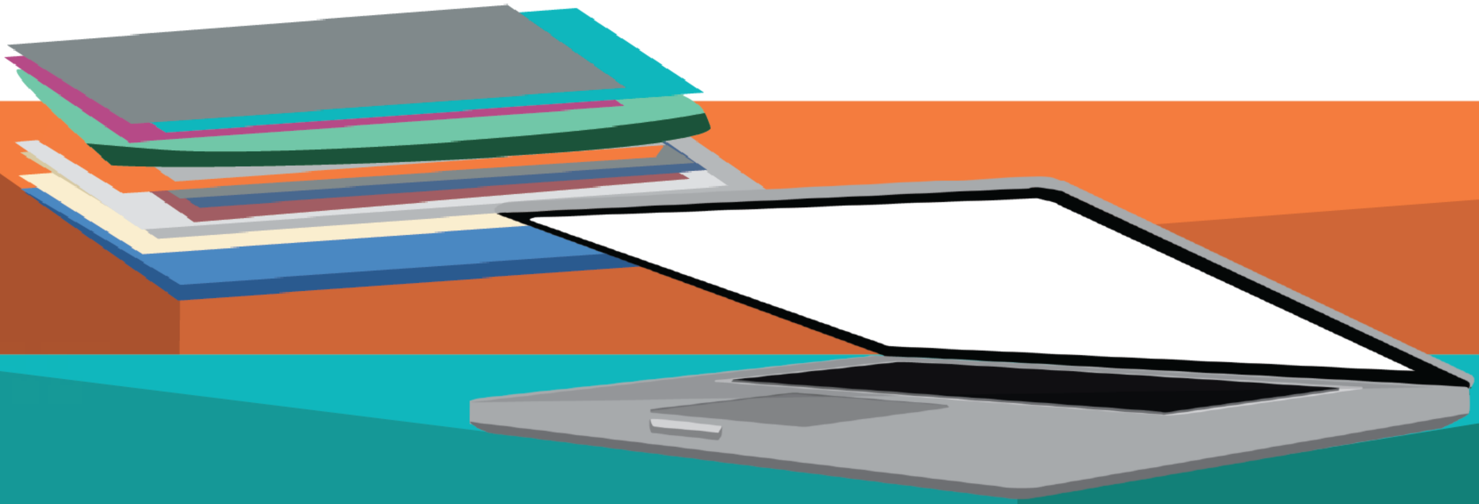
Three Studies of Comparability Between Paper-Based and Computer-Based Testing for the ACT

Jeffrey Steedle, PhD, Peter Pashley, PhD, and YoungWoo Cho, PhD



Authors

Jeffrey Steedle, PhD

Jeffrey Steedle is a lead psychometrician in Assessment Transformation directing the team responsible for statistical analyses for the ACT test and guiding research studies related to maintaining measurement quality while making changes to the assessment program. Jeff holds advanced degrees in education, statistics, and educational psychology, and his research interests include assessment validation and motivation on achievement tests.

Peter Pashley, PhD

Peter Pashley is a principal research scientist in Assessment Transformation at ACT. In that role, he leads research to investigate the psychometric properties of the ACT test administered in different conditions.

YoungWoo Cho, PhD

YoungWoo Cho is a lead psychometrician in Assessment Transformation at ACT specializing in scaling and equating, IRT, and classical test theory.

Conclusions

As in prior ACT mode comparability research, three recent studies demonstrated that paper and online tests measure the same knowledge and skills, but students who test online tend to perform slightly better than students who test on paper, especially on the English, reading, and writing tests.

So What?

ACT will equate scores across modes as needed to ensure that ACT scores can be treated as interchangeable regardless of testing mode.

Now What?

ACT will continue to monitor comparability between paper and online testing and investigate reasons for observed mode effects on the ACT test.

Acknowledgments

The authors gratefully acknowledge the following ACT Assessment Transformation team members for their contributions to the planning and implementation of the studies reported here: Ann Arthur, Benjamin Deonovic, Dongmei Li, Xin Li, Yang Lu, Shalini Kapoor, Lu Wang, and Shichao Wang.

R1847

Table of Contents

Executive Summary

Over the past year, ACT has been investigating how best to offer online testing during a Saturday national testing event. Prior research on the comparability between paper and online testing for the ACT® test indicated that students testing on computers tended to perform slightly better than students testing on paper, especially on the English and reading tests (Li et al., 2017). The studies conducted in 2014 and 2015 used the Pearson TestNav online testing platform, which is currently used for state and district online testing. However, on national testing dates, online testing will occur on the TAO platform developed by OAT. In part due to concerns that test scores from different online testing platforms might exhibit different mode effects (e.g., due to differences in the interface and item rendering), a series of mode comparability studies was conducted during the 2019–2020 academic year.

The three studies took place on the national testing dates in October 2019, December 2019, and February 2020. Only the February 2020 study included writing as an optional component. In each study, the same form was administered on paper and online, but a different form was used for each study. As in earlier mode comparability studies, students were randomly assigned to test on paper or online, and all participants received college-reportable scores. Since the paper and online testing groups were randomly equivalent in terms of academic ability and demographics, observed differences in item and test performance should be attributable to mode effects.

In general, the results were quite consistent across studies and with prior ACT mode comparability studies. Item-level analyses indicated that students who tested online were more likely to respond correctly to most items and they were less likely to leave items blank, in particular near the end of the English and reading tests. Score equivalence analyses revealed that online scores were higher than paper scores on average, especially on the reading and English tests. Across studies, the mode effect ranged from 0.16 to 0.22 standard deviations in reading and from 0.10 to 0.13 in English. The mode effects ranged from 0.04 (nonsignificant) to 0.12 in science, and they ranged from -0.01 (nonsignificant) to 0.06 in math. In the February 2020 study, the average online writing score was 0.39 standard deviations higher than the average paper score. The construct equivalency analyses indicated that paper and online testing appeared to be comparable in terms of correlations among the subject areas, effective weights, internal consistency reliability, and confirmatory factor analysis model fit and average factor loadings. In all cases, the online test was equated to the paper test to ensure that scores reported from this study would be comparable regardless of testing mode.

Introduction

Three mode comparability studies were conducted on the following Saturday national ACT test dates: October 26, 2019, December 14, 2019, and February 8, 2020. The primary goal of these studies was to evaluate whether ACT scores exhibited mode effects between paper and online testing that would necessitate statistical adjustments to the online scores to make them comparable to paper scores. In all three studies, online testing was administered through the TAO Unity platform on non-convertible Chromebooks. All participants took the full ACT multiple-choice test (English, math, reading, and science), and participants in the February study had the option to take the writing test. The studies each employed an experimental, randomly equivalent groups design. All study participants received college-reportable scores, so their motivation to perform well on the ACT was expected to be like typical examinees during a national testing event.

Analyses of mode comparability between the multiple-choice subject tests were related to three broad categories: item-level equivalency, score equivalency, and construct equivalency. The analyses of the writing test centered mainly on score equivalency. Analyses were conducted on both raw score (number correct) and scale scores (on the 1–36 ACT scale), though results reported here focus on scale scores. Initially, the scale scores for all participants (paper and online) were generated using raw-to-scale score conversion tables appropriate to the paper version of the form (established previously when the form was equated). Applying the same conversion table to the paper and online examinees allowed for direct observation of mode effects (e.g., differences in performance between paper and online). In general, the hypotheses considered in this study regarding mode comparability for each subject test were:

- Null Hypothesis (H_0): There is no mode effect.
- Alternative Hypothesis (H_A): There is a mode effect that requires a statistical adjustment to the online scores.

The decision to reject the null hypothesis for a certain subject test was based on consideration of the totality of the statistical evidence gathered. The following sections provide detailed descriptions of the data selection and cleaning activities, item-level equivalency analyses, score equivalency analyses, construct equivalency analyses, conversion table comparisons, and writing analyses.

Data Selection and Cleaning

Random Assignment

All participants in the mode studies took the four subject tests that make up the full ACT multiple-choice test (English, math, reading, and science). To form randomly equivalent groups within each test center, the following steps were implemented with the test registration information prior to test administration:

1. A master list of all eligible participants was created, grouped by test centers.
2. The order of student records was randomized within test centers.
3. For the February study only, student records were ordered within test centers according to which tests they registered to take (full multiple-choice test, followed by full multiple-choice test plus writing).
4. From the reordered lists, odd-numbered students were assigned to paper testing, and even-numbered students were assigned to online testing.

Data Preparation

The following rules were applied to student records after the test event:

1. If a record was flagged in an irregularity report (e.g., due to some interruption of testing), psychometricians applied a set of guidelines to decide whether the record should be analyzed. Any exceptions to the guidelines were resolved through group discussion. As needed, records were deleted after the irregularity review.

2. Records missing any subject score were deleted.

3. Duplicate records were resolved when possible and deleted when not.

4. Records were deleted when the examinees tested in a different mode than they were assigned.

5. Entire test centers were removed if random assignment was not implemented (e.g., all examinees tested in one mode).

6. For both paper and online examinees, scale scores on the 1–36 ACT scale were added to the data file based on the paper raw-to-scale score conversion table.

Table 1 shows the number of records remaining in the data after each data cleaning step. The last row of the table shows the final multiple-choice analysis sample sizes from each study. Table 2 provides the demographic breakdown of the samples by gender and race/ethnicity. The distributions were quite similar across the three studies, though the percentage of Black/African American participants was slightly higher in October than other months. Overall, these percentages were similar to a typical national testing sample. The near-zero percentage differences shown in the last column of Table 3 illustrate effective randomization of participants into the paper and online conditions.

Table 1. Number of Records After Data Cleaning Steps

	October		December		February	
	Paper	Online	Paper	Online	Paper	Online
Original sample size	1,893	1,937	3,204	3,431	3,328	3,453
Step 1 - Irregularity Review	1,893	1,893	3,193	3,421	3,328	3,372
Step 2 - Raw score = 0 or blank	1,893	1,883	3,191	3,415	3,317	3,348
Step 3 - Duplicate	1,893	1,883	3,191	3,415	3,317	3,348
Step 4 - Wrong Mode	1,850	1,883	3,164	3,414	3,297	3,348
Step 5 - Failed Randomization	**1,807**	**1,776**	**3,147**	**3,205**	**3,297**	**3,348**

Table 2. Sample Demographic Distributions

	October		December		February		All Studies		
Group	Paper	Online	Paper	Online	Paper	Online	Paper	Online	Online-Paper
Male	43.2%	41.5%	40.9%	41.7%	40.2%	40.6%	41.1%	41.2%	0.1%
Female	55.9%	58.0%	58.4%	57.4%	59.2%	58.7%	58.2%	58.1%	-0.1%
Black/African American	20.0%	18.6%	12.9%	13.0%	11.6%	12.4%	14.0%	14.0%	0.0%
American Indian/ Alaska Native	0.6%	0.5%	0.8%	0.7%	0.8%	1.1%	0.8%	0.8%	0.0%
White	54.3%	55.7%	58.4%	57.7%	59.4%	59.0%	57.9%	57.8%	-0.1%
Hispanic/Latino	13.7%	13.3%	15.6%	16.5%	17.1%	17.1%	15.8%	16.0%	0.2%
Asian	4.8%	4.8%	4.9%	4.4%	3.0%	2.7%	4.1%	3.8%	-0.3%
Native Hawaiian/ Other Pacific Islander	0.0%	0.1%	0.1%	0.2%	0.3%	0.2%	0.1%	0.2%	0.1%
Two or More Races	3.7%	4.2%	5.1%	4.9%	4.9%	4.5%	4.7%	4.6%	-0.1%
Prefer Not to Respond	2.8%	2.8%	2.2%	2.5%	2.9%	3.0%	2.6%	2.8%	0.2%

Item-level Equivalency Analyses

If there are mode effects, differences in performance between paper and online testing should be apparent on individual items. This section focuses on item-level mode effects by comparing paper and online testing in terms of item-level proportions correct and omit rates. In addition, differential item functioning methods were applied to detect items with unusual differences in performance between modes.

Proportion Correct

Figures 1–4 show plots of the difference in proportion correct (p-values) between paper and online testing. On average, the differences were greater than 0.00, which is apparent from the dashed regression lines. This result indicated that p-values tended to be higher for online testing (i.e., more examinees responded correctly when testing online). This effect was particularly small on the math test, and the regression line was even below zero in the February plot (see Figure 2). On the English and reading tests, the differences in p-values tended to increase with item position, meaning that items appearing later in the test tended to exhibit greater mode effects than items earlier on the test. The effect could, in part, reflect differences in item omit rates, which are presented in the following section.

