ACT WORKING PAPER 2016-02

Evidence for Paper and Online ACT Comparability

Spring 2014 and 2015 Mode Comparability Studies

Dongmei Li, PhD
Qing Yi, PhD
Deborah Harris, PhD

May, 2016

Dongmei Li is a principal psychometrician at ACT, specializing in test equating, scaling, and growth modeling.

Qing Yi is a principal psychometrician at ACT, specializing in computerized adaptive testing, test equating, and educational measurement theories.

Deborah J. Harris is vice president of Measurement Research at ACT, specializing in equating, linking, and comparability of reported scores.

Table of Contents

List of Tables

List of Figures

Executive Summary

In preparation for online administration of the ACT, ACT has conducted studies to examine the comparability of scores between online and paper administrations, including a timing study in fall 2013, a mode comparability study in spring 2014, and a second mode comparability study in spring 2015. This report presents major findings from these studies, focusing on the mode comparability studies.

Fall 2013 Timing Study

Standard paper administration of the ACT allows 45, 60, 35, and 35 minutes for the English, mathematics, reading, and science tests, respectively. The purpose of the timing study was to evaluate whether online administration of the ACT would require different time limits than the paper administration.

The four tests were administered online to approximately 3,000 examinees, with each examinee responding to one test. Students were randomly assigned to take the test under one of the three timing conditions: the current paper time limit, the current time limit plus five minutes, and the current time limit plus ten minutes. At the end of the test, the students were also given a survey with questions regarding their testing experience, including whether or not they felt they had enough time to finish the test.

Students' item and test level scores, item omission rates, item and test latency information, and student survey results were analyzed using a variety of methods, both descriptive and inferential. Results suggested that online scores on the reading and science tests would be more likely to be comparable to paper administration scores with an increase in testing time, given the delivery system and conditions at the time. Because of the potential confounding of motivation and familiarity with the online testing format in the timing study, a decision was made to

tentatively increase online testing time for the reading and science tests by five minutes and continue to evaluate the timing issue in the subsequent mode comparability studies.

Spring 2014 Mode Comparability Study

To gather additional information about the differences across modes and to learn about administration issues, ACT conducted a mode comparability study in an operational testing environment where participating students received college-reportable scores. Therefore, it was imperative that scores reported across modes be comparable. To ensure this was the case, a randomly equivalent groups design was implemented, allowing equating methodology to be used to adjust for score differences across modes. The purposes of the mode comparability study were to (1) investigate the comparability of the scores from the two testing modes; (2) obtain interchangeable scores across modes for operational score reporting; (3) re-evaluate the timing decisions for the online administration of the reading and science tests; and (4) gain insights into the online administration process.

Students participating in the spring 2014 study could choose to register for the ACT with or without the writing test. Within the group of students taking the ACT with writing and within the group taking it without writing, students were randomly assigned to take one of the three forms (two online and one paper) that were administered in the study. The assignment was similar to distributing spiraled paper booklets. After the administration, survey questions were sent to students who participated in the study to ask for their comments and feedback on their testing experience.

More than 7,000 students from about 80 schools across the country signed up for this study. Data were cleaned based on reviews of the proctor comments, phone logs, irregularity reports, latency information, and an examination of the random assignment. Students with invalid scores

and test centers with large discrepancies in form counts across modes were excluded from further analyses.

Analyses were conducted to investigate mode comparability at two levels: score equivalency and construct equivalency. These two levels were differentiated by some researchers (e.g. Lottridge, Nicewander, Schulz, & Mitzel, 2008), but were used here mainly for the convenience of organizing analyses. Score equivalency was examined in terms of the similarity of test score distributions between the two modes, such as means, standard deviations, and relative cumulative frequency distributions. For the English, mathematics, reading, and science tests, the similarity of item score distributions, such as the item *p*-values, item response distributions, and item omission rates were compared. In addition, measurement precision (reliability and conditional standard errors of measurement) was compared across modes, and the item latency information for the online test items was also examined. The ACT writing scores were examined conditioning on examinees' English scores. Construct equivalency was examined by comparing the dimensionality and factor loadings, and by examining differential item functioning (DIF) between online and paper scores.

Results showed that although little difference was found between the two modes in terms of test reliability, correlations among tests, effective weights, and factor structure, item scores and test scores tended to be higher and omission rates tended to be lower for the online group than for the paper group, especially for the reading and science tests. Equating methodology was used for all four multiple-choice tests to adjust for the differences to ensure that the college reportable scores of students participating in the mode comparability study were comparable to national test takers, regardless of the testing mode. Based on the findings from the spring 2014 study, a decision was made to eliminate the extra five minutes for the online reading and science tests in the spring

2015 mode comparability study. Refinements in the delivery of the online assessments may be one of the factors contributing to the different recommendations of online test time limits for the two mode comparability studies.

Spring 2015 Mode Comparability Study

The mode comparability study in spring 2015 used the same data collection design as the spring 2014 study. As stated above, the main purposes of this second mode comparability study were to further examine the comparability between online and paper scores and the impact of eliminating the extra five minutes for the reading and science online tests. More than 4,000 students from more than 40 schools signed up to participate in this study. Two online forms and one paper form were administered. Students who participated in the 2015 study all took the re-designed ACT writing test, which was to be launched in fall 2015. Since the spring 2015 study followed the same design as the 2014 study, similar analyses were conducted for the four multiple-choice tests.

Results showed that students performed similarly across modes on the science test but still higher on the online reading test even without the extra five minutes. Equating methodology was applied to produce comparable scores regardless of the testing mode. For the two prompts included in the writing mode study, students performed similarly across modes on one prompt but differentially on the other. Score distributions of randomly equivalent groups from a subsequent online administration, also conducted in spring 2015, were examined as a further validation of online form conversions obtained from the mode comparability study.

Acknowledgement

This report is a summary of the work of many. The authors are in debt to (in alphabetical order by last name) Benjamin Andrews, Zhongmin Cui, Yu Fang, Yong He, Yvette Hoffmann, JP Kim, Tianli Li, Yang Lu, Wei Tao, Tony Thompson, Lu Wang, David Woodruff, Qing Xie, and many others in the Measurement Research Department who participated in different components of the studies and/or reviewed various versions of this report. Their unique contributions and diligent work made the studies and report possible. The authors would also like to thank all other ACT departments that were involved in the studies.

Evidence for Paper and Online ACT Comparability
Spring 2014 and 2015 Mode Comparability Studies

As part of the initial development process of delivering the ACT online, ACT conducted several special studies. A timing study was conducted in fall 2013 to help inform the time limits for online administration, followed by a mode comparability study conducted in spring 2014, and a second mode comparability study in spring 2015. This report presents the designs, statistical analyses, and major findings from these studies, focusing on the mode comparability studies.

Transferring test items from paper booklets to computer for online delivery is more complicated than it might appear to be (Leeson, 2006; Mutler, 1996; Parshall, Spray, Kalohn, & Davey, 2002; Pommerich, 2004; Schroeders & Wilhelm, 2011). If score equivalence is sought between online and paper versions of a test, careful decisions need to be made not only to optimize the presentation of items, but also to minimize mode effects so that potential interference with students' performance can be eliminated to the extent possible, and the differential test taker performance between online and paper versions of the test can be negated.

To best achieve both maximum comparability to the paper version and optimal online interface and delivery, an iterative process was adopted by ACT when developing the online delivery system for the ACT. That is, to aid the online version development process, studies were conducted to evaluate the comparability of scores from online and paper delivery of the ACT under various conditions of the online delivery system/design and to inform decisions about revisions of the online version to be evaluated in further studies. In addition, due to the intended high-stakes uses of the ACT test scores, if a study involved operational score reporting, scores were adjusted for students participating in the study using equating methodology.

Fall 2013 Timing Study

Standard paper administration of the ACT allows 45, 60, 35, and 35 minutes for the English, mathematics, reading, and science tests, respectively. To inform timing decisions about the online administration, a study was undertaken in fall 2013 to evaluate the online experience, such as whether scrolling passages would require more time.

Data and Design

Online versions of the four multiple-choice tests were administered to approximately 3,000 examinees from 58 schools, with each examinee taking one of the tests. Each test was administered under three conditions: the current paper time limit, the current time limit plus five minutes, and the current time limit plus ten minutes. The tests with the different time limits were randomly assigned to students. At the end of the test, students were also given a survey with questions regarding their testing experience, including whether or not they felt they had enough time to finish the test. Depending on their testing time limit they would receive different amounts of time for the survey with a different number of questions so that all students were engaged in the task for the same amount of time. The three testing time limits and the four tests produced 12 different combinations of study conditions with about 250 examinees in each condition.

Statistical Analyses and Results

The representativeness of the schools participating in the timing study was evaluated by comparing these schools' earlier ACT test scores with other samples. Table 1 presents the means and standard deviations (SDs) of the ACT scores for three different samples: all operational data from the 2012-13 ACT testing year, 2012-13 ACT operational data from only those schools participating in the fall 2013 timing study, and all data from the 2013 ACT equating study. The fact that these schools' average performance on the ACT in the previous year was just slightly

higher than the national average and similar to the equating sample average provided some support for the representativeness of the timing study samples in terms of overall academic achievement.

Table 1

Timing Study Participating Schools Compared with National and Equating Samples on the ACT

Test	All 2012-13 Operational Data			Only Students from Timing Study Schools in 2012-13 Operational Data			All 2013 Equating Data		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
English	3,342,127	20.66	6.37	9,656	21.51	6.05	31,553	21.26	5.57
Mathematics	3,342,422	21.09	5.25	9,655	21.70	4.87	31,553	21.79	4.71
Reading	3,340,291	21.36	6.12	9,653	21.89	6.00	31,553	22.28	5.57
Science	3,338,369	20.99	5.18	9,652	21.41	4.93	31,553	21.93	4.42

Item and test level scores, item omission rates, item and test latency information, and student survey results were analyzed using a variety of methods, both descriptive and inferential. The results from a few of the timing study analyses are presented below.

Table 2 contains the percentages of students omitting zero to three or more items under the three timing conditions (i.e., current, plus 5 minutes, or plus 10 minutes). More students omitted three or more items for the reading and science tests under the current timing condition; however, with five or ten more minutes added, the percentage of students omitting three or more items was substantially reduced.

Figure 1 presents the item *p*-values, that is, the percentages of students who answered each item correctly, under each of the timing conditions for each test. The items are ordered along the horizontal axis by their position in the test. For the English and mathematics tests, the *p*-values were similar across the three timing conditions, indicating that extending testing time did not have much effect on students' performance on the test items. However, for the reading and science tests,

higher *p*-values were observed for tests with additional time, especially for items near the end of the test.

Table 2

Percentage of Students Omitting Zero to Three or More Items for Fall 2013

Test	# of Omissions	Timing		
		Current	Plus 5 Minutes	Plus 10 Minutes
English (75 items)	0	67.53	73.88	71.75
	1	13.28	13.81	16.73
	2	2.95	3.36	4.09
	3 +	16.28	8.95	7.41
Mathematics (60 items)	0	63.43	69.61	70.18
	1	13.43	14.71	13.30
	2	2.78	3.43	3.21
	3 +	20.40	12.25	13.33
Reading (40 items)	0	54.19	63.64	71.72
	1	7.88	7.39	8.08
	2	1.48	1.70	2.02
	3 +	36.46	27.27	18.24
Science (40 items)	0	61.03	75.00	82.89
	1	9.93	10.45	9.89
	2	1.84	1.49	2.28
	3 +	27.25	13.04	4.94

Figure 2 presents the percentage of omissions for each item of the four tests. Again, the items are ordered by their position in the test, and the percentage of examinees not responding to the item is given on the vertical axis. The graphs show that the percentage of examinees omitting items near the end of the tests was much higher for the reading and science tests than those for the English and mathematics tests. This was especially true for the reading test, where the omission rate reached 40% for the last item under the current paper timing limit.

