	100
2	41	59	42	58	35	65	NA	100
3	34	66	35	65	29	71	NA	100
4	37	63	41	59	33	67	17	83
5	37	63	41	59	33	67	17	83
6	36	64	37	63	28	72	30	70
7	34	66	36	64	27	73	33	67
8	33	67	36	64	26	74	33	67
9	31	69	35	65	27	73	42	58
10	35	65	40	60	31	69	41	59
11	37	63	44	56	34	66	39	61
12	36	64	42	58	33	67	41	59

Part 1: Population performance and distribution

After removing interrupted cases, each domain's proficiency level distribution was compared between the two modes.

Table 4 displays Listening domain test comparisons. Online test takers tend to outperform paper-based takers in Grades 1-3, especially having higher percentages of students at proficiency level 6. Figure 1 graphically shows the average scale score performance per grade in Listening. At all grades, online test takers score higher than their paper-based counterparts. In some cases, the difference is very small (e.g., Grades 5-12); in others, it is quite noticeable (Grades 1-3).

Table 4

Population Distribution by Proficiency Level and Test Mode: Listening

Proficiency Level	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online
	Percent in G1		Percent in G2		Percent in G3		Percent in G4		Percent in G5		Percent in G6	
1	5	5	4	2	3	2	1	2	2	3	3	2
2	5	8	4	2	5	3	5	3	6	4	12	8
3	15	19	9	10	12	9	10	8	11	10	17	24
4	24	9	11	7	15	11	18	15	19	18	19	23
5	41	22	48	29	38	29	36	38	29	36	28	23
6	10	38	24	49	28	47	30	33	33	30	21	20
Proficiency Level	Percent in G7		Percent in G8		Percent in G9		Percent in G10		Percent in G11		Percent in G12	
1	5	3	7	4	13	6	13	7	15	9	15	12
2	14	10	16	11	18	19	16	22	13	19	12	17
3	17	23	11	23	14	25	18	31	17	29	22	26
4	18	21	19	20	20	23	22	19	26	23	27	25
5	26	20	22	17	25	13	20	11	18	10	13	9
6	19	23	25	24	10	14	11	9	12	10	12	12

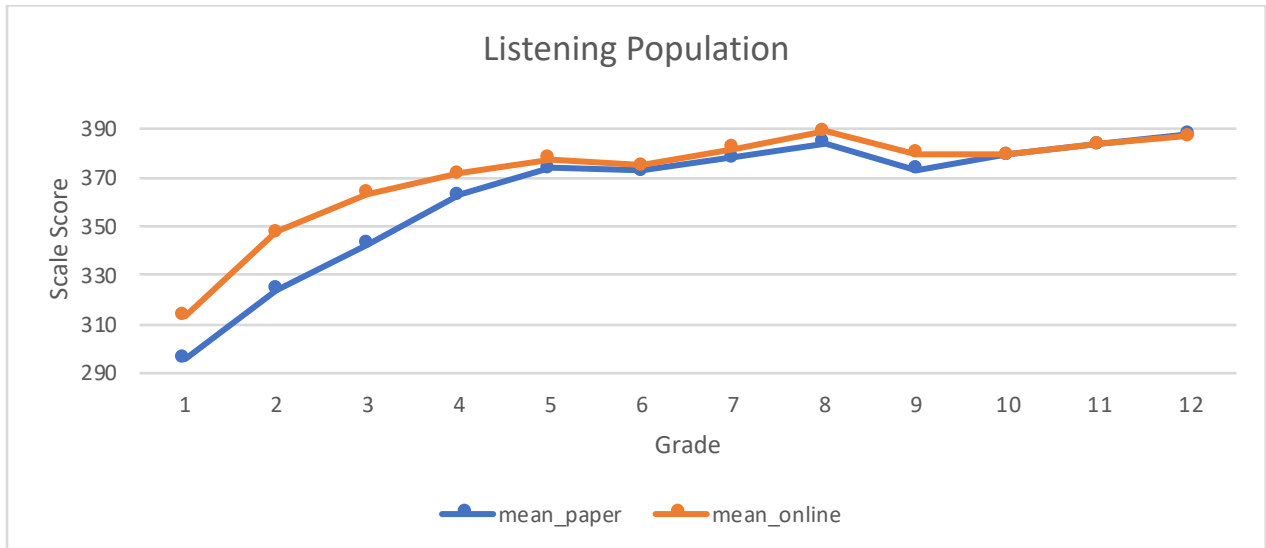


Figure 1. Population performance by grade and test mode: Listening.

In Table 5, the reading tests population proficiency level distribution is provided. It shows that there is a slightly higher percentage of students at proficiency level 6 in the online test except

for Grades 9-12. Average scores by grade are shown in Figure 2. There is very little difference between the average scale score for the two modes. Grade 1-3 online test takers were slightly higher in their average reading scale scores. Conversely, Grade 9-12 students have somewhat higher average reading scores.