Figure 5 provides a different illustration of differences in proportion correct between paper and online testing. The cumulative difference in p-values illustrates the accumulation of mode effects over the course of a test. For reading and English, much of the mode effect occurred in later items on the test—a result consistent with differential speededness between paper and online testing. On the reading test, for example, the cumulative difference was less than 0.5 for approximately the first 30 items. In the last 10 items, the difference increased to approximately 1.5 (i.e., online examinees answered an average of 1.5 more items correctly than paper examinees). A similar effect was apparent on the English test, especially in December and February. Mode effects tended to be smaller on the math and science tests, though the October item scores exhibited greater mode effects than December or February.

Figure 1. Differences in English Item Proportion Correct (p-value)

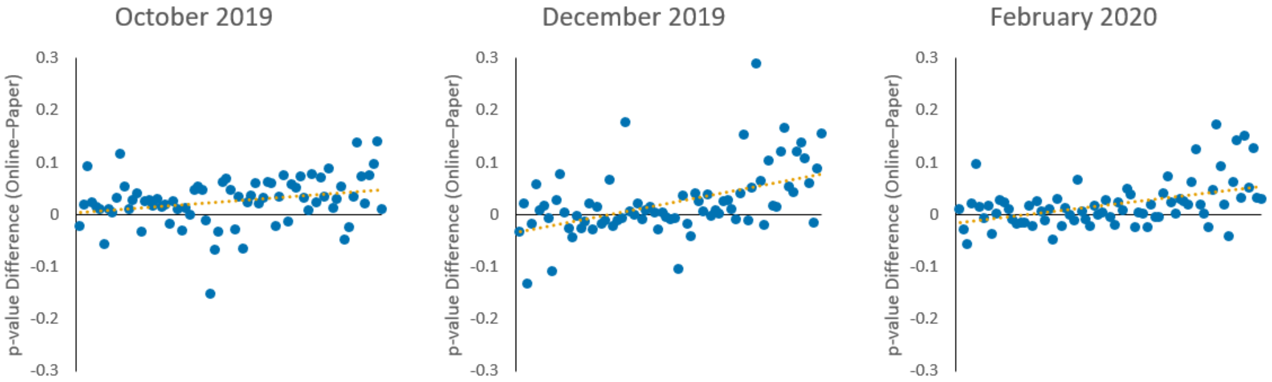


Figure 2. Differences in Math Item Proportion Correct (p-value)

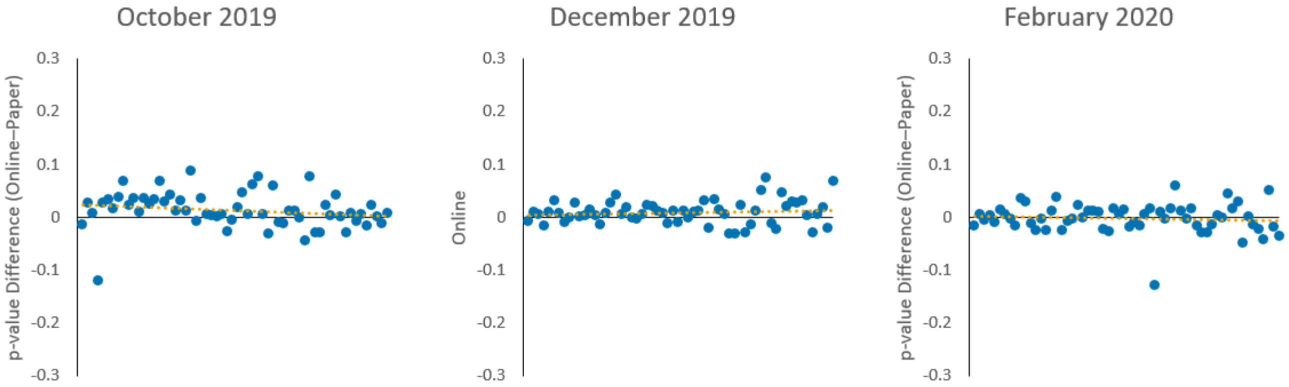


Figure 3. Differences in Reading Item Proportion Correct (p-value)

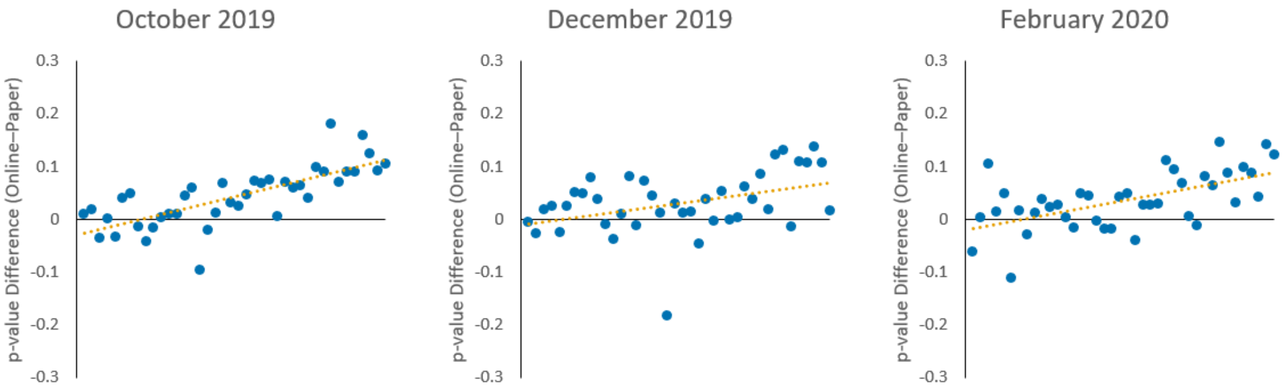


Figure 4. Differences in Science Item Proportion Correct (p-value)

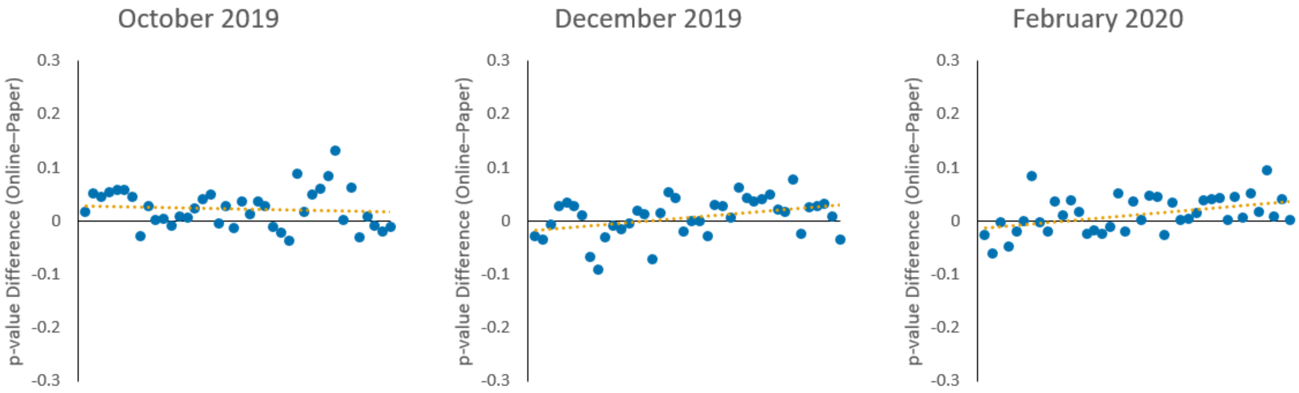
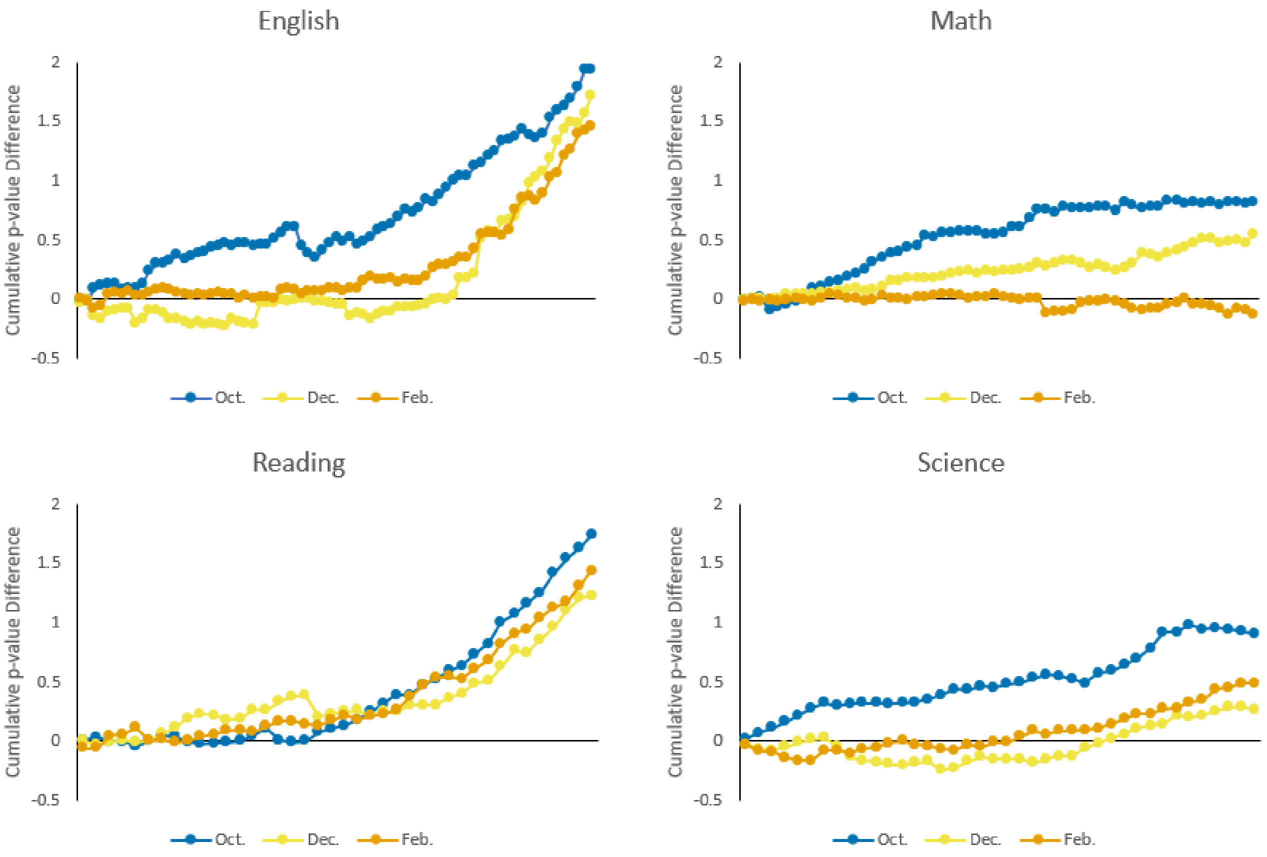


Figure 5. Cumulative Differences in Item Proportion Correct (p-value)



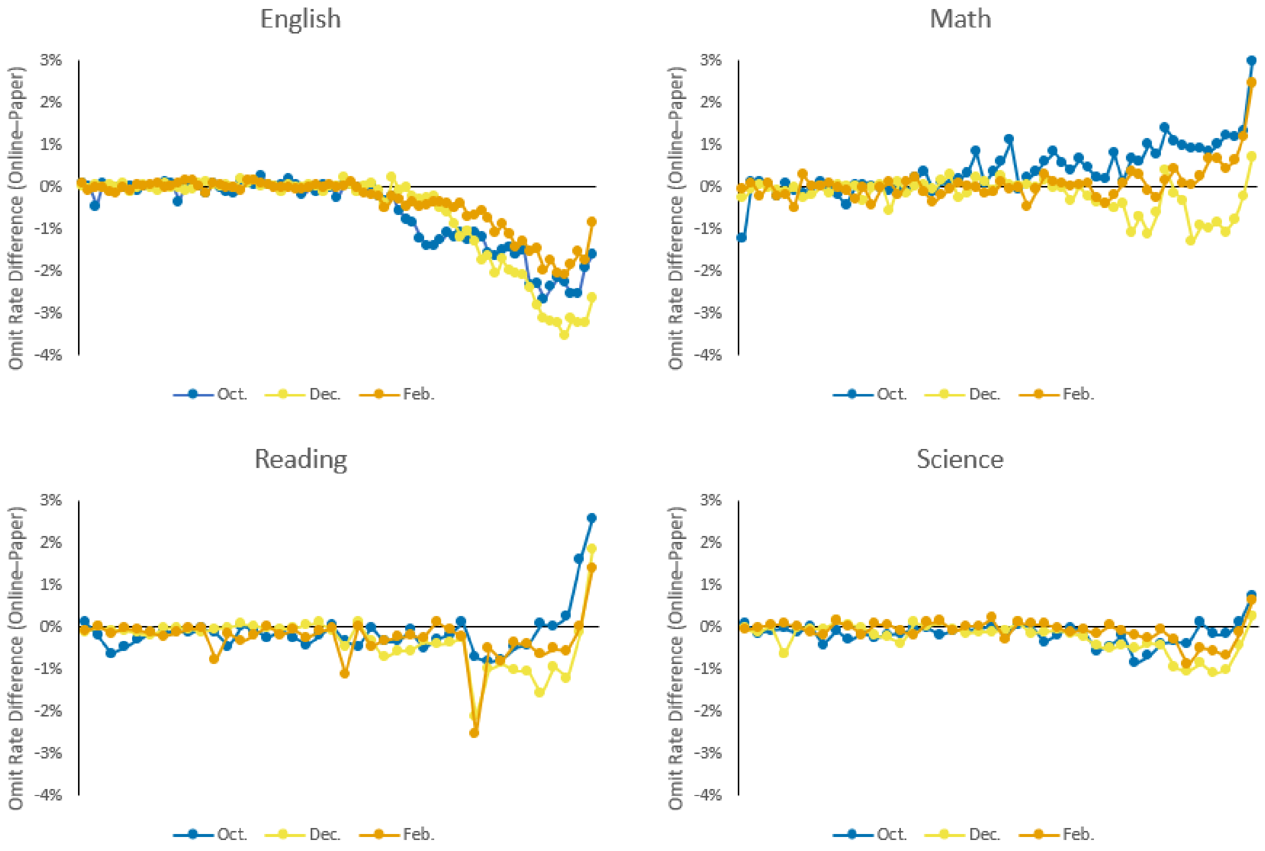
Item Omit Rates

The majority of students responded to every item on each test, but the number of students omitting any items was consistently higher for paper testing than for online testing (13.8% vs. 10.2% for English, 15.9% vs. 13.1% for math, 11.1% vs. 8.3% for reading, and 9.4% vs. 6.9% for science). This result partly explains why paper examinees exhibited lower item proportions correct (p-values) than online examinees. Figure 6 plots differences in omit rates (online minus paper), with negative values indicating that paper examinees were more likely to omit an item than online examinees.