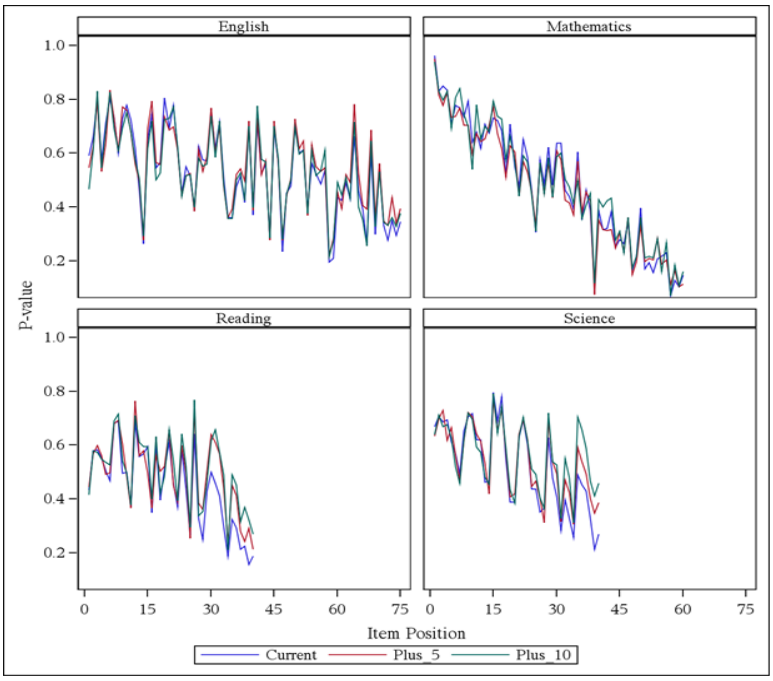


Figure 1. Item *p*-values by item position for the three timing conditions for fall 2013.

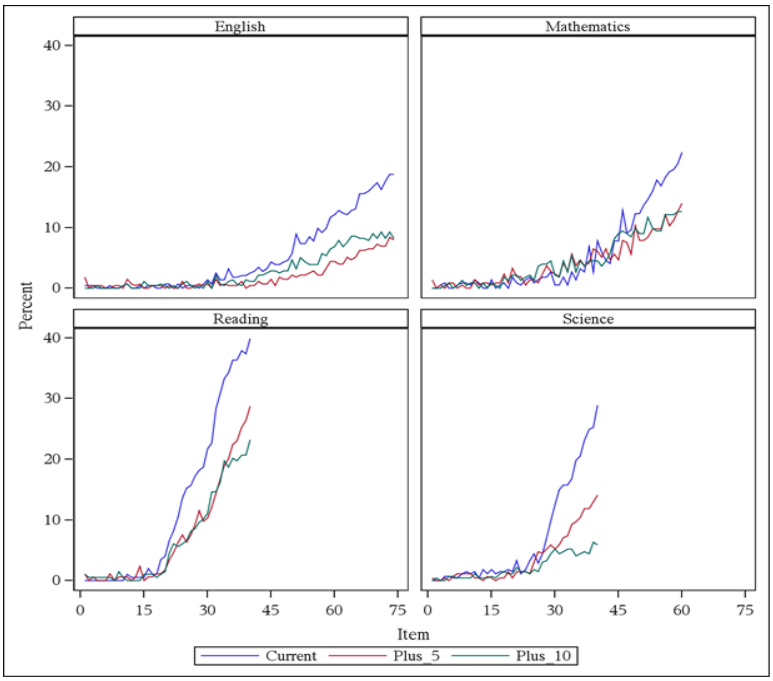


Figure 2. Percentage of students omitting items under the three timing conditions for fall 2013.

Figure 3 shows the percentages of responses to the survey question regarding their level of agreement with the statement that they had enough time to finish the test for each timing condition of each test. The response rates are shown at the top of each plot (e.g., response rate RR=82.46% for the English test with current time limit). The "Other" category in the pie charts included the percentage of students who strongly disagreed or disagreed with the statement. In general, students who took the reading and science tests had larger percentages of disagreements on the statement that they had enough time to finish the test. As testing time increased, those disagreement percentages were reduced. However, they were still higher than those for the English and mathematics tests under the same timing condition.

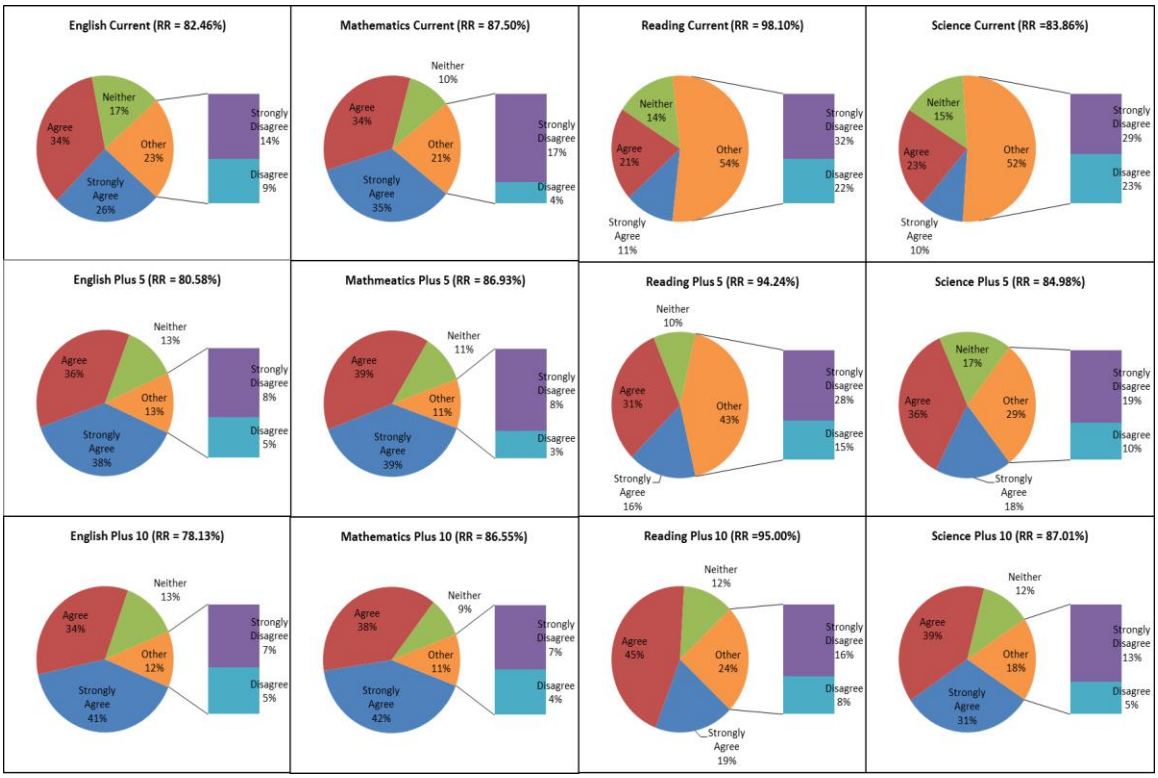


Figure 3. Student responses to survey question about if they had enough time under the three timing conditions for fall 2013.

Online Timing Recommendations and Concerns

Though results from the fall 2013 timing study suggested that online administration might require more time for students to complete the reading and science tests, acting upon these results to establish timing recommendations was confounded with issues such as motivation and familiarity with the online testing format. For example, while the reading and science tests showed speededness in these analyses, it was also true that fewer examinees taking those two tests reported watching the orientation videos about how the online testing worked based on the responses to a survey question (See Table 3).

Table 3

Survey Results for the Question Regarding Online Tutorial Video for Fall 2013

Before taking the test, did you watch the online video about learning to use the online testing system?				
%	English	Mathematics	Reading	Science
Yes	41	39	14	15
No	59	61	86	85

The final decision was to tentatively increase testing time for the reading and science online tests by five minutes and to revisit the time limit issue in the mode comparability studies. Because equating methodology had been planned in case of evidence suggesting mode differences, comparable scores for the examinees in the mode comparability studies could be ensured regardless of possible changes in administration time in the future.

Mode Comparability Studies

Two mode comparability studies for the ACT were conducted in spring 2014 and spring 2015. As in the fall 2013 timing study, the content of the test items for the online version and the

paper version of a test form was intended to be exactly the same. However, there were some differences (for example, to assist students in marking the correct row of an answer document on the paper administration, some item choices run A, B, C, D and some run F, G, H, J. Online, these item choices may all run A, B, C, D). In addition, some improvements had been made to the online test delivery system based on experiences and feedback from the fall 2013 timing study and the spring 2014 mode comparability study. In the spring 2014 study, the testing time for the online and paper administrations was the same for the English and mathematics tests, but it was different for the reading and science tests. Five additional minutes were added to the online versions of the reading and science tests based on the recommendation from the fall 2013 timing study. However, for the spring 2015 study, the testing time for both online and paper administration was kept the same based on the findings from the spring 2014 mode comparability study.

The purposes of the mode comparability studies were to (1) investigate the comparability of the ACT scores from the online and paper testing modes; (2) obtain interchangeable scores across modes for operational score reporting; (3) re-evaluate the timing decisions for the online administration of the ACT; and (4) gain additional insights about the online administration process.

Design

A randomly equivalent groups design was used for the ACT mode comparability studies. Students were randomly assigned to take one of the three ACT forms that were administered in each study—two online forms (Online_1 and Online_2) and the paper version of one of the two online forms (Paper_1). One purpose of having an additional online form was to help evaluate the extent of the mode effect relative to form differences.

These studies took place in operational testing environments on one of the ACT national test dates. Schools with sufficient numbers of computers that met ACT requirements for online

testing were recruited to participate in the studies. Participating schools were also required to meet all the other requirements of ACT test centers. Online testing occurred on school-provided desktop or laptop computers (Windows/Macintosh). Tablets did not meet the requirements for the mode comparability studies.

Students from participating schools registered for testing via the normal ACT registration process and were randomly assigned to take the online or paper version of the ACT test. Since students did not know which mode they were going to be assigned until the day of testing, a student tutorial video and practice test intended to help students navigate the online testing system was made available to all participating students, in addition to the standard paper practice test. After testing, survey questions were sent to students who participated in the study to ask for their comments and feedback on their testing experiences.

Procedure

For the multiple-choice tests, though the content of items on the online versions was intended to be exactly the same as the paper ones, there could still be differences in the appearance of items between the two modes. These differences could include text font, page size, page layout, graphics, and others. Before conducting data analyses, ACT first examined comparability of the online and paper versions of the tests through a qualitative comparison of the items in the paper booklets and those in the online version. All substantial differences were documented.

Mode comparability was examined at two levels for the multiple-choice tests: score equivalency and construct equivalency. Score equivalency indicates that observed score distributions from the two modes are very similar for the two randomly equivalent groups. Construct equivalency indicates that the two modes are measuring the same underlying abilities or attributes. If score equivalency holds, construct equivalency is partly supported, especially with

the fact that items are the same on the two modes. However, score equivalency cannot guarantee construct equivalency, so additional comparisons to determine whether the two modes measure the same construct are still needed.

Analyses to evaluate mode comparability were carried out in two phases for the multiple-choice tests. Phase I analyses focused more on score equivalency, examining the similarity of test score distributions between the two modes, such as means, standard deviations, and relative cumulative frequency distributions. The similarity of item score distributions, such as the item *p*-values, item response distributions, and item omission rates, were also compared.

Equating methodology was used to ensure that the college reportable scores for students participating in the studies across modes were comparable. Timing decisions were also re-evaluated based on the new evidence gathered from the previous timing study (i.e., spring 2014 based on fall 2013 findings and spring 2015 based on spring 2014 results).

Phase II mode comparability analyses focused more on construct equivalency as well as some additional analyses, including item and test comparisons based on the item response theory (IRT), factor analysis, differential item functioning (DIF), generalizability analysis, and evaluation of measurement precision after any mode effect was adjusted through equating methodology.

The following sections present the specifics and results from the two mode comparability studies. Results from the Phase I and Phase II analyses are shown first, after which results for the ACT writing test are discussed. Presented at the end are some results of the online forms obtained from an additional online administration, which occurred shortly after the spring 2015 mode comparability study.

Spring 2014 Comparability Study

Students could register to take either the ACT with writing or the ACT without writing for the 2014 study. Random assignment of students to the online and paper forms was done separately for students taking the ACT with writing and those without writing so the groups taking the writing test would also be randomly equivalent.