Table 5

Population Distribution by Proficiency Level and Test Mode: Reading

Proficiency Level	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online
	Percent in G1		Percent in G2		Percent in G3		Percent in G4		Percent in G5		Percent in G6	
1	11	8	7	2	5	4	5	3	8	8	7	11
2	10	9	8	11	10	13	9	9	12	12	24	24
3	15	23	16	26	13	19	15	24	20	30	31	32
4	24	26	13	11	14	9	15	11	10	7	13	9
5	30	19	34	26	38	21	29	24	26	19	18	16
6	9	14	22	23	21	34	26	28	24	24	7	8
Proficiency Level	Percent in G7		Percent in G8		Percent in G9		Percent in G10		Percent in G11		Percent in G12	
1	11	16	14	18	14	24	14	24	13	22	13	23
2	27	24	28	30	24	25	27	32	26	31	24	30
3	28	30	24	23	17	18	13	14	11	11	11	10
4	13	9	9	6	11	5	14	6	13	8	11	8
5	16	13	18	13	16	13	15	12	16	12	22	11
6	6	9	7	11	17	15	17	12	21	16	19	17

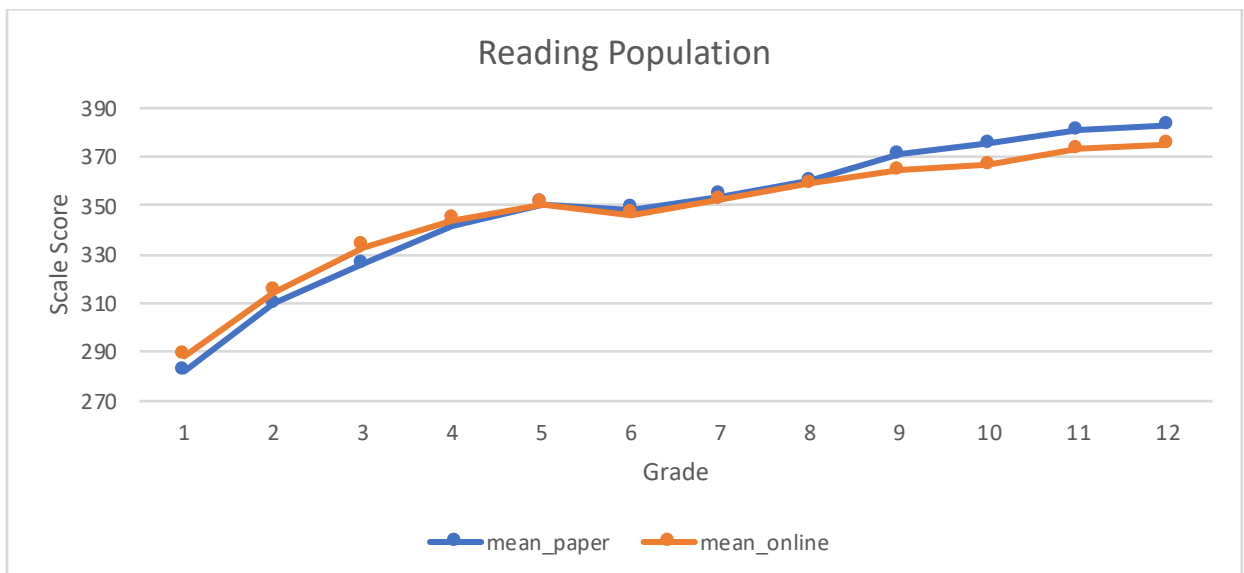


Figure 2. Population performance by grade and test mode: Reading.

Table 6 shows the Speaking test distribution. Here we see much larger differences between online and paper forms of the test than in any other domains. Of note are the large numbers of paper test takers that are classified at proficiency level 6. Figure 3 displays Speaking average scale scores by grade. With the exception of Grades 4 and 5, large differences between modes are clearly observed. Operationally, the distinction between the paper and online forms is that the paper test is administered and scored locally, using training and rating materials provided by WIDA and following training requirements directed by the State Educational Agency. The online Speaking test is administered online and centrally scored by DRC raters using scorer training and rating materials provided by WIDA. All other administrative and content characteristics of the Speaking test are designed to be similar between modes of assessment. WIDA also frequently monitors the Speaking test scoring at DRC. Thus, the observed differences on this year's assessment may be due in part to unique characteristics between locally and centrally scored Speaking tests.

Table 6

Population Distribution by Proficiency Level and Test Mode: Speaking

Proficiency Level	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online
	Percent in G1		Percent in G2		Percent in G3		Percent in G4		Percent in G5		Percent in G6	
1	14	10	8	7	10	8	13	10	13	10	12	12
2	19	25	8	15	10	21	13	12	11	11	4	8
3	10	16	5	11	8	13	6	9	6	8	8	16
4	6	13	4	13	5	13	6	12	5	10	10	26
5	5	12	7	15	8	13	8	16	7	15	5	12
6	47	25	67	40	59	31	53	41	57	47	60	25
Proficiency Level	Percent in G7		Percent in G8		Percent in G9		Percent in G10		Percent in G11		Percent in G12	
1	14	14	13	14	25	29	16	18	11	11	5	10
2	6	11	7	14	4	5	11	15	11	12	11	12
3	8	19	6	12	2	6	6	15	8	17	7	15
4	6	14	7	13	7	21	4	13	4	15	5	18
5	4	12	4	11	2	6	3	6	2	5	2	1
6	61	30	63	37	59	33	60	34	64	40	70	43