There were some notable trends consistent across subject areas. The difference in item omit rates was nearly zero for approximately the first 75% of each test (the paper and online omit rates were both close to zero on all those items). Near the end of the test, the difference turned negative, which indicated that paper examinees were more likely to omit items. However, at the very end of the test (the last 1–2 items), the difference in omit rates consistently shifted toward the positive. Indeed, on the math, reading, and science tests, online examinees were more likely to omit items at the very end of the test. Perhaps, in the very last moments of the testing session, it was easier for examinees testing on paper to bubble in guesses on the last few items, whereas examinees testing on computers would have had to navigate to each item and click an answer.

On the English test, paper examinees were consistently more likely to omit items. The difference in omit rates was typically less than 1%, but the difference grew as the students approached the end of the test. There, approximately 2–3% more students testing on paper omitted items than students testing on computers. On the math test, differences in item omit rates varied somewhat across studies. The October and December studies showed opposite patterns, but the differences were almost always less than 1%. For much of the reading test, the differences in omit rates were negligible, but paper examinees were about 1–2% more likely to omit near the end of the test (from items 31–38). A similar pattern was observed on the science test, but the differences near the end of the test were smaller than those on the reading test.

Figure 6. Differences in Item Omit Rate



Differential Item Functioning

Differential item functioning (DIF) analyses were conducted on data from each of the mode comparability studies. An item is said to exhibit DIF when there is a significant difference between the probability of getting the item correct for an examinee group of concern (the “focal” group) and the probability for the comparison group (the “reference” group) after controlling for group differences in achievement with respect to the content being tested. This study implemented the Mantel-Haenszel common odds-ratio (MH) procedure (Holland & Thayer, 1988). Using pre-established criteria, items with MH statistics exceeding the tolerance levels were flagged. Table 3 shows the DIF categorization criteria based on MH statistics for dichotomously scored (0/1) items.

Table 3. Criteria for the A, B, and C DIF Categories for Mantel-Haenszel DIF Procedure for Items Scored Dichotomously

Category	Description	Criterion
A	Negligible DIF	Nonsignificant MH-CHISQ ($p \geq .05$) or |MH-D| < 1.0
B	Moderate DIF	Significant MH-CHISQ ($p < .05$) and 1.0 ≤ |MH-D| < 1.5
C	Large DIF	Significant MH-CHISQ ($p < .05$) and |MH-D| ≥ 1.5

Note: MH-CHISQ is the Mantel-Haenszel chi-squared statistic. MH-D is the Mantel-Haenszel D-DIF statistic, which is a transformation of the Mantel-Haenszel common odds ratio (α_{MH}). MH-D indicates the magnitude and direction of DIF.

Whereas DIF analyses typically investigate differences in item performance between demographic groups (e.g., female vs. male, Black vs. White), items flagged for DIF in this study exhibited differences in performance between test modes (online vs. paper). The online test group was treated as the focal group for DIF analyses. Table 4 shows the numbers of flagged items and which groups were favored by the DIF. For example, on the English test in October, one item was flagged with B- DIF (favoring paper), and three items were flagged with B+ DIF (favoring online testing). Across subject areas, English items were most commonly flagged for mode DIF, and the DIF tended to favor online testing (see B+ and C+). Reading items were the next most likely to exhibit mode DIF, though there was more balance between items favoring paper and online testing. Only two out of 180 total math items were flagged for mode DIF, and only five out of 120 total science items were flagged.

Items flagged for DIF were examined by content experts, but there were no apparent explanations for why testing on paper or online might have offered some advantage. Note that false-positive DIF flags were expected to occur at a rate of approximately 5%. A further investigation of English DIF results revealed that nearly all English items flagged for DIF favoring online testing appeared in the last 20 items of the test. This result is consistent with the notion that the DIF was a reflection of differential speededness (related to mode). That is, students testing on computers experienced less speededness near the end of the English test, which allowed them to perform particularly well on those items relative to paper examinees with similar total scores.

Table 4. Number of Items Flagged for Mode DIF Using the Mantel-Haenszel Procedure

Subject	Study	Total	A	B-	B+	C-	C+
English	October 2019	75	70	1	3	1	--
	December 2019	75	63	1	4	2	5
	February 2020	75	66	1	4	--	4
Math	October 2019	60	59	--	--	1	--
	December 2019	60	60	--	--	--	--
	February 2020	60	59	--	--	1	--
Reading	October 2019	40	36	1	1	1	1
	December 2019	40	32	2	5	1	--
	February 2020	40	33	3	3	1	--
Science	October 2019	40	39	--	1	--	--
	December 2019	40	37	3	--	--	--
	February 2020	40	39	1	--	--	--

Note: - sign indicates DIF favoring paper testers (the reference group), and + sign indicates DIF favoring online testers (the focal group)

Score Equivalency Analyses

Whereas the previous section focused on item-level mode effects, the following sections describe analyses comparing paper and online testing in terms of test scores. Mode effects observed at the item level should manifest in test-level mode effects.

Mean Differences

Table 5 provides descriptive statistics for the ACT scale scores of paper and online examinees. In all cases, the scale scores were generated using the paper raw-to-scale score conversion tables (i.e., with no mode adjustment). Thus, the mean differences reported in Table 5 potentially reflect mode effects. With only one exception, all the mean differences were positive, which indicated higher test scores for examinees testing online. For the English test, all the mean differences were statistically significant (see *t* statistics in Table 5). The differences corresponded to effect sizes ranging from 0.10 to 0.13 standard deviations. The math mode effects were closest to zero on average, and only one of the three math mode differences was significantly different from zero. The largest mode effects were observed on the reading tests. The average reading mode effect ranged from 0.16 to 0.22 standard deviations, and all were significantly greater than zero. The average mode effects on the science test ranged from 0.04 (nonsignificant) to 0.12, and two of the three were significantly greater than zero.

Note that the average mode effect varied slightly between male and female examinees and between race/ethnicity groups. However, an in-depth analysis of differential mode effects detected no statistically significant mode by gender or mode by race/ethnicity interactions (Wang & Steedle, 2020). Results from that study suggest that general mode adjustments are appropriate for students in different demographic groups. That is, no examinee groups would be unfairly advantaged or disadvantaged by a mode adjustment applied to online scores. The study detected evidence that mode effects are slightly greater for higher ability examinees, but mode adjustments based on equating account for this interaction.

Table 5. Scale Score Descriptive Statistics and Comparisons

		Paper		Online		Mean	Effect		
Subject	Study	Mean	SD	Mean	SD	Difference	Size	*t*	*p*
English	October 2019	18.37	6.08	19.17	6.02	0.80	0.13	3.97***	< .0001
	December 2019	19.64	6.02	20.34	6.08	0.70	0.12	4.65***	< .0001
	February 2020	19.31	5.90	19.94	6.12	0.63	0.10	4.25***	< .0001
Math	October 2019	19.09	4.99	19.37	4.82	0.29	0.06	1.74	.0818
	December 2019	20.05	4.93	20.30	5.20	0.25	0.05	2.00*	.0452
	February 2020	19.83	4.88	19.76	4.94	-0.07	-0.01	-0.55	.5830
Reading	October 2019	20.00	6.53	21.50	6.81	1.50	0.22	6.74***	< .0001
	December 2019	21.39	6.42	22.45	6.52	1.06	0.16	6.51***	< .0001
	February 2020	20.73	6.39	21.92	6.49	1.19	0.18	7.52***	< .0001
Science	October 2019	19.61	5.24	20.23	5.35	0.62	0.12	3.51***	.0005
	December 2019	20.59	5.00	20.78	5.45	0.19	0.04	1.46	.1440
	February 2020	20.31	5.06	20.71	5.52	0.39	0.07	3.03**	.0020

* *p* < .05, ** *p* < .01, *** *p* < .001

Distributional Differences

Kolmogorov-Smirnov (KS) tests were used to determine whether there were statistically significant differences between the distributions of scale scores (without a mode adjustment) for the paper and online testers. Specifically, the KS test evaluates whether two sets of scores might have been drawn from the same distribution. As shown in Table 6, the KS test results indicated that the scale score distributions were significantly different for all comparisons except for math, which was significant only in the October study. Cumulative distributions of scale scores for the four subjects were compared for online and paper testing (Figures 7–10). As expected, the online score distributions appear slightly shifted toward higher scores compared to the paper score distributions, especially on the English and reading tests. When comparing the October, December, and February studies, there was notable consistency in the distributions and the gaps between paper and online testing.

Table 6. Kolmogorov-Smirnov Test Statistics

	October 2019		December 2019		February 2020	
Subject	*D*	*p*	*D*	*p*	*D*	*p*
English	0.069***	.0004	0.063***	< .0001	0.062***	< .0001
Math	0.048*	.0322	0.030	.1183	0.020	.5091
Reading	0.100***	< .0001	0.076***	< .0001	0.074***	< .0001
Science	0.064**	.0012	0.044**	.0040	0.041**	.0071