Data. More than 7,000 students from about 80 schools across the country signed up for the spring 2014 mode comparability study. As expected, not everyone who signed up actually showed up on the day of testing. Computer issues, power outages, and other problems also prevented some students from testing on the scheduled dates. Those students were rescheduled to take the test on paper on another date, and their scores were not included in the analyses. Data were also cleaned based on reviews of proctor comments, phone logs, irregularity reports, and latency information. Centers with large discrepancies in form counts were deleted from the analyses.

All subsequent analyses were based on the final cleaned data. More than 5,500 students with at least 1,800 students for each form were included in the spring 2014 cleaned data. Among these students, over 2,000 students took the writing test, either paper or online.

Phase I mode comparability analyses and results for multiple-choice tests. Phase I comparability analyses for the multiple-choice tests included an examination of the test and item level score distributions, test score reliabilities, and item omission rates across modes. Table 4 presents the samples size of each test form in the spring 2014 study as well as the means, standard deviations (SD), minimums, and maximums of the observed total raw scores and scale scores for all three forms. Online Form 1 and paper Form 1 contained the same items, and were the focus of most analyses. They are simply referred to as the online and paper form when online Form 2 is not involved in the comparison. ***Note that the scale scores mentioned in the Phase I analyses refer***

to scale scores obtained by applying the paper raw to scale score conversions regardless of testing mode. The purpose of doing so was to examine the mode effect on the scale scores. For example, in Table 4, the scale score descriptive statistics for Online_1 and Online_2 were obtained by applying the paper version conversions of Form 1 and Form 2, respectively. However, final reported scale scores for the online forms were based on the conversions obtained through using equating methodology that is discussed later. For the spring 2014 data, on average the online Form 1 scores tended to be higher than the paper Form 1 scores for all tests. Though online Form 2 had higher raw score means than those of online Form 1 (except reading), their scale score means were similar.

Table 4

Descriptive Statistics of Raw and Scale Scores of all Test Forms for Spring 2014

Form	Test	Raw Score					Scale Score			
		N	Mean	SD	Min	Max	Mean	SD	Min	Max
Online_1	English	1801	45.04	13.94	7	75	21.39	5.95	5	36
	Mathematics	1801	31.38	11.63	7	60	21.30	5.26	11	36
	Reading	1801	25.17	7.65	2	40	23.56	6.43	4	36
	Science	1801	22.57	7.46	4	40	22.12	5.23	8	36
Paper_1	English	1987	42.87	14.50	10	75	20.47	6.12	7	36
	Mathematics	1987	30.80	11.49	7	60	21.02	5.15	11	36
	Reading	1987	22.62	7.89	3	40	21.47	6.43	5	36
	Science	1987	21.14	7.26	2	40	21.14	5.03	5	36
Online_2	English	1805	49.57	14.27	6	75	21.01	5.95	4	36
	Mathematics	1805	34.04	12.40	5	60	21.46	5.06	10	36
	Reading	1805	24.32	7.66	4	40	23.36	6.26	7	36
	Science	1805	22.86	7.83	4	40	21.95	5.39	8	36

Raw and scale score mean differences, effect sizes, and t-tests of mean differences. Scores across modes can be compared either on raw scores or scale scores, by applying the paper version conversions to both the paper and online forms. Since there are more raw score points than

scale score points for the ACT, comparability at the raw score level is a more stringent requirement than comparability at the scale score level. However, only differences at the scale score level could have any practical impact because decisions are usually made based on scale scores.

Figure 4 shows graphical presentations of the raw score and scale score mean differences across modes. Mean differences, effect sizes, and *p*-values from t-tests of mean differences for raw and scale scores are presented in Table 5. The effect sizes were calculated by dividing the mean differences by the pooled standard deviations across modes for each test. As shown in Table 5, in spring 2014, for all tests the online group tended to have higher mean scores than the paper group. The mean differences were all statistically significant, for both raw scores and scale scores, except for the mathematics test.

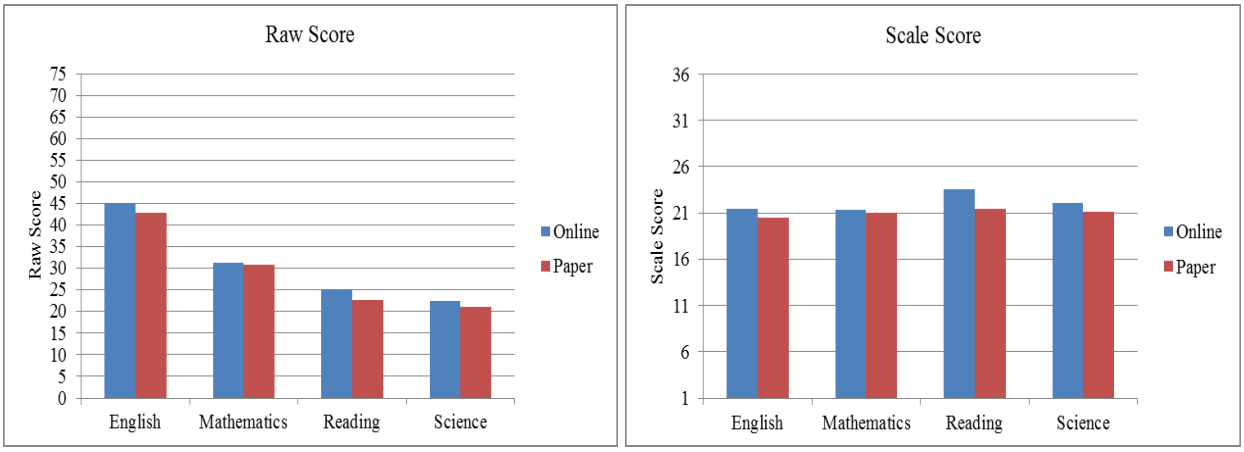


Figure 4. Raw and scale score mean comparisons across modes for spring 2014.

Table 5

Raw and Scale Score Mean Differences across Modes (Online minus Paper) for Spring 2014

	Raw Score Comparison			Scale Score Comparison		
Test	Mean Difference	Effect Size	t-test *p*	Mean Difference	Effect Size	t-test *p*
English	2.17	0.15	<.0001	0.93	0.15	<.0001
Mathematics	0.58	0.05	0.1204	0.28	0.05	0.0942
Reading	2.56	0.33	<.0001	2.09	0.32	<.0001
Science	1.43	0.19	<.0001	0.98	0.19	<.0001

Raw and scale score cumulative frequency distributions and Kolmogorov-Smirnov test of equivalency of distributions. Raw score and scale score frequency distributions and cumulative frequency distributions were also compared across modes. The plots of the relative cumulative frequency distributions of proportion correct raw scores and scale scores are shown in Figure 5. For spring 2014, scores tended to be higher for the online group than for the paper group for all tests except the mathematics test.

The Kolmogorov-Smirnov (KS) test of equivalency of distributions was conducted for the raw and scale scores for each test. Similar with the results of the t-tests of mean differences, for the spring 2014 study, the KS tests showed that the across-mode raw and scale score distributions were statistically significant for all tests except for the mathematics test.

Correlations, effective weights, and Cronbach's alpha. Correlations among tests and effective weights of each test were also calculated to examine whether relationships between tests were consistent across modes. Measurement precision of scores from the two modes was examined by calculating Cronbach's alpha. Reported in Table 6 are the scale score correlations, effective weights, and Cronbach's alpha. These values were all very similar across modes.

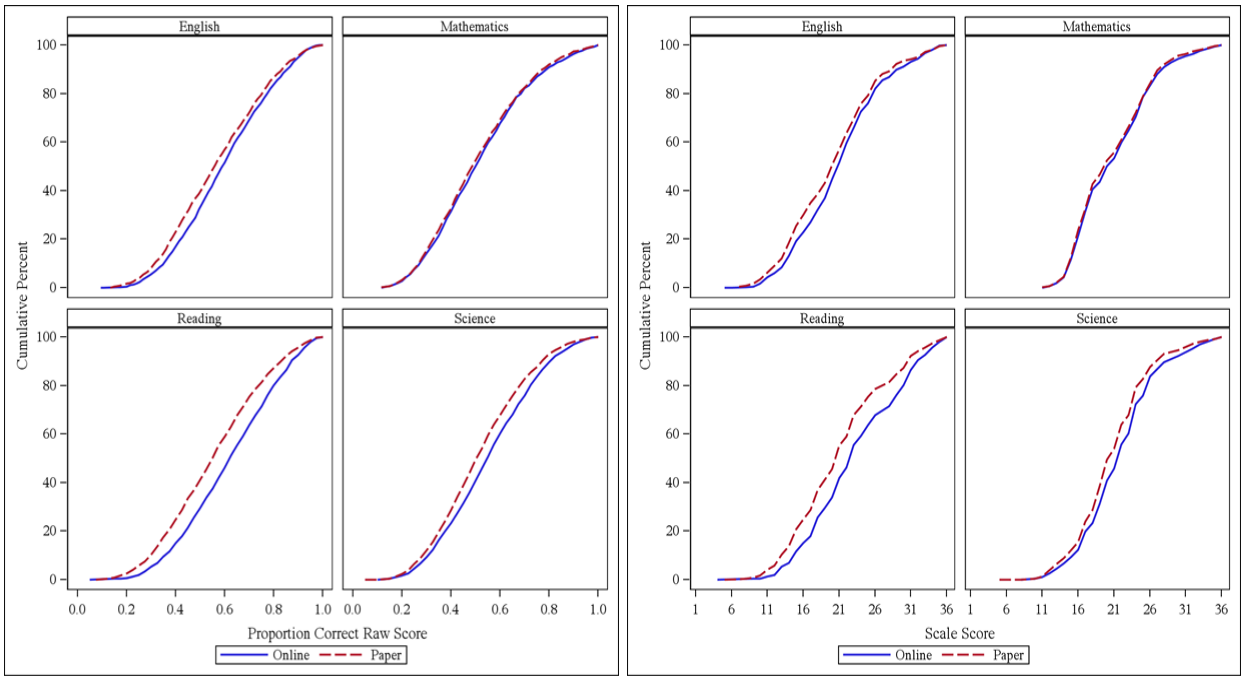


Figure 5. Relative cumulative frequency distributions of proportion correct raw scores and scale scores for spring 2014.

Table 6

Scale Score Correlations, Effective Weights, and Cronbach's Alpha for Spring 2014

		Online				Paper			
		English	Mathematics	Reading	Science	English	Mathematics	Reading	Science
Correlation	English	1.00	.74	.82	.77	1.00	.75	.81	.76
	Mathematics	.74	1.00	.69	.80	.75	1.00	.67	.79
	Reading	.82	.69	1.00	.76	.81	.67	1.00	.74
	Science	.77	.80	.76	1.00	.76	.79	.74	1.00
Effective Weight		.26	.22	.28	.23	.28	.22	.28	.22
Cronbach's Alpha		.93	.92	.87	.86	.93	.92	.87	.86

P-values, omission rates, and option analyses. Item difficulties (*p*-values) were compared across modes. The graphs on the left side of Figure 6 present the proportion of correct responses for each item (item *p*-values) by item position across modes, with smaller values indicating harder items. The graphs on the right side show the *p*-value differences across modes, with positive

differences indicating that the item was easier for the online administration. Figure 6 shows that while later items tended to be harder compared to earlier items for each test regardless of mode, the items tended to be easier for the online administration, especially for items that appeared later in the test.

Consistent with the effect size differences observed in Table 5, the *p*-value differences were the smallest for the mathematics test, and the largest for the reading test in spring 2014. For the mathematics test, the item *p*-value differences were mostly within the range of -0.05 to 0.05, and the direction of the differences seemed to vary randomly. For English and science, items in the latter part of the test were consistently easier for the online administration, and for reading, almost all items were easier for the online administration.