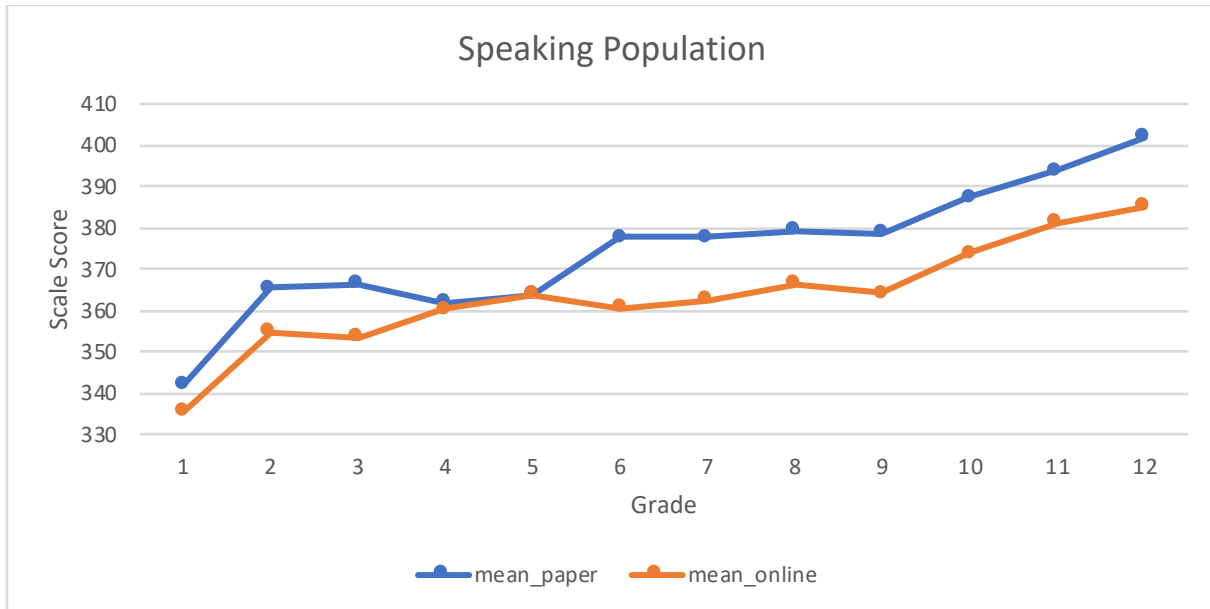


Figure 3. Population performance by grade and test mode: Speaking.

The next table shows Writing proficiency level distributions for both modes. Note that Grades 1-3 are not shown since only the paper form of the Writing test is provided in those grades. Across all grades, the paper mode has higher percentages of students classified at proficiency levels 4 and 5. Figure 4 displays the average Writing scale score by grade. The distinction between online and paper test takers' average writing scores is clear. In aggregate, students taking the paper version of the Writing test have higher scores. Unlike the Speaking test, all Writing tests (paper or online) are centrally scored. Thus, observed differences between paper and online scores are likely due to unique test mode characteristics.

Table 7

Population Distribution by Proficiency Level and Test Mode: Writing

Proficiency Level	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online
	Percent in G4		Percent in G5		Percent in G6		Percent in G7		Percent in G8		Percent in G9	
1	2	5	4	7	5	7	7	10	10	14	7	8
2	4	9	4	9	11	20	12	20	16	26	14	19
3	12	25	22	36	37	48	57	56	67	55	22	28
4	69	56	65	46	47	25	23	13	7	4	33	31
5	12	5	5	2	0	0	0	0	0	0	23	13
6	0	0	0	0	0	0	0	0	0	0	1	1
Proficiency Level	Percent in G10		Percent in G11		Percent in G12							
1	9	10	9	10	10	12						
2	7	12	9	12	7	12						
3	28	38	32	41	38	45						
4	44	33	44	33	43	30						
5	11	6	6	4	2	2						
6	0	0	0	0	0	0						