* *p* < .05, ** *p* < .01, *** *p* < .001

Figure 7. Relative Cumulative Frequency Distributions of English Scale Scores

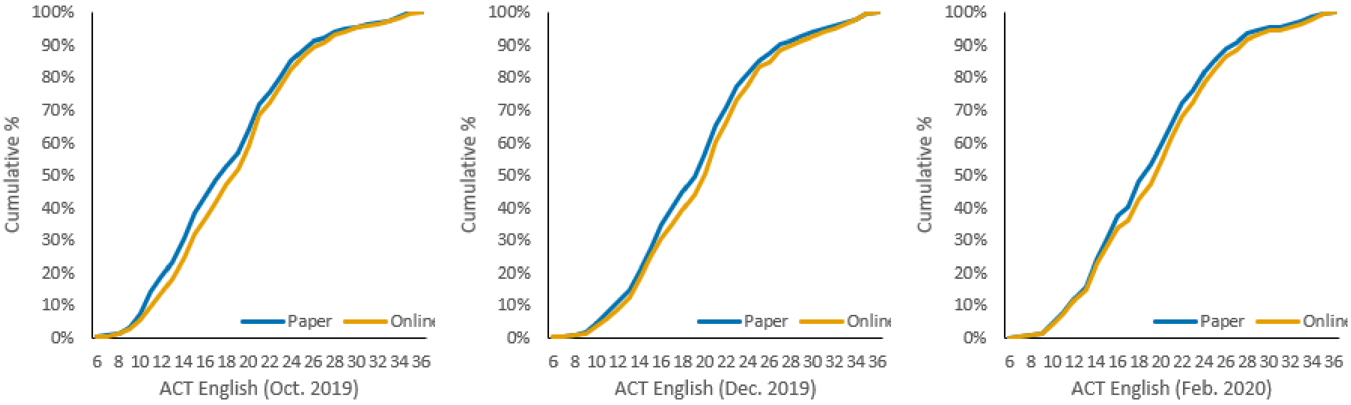


Figure 8. Relative Cumulative Frequency Distributions of Math Scale Scores

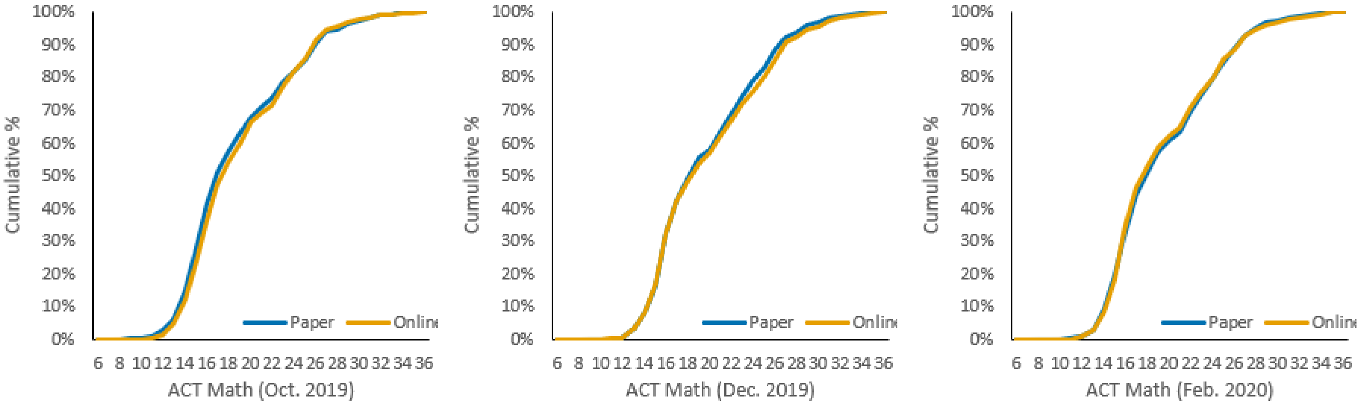


Figure 9. Relative Cumulative Frequency Distributions of Reading Scale Scores

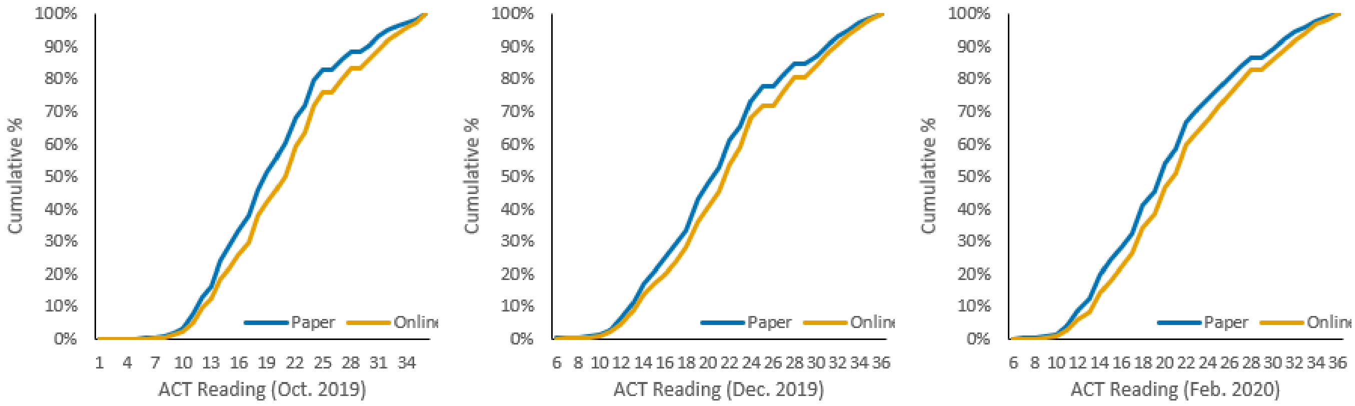
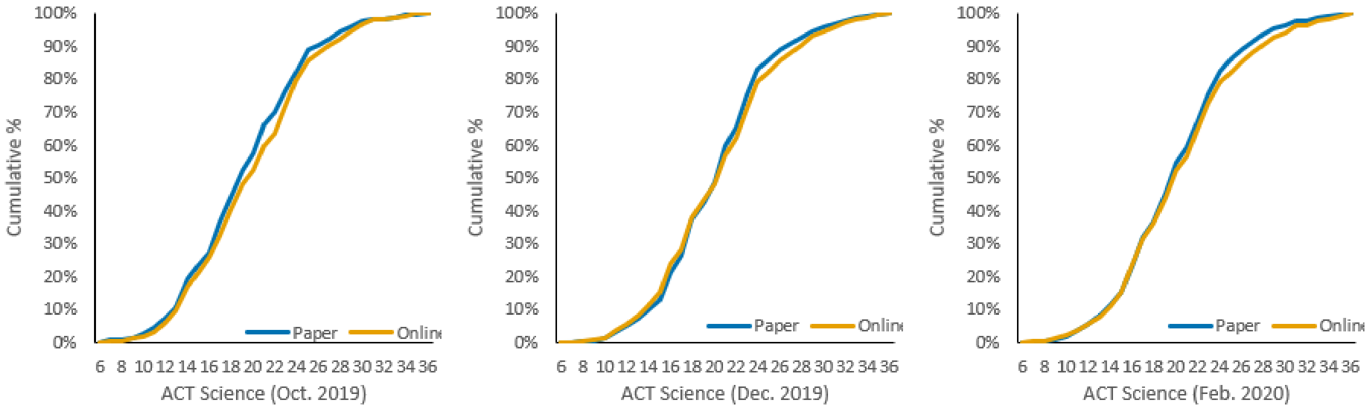


Figure 10. Relative Cumulative Frequency Distributions of Science Scale Scores



Reliability and Standard Error of Measurement

The psychometric properties of scale scores were examined after equating the online forms to the paper forms. This analysis included estimating scale score reliability, standard error of measurement (SEM), and conditional standard error of measurement (CSEM) based on the four-parameter beta compound binomial model for raw scores (Lord, 1965). Table 7 presents the scale score reliability (α) and SEM for each testing mode, and Figures 11–14 show plots of the CSEM at each true scale score point for the paper forms and the online forms after applying a mode adjustment. In general, reliability coefficients and SEMs indicate the precision of test scores (or lack thereof). Imprecision (or random measurement error) is related to the fact that examinees would not necessarily earn the same test scores if they took the test repeatedly. Higher reliability coefficients and lower SEMs indicate greater measurement precision.

The estimated reliability coefficients from online testing (before and after adjustment) were very similar to those from paper testing (Table 7). For online testing, the SEM values changed slightly after applying a mode adjustment (often decreasing), but applying the mode adjustment did not change the reliability estimates except in one case (science in the December 2019 study). As shown in Figures 11–14, the CSEMs were generally similar for paper and online testing. When there were differences, they did not consistently favor one mode or the other.

Table 7. Coefficient Alpha and Standard Error of Measurement

		Paper		Online		Online Adjusted	
Subject	Study	α	SEM	α	SEM	α	SEM
English	October 2019	.92	1.70	.92	1.72	.92	1.74
	December 2019	.92	1.72	.92	1.73	.92	1.70
	February 2020	.92	1.70	.92	1.73	.92	1.68
Math	October 2019	.89	1.63	.88	1.67	.88	1.72
	December 2019	.89	1.62	.90	1.62	.90	1.55
	February 2020	.88	1.70	.88	1.68	.88	1.65
Reading	October 2019	.85	2.50	.86	2.54	.86	2.41
	December 2019	.85	2.47	.85	2.50	.85	2.48
	February 2020	.86	2.38	.86	2.40	.86	2.33
Science	October 2019	.85	2.05	.85	2.06	.85	2.07
	December 2019	.80	2.26	.83	2.26	.82	2.08
	February 2020	.83	2.11	.85	2.13	.85	1.97

Figure 11. Conditional Standard Error of Measurement for English Scale Scores

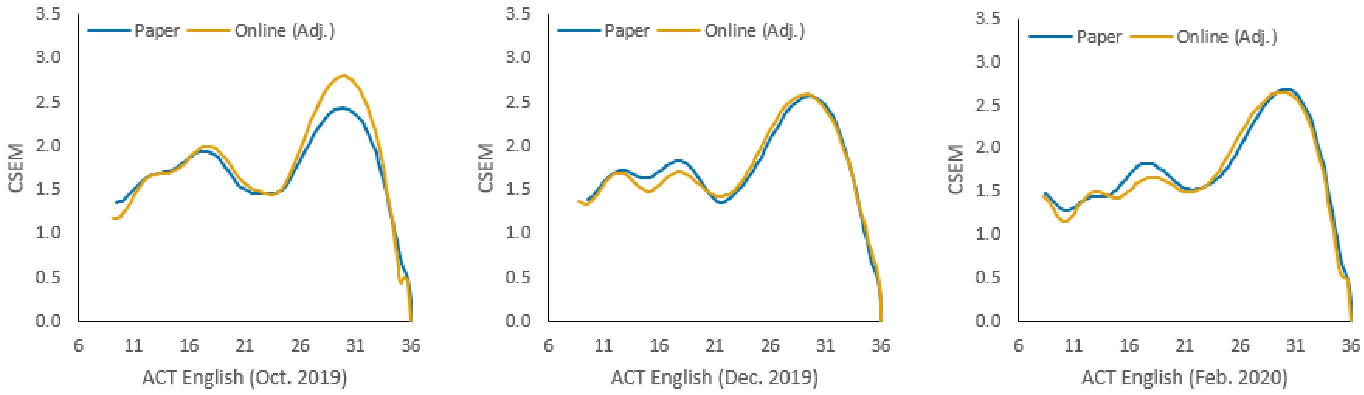


Figure 12. Conditional Standard Error of Measurement for Math Scale Scores

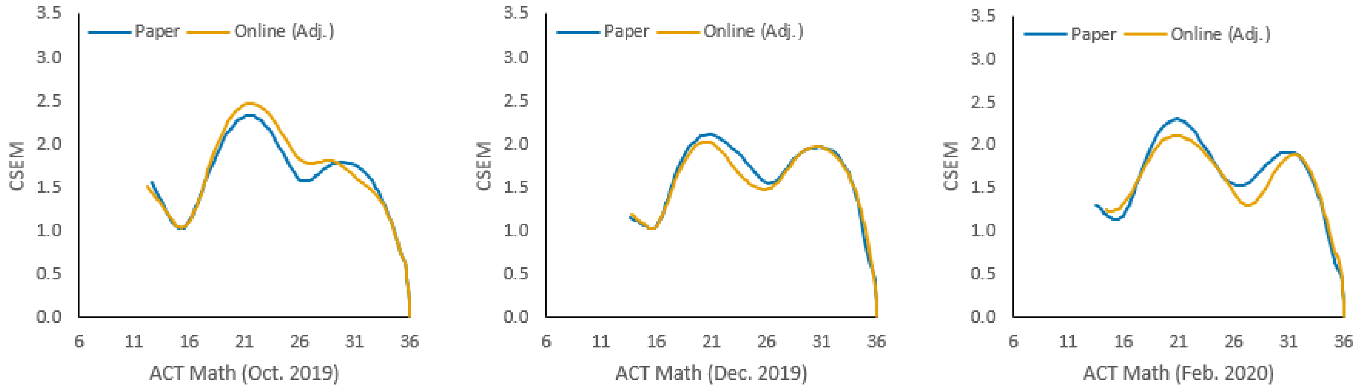


Figure 13. Conditional Standard Error of Measurement for Reading Scale Scores

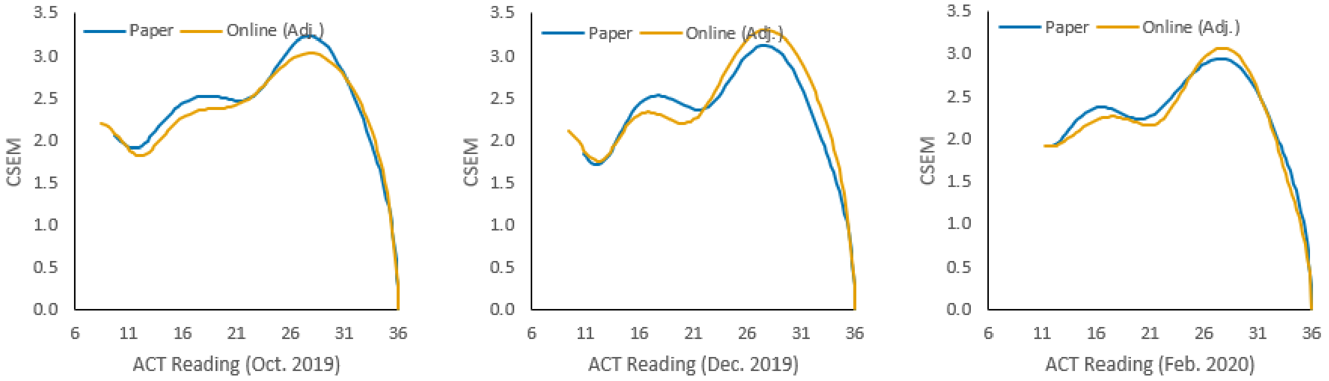
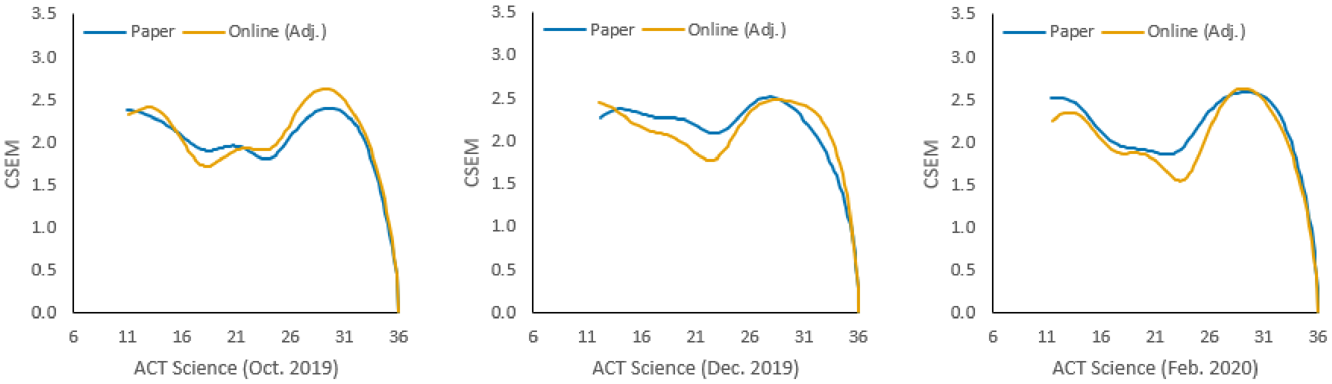