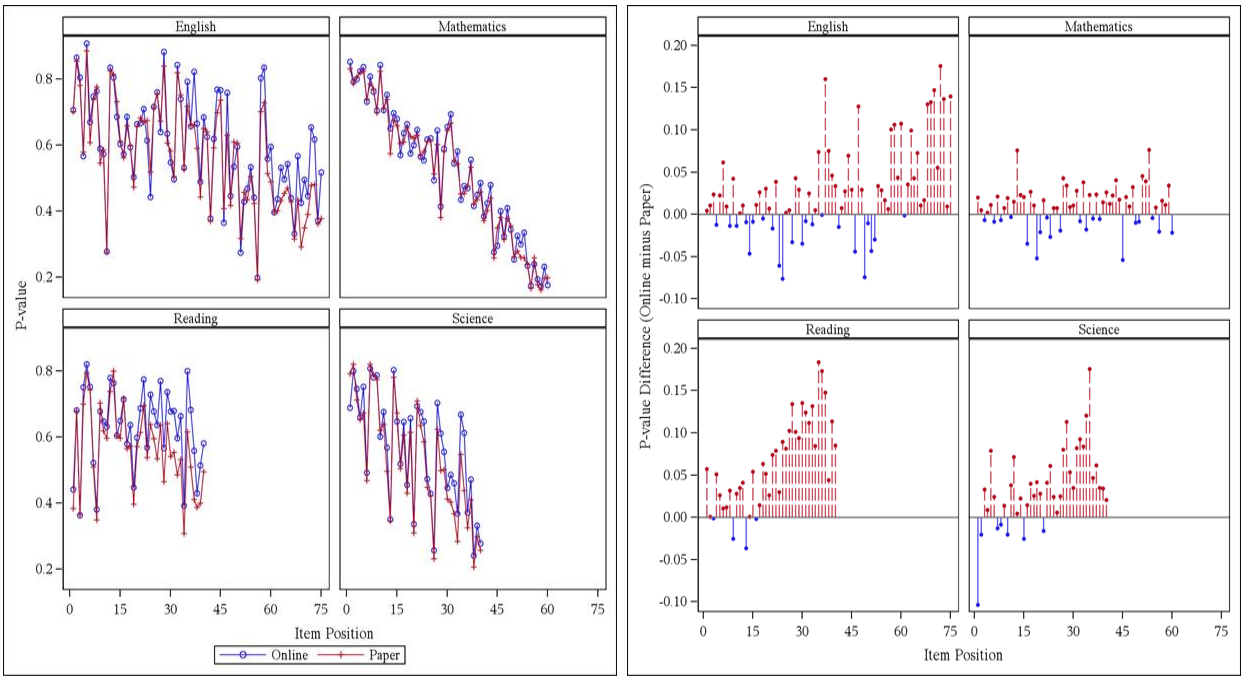


Figure 6. Scatter plots of item *p*-values and needle plots of *p*-value differences across modes for spring 2014.

The omission rate (i.e., the proportion of missing responses) for each item was also compared across modes. As shown in Figure 7, across all four tests, the paper group consistently had a higher omission rate than the online group for the latter half of the tests, except the last few items of the mathematics test. In addition, the proportion of examinees choosing the incorrect options was also examined for each item across mode, but no obvious patterns were found.

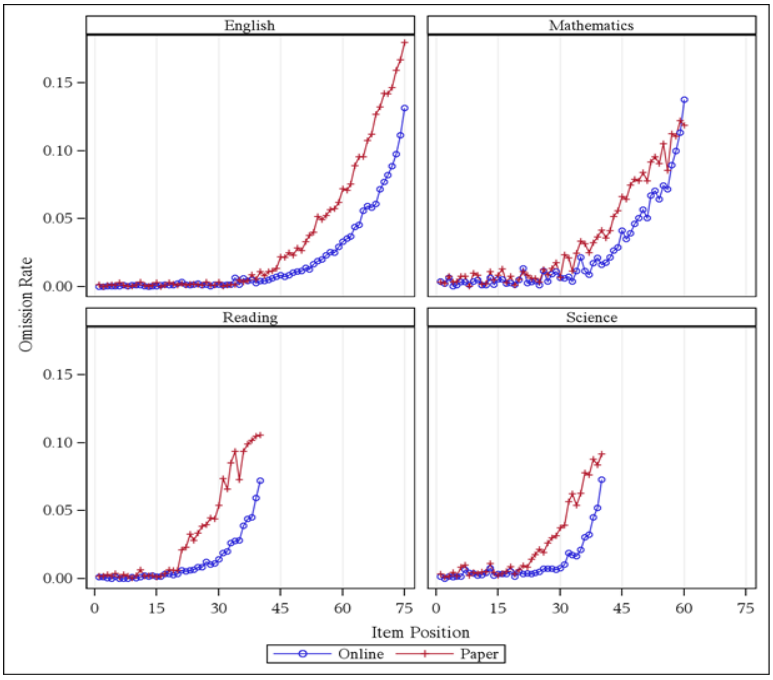


Figure 7. Item omission rates by item position for spring 2014.

Adjustments to score differences. Due to the differences observed between the online and paper scores as discussed above, equating methodology was used for the multiple-choice tests so that the college reportable scores are comparable regardless of conditions (i.e., mode and time limit) under which students took the tests. Consistent with the methodology used for equating paper forms of the ACT, the equipercentile method with post smoothing was used to "equate" the online test scores to the paper form scores, using the randomly equivalent groups design.

Equating methodology adjusted for the potential mode effect for each test and created raw to scale score conversion tables for the online forms that were different from the corresponding paper conversions. These conversions are referred to as online conversions or adjusted conversions to differentiate them from the paper conversions. The left side of the graphs in Figure 8 shows the raw to scale score conversions for the two online forms together with their counterpart paper conversions, and the right side of the graphs displays the differences between the online and paper conversions at each raw score point for the two online forms, with negative values indicating that the same raw score was converted to a lower scale score in the online conversion than in the paper conversion. For spring 2014, except for a few raw score points for Form 2 English and mathematics, the online conversions after adjusting for mode effect resulted in equal or lower scale scores than the paper conversions.

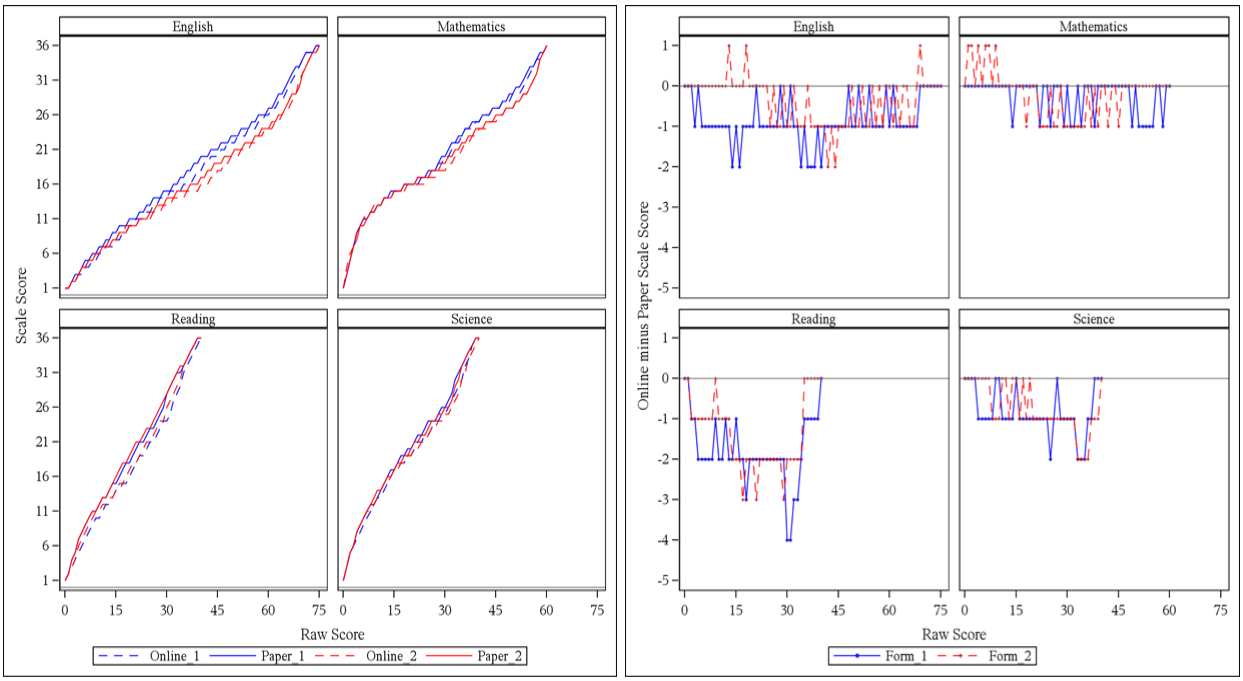


Figure 8. Raw to scale score conversions and scale score differences for spring 2014.

Two sets of scale scores were calculated for the students who took the online tests by applying both the online conversions and the paper conversions. The differences between these scale scores (online minus paper scale score conversion) were also calculated. Figure 9 presents the distributions of these difference scores. For spring 2014, the majority (around 50% to nearly 100%) of the differences were zero or one score point for English, mathematics, and science. For reading, however, more than half of the difference scores were two scale score points or more.

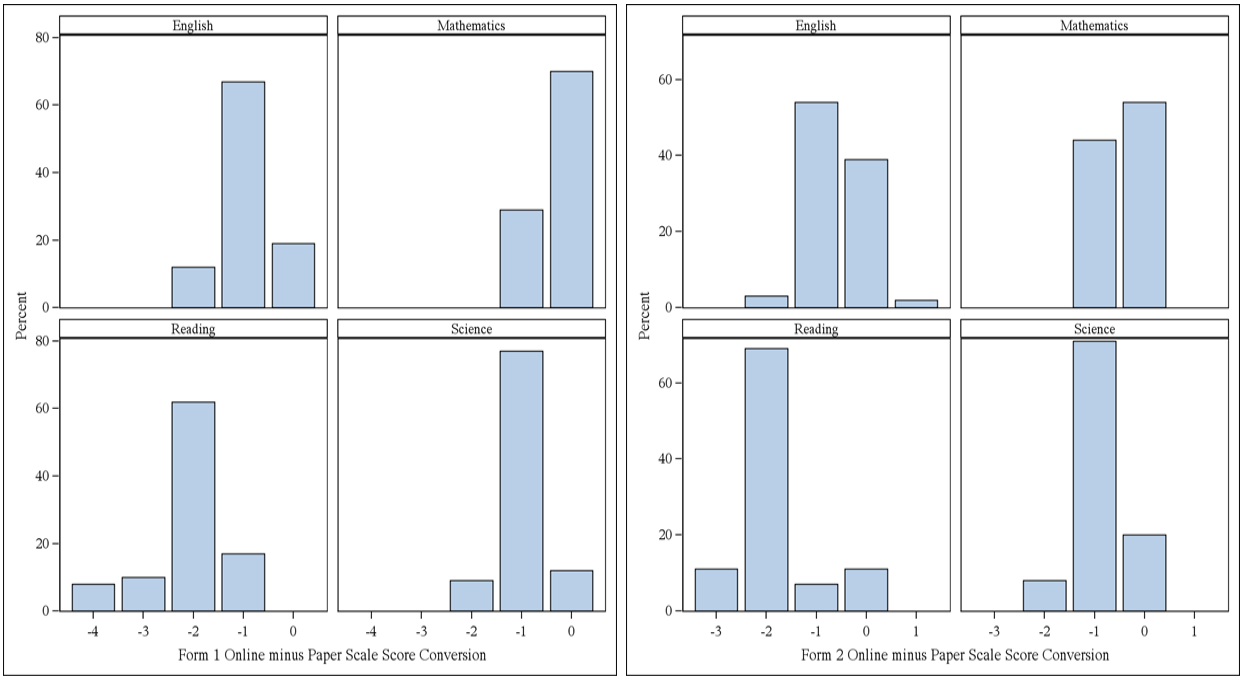


Figure 9. Distributions of scale score adjustment for the two online forms for spring 2014.

Online timing re-evaluation. As mentioned earlier, in the spring 2014 mode comparability study the online administration added five minutes to the current paper administration time for the reading and science tests based on the recommendations from the fall 2013 timing study. However, the limitations of that timing study made it necessary to continue to gather information to inform the timing decisions. Since the mode comparability studies were conducted in an operational testing environment with a paper control group, the studies provided

information for timing decisions that were less confounded than that of the fall 2013 timing study. Results from analyses presented in previous sections were considered together with the following additional information from the student survey and online item latency information to inform the online timing decisions in spring 2014.