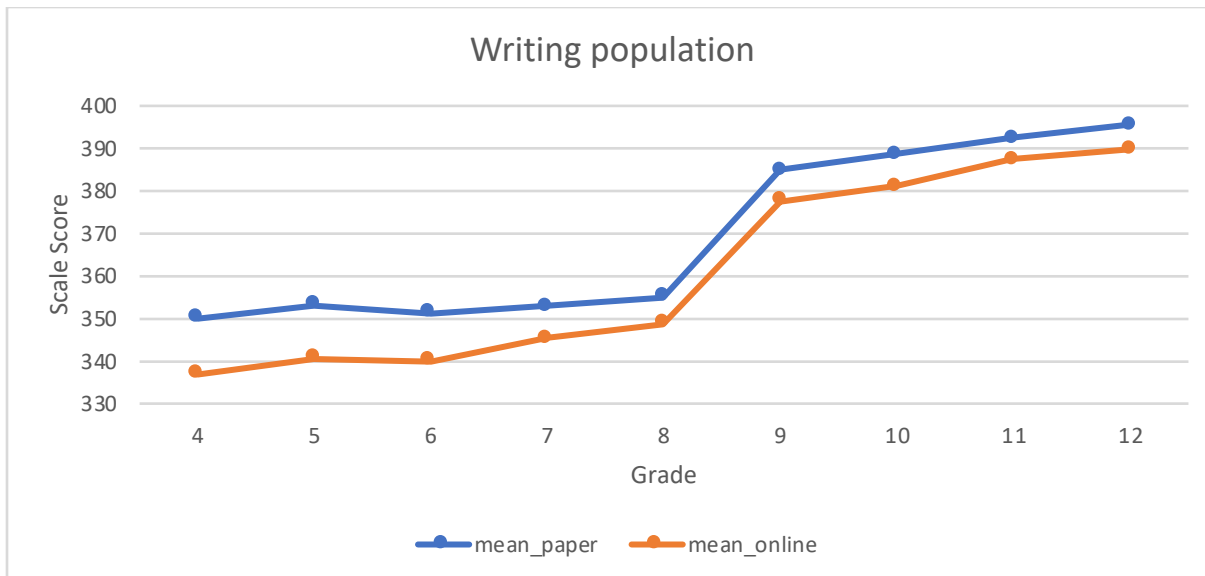


Figure 4. Population performance by grade and test mode: Writing.

What might be contributing to observed differences between domain tests? It could be that tier placement protocols, local vs. centralized scoring procedures, or HW vs. KB responses account for what is observed. It could be that, in general, students taking the paper version of ACCESS are at higher proficiency levels in Speaking and Writing. The concern with using population data is the difficulty in determining whether observed differences are *true* or *artifacts* of the test mode. One way to ameliorate this is to match students using a common metric. Given that ACCESS online and paper forms have been psychometrically connected, matching students who have similar domain tests and then looking at other domain tests of interest should provide *truer* differences between modes. For example, if students with the same reading score on the online and paper tests were matched and differences between the writing test modes were still observed, there would be a stronger argument for a mode effect. The correlations presented earlier provide support for using domain tests as matching criteria. The next section of this report presents results from an analysis comparing modes with students matched by domain scores.

Part 2: Matched sample performance and distribution

Table 8 presents differences, in percentages, between paper and online test takers' Listening proficiency level after matching on reading scores. After matching, students in Grades 1-2 still had a higher concentration of students reaching proficiency levels 5 and 6 for the online test. Between Grades 3 and 7, paper-based and online test takers showed generally similar proportions of students reaching levels 5 and 6. Beyond Grade 7, paper test takers had greater proportions of students at levels 5 and 6. Figure 5 displays listening average scale scores, and, in general, online grade scores are higher in Grades 1-3 and similar in the later grades.

Table 8

Matched Sample Distribution by Proficiency Level and Test Mode: Listening

Proficiency Level	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online
	Percent in G1		Percent in G2		Percent in G3		Percent in G4		Percent in G5		Percent in G6	
1	2	6	2	3	1	3	1	2	2	4	3	1
2	5	9	6	3	9	4	4	3	7	4	12	7
3	19	22	14	13	18	13	9	8	13	10	19	27
4	31	10	15	8	17	15	16	15	21	20	21	25
5	34	22	33	32	32	33	32	40	31	39	26	24
6	10	31	30	42	22	32	38	31	27	23	19	15
Proficiency Level	Percent in G7		Percent in G8		Percent in G9		Percent in G10		Percent in G11		Percent in G12	
	1	5	3	8	4	7	3	10	4	18	7	20
2	16	11	18	13	20	17	19	22	17	21	18	20
3	19	27	11	27	24	36	26	39	22	36	28	34
4	19	23	20	22	26	27	25	22	28	25	23	26
5	23	20	21	17	22	11	17	9	12	7	7	6
6	18	17	22	17	3	5	4	3	4	4	3	4