Figure 14. Conditional Standard Error of Measurement for Science Scale Scores



Construct Equivalency Analyses

In order to treat paper and online scores as comparable, it must be true that paper and online testing measure the same constructs. The construct equivalency analyses were designed to compare paper and online testing in terms of the measured constructs. If paper and online testing measure the same constructs, then their results in the following analyses should be similar.

Correlations and Effective Weights on Composite Scores

The first step in the construct equivalency analyses was calculating the correlations among scale scores on the four subject tests and proportional effective weights of subject test scores. An effective weight ($e\omega_i$) is the statistical contribution of the subject test scores to the variance of the composite. The proportional effective weight was computed as follows, given composite scores defined by an equally weighted mean of the four subject tests (i.e., the average of English, math, reading, and science scores):

$$e\omega_i = \frac{\sigma_i^2 + \sum_{j \neq i} \sigma_{ij}}{\sum_i [\sigma_i^2 + \sum_{j \neq i} \sum_{j \neq i} \sigma_{ij}]},$$

where σ_i^2 is the variance of scale scores on test *i*, and σ_{ij} is the covariance between scale scores on tests *i* and *j* (e.g., English and math, or reading and science). Results were then averaged across studies. Average correlations are shown in Table 8, and average effective weights are shown in Table 9.

The correlation coefficients for online testing were higher than those for paper testing except for the correlation between English and math (Table 8). However, differences in the correlations between the two modes were very small (.00–.02). Moreover, the pattern of the correlations in terms of relative magnitude were similar for online and paper tests. For example, the correlation between English and math was lower than the correlation between English and reading for both testing modes. Like the correlation results, there were differences in effective weights between two modes, but their patterns were very similar. For both test modes, the effective weights for reading and English scores were relatively large compared to math and science scores, and math scores had the smallest effective weight. In summary, results of correlation and effective weight analyses supported the construct equivalency of the two modes.

Table 8. Average Correlations Between Subject Test Scale Scores

Mode	Subject	English	Math	Reading	Science
Paper	English	--	.73	.80	.76
	Math		--	.67	.78
	Reading			--	.75
	Science				--
Online	English	--	.73	.82	.78
	Math		--	.68	.80
	Reading			--	.76
	Science				--

Table 9. Average Proportional Effective Weights

Mode	English	Math	Reading	Science
Paper	.27	.21	.29	.23
Online	.27	.21	.29	.24

Confirmatory Factor Analysis

This section presents evidence showing the extent to which the scoring and reporting structure of both paper and online testing were consistent with the internal structure reflected in the observed data. In addition to an overall score, each ACT subject test provides subscores associated with content-related reporting categories (i.e., reporting category scores). A series of confirmatory factory analyses (CFA) were conducted to test the hypothesis that the subscore reporting structure was consistent with the internal structure shown in observed data for both paper and online testing. CFA is an approach to test whether a theoretical model of internal structure is consistent with observed data for a test or measure. For each ACT subject test, a model of the internal structure was theorized based on the test blueprint and reporting category classifications (see ACT, 2019). For example, the structural model for Science consisted of three latent factors, which represented three content-related reporting categories: Interpretation of Data, Scientific Investigation, and Evaluation of Models, Inferences, and Experimental Results. Each of these latent factors was manifested by a certain number of observed variables (i.e., test item responses), and each item was associated with only one latent factor. The theoretical model allowed the three latent factors to correlate with each other. Similar models were built for English, math, and reading based on the test blueprints.

The CFA results were evaluated and compared based on model-data fit statistics and factor loadings of items on latent factors for all four subject tests in both modes. Table 10 presents the model fit statistics for the four subjects (English, math, reading, and science) by mode. First, a chi-squared (χ^2) statistic—the most frequently cited index of absolute model-data fit—was examined. The χ^2 test compares the observed covariance matrix with the theoretically proposed covariance matrix. However, the result of a χ^2 test is known to be greatly influenced by the sample size, so that well-fitting models can sometimes produce statistically significant χ^2 test results when sample sizes are large. Thus, other fit statistics were included when interpreting model-data fit results. The additional fit statistics included Goodness of Fit (GFI), Standardized Root Mean Square Residual (SRMR), and Root Mean Square Error of Approximation (RMSEA). A GFI value of larger than 0.9 indicates satisfactory fit, and a cutoff SRMR value of 0.08 suggests relatively good fit between the hypothesized model and the observed data. RMSEA is a measure of “discrepancy per degree of freedom” in a model when taking parsimony into account in model comparison (Browne & Cudeck, 1993) Values less than 0.08 suggest reasonable model fit, and values less than 0.05 suggest good model fit. Based on the suggested cutoff values and interpretation of the fit statistics (Table 10), the subscore reporting structure is consistent with the internal structure shown in observed data for both paper and online testing, though the online data exhibited slightly better model-data fit than the paper data.

Table 10. Confirmatory Factor Analysis Model-Data Fit Indices

Subject	Mode	Study	χ^2	*df*	χ^2 *p*-value	RMSEA	SRMR	GFI
English	Paper	October 2019	6898.6	2697	< .0001	0.029	0.034	0.889
		December 2019	85236.8	2775	< .0001	0.031	0.032	0.905
		February 2020	46742.9	2775	< .0001	0.031	0.034	0.888
	Online	October 2019	6048.4	2697	< .0001	0.027	0.032	0.900
		December 2019	44691.1	2775	< .0001	0.029	0.032	0.907
		February 2020	50384.0	2775	< .0001	0.030	0.032	0.897
Math	Paper	October 2019	3963.0	1695	< .0001	0.027	0.033	0.919
		December 2019	60774.6	1770	< .0001	0.023	0.026	0.947
		February 2020	26125.4	1770	< .0001	0.024	0.028	0.943
	Online	October 2019	3596.7	1695	< .0001	0.025	0.032	0.924
		December 2019	33559.1	1770	< .0001	0.024	0.029	0.933
		February 2020	26671.5	1770	< .0001	0.024	0.027	0.943
Reading	Paper	October 2019	2438.4	737	< .0001	0.036	0.040	0.916
		December 2019	31532.2	780	< .0001	0.029	0.028	0.956
		February 2020	18241.9	780	< .0001	0.035	0.352	0.934
	Online	October 2019	1691.9	737	< .0001	0.027	0.031	0.947
		December 2019	15968.8	780	< .0001	0.024	0.025	0.965
		February 2020	17729.4	780	< .0001	0.023	0.025	0.967
Science	Paper	October 2019	2124.8	737	< .0001	0.032	0.034	0.936
		December 2019	27192.9	780	< .0001	0.031	0.030	0.954
		February 2020	16420.7	780	< .0001	0.029	0.031	0.952
	Online	October 2019	1945.8	737	< .0001	0.030	0.033	0.940
		December 2019	15485.1	780	< .0001	0.030	0.030	0.952
		February 2020	18707.9	780	< .0001	0.025	0.026	0.963

Next, adequacy of the factor model structure was evaluated by examining the factor loadings. Factor loadings represent the strength of association between a latent factor and its observed indicators. In the current analysis, factor loadings indicated the strength of association between a reporting category and scores on the items associated with that reporting category. Table 11 shows the average of the items' standardized factor loadings for each reporting category. For all four subject tests, most of the factor loadings suggested a moderate association (approximately .40) between the items and their respective reporting categories. Across subjects, the average difference between the paper and online averages was only .007, indicating substantial consistency between the factor loadings for paper and online testing.

Table 11. Average Factor Loadings by Subject Test and by Mode

Subject	Reporting Category	October 2019		December 2019		February 2020		Avg. Difference
		Paper	Online	Paper	Online	Paper	Online	
English	Production of Writing (PoW)	.393	.398	.446	.454	.423	.446	.012
	Knowledge of Language (KLA)	.415	.384	.388	.387	.433	.452	-.004
	Conventions of Standard English (CoE)	.395	.399	.361	.364	.370	.383	.007
Math	Number & Quantity (PHM)	.427	.400	.433	.442	.310	.297	-.010
	Algebra (ALG)	.387	.377	.400	.419	.415	.436	.010
	Functions (FUN)	.301	.298	.387	.388	.376	.389	.004
	Geometry (GEO)	.362	.352	.340	.354	.376	.389	.006
	Statistics & Probability (SAP)	.373	.366	.312	.311	.330	.340	.001
	Integrating Essential Skills (IES)	.377	.373	.411	.424	.349	.345	.002
Reading	Key Ideas & Details (KID)	.362	.384	.375	.378	.381	.387	.010
	Craft & Structure (CAS)	.403	.388	.362	.358	.391	.391	-.006
	Integration of Knowledge & Ideas (IOK)	.341	.376	.391	.394	.368	.380	.017
Science	Interpretation of Data (IOD)	.361	.373	.368	.380	.340	.358	.014
	Scientific Investigation (SIN)	.427	.430	.357	.375	.351	.383	.018
	Evaluation of Models, Inferences & Experimental Results (EMI)	.383	.398	.308	.326	.382	.410	.020

Invariance Testing

The extent to which item factor models measuring each multiple-choice subject exhibited measurement invariance between paper and online testing was evaluated with a series of multiple-group invariance models. All models were estimated using Mplus 7.4 (Muthén & Muthén, 2017) with a weighted least square mean and variance adjusted (WLSMV) limited-information estimator, the probit link, and the theta parameterization. At each step, equality constraints were tested with nested model comparisons. English, reading, and science are passage-based tests where several items reference a common passage. For these subjects, bi-factor models (Holzinger & Swineford, 1937) were used to specify nuisance factors that account for residual correlations due to the passages.

Table 12 provides a summary of results. All analyses detected configural invariance, which indicates that the model form is invariant across paper and online testing. Put another way, the same latent constructs are manifested by the same observations (i.e., test item scores) for paper and online testing. This supports the notion that paper and online testing measure the same constructs. Beyond configural invariance, the strength of evidence supporting invariance differed by subject area and across the three studies. Evidence of construct equivalency was very strong for the math test, which exhibited residual invariance (i.e., equivalence of loadings, intercepts, and residuals). The English and reading tests each exhibited some degree of metric invariance (i.e., equivalence of loadings), which is relatively weak evidence of invariance. Like English and reading, the October science test data exhibited partial metric invariance. However, results from the December and February science test data indicated relatively strong evidence of invariance: partial residual invariance (i.e., equivalence of loadings, intercepts, and some residuals).

Table 12. Summary of Invariance Testing Results

Study	English	Math	Reading	Science
October 2019	Metric invariance (weaker evidence of comparability)	Residual invariance (strong evidence of comparability)	Partial metric invariance (weaker evidence of comparability)	Partial metric invariance (weaker evidence of comparability)
December 2019	Partial metric invariance (weaker evidence of comparability)	Residual invariance (strong evidence of comparability)	Partial metric invariance (weaker evidence of comparability)	Partial residual invariance (moderate evidence of comparability)
February 2020	Partial metric invariance (weaker evidence of comparability)	Residual invariance (strong evidence of comparability)	Partial metric invariance (weaker evidence of comparability)	Partial residual invariance (moderate evidence of comparability)

Conversion Table Comparisons

Raw-to-scale score conversion tables were generated for paper and online testing using equipercentile equating. Considering that online testers tended to earn higher scores, the scale scores associated with certain raw score were likely to be higher for paper testers (to adjust for the fact that their test was more difficult). Indeed, that was the generally the case, as shown in Figures 15–18, which present plots of differences between the raw-to-scale score conversion tables (online minus paper). Consider, for example, a paper tester who answered 35 correct (out of 75) on the English test and an online tester who answered the same number correct. The paper tester would have a reported score of 17 (on the 1–36 scale), and the online tester would have a reported score of 16. This is reflected by a difference of -1 in Figure 15 at the raw score 35.