Survey results on timing-related questions. In the student survey, students were asked whether they felt they had enough time to finish each of the tests. About 1,500 students, approximately two thirds of which took the online versions of the tests, completed the survey in spring 2014. Table 7 presents the spring 2014 survey results related to this question. Except for writing, a higher percentage of students either agreed or strongly agreed that they had enough time to finish the test for the online administration compared to the paper administration.

Table 7

Percentage of Responses to the Timing Related Question for Spring 2014

I felt I had enough time to finish the…test	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Either agree or strongly agree	Either disagree or strongly disagree
Online							
English	40	39	7	9	4	79	12
Mathematics	21	32	13	24	9	53	33
Reading	21	34	12	21	10	55	31
Science	15	32	17	23	11	47	34
Writing	19	28	16	18	14	48	32
Paper							
English	27	36	11	18	7	62	25
Mathematics	14	34	12	25	14	49	38
Reading	11	26	12	30	18	37	48
Science	9	29	15	29	17	38	45
Writing	29	42	11	9	7	71	17

Online form item response time. Item latency information was examined for the two online forms. Figure 10 presents the average time spent on each item for all four multiple-choice tests. If

the time spent on the last few items of each test was significantly less than on the other items, the test may be speeded. However, no such evidence was found for the online tests. Note that the peaks showed in the graphs are usually the first item associated with a passage, which included the time spent reading the passage.

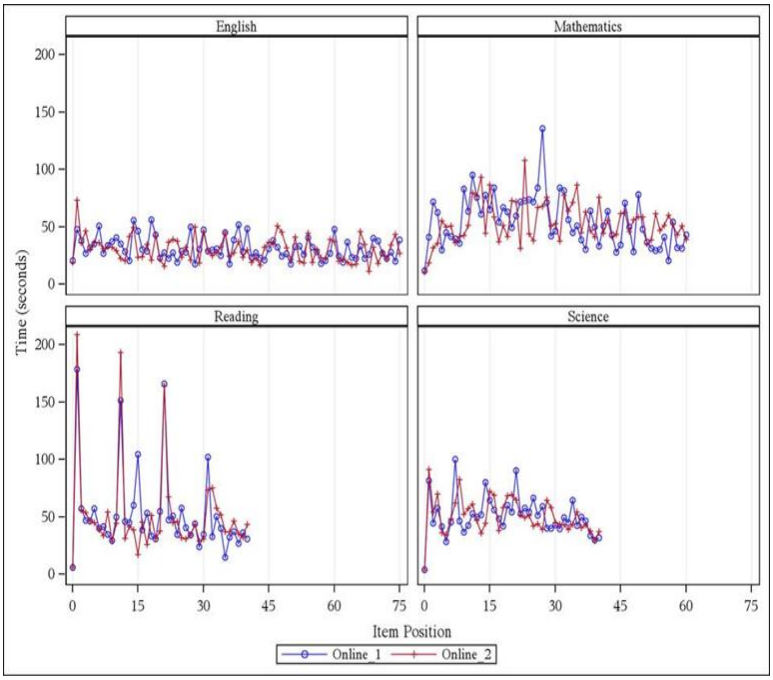


Figure 10. Average time spent on each item for spring 2014.

Online timing decision. Based on the results from the spring 2014 mode comparability analyses, student survey, and item latency information, it was decided that the extra five minutes for the online administration of the reading and science tests should be removed, resulting in the same testing time regardless of whether the test is administered online or on paper.

Phase II mode comparability analyses and results for multiple-choice tests

IRT analysis. Mode effects were examined under item response theory (IRT) at both the test and item level by comparing the test characteristics curves (TCCs) and item parameters across

modes, using the three-parameter logistic IRT model. Figure 11 contains plots of the TCCs across modes for each subject. Consistent with the patterns observed in Figure 5 for the raw and scale score relative cumulative frequency distributions, the across-mode TCC difference is the smallest for the mathematics test, but largest for reading in the spring 2014 study.

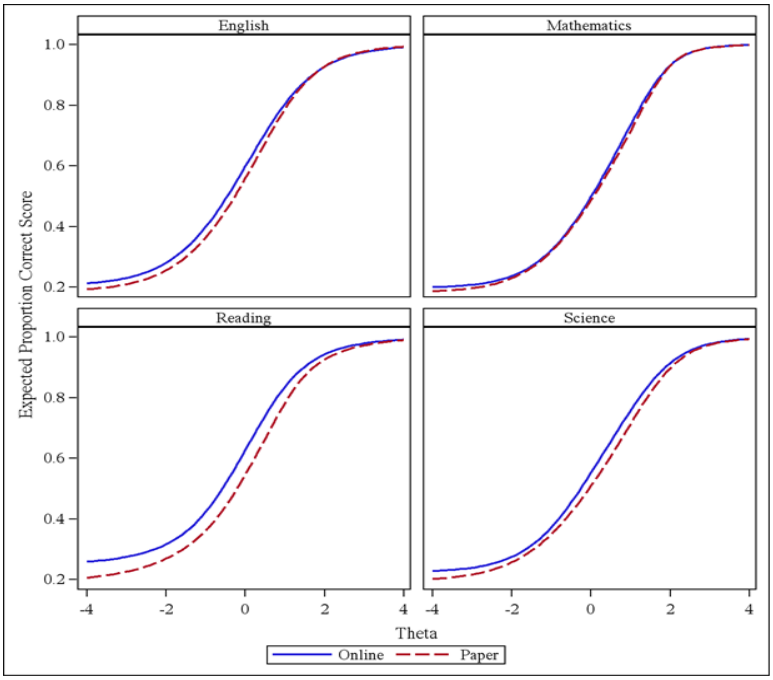
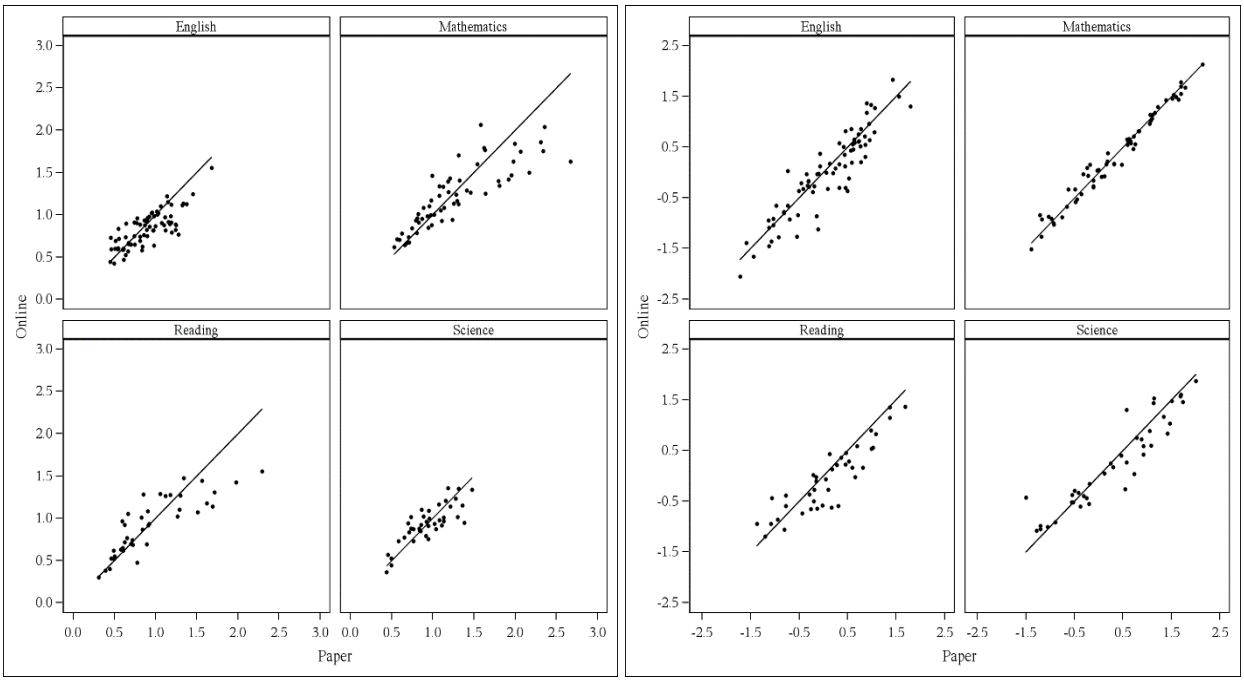
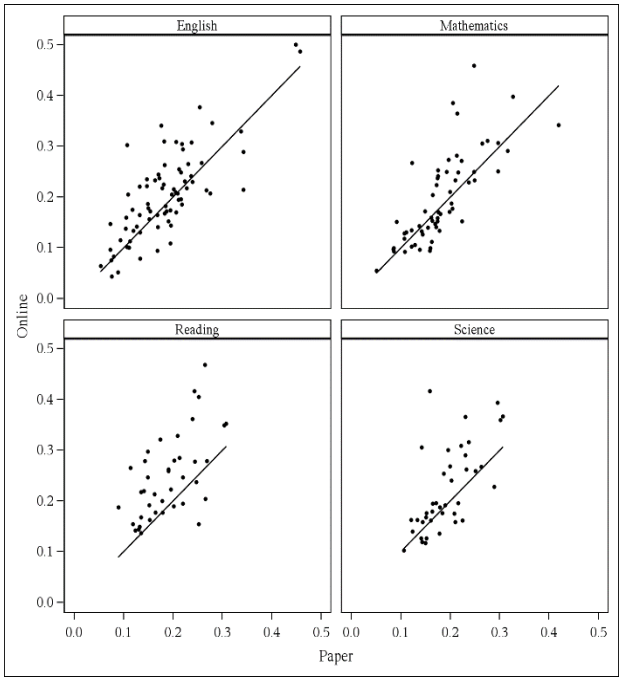


Figure 11. Test characteristic curves across modes for spring 2014.

Scatter plots of item parameter estimates from online and paper are presented in Figure 12. Consistent with the comparison of item *p*-values for spring 2014, the *b*-parameter comparison showed that the online items tended to be easier than the paper items, especially for the reading and science tests. In addition, the *c*-parameters tended to be higher for the online items, which indicated that low-performing students had a higher chance of answering the online items correctly than the paper items.



a-parameter *b*-parameter



c-parameter

Figure 12. IRT parameter comparison across modes for spring 2014.

Factor analysis. Exploratory factor analysis was conducted to explore the dimensionality and construct equivalency of the online and paper tests. Eigenvalue scree plots for each test were examined across modes, as shown in Figure 13.

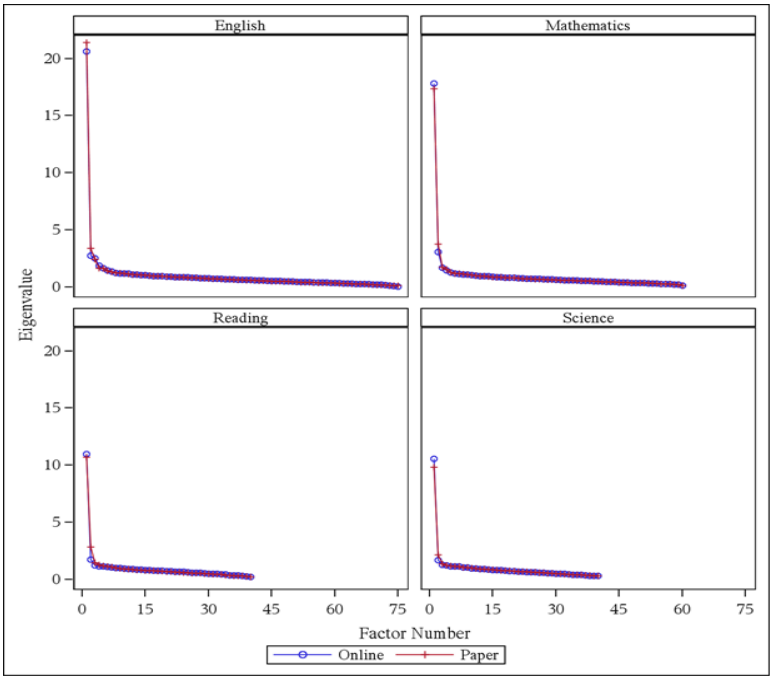


Figure 13. Eigenvalue scree plot for spring 2014.