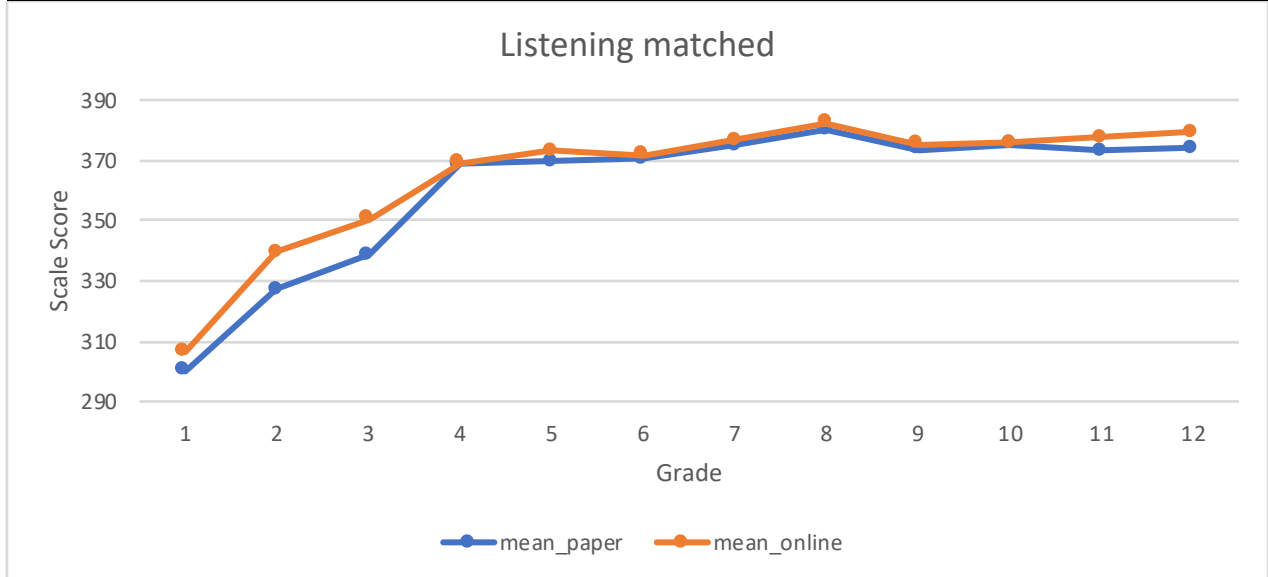


Figure 5. Matched sample performance by grade and test mode: Listening.

The Reading proficiency level percentages is shown Table 9. In general, students taking the online tests have noticeably higher percentages of students at proficiency levels 5 and 6 in Grades 2, 3, and 4 while paper test takers have noticeably higher percentages of students at

levels 5 and 6 in Grades 11 and 12. Figure 6 shows average scale scores by grade. There is fluctuation in scale score averages between mode of administration in reading in Grades 1, 2, 4, and 5, but the observed difference does not seem to be systematic. That is, there seems to be no obvious reason why Grade 1 and 2 online test takers would outperform their paper test taker peers, while the opposite is true for Grades 4 and 5. The observed differences may very well be attributed to sample fluctuation rather than systematic differences between the two modes of administration in these grades.

Table 9

Matched Sample by Proficiency Level and Test Mode: Reading

Proficiency Level	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online
	Percent in G1		Percent in G2		Percent in G3		Percent in G4		Percent in G5		Percent in G6	
1	8	3	6	1	6	1	9	4	17	10	12	14
2	9	4	7	5	7	5	14	10	18	15	31	27
3	13	14	14	12	14	12	20	26	22	31	31	31
4	20	28	13	11	13	11	15	12	9	7	10	8
5	38	30	35	38	35	38	26	23	19	17	13	14
6	12	21	24	33	24	33	17	25	15	20	4	7
Proficiency Level	Percent in G7		Percent in G8		Percent in G9		Percent in G10		Percent in G11		Percent in G12	
1	17	19	22	21	20	26	20	27	17	25	18	26
2	32	27	32	32	29	28	33	32	32	32	29	30
3	26	29	21	22	17	17	12	13	11	11	11	10
4	10	8	7	6	10	5	11	6	12	7	11	8
5	12	11	13	11	12	12	12	12	14	12	18	11
6	4	7	5	8	12	12	12	10	15	13	14	15