Recall that the largest average mode effects were observed on the English and reading tests. With very few exceptions, the size of the mode adjustment was 0 or -1 point for the English tests. For the reading tests, the mode adjustment was most often -1 point, but it reached -2 in several cases and even -3 near the high end of the score scale in the October study. Consistent with the small magnitude of the average math mode effect, the math mode adjustment was most often 0, but it ranged from -1 to +1. For the science test in the October study, the mode adjustment was most often -1. In contrast, the conversion table comparisons indicated a mode effect favoring paper testing for lower ability examinees but a mode effect favoring online testing for higher ability examinees.

Table 13 summarizes the effects of applying mode-adjusted conversion tables for online examinees. As expected, the average effects of the adjustments were negative, except on the math test. That is, average online scores were lower when applying the mode-adjusted conversion tables. Table 13 also shows the mean difference between paper and online scores with and without mode adjustments. If the mode adjustment works as intended, the mean differences between paper and online testing should be approximately zero after adjusting. In general, application of the mode-adjusted conversion tables resulted in mean differences close to zero. In only one case did the difference increase slightly in magnitude (math in February 2020), but the difference was very small to begin with.

Figure 15. Difference between English Scale Scores Corresponding to the Same Raw Score (Online Minus Paper)

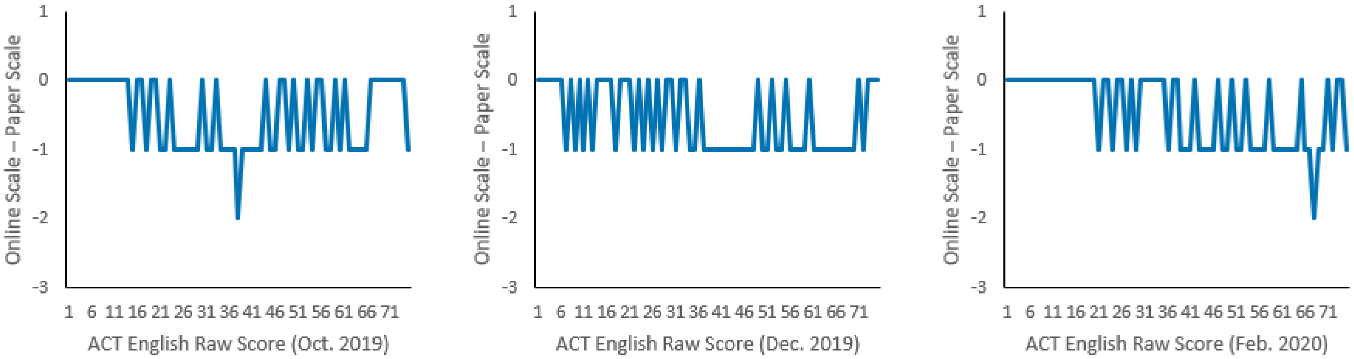


Figure 16. Difference between Math Scale Scores Corresponding to the Same Raw Score (Online Minus Paper)

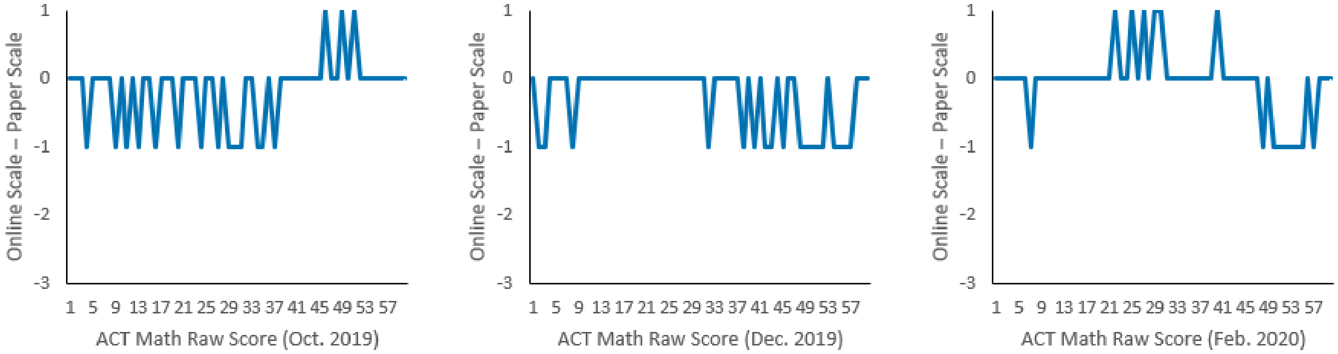


Figure 17. Difference between Reading Scale Scores Corresponding to the Same Raw Score (Online Minus Paper)

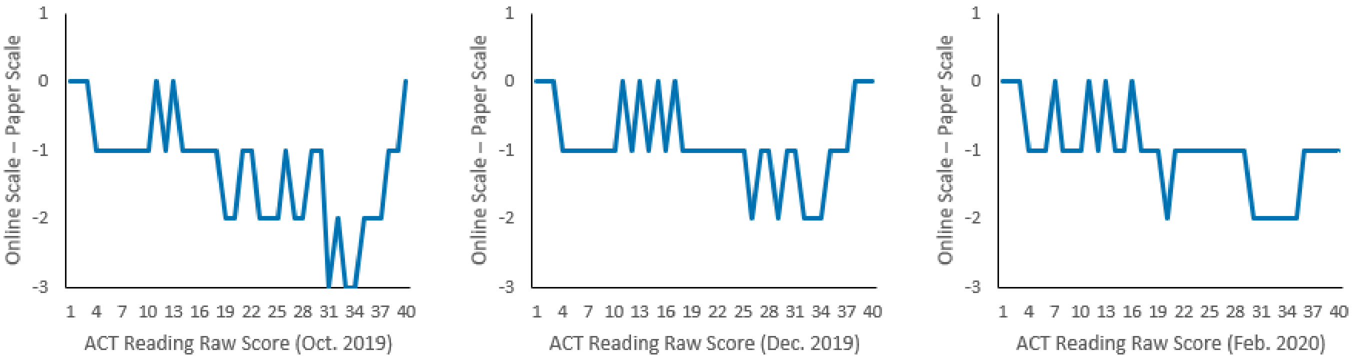


Figure 18. Difference between Science Scale Scores Corresponding to the Same Raw Score (Online Minus Paper)

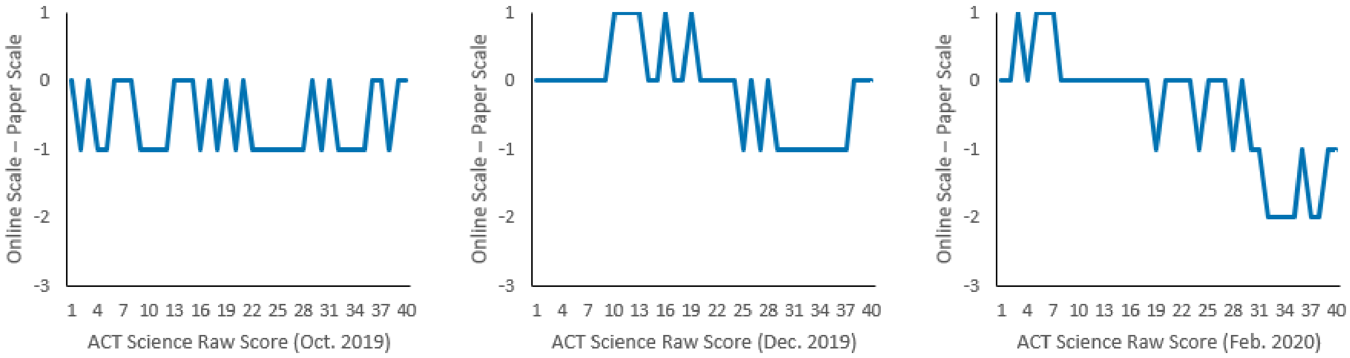


Table 13. Mean Online Scores With and Without Mode Adjustment

Subject	Study	Paper Mean	Online Mean (Unadjusted)	Online Mean (Adjusted)	Online-Paper (Unadjusted)	Online-Paper (Adjusted)
English	October 2019	18.37	19.17	18.46	0.80	0.09
	December 2019	19.64	20.34	19.62	0.70	-0.02
	February 2020	19.31	19.94	19.35	0.63	0.04
Math	October 2019	19.09	19.37	19.08	0.29	-0.01
	December 2019	20.05	20.30	20.10	0.25	0.05
	February 2020	19.83	19.76	19.92	-0.07	0.09
Reading	October 2019	20.00	21.50	19.97	1.50	-0.02
	December 2019	21.39	22.45	21.38	1.06	-0.01
	February 2020	20.73	21.92	20.74	1.19	0.01
Science	October 2019	19.61	20.23	19.60	0.62	-0.01
	December 2019	20.59	20.78	20.70	0.19	0.11
	February 2020	20.31	20.71	20.28	0.39	-0.03

Writing Analyses

The ACT writing test was included in the February 2020 study to evaluate differences between writing performance across testing modes. Students at participating schools could choose to take the ACT with or without writing. To obtain randomly equivalent groups of students taking the ACT test on paper and online for both the multiple-choice tests and the writing test, random assignment was conducted separately for students who did and did not choose to take the writing test within each school.

To ensure that scores were comparable regardless of testing mode, test equating was applied to adjust for mode differences. Specifically, equipercentile equating was conducted using the random groups design, and scores for students who took the writing test online were adjusted to be comparable to students who took the test on paper.

Besides test equating, other analyses were conducted to examine the representativeness of the writing mode study sample and the writing test mode effects. The following sections provide more details about these analyses and results. For analyses conducted without a mode adjustment, the raw-to-scale score conversion for paper testing was applied to both paper and online testers. Note that writing scores are reported on a 2–12 scale, but writing raw scores are on an 8–48 scale, which is the sum of four domain scores (each scored on a 1–6 rubric scale) from two raters.

Sample Characteristics and Representativeness

The multiple-choice tests and the writing test forms administered in the February 2020 mode comparability study were the same as those administered to the large sample of students who participated in the Saturday national test administration on the same day. The mode comparability study sample was compared to the national test sample in terms of demographics, average ACT scores, and percentage of students who chose to take the writing test.

Table 14 shows how the February 2020 writing mode study sample compared with the national testing sample in terms of gender and race/ethnicity percentages. In the study sample, there were proportionately fewer males, more Hispanic/Latino students, and fewer Black/African American students.

Table 14. Demographic Comparison between the Mode Comparability Study Sample and the Saturday National Test Sample in February 2020

		N		Percentage	
Variable	Group	Study	National	Study	National
	All	2,581	234,976		
Gender	Male	990	102,881	38.4%	43.8%
	Female	1,571	130,957	60.9%	55.7%
	Another Gender	5	144	0.2%	0.1%
	Prefer not to respond	15	994	0.6%	0.4%
Race/ Ethnicity	Black/African American	233	33,108	9.0%	14.1%
	American Indian/Alaska Native	17	1,624	0.7%	0.7%
	White	1,370	127,815	53.1%	54.4%
	Hispanic/Latino	661	28,604	25.6%	12.2%
	Asian	87	12,842	3.4%	5.5%
	Native Hawaiian/Other Pacific Islander	10	521	0.4%	0.2%
	Two or more races	113	9,453	4.4%	4.0%
	Prefer not to respond	90	10,500	3.5%	4.5%

Table 15 shows how the percentages of student who chose to take the writing test in the mode study compared with the percentages in the Saturday national test sample, in total and by gender and race/ethnicity. Students were considered as having chosen to the take the ACT writing test if they had a valid writing form code in the data, whether they had a valid writing score or not. A higher percentage of students chose to take the writing test in the mode study than the national sample (39.1% vs. 25.7%). Students may have been encouraged to take writing in the mode study by their schools or because it was provided free of charge. This was also true within most gender and race/ethnicity groups.

Note that not all students who chose to take the writing test obtained valid writing scores (e.g., some students provided no response to the writing prompt). The percentages of students who had valid writing scores among those who chose to take the writing test were similar between the study (95.7%) and the national testing sample (96.6%).