The data were fit with both a one-factor and a two-factor model. Table 8 presents the criteria used for evaluating model fit, and Table 9 contains several fit indices resulting from fitting the one- and two-factor models for the four multiple-choice tests across modes. Table 9 also includes the fit statistic differences (DIFF) in fitting two- or one-factor models. The bolded numbers in Table 9 are values that did not meet the criteria presented in Table 8. As shown in Table 9, all statistics for the spring 2014 data indicated good model fit for the one-factor model except for a couple of statistics for the online reading test. Compared with the one-factor model, the use of the two-factor model did not seem to improve the model fit substantially except for the online reading test. However, based on the principle of parsimony, the one-factor model was

considered to be adequate and the factor loadings of each test for the one-factor model were compared across modes. Table 10 presents the descriptive statistics of the factor loadings of each mode and the correlations of the factor loadings across the two modes.

Table 8

Criteria for Good Model Fit

Fit Statistic	Value
CFI	>=0.95
TLI	>=0.95
RMSEA	<=0.06
SRMR	<=0.08

Table 9

Fit Statistics of One- and Two-Factor Models for Spring 2014

Test	Fit Statistic	Online			Paper		
		One Factor	Two Factors	DIFF	One Factor	Two Factors	DIFF
English	CFI	0.97	0.98	0.01	0.96	0.98	0.02
	TLI	0.97	0.98	0.01	0.96	0.98	0.02
	RMSEA	0.03	0.03	0.01	0.04	0.03	0.01
	SRMR	0.05	0.05	0.01	0.06	0.05	0.01
Mathematics	CFI	0.97	0.99	0.02	0.95	0.99	0.03
	TLI	0.97	0.99	0.02	0.95	0.99	0.03
	RMSEA	0.03	0.02	0.01	0.04	0.02	0.02
	SRMR	0.06	0.04	0.02	0.06	0.04	0.02
Reading	CFI	0.98	0.99	0.02	**0.93**	0.98	0.05
	TLI	0.98	0.99	0.02	**0.92**	0.98	0.06
	RMSEA	0.03	0.01	0.01	0.05	0.03	0.02
	SRMR	0.04	0.03	0.01	0.07	0.04	0.03
Science	CFI	0.98	0.99	0.01	0.96	0.98	0.03
	TLI	0.98	0.99	0.01	0.95	0.98	0.03
	RMSEA	0.02	0.02	0.01	0.03	0.02	0.01
	SRMR	0.05	0.04	0.01	0.05	0.04	0.01

Table 10

Descriptive Statistics and Correlation of Factor Loadings across Modes for Spring 2014

Test	Mode	Mean	SD	Minimum	Maximum	Correlation
English	Online	0.51	0.11	0.25	0.70	.88
	Paper	0.52	0.11	0.26	0.75	
Mathematics	Online	0.53	0.10	0.25	0.72	.90
	Paper	0.52	0.11	0.26	0.70	
Reading	Online	0.50	0.12	0.22	0.73	.87
	Paper	0.49	0.11	0.26	0.76	
Science	Online	0.48	0.13	0.23	0.71	.87
	Paper	0.47	0.11	0.26	0.69	

Generalizability analysis. Raw score reliability was further examined based on the results from a multivariate generalizability analysis under a person-crossed-with-item design, treating the different content categories as different variables. The generalizability coefficients ($E\rho^2$) and dependability indices or phi coefficients (Φ) are reported in Table 11, together with the Cronbach's alpha reliability already reported in Table 6 to facilitate comparison. The Φ coefficients were slightly lower than the generalizability coefficients, and these two coefficients were both very close to the alpha estimates. Similar with alpha, reliability indices from the generalizability analyses showed barely any differences across modes.

In addition, correlations between the content areas, the variance components of each facet, and the contribution of each content category to the total variance from the generalizability analysis results were also compared across modes. No noteworthy differences were found except that the correlations among the online tests tended to be higher than those of the paper versions for reading.

Table 11

Raw Score Generalizability Coefficient, Phi Coefficient, and Alpha for Spring 2014

	Online				Paper			
	English	Mathematics	Reading	Science	English	Mathematics	Reading	Science
$E\rho^2$	0.93	0.92	0.88	0.86	0.93	0.92	0.88	0.86
Φ	0.92	0.91	0.87	0.85	0.93	0.91	0.87	0.84
Alpha	0.93	0.92	0.87	0.86	0.93	0.92	0.87	0.86

Differential item functioning. The purpose of conducting differential item functioning (DIF) analyses was to examine whether some items function differently across modes for examinees at the same overall proficiency level on the test and, if so, whether sources of that difference can be identified. Recall that a qualitative content comparison was made for items across modes, which was used as a basis for judging the practical significance of the statistically identified items.

The qualitative comparison documented differences across modes that may affect student performance. For example, one general difference was that the online version line breaks of passages, stems, and options were usually different from the paper version, but this probably would not have any effect on students' performance. Other differences might or might not affect performances. For example, the paper version might have the entire passage or entire set of tables and figures visible on a single page whereas online might need scrolling. The online version used highlighting, but paper used underlying or reference to line numbers. Items that were potentially affected by these differences were identified.

The item *p*-value differences as presented in Figure 6 as well as omission rate comparisons in Figure 7 all contributed to the evaluation of item DIF, because the groups were randomly

equivalent. In addition, DIF was examined by using the Mantel-Haenszel procedure (Camilli & Shepard, 1994; Mantel & Haenszel, 1959).

The Mantel-Haenszel procedure calculates the weighted average of the odds-ratios across all score levels. In this study, items with odd ratio values smaller than 0.5 or larger than 2 were flagged for further review. When controlling for raw or scale scores before applying the equating methodology, two English, one reading, and one science items were flagged, and a few more items were flagged when controlling for scale scores after using the equating methodology to adjust for mode effects. These items were those with the largest *p*-value differences, almost always favoring the online mode. A comparison of the statistically identified items and what was documented in the qualitative comparison did not reveal any concrete sources of DIF for these items.

Scale score moments and measurement precision after applying equating methodology. Scale score properties, after using the equating methodology to adjust for any mode effects, the online forms were also examined across modes and across the online forms, including scale score reliability, standard error of measurement (SEM), and conditional SEM based on Lord's (1965) four parameter beta compound binomial model for raw scores (Kolen, Hanson, & Brennan, 1992). Table 12 presents the scale score moments, SEM, and reliability of each form. Figure 14 contains plots of the conditional SEM of each true scale score point for all three forms.

Table 12

Scale Score Moments, Standard Error of Measurement (SEM), and Reliability for Spring 2014

Test		Mean	SD	Skewness	Kurtosis	SEM	Reliability
	Online_1	20.47	6.13	0.32	2.63	1.76	0.92
English	Paper_1	20.47	6.12	0.29	2.62	1.71	0.92
	Online_2	20.43	6.12	0.26	2.61	1.72	0.92
	Online_1	21.01	5.21	0.58	2.60	1.57	0.91
Mathematics	Paper_1	21.02	5.15	0.56	2.59	1.58	0.91
	Online_2	21.02	5.18	0.57	2.57	1.45	0.92
	Online_1	21.48	6.46	0.30	2.35	2.35	0.87
Reading	Paper_1	21.47	6.43	0.32	2.37	2.29	0.87
	Online_2	21.56	6.49	0.32	2.36	2.41	0.86
	Online_1	21.15	5.12	0.41	3.28	2.05	0.84
Science	Paper_1	21.14	5.03	0.4	3.29	2.11	0.82
	Online_2	21.07	5.07	0.4	3.24	1.94	0.85

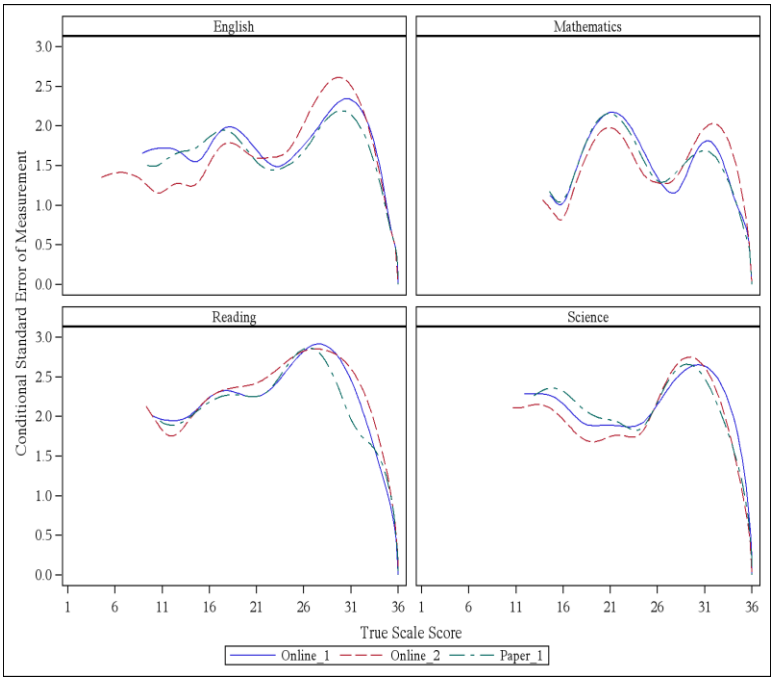


Figure 14. Conditional standard errors of measurement for spring 2014.

ACT writing test. In spring 2014, the writing test was holistically rated with a score of 2 to 12 and a testing time limit of 30 minutes. More than half of the students in the spring 2014 final

data took the ACT with writing. The mode effect for the writing test was examined by comparing the online and paper writing mean scores and by comparing the conditional writing scores after controlling for the English scale scores.

Table 13 presents the descriptive statistics, mean differences, effect sizes, and t-test *p*-values not only for writing but also for English, for students who took the ACT Form 1 writing test. Though random assignment of the online and paper forms was done within the group of students who registered for the ACT with writing, group equivalency might be affected by data cleaning. The purpose of including English scale scores (based on the online conversions) was to obtain additional evidence for the equivalency of the two groups taking the online versus the paper writing test. The effect size of between-mode group difference for English was small, and the t-test of mean difference was not statistically significant at the .05 level, providing additional evidence for the equivalency of the two groups for the writing test mode comparison. The small effect size and the relatively large t-test *p*-value for writing indicated that mode effect was not significant for the writing test in the spring 2014 special study.

Table 13

Across Mode Comparisons for Students Taking the ACT Writing Test for Spring 2014

	Online			Paper					
	N	Mean	SD	N	Mean	SD	Mean Difference	Effect Size	t-test *p*
English	1059	21.58	6.37	1255	21.31	6.24	0.27	0.04	0.29
Writing	1059	7.26	1.74	1255	7.22	1.57	0.04	0.02	0.57

The ACT writing scores were also examined by comparing the score distributions across modes conditioning on the English scale scores after adjusting for the mode effect, that is, the paper form applying the paper conversions and the online form applying the online conversions.

Figure 15 includes a scatter plot of the online and paper writing scores against students' English scale scores (left side of the graph), and the conditional mean writing scores for each mode (right side of the graph). Though there seemed to be a weak trend that the conditional online mean scores were slightly lower than the paper mean scores for lower English scores but slightly higher than paper means for higher English scores, the magnitude of the differences was small for most of the English scale score points. Since no evidence of significant mode effect for the writing test was found, no adjustment was made for mode to the ACT writing scores in spring 2014.