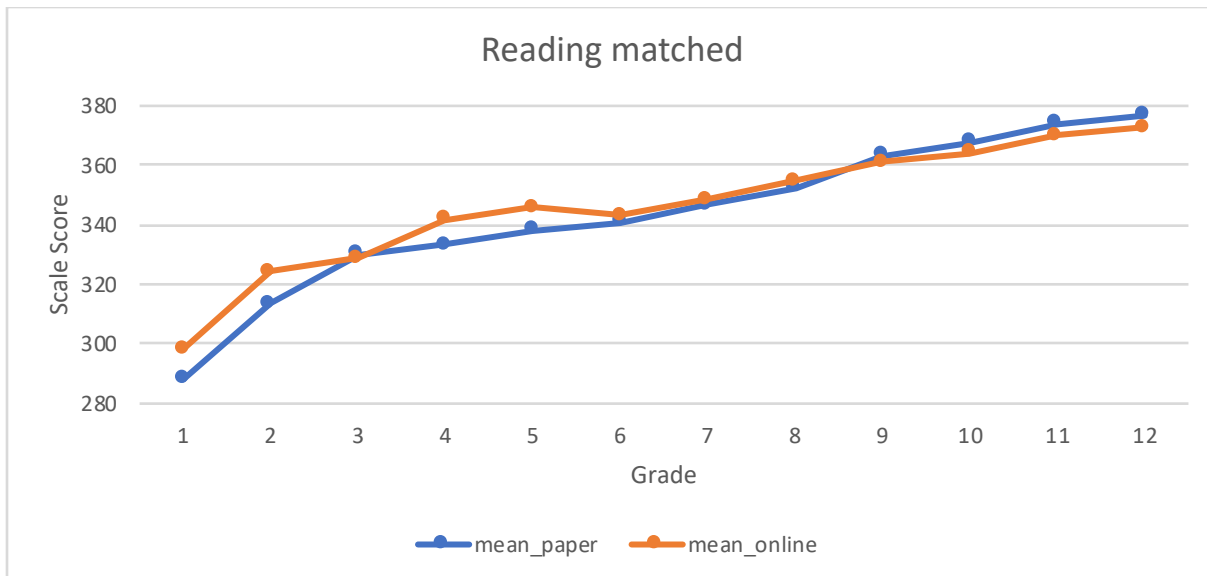


Figure 6. Matched sample performance by grade and test mode: Reading.

Table 10 presents the matched sample differences between the paper and online Speaking tests. Figure 7 displays the average Speaking scale score difference between modes of assessments. At every grade, the percent of students attaining at or above a level 5 on the paper assessment is greater (in some cases substantially) than that of the online mode. Similarly, at every grade, save Grade 5, the average Speaking scale score is greater on paper than it is online. Across grades, the paper-based mode averages 20 score points higher than online mode. The main distinction between the paper and online modes of the Speaking is paper mode tests are scored locally and online tests are scored centrally by DRC. It is likely that the observed differences are due in part to how and where Speaking tests are scored. There are substantial training materials and suggested requirements made available for states and local scorers by WIDA; however, the application of WIDA recommendations varies greatly by state. After matching students, there are differences between centrally and locally scored Speaking tests. Observed differences may, in part, be attributed to unique scorer characteristics between modes.

Table 10

Matched Sample Distribution by Proficiency Level and Test Mode: Speaking

Proficiency Level	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online	
	Percent in G1		Percent in G2		Percent in G3		Percent in G4		Percent in G5		Percent in G6		
1	16	12	11	11	16	16	14	16	15	11	25	28	
2	26	32	11	24	15	32	17	21	18	17	7	11	
3	10	19	8	14	10	16	9	13	8	12	11	17	
4	6	10	4	14	4	10	8	14	5	11	12	23	
5	5	8	7	14	10	10	9	15	10	16	5	8	
6	37	19	58	23	44	15	44	21	44	33	41	13	
Proficiency Level	Percent in G7		Percent in G8		Percent in G9		Percent in G10		Percent in G11		Percent in G12		
	1	24	28	27	30	30	44	11	22	10	15	5	14
2	7	12	9	14	7	3	19	15	15	15	15	15	12
3	11	19	10	13	3	4	8	18	9	18	10	16	
4	7	14	7	12	8	19	6	9	5	13	5	19	
5	4	8	3	7	2	8	2	10	3	8	4	2	
6	47	18	43	23	49	22	54	26	58	32	61	37	

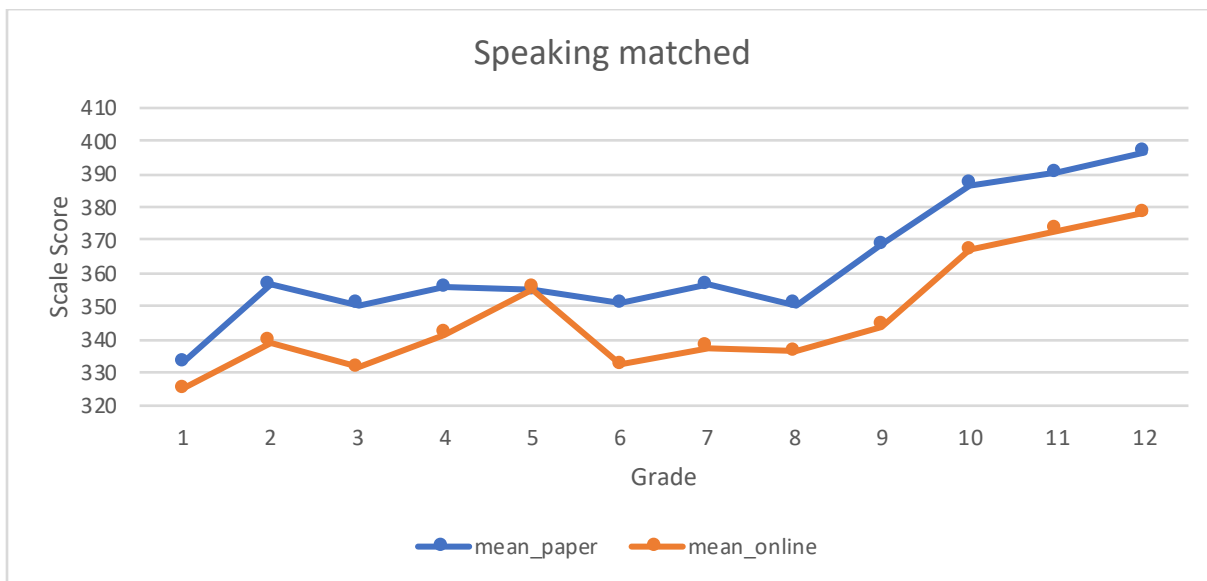


Figure 7. Matched sample performance by grade and test mode: Speaking.