Table 15. Percentages of Students Who Chose to Take Writing in February 2020

Variable	Group	Study	National
	All	39.1%	25.7%
Gender	Female	37.2%	25.0%
	Male	40.3%	26.2%
	Another Gender	45.5%	43.8%
	Prefer not to respond	47.1%	39.8%
Ethnicity	Black/African American	29.9%	18.5%
	American Indian/Alaska Native	27.7%	19.5%
	White	35.1%	24.7%
	Hispanic/Latino	57.8%	33.4%
	Asian	46.0%	47.6%
	Native Hawaiian/Other Pacific Islander	63.2%	43.4%
	Two or more races	36.3%	29.5%
	Prefer not to respond	45.6%	35.2%
	Blank	0.0%	0.3%

Table 16 provides sample sizes and relevant score means for the February 2020 study sample and national testing sample that took writing. The study sample had relatively low average scores in writing, English, and reading compared to the February 2020 national sample.

Table 16. Score Means for February 2020 Study Sample (N = 1,329) and National Test Sample (N = 57,837) That Took Writing

Score	Study Mean	National Mean
Ideas and Analysis (2–12)	6.42	7.13
Development and Support (2–12)	5.89	6.59
Organization (2–12)	6.31	7.02
Language Use and Conventions (2–12)	6.78	7.45
Writing Raw Score (8–48)	25.40	28.18
English Raw (0–75)	41.90	51.29
Reading Raw (0–40)	22.59	27.03

Writing Scores and Correlations

Score distributions of the writing test scores and correlations among the writing test scores were compared across modes without any mode adjustments applied. Table 17 presents descriptive statistics for the writing test scores for each mode, mean differences, effect sizes, and *t*-test results. All online writing scores were significantly higher than corresponding paper scores on average. Figure 19 shows the relative cumulative frequency distributions of each domain score by test mode, and Figure 20 shows the same for writing raw scores and scale scores. As a reminder, the paper raw-to-scale score conversion table was applied to paper and online raw scores, which permits the observation of mode effects. Results for the four domain scores, raw scores, and scale scores (without adjustment) consistently indicate that the online mode was easier than the paper mode.

Table 17. Score Descriptive Statistics and Comparisons for Writing Analysis

	Paper		Online					
Score	Mean	SD	Mean	SD	Mean Difference	Effect Size	*t*	*p*
Ideas and Analysis (2–12)	6.42	1.72	7.06	1.65	0.64	0.38	9.64***	< .0001
Development and Support (2–12)	5.89	1.72	6.57	1.63	0.68	0.41	10.30***	< .0001
Organization (2–12)	6.31	1.70	6.89	1.59	0.58	0.35	8.94***	< .0001
Language Use and Conventions (2–12)	6.78	1.50	7.35	1.54	0.57	0.38	9.53***	< .0001
Writing Raw Score (8–48)	25.4	6.45	27.88	6.21	2.48	0.39	9.94***	< .0001
Writing Scale Score (1–36)	16.93	5.64	19.02	5.13	2.09	0.39	9.84***	< .0001

* $p < .05$, ** $p < .01$, *** $p < .001$

Figure 19. Relative Cumulative Frequency Distributions of Writing Domain Scores

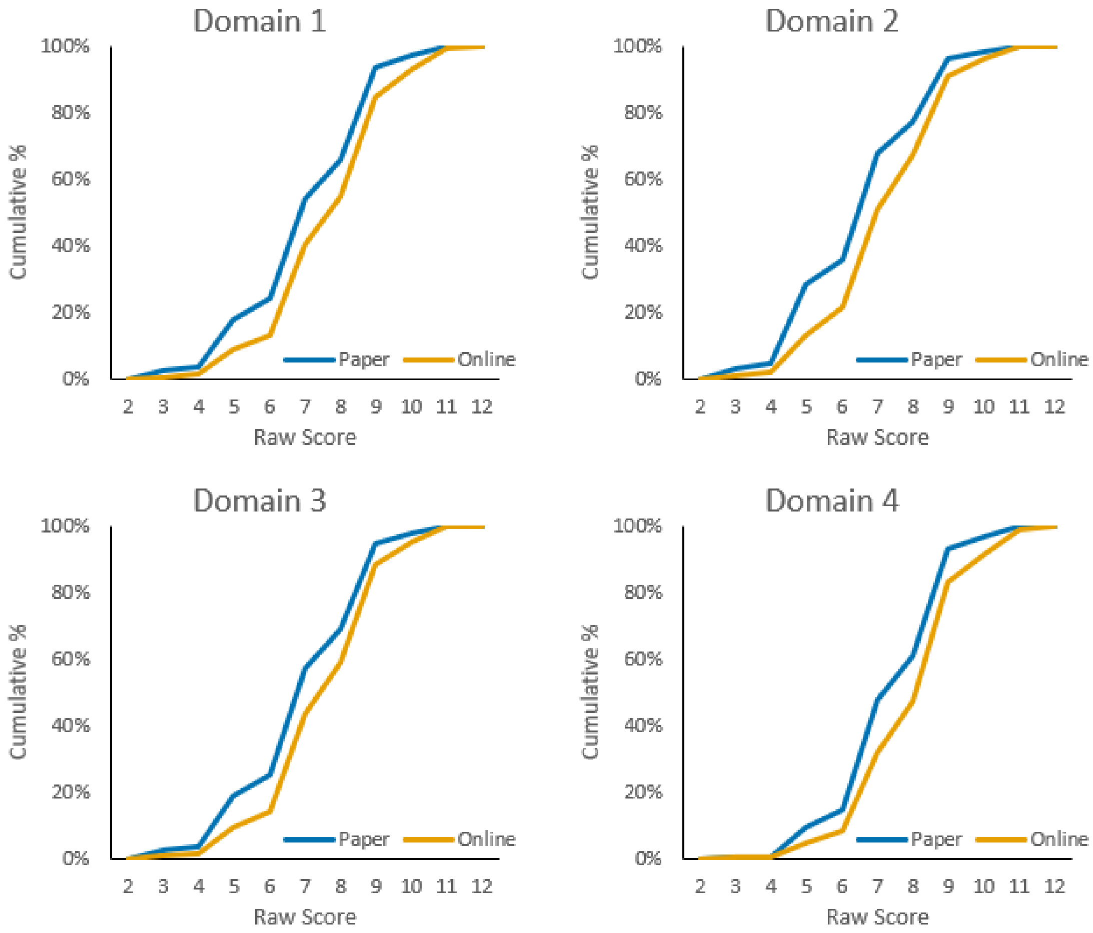
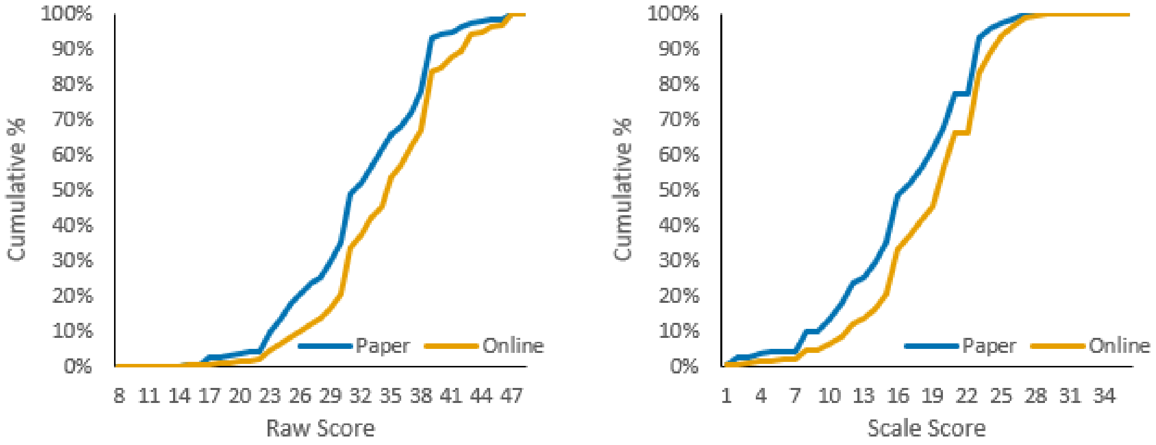


Figure 20. Relative Cumulative Frequency Distributions of Writing Raw Scores and Scale Scores



In addition, writing mean raw scores conditional on English and reading scale scores were examined across mode. Examination of raw scores permits observation of the average mode effect conditional on a related measure of ability. Note that the English and reading scale scores analyzed incorporated the mode adjustment (to make the x-axis values comparable for paper and online testing). Figure 21 shows the average writing raw scores conditional on English scale scores, and Figure 22 shows the average writing raw scores conditional on reading scale scores. On average, the writing scale scores were slightly higher for the online group than the paper group at almost all English and reading scale score levels.

Figure 21. Average Writing Scale Scores Conditional on English Scale Scores

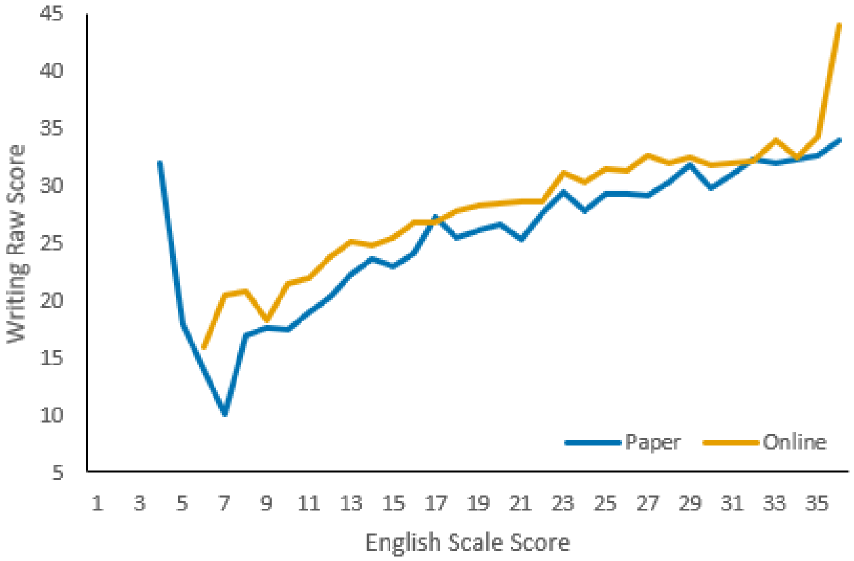
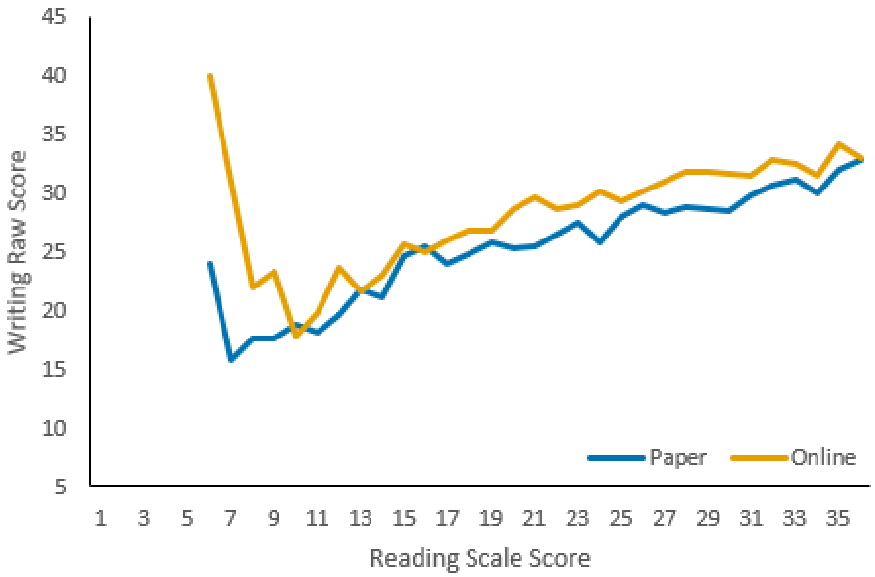


Figure 22. Average Writing Scale Scores Conditional on Reading Scale Scores



Correlations among the domain scores, raw and scaled writing scores, and raw English and reading scores across modes are presented in Table 18. There correlations were highly similar across modes, with differences never exceeding .02 in magnitude.