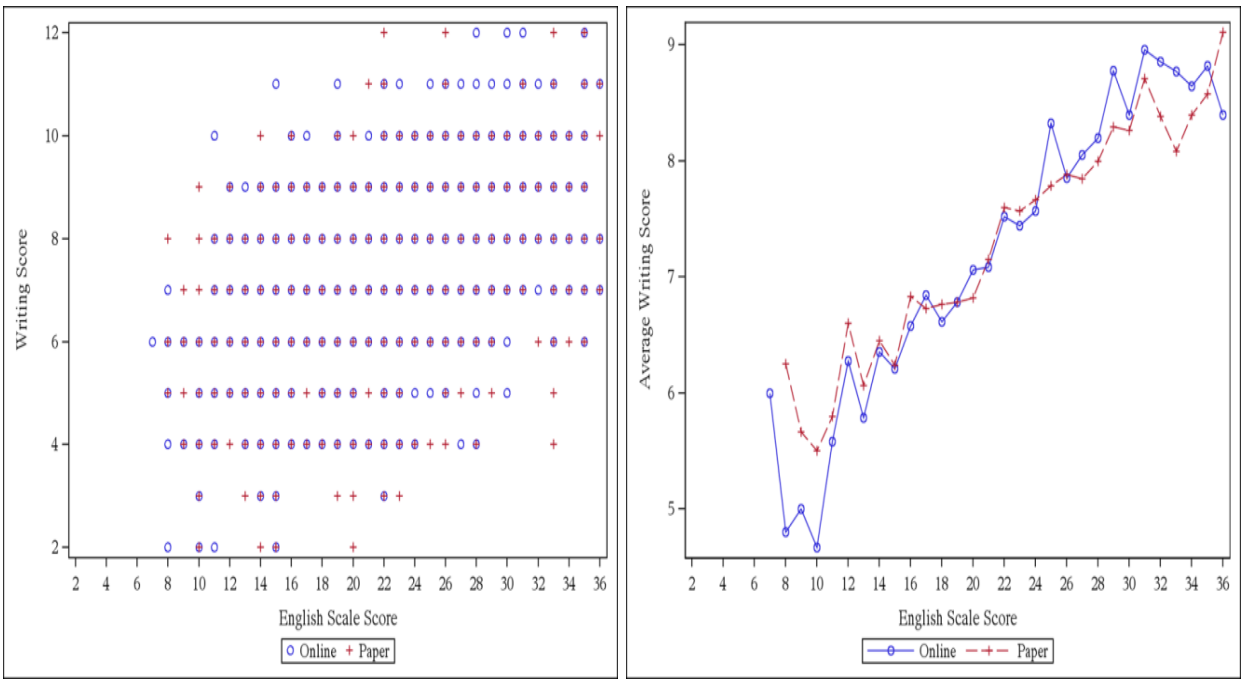


Figure 15. Writing scores conditioning on English scale scores for spring 2014.

Spring 2015 Mode Comparability Study

A second mode comparability study was conducted in spring 2015. A similar data collection design and procedure was used for the spring 2015 study as for the spring 2014 study. Students were randomly assigned to take one of the three ACT forms (Online_1, Online_2, or

Paper_1) that were administered in an operational testing setting on one of the ACT national test dates. Online testing was conducted on school-provided desktop or laptop computers.

There were two main differences between the spring 2014 and 2015 mode comparability studies. For the spring 2015 study, all the online tests had the same time limits as their paper counterparts. The extra five minutes added to the online reading and science tests in the spring 2014 study were eliminated. The second difference was that different versions of the ACT writing test were administered in these two mode comparability studies. The writing test administered in the spring 2015 study was the enhanced version that was to be operationally launched in fall 2015. The writing test is analytically scored with the administration time changed from 30 to 40 minutes. The enhanced version of the writing test launched in September 2015 reported four domain scores and a writing scale score ranging from 1 to 36, though only students' domain scores were reported back to schools in this special study. The participants in the 2015 study only came from those who registered for the ACT without writing but agreed to participate in a special writing study without receiving college reportable scores. Students took the writing test either online or on paper. Similar to the multiple-choice tests, two online and one paper prompts were randomly assigned to students participating in the study.

Data. More than 4,000 students from about 40 schools signed up for the spring 2015 study. After data cleaning, more than 3,000 students with at least 1,000 cases for a form were included in the spring 2015 final data. Similar two-phase analyses as for the spring 2014 study were conducted for the spring 2015 mode comparability study. The following sections present the results for the multiple-choice tests then the writing test.

Phase I mode comparability analyses and results for multiple-choice tests. Phase I analyses focused on test score distributions and the similarity of item level scores across modes.

Table 14 contains the sample size of each test form in the spring 2015 study as well as the means, standard deviations (SDs), minimums, and maximums of the observed total raw scores and scale scores for all three forms. The scale score descriptive statistics for Online_1 and Online_2 were obtained by applying the paper version conversions of Form 1 and Form 2, respectively. On average the online Form 1 scores tended to be higher than the paper Form 1 scores for all tests. Though online Form 2 had higher raw score means than those of online Form 1 (except reading and science), their scale score means were similar.

Table 14

Descriptive Statistics of Raw and Scale Scores of all Test Forms for Spring 2015

Form	Test	Raw Score					Scale Score			
		N	Mean	SD	Min	Max	Mean	SD	Min	Max
Online_1	English	1092	43.62	14.10	9	74	20.79	5.98	6	36
	Mathematics	1092	30.02	11.76	5	60	20.69	5.20	10	36
	Reading	1092	23.28	7.59	4	40	21.99	6.24	7	36
	Science	1092	20.73	7.43	2	40	20.86	5.17	5	36
Paper_1	English	1056	41.26	14.43	5	74	19.79	6.03	4	36
	Mathematics	1056	29.74	11.78	5	60	20.58	5.16	10	36
	Reading	1056	22.00	7.57	2	40	20.91	6.08	4	36
	Science	1056	20.72	7.20	3	40	20.80	4.96	6	36
Online_2	English	1044	44.54	14.83	4	75	20.46	6.27	3	36
	Mathematics	1044	30.24	11.46	7	58	20.82	5.05	12	35
	Reading	1044	22.66	8.00	3	40	21.77	6.14	5	36
	Science	1044	20.30	7.03	3	39	21.36	5.00	6	36

Figure 16 provides graphical presentations of the raw score and scale score mean differences across modes. Mean differences, effect sizes, and *p*-values from t-tests of mean differences for raw and scale scores are presented in Table 15. The effect sizes were calculated by dividing the mean differences by the pooled standard deviations across modes for each test. In the spring 2015 study, students who took the online tests still performed better than the students who

took the paper tests. However, the mean differences were only statistically significant for the English and reading tests.

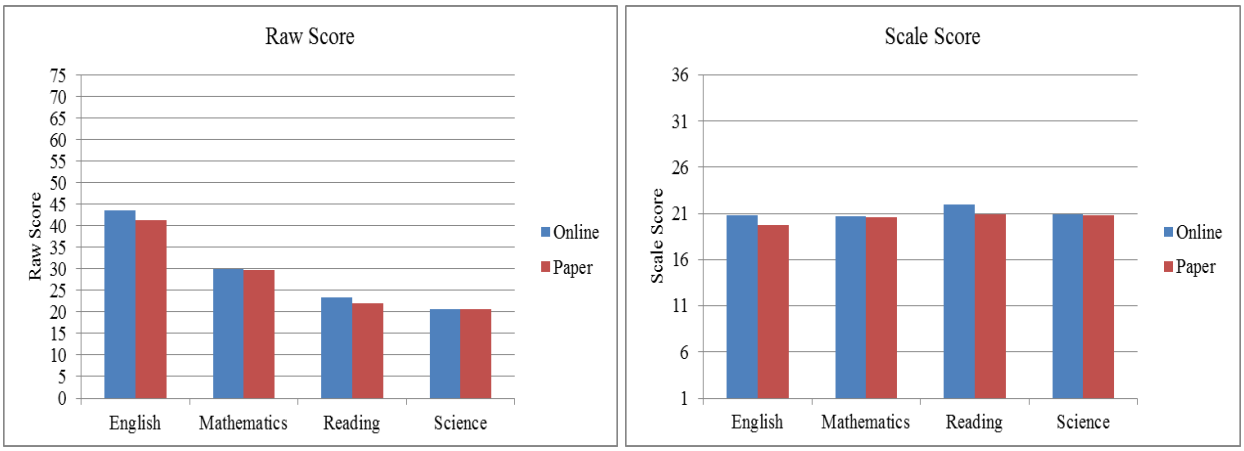


Figure 16. Raw and scale score mean comparisons across modes for spring 2015.

Table 15

Raw and Scale Score Mean Differences across Modes (Online minus Paper) for Spring 2015

Test	Raw Score Comparison			Scale Score Comparison		
	Mean Difference	Effect Size	t-test *p*	Mean Difference	Effect Size	t-test *p*
English	2.36	0.17	0.0001	1.00	0.17	0.0001
Mathematics	0.28	0.02	0.5808	0.11	0.02	0.6199
Reading	1.28	0.17	<.0001	1.08	0.18	<.0001
Science	0.01	0.00	0.9742	0.06	0.01	0.7717

The plots of the relative cumulative frequency distributions of proportion correct raw scores and scale scores are shown in Figure 17. For spring 2015, only for the English and reading tests, the online group seemed to score higher than the paper group. The differences between the online and paper groups for the English test appeared to be similar for the spring 2014 and 2015 studies, while the differences for the reading test tended to be smaller for the spring 2015 study than that for spring 2014. The Kolmogorov-Smirnov (KS) test of equivalency of distributions was

conducted for the raw and scale scores for each test. For the spring 2015 study, the KS tests were statistically significant for the English and reading tests, but not for the mathematics and science tests.

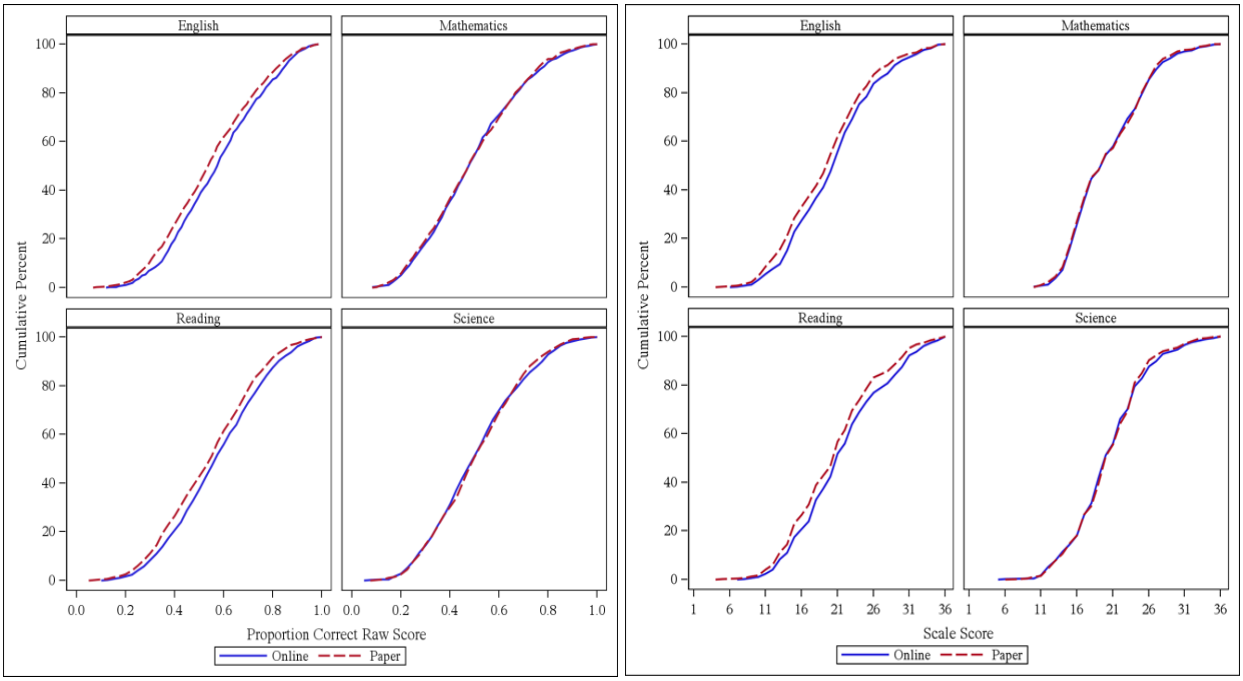


Figure 17. Relative cumulative frequency distributions of proportion correct raw scores and scale scores for spring 2015.

Table 16 contains the scale score correlations, effective weights, and Cronbach's alpha. These values were all very similar across modes. Item difficulties (*p*-values) were compared across modes. The graphs on the left side of Figure 18 presents the proportion of correct responses for each item (item *p*-value) by item position across modes, with smaller values indicating harder items. The graphs on the right side show the *p*-value differences across modes, with positive difference indicating that the item was easier for the online administration. Figure 18 shows that while later items tended to be harder compared to earlier items for each test regardless of mode, the items tended to be easier for the online administration, especially for items that appeared later

in the test. Consistent with the results shown in Table 15, the mathematics and science tests showed smaller mode differences in item *p*-values compared with the English and reading tests. The items toward the end of the English and reading tests tended to be easier for online. Compared with spring 2014 results, the mode differences on the reading and science tests were smaller in the spring 2015 study, probably due to the removal of the extra five minutes for these two tests.

Table 16

Scale Score Correlations, Effective Weights, and Cronbach's Alpha for Spring 2015

		Online				Paper			
		English	Mathematics	Reading	Science	English	Mathematics	Reading	Science
Correlation	English	1.00	.75	.81	.75	1.00	.75	.80	.76
	Mathematics	.75	1.00	.68	.80	.75	1.00	.66	.79
	Reading	.81	.68	1.00	.74	.80	.66	1.00	.73
	Science	.75	.80	.74	1.00	.76	.79	.73	1.00
Effective Weight		.27	.23	.28	.23	.28	.23	.27	.22
Cronbach's Alpha		.93	.92	.86	.86	.93	.92	.86	.85

Figure 19 has the omission rate (i.e., the proportion of missing responses) for each item across modes. The paper group consistently had a higher omission rate than the online group for the latter half of the tests, except for the last few items of the mathematics test, across all four tests. The omission rate was slightly higher for spring 2015 than that of spring 2014 across all four tests. In addition, the proportion of examinees choosing the incorrect options was also examined for each item across mode, but no obvious patterns were found.

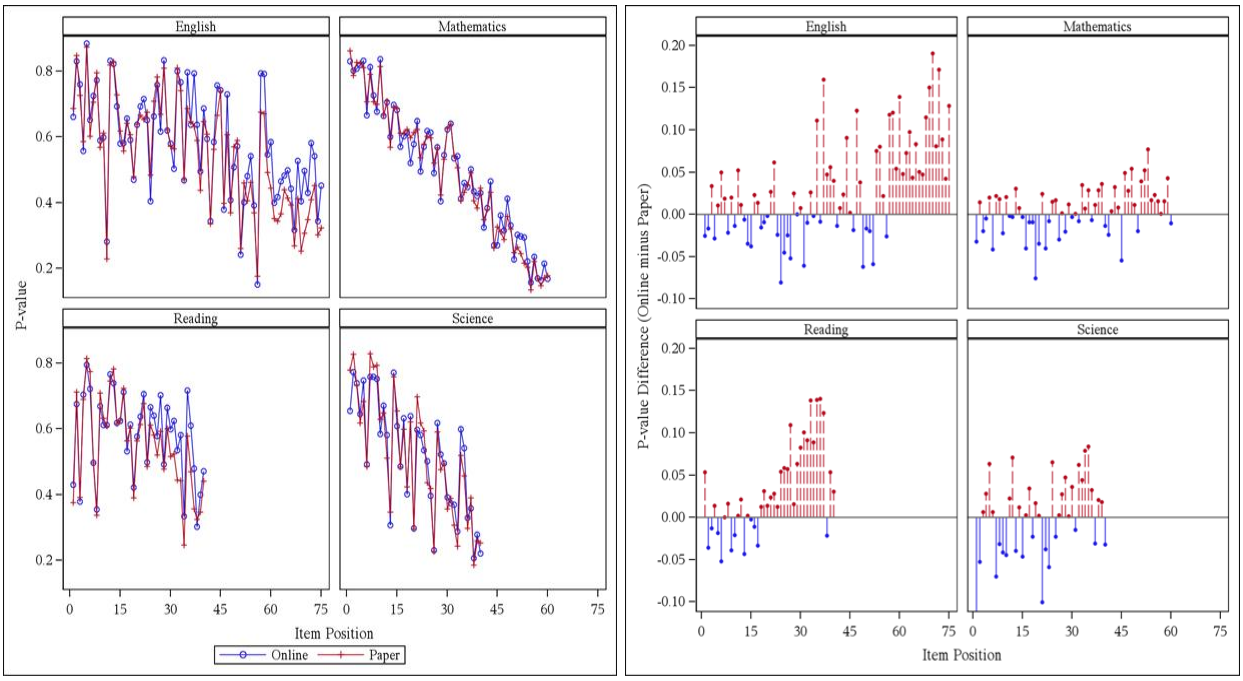


Figure 18. Scatter plots of item *p*-values and needle plots of *p*-value differences across modes for spring 2015.

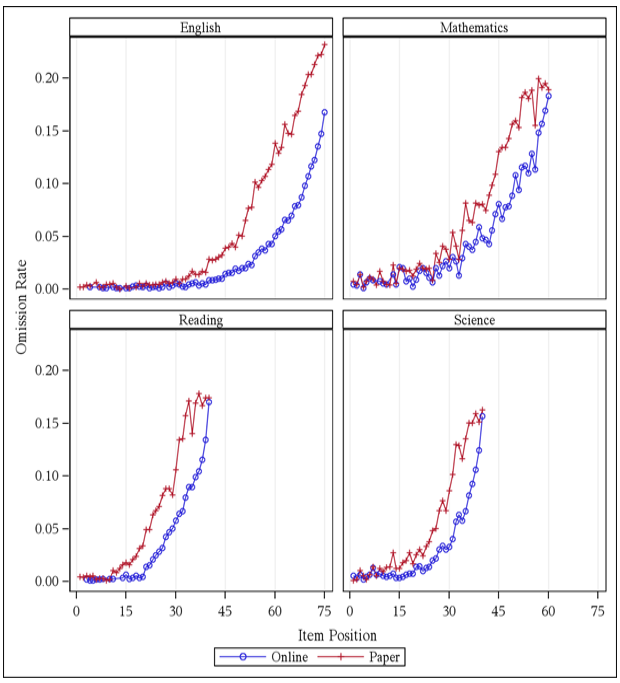


Figure 19. Item omission rates by item position for spring 2015.

Adjustments to score differences. Equating methodology was used to generate raw to scale score conversion tables for the online forms that were different from the corresponding paper conversions. The raw to scale score conversions for the two online forms together with their counterpart paper conversions are shown on the left side of Figure 20, and the right side displays the differences between the online and paper conversions at each raw score point for the two online forms, with negative values indicating that the same raw score was converted to a lower scale score in the online conversion than in the paper conversion. For spring 2015, a similar pattern was observed as seen for spring 2014, except for a few raw score points for Form 1 mathematics and science, and for Form 2 English and mathematics.

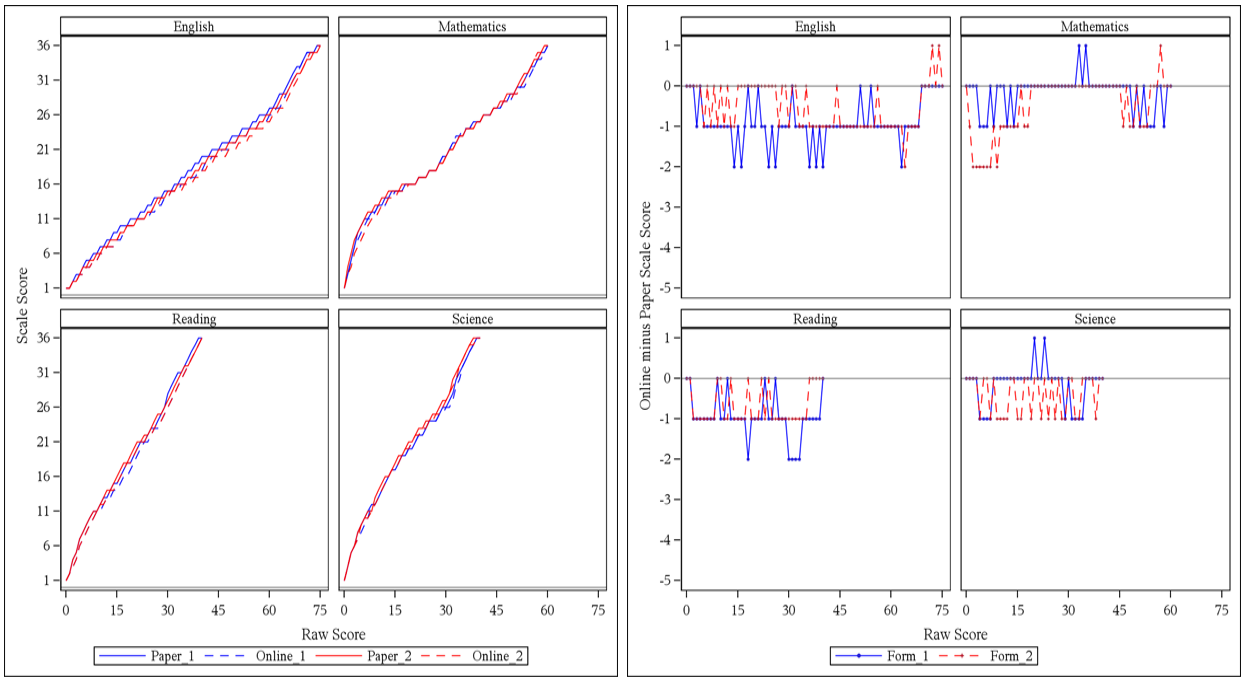


Figure 20. Raw to scale score conversions and scale score differences for spring 2015.

Scale score differences were computed as online minus paper scores by applying online or paper conversions. Figure 21 presents distributions of the difference scores. For spring 2015, only

a small portion of the difference scores surpassed one score point for the English and reading tests. Almost all the differences were zero or one score point for mathematics and science.

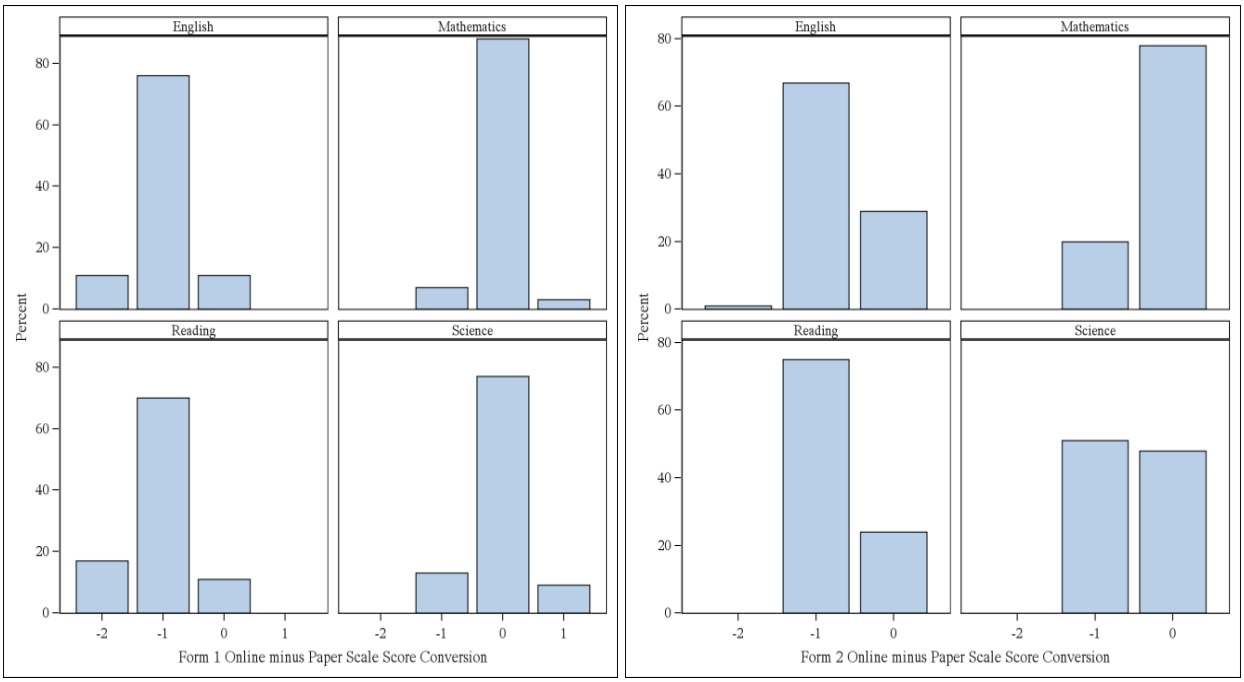


Figure 21. Distributions of scale score adjustment for the two online forms for spring 2015.

Online timing re-evaluation. Time limits for the online tests were re-evaluated