Table 11 and Figure 8 show differences in Writing test performances for matched students. Only Grades 4 through 12 are displayed. All Writing tests are centrally scored. Thus, the chief difference between Writing modes of assessment are paper-based writing is handwritten and online writing is keyboarded. Across all grades, the percent of students receiving a 4 or higher is greater for the paper-based test. Similarly, average scale scores by grade are higher for the paper-based test than for the online test.

Table 11

Matched Sample Distribution by Proficiency Level and Test Mode: Writing

Proficiency Level	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Paper	Online
	Percent in G4		Percent in G5		Percent in G6		Percent in G7		Percent in G8		Percent in G9	
1	2	5	5	9	6	6	10	11	14	16	5	6
2	5	9	7	11	15	22	16	24	23	30	11	15
3	13	25	30	37	41	50	57	55	59	52	30	37
4	70	57	56	41	38	21	17	9	4	2	40	34
5	11	4	3	2	0	0	0	0	0	0	14	8
6	0	0	0	0	0	0	0	0	0	0	0	0
Proficiency Level	Percent in G10		Percent in G11		Percent in G12							
1	7	7	8	7	11	10						
2	8	11	13	11	14	13						
3	35	45	41	53	49	55						
4	45	34	36	27	26	21						
5	5	3	2	1	0	1						
6	0	0	0	0	0	0						

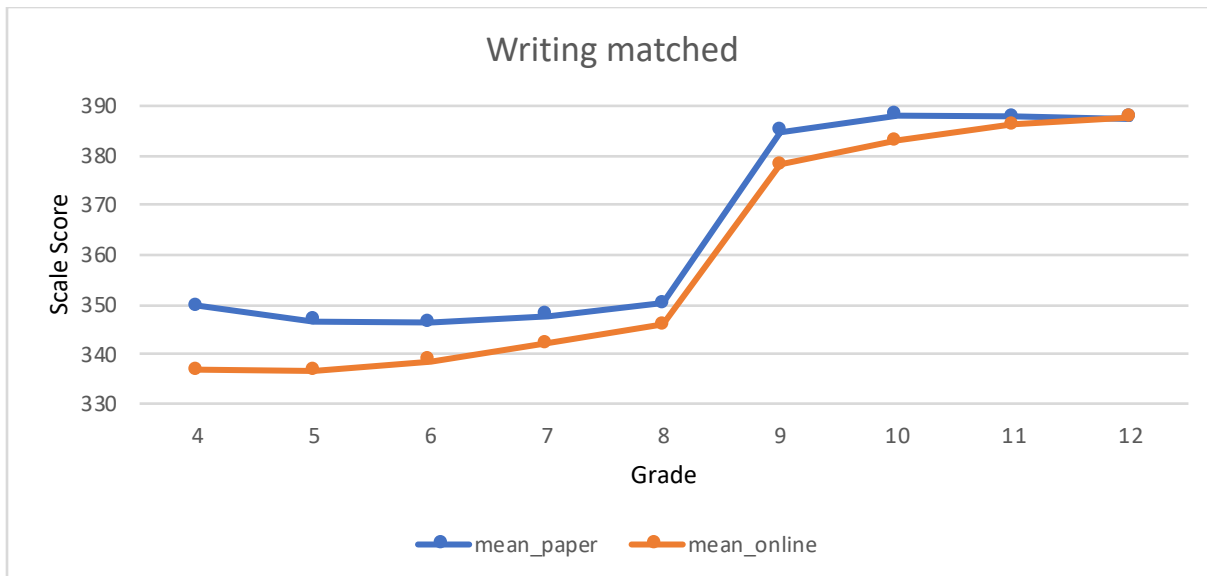


Figure 8. Matched sample performance by grade and test mode: Writing.

Table 12 presents the analysis results of Cohen’s d (1988)² effect sizes between paper-based test and online test. Score difference shown in the earlier tables in each domain by mode were unstandardized measures of mean effects. Using Cohen’s d , mean differences are standardized, so the score differences of each mode are comparable, on the same scale and interpretable by size.

As rule of thumb, small differences are $d \geq |0.2|$; moderate differences are $d \geq |0.5|$, and large differences are $d \geq |0.8|$. In Listening and Reading domains across most grades, small to moderate negative effect sizes observed, meaning online takers performed better than the paper-based takers. On the other hand, there are moderate positive effect sizes in the Speaking and Writing across most grades.

² Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Table 12

*Cohend's d Effect Sizes Between Paper-based and Online Tests**

Grade	Listening	Reading	Speaking	Writing
1	-0.19	-0.39	0.15	
2	-0.33	-0.38	0.34	
3	-0.31	0.05	0.35	
4	-0.02	-0.28	0.27	0.48
5	-0.09	-0.22	-0.01	0.33
6	-0.04	-0.08	0.26	0.28
7	-0.04	-0.06	0.27	0.19
8	-0.04	-0.07	0.18	0.16
9	-0.06	0.06	0.35	0.22
10	-0.02	0.11	0.37	0.17
11	-0.12	0.13	0.32	0.05
12	-0.15	0.12	0.35	0.01

Note. *Positive effect size indicates paper-based test scores are higher than online. Negative effect means online test scores are higher than paper-based test scores.

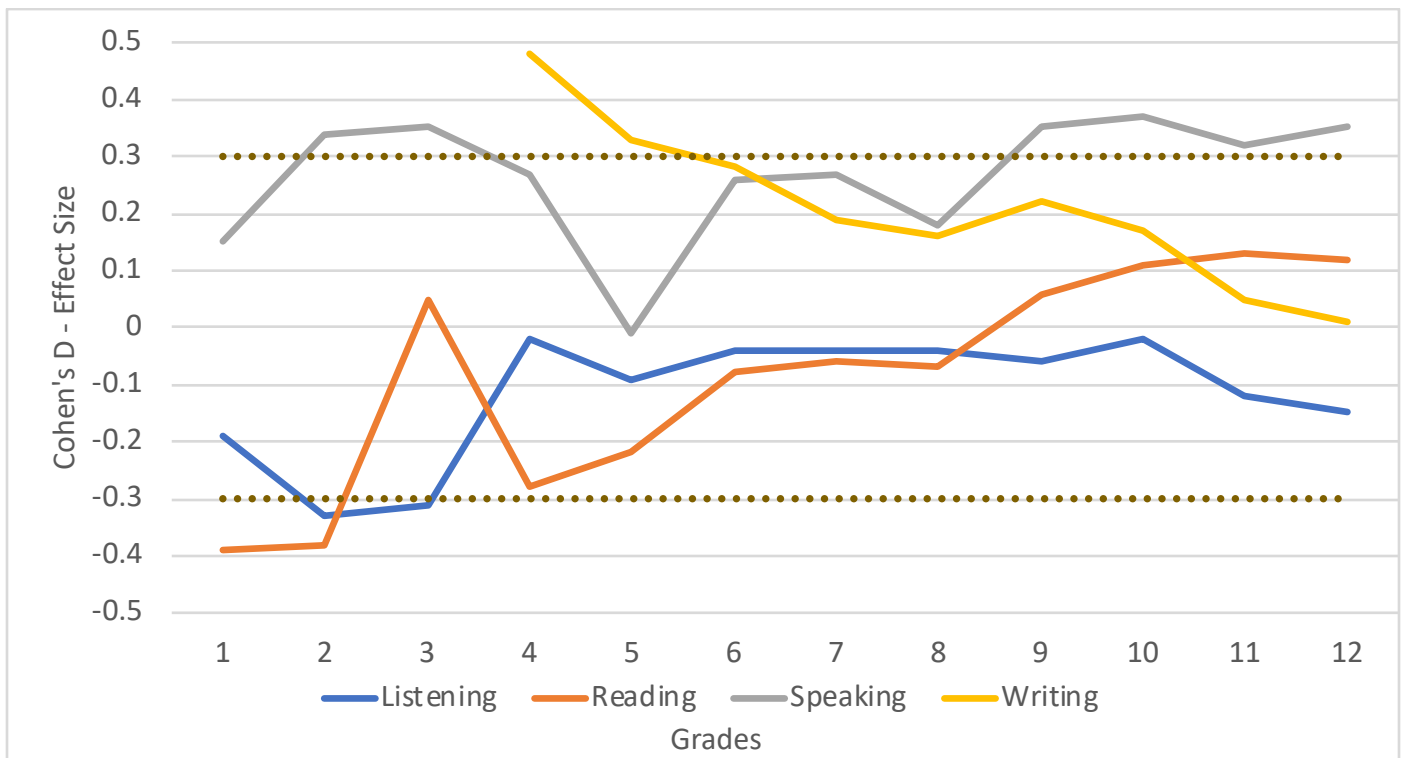


Figure 9. Effect size differences between paper and online ACCESS 2.0 for matched students' scores.

Conclusion

In the matched sample of students, the 400 ACCESS, paper-based Speaking and Writing tests had higher average performance. This result was seen across most grades. The Listening and Reading tests in earlier grades (i.e. Grades 1-5) tended to have higher performance in the online mode. In later grades, the average scores of the paper-based tests did not differ much. There are a variety of potential explanations for observed differences. For example, higher observed online Listening and Reading test performances may be due to the concentration of students at proficiency levels 5 and 6 or the online versions of ACCESS not having capping rules on scale scores. The online takers performances might be closer to students' true underlying latent ability. As stated earlier, differences in Speaking performances across modes may be due, in part, to unique difference between raters. The paper-based Writing score performances were higher across all grades, which might be influenced by age and keyboarding capacity of students. Ultimately, the observed differences between modes of assessments on ACCESS should be carefully monitored in subsequent test administrations. If this trend of differences continues, WIDA will need to re-examine the scaling, scoring and administration methodology for the paper and online ACCESS assessments.