Table 18. Correlations of writing scores and English/reading raw scores across modes before writing mode adjustment

Mode	Score	Domain 1	Domain 2	Domain 3	Domain 4	Writing Raw	Writing Scale	English Raw	Reading Raw
Paper	Domain 1	--	.92	.98	.93	.99	.98	.52	.49
	Domain 2		--	.93	.88	.96	.96	.47	.45
	Domain 3			--	.91	.98	.97	.50	.48
	Domain 4				--	.95	.95	.55	.52
	Writing Raw					--	.99	.52	.50
	Writing Scale						--	.51	.49
	English Raw							--	.84
	Reading Raw								--
Online	Domain 1	--	.92	.98	.93	.99	.98	.52	.49
	Domain 2		--	.93	.88	.96	.96	.47	.45
	Domain 3			--	.91	.98	.97	.50	.48
	Domain 4				--	.95	.95	.55	.52
	Writing Raw					--	.99	.52	.50
	Writing Scale						--	.51	.49
	English Raw							--	.84
	Reading Raw								--

Writing Conversion Table Comparisons

As a result of equating, the online raw-to-scale score conversion table was different for the paper conversion for the writing prompt administered in the February 2020 mode comparability study. The differences accounted for mode effects between paper and online testing to make the scale scores comparable across modes. Figure 23 plots the paper and online raw-to-scale score conversions from the study. Results in the preceding sections indicated that online was relatively easy compared with paper testing (i.e., online writing scores were higher than paper writing scores), so it was expected that paper scale scores would tend to be higher than online scale scores at many raw scores. Figure 24 shows the differences between paper and online scale scores at each raw score point. Most differences were between two and three points. The largest difference (four points) occurred at the raw score 16.

Applying the mode adjustment reduced the mean online writing scale score from 19.02 to 16.90, which was very close to the paper mean of 16.93. Since the paper and online groups were randomly equivalent in ability, they should have the same average reported scores, and the mode adjustment achieved this desirable outcome.

Figure 23. ACT Writing Raw-to-Scale Score Conversions

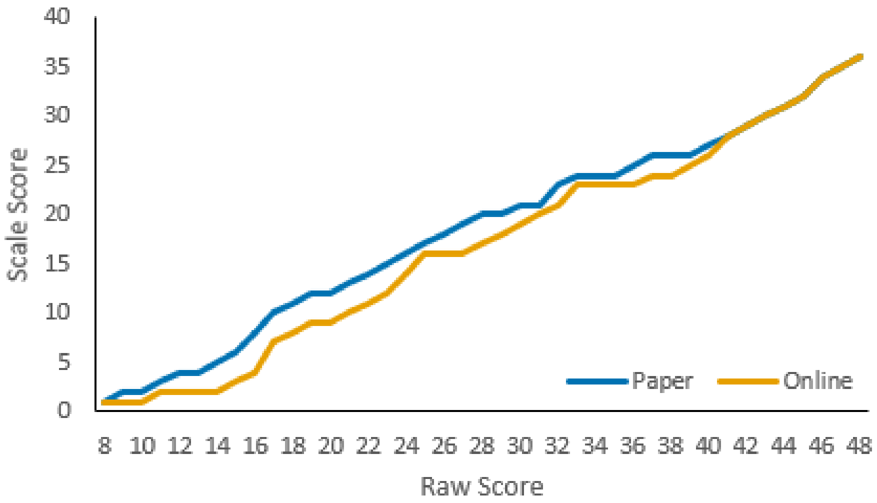
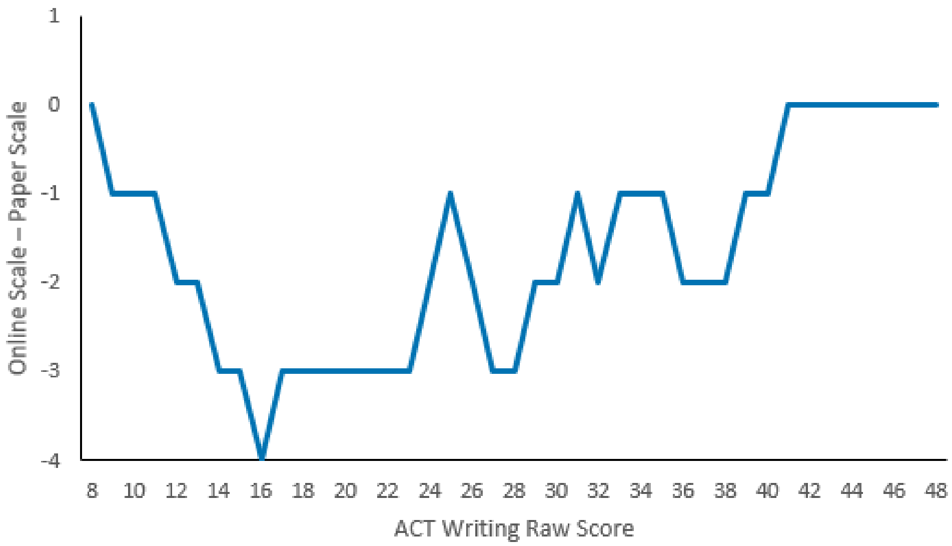


Figure 24. Difference between Writing Scale Scores Corresponding to the Same Raw Score (Online Minus Paper)



Summary and Conclusions

Table 19 provides a summary of the results from mode comparability analyses for the English, math, reading, science, and writing tests across the three studies. Item-level analyses indicated that proportions correct (p-values) tended to be higher for online testing on the English, reading, and science tests, especially near the end of the test. That is, items tended to be easier in the online testing condition compared with paper testing. Consistent with those results, item omit rates tended to be lower for online testing near the end of the English, reading, and science tests, and differential item functioning analyses detected more items favoring online testing than paper testing on the English and reading tests. Many of those items were positioned near the end of their respective tests.

Item-level differences manifested in differences between the score distributions for paper and online testing. Average performance was consistently greater on the English, reading, and science tests, with reading and English exhibiting the largest mode effects among the multiple-choice tests. The mode effect on the writing test in the February study was substantially larger in magnitude (approximately twice the reading mode effect and three times the English mode effect). Note that many of the mode effects observed in the three studies would be considered negligible or small by social science conventions (e.g., Cohen, 1988). However, even a small mode effect can impact individual student scores and interfere with aggregate score trends (e.g., for a school, district, or state). Subsequent analyses detected statistically significant differences between the distributions of paper and online scores for the English, reading, and science tests. There was not consistent evidence of a mode effect for the math test.
In terms of measurement precision (i.e., reliability and SEM), paper and online testing were comparable in all subject areas and all studies.

Across these four subject tests, several evaluations of construct equivalency provided evidence of comparability between paper and online testing. This was particularly true for the analyses of correlations, effective weights, and factor analysis models. In all cases, paper and online results were very similar. Invariance testing indicated that paper and online testing measured the same constructs (configural invariance). Analyses provided strong evidence supporting measurement invariance for the math test and moderately strong evidence for the science test. Evidence supporting measurement invariance was weaker for the English and reading tests, which might have been expected considering that those tests exhibited the largest mode effects.

Generally, results were consistent with prior ACT mode comparability studies conducted on a different online platform (Li et al., 2017). That is, the strongest evidence of mode effects was observed for the reading and writing tests, and statistical evidence consistently indicated mode effects for English. The evidence of a mode effect was weaker for science and weakest of all for math. Among mode comparability studies conducted in the last decade, results from ACT studies tend to stand out (Arthur et al., 2020). That is, most studies support comparability between paper and online testing, and when they do not, test performance tends to be better on paper. In contrast, examinees tend to answer more items correctly when taking the ACT online. This effect is possibly related to speededness on the ACT—especially on the English, reading, and writing tests—because speededness is known to moderate mode effects (Mead & Drasgow, 1993). For example, students testing online may pace themselves more effectively (with the aid of an on-screen timer) and enter answers more quickly (with mouse or touchpad clicks rather than bubbling answers with a pencil). Pommerich (2004) referred to the latter as the “no-bubble effect.” On the writing test, students testing online may have been advantaged by entering and editing their responses on a computer.

Considering results from each of the mode comparability studies, ACT determined that it was appropriate to apply mode-adjusted score conversion tables to obtain English, math, reading, and science scale scores for online participants. In the February 2020 study, the mode-adjusted writing score conversion table was also applied to obtain scale scores that contribute to the ACT English Language Arts (ELA) score. This ensured that online participants received college-reportable scores that were comparable to paper participants. Moreover, this procedure is consistent with typical ACT equating applied to different forms. In general, equating adjusts for slight differences in difficulty between forms (with different items) to ensure comparability. In the context of these studies, the paper and online tests comprised the same items, but they were treated like different test forms with different levels of difficulty. In all cases, use of the mode-adjusted conversion table produced online mean scale scores nearly identical to paper mean scale scores, which is the desired result when administering a test to randomly equivalent groups.

The decision to apply mode-adjusted score conversion tables for online testing in this study does not preclude the possibility that no mode adjustment will be applied to online scores in the future, particularly on the math test, which exhibited the weakest mode effects. ACT will continue monitoring mode effects as online testing becomes more widespread, though this will likely occur through nonexperimental methods applied to operational data (e.g., matched samples) rather than randomized controlled trials. As needed, ACT will plan and conduct additional mode comparability studies when significant changes are introduced to the ACT program.

Table 19. Summary of Results from Mode Comparability Analyses

Analysis	English	Math	Reading	Science	Writing
Item Proportion Correct	Later items were relatively more difficult for paper	On average, items were similarly difficult across the modes	Later items were relatively more difficult for paper	Later items were relatively more difficult for paper (weak effect)	
Omit Rate	Omit rate was ~2–3% higher for paper near the end of the test	No consistent trend in results	Omit rate was ~1–2% higher for paper near the end of the test	Omit rate was ~1% higher for paper near the end of the test	
Mantel-Haenszel DIF	1.3% B-, 4.9% B+, 1.3% C-, 4.0% C+	1.1% C-	5.0% B-, 7.5% B+, 2.5% C-, 0.8% C+	3.3% B-, 0.8% B+	
Effect size and t-test	$0.10 \le d \le 0.13$, all statistically significant	$-0.01 \le d \le 0.06$, 1 out of 3 statistically significant	$0.16 \le d \le 0.22$, all statistically significant	$0.04 \le d \le 0.12$, 2 out of 3 statistically significant	$d = 0.39$, statistically significant
Kolmogorov-Smirnov Test	Significant differences in distributions	1 out of 3 with significant differences in distributions	Significant differences in distributions	Significant differences in distributions	
Reliability and SEM	Coefficient alpha, SEM, and CSEM were comparable	Coefficient alpha, SEM, and CSEM were comparable	Coefficient alpha, SEM, and CSEM were comparable	Coefficient alpha, SEM, and CSEM were comparable	
Correlations and Effective Weights	Pattern of correlations among tests and effective weights were comparable	Pattern of correlations among tests and effective weights were comparable	Pattern of correlations among tests and effective weights were comparable	Pattern of correlations among tests and effective weights were comparable	
Confirmatory Factor Analysis	Model-data fit was satisfactory, factor loadings were comparable	Model-data fit was satisfactory, factor loadings were comparable	Model-data fit was satisfactory, factor loadings were comparable	Model-data fit was satisfactory, factor loadings were comparable	
Invariance Testing	Metric or partial metric invariance (weak evidence of comparability)	Residual invariance (strong evidence of comparability)	Partial metric invariance (weak evidence of comparability)	Partial metric or partial residual invariance (weak–moderate evidence)	
Raw-to-Scale Score Conversion Tables	0 to 2 point difference, always lower for online testing	0 to 1 point difference, more often lower for online testing	0 to 3 point difference, always lower for online testing	0 to 2 point difference, more often lower for online testing	0 to 4 point difference, always lower for online testing

References

ACT. (2019). *The ACT technical manual*. Iowa City, IA: ACT. http://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf

Arthur, A., Kapoor, S., & Steedle, J. (2020). *Paper and online testing mode comparability: A review of research from 2010–2020*. Iowa City, IA: ACT. https://www.act.org/content/dam/act/unsecured/documents/R1842-paper-online-testing-modes-2020-12.pdf

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 135–162). Newbury Park, CA: Sage.

Cohen, J. (1988). S*tatistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Lawrence Erlbaum Associates, Inc .

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). New York, NY: Lawrence Erlbaum Associates, Inc.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*(1), 41–54. https://doi.org/10.1007/BF02287965

Li, D., Yi, Q., & Harris, D. (2017). *Evidence for paper and online ACT® comparability: Spring 2014 and 2015 mode comparability studies*. Iowa City, IA: ACT. https://www.act.org/content/dam/act/unsecured/documents/Working-Paper-2016-02-Evidence-for-Paper-and-Online-ACT-Comparability.pdf

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, 30*, 239–270. https://doi.org/10.1007/BF02289490

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Muthén, L. K., & Muthén, B. O. (2017). *Mplus* (Version 7.4) [Computer software].

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *The Journal of Technology, Learning and Assessment, 2*(6), Article 6. https://ejournals.bc.edu/index.php/jtla/article/view/1666

Wang, L., & Steedle, J. (2020). *An investigation of differential mode effects when comparing paper-based and computer-based ACT testing*. Iowa City, IA: ACT. https://www.act.org/content/dam/act/unsecured/documents/R1838-differential-mode-effects-paper-online-2020-11.pdf

About ACT

ACT is an independent, nonprofit organization that provides assessment, research, information, and program management services in the broad areas of education and workforce development. Each year, we serve millions of people in high schools, colleges, professional associations, businesses, and government agencies, nationally and internationally. Though designed to meet a wide array of needs, all ACT programs and services have one guiding purpose—helping people achieve education and workplace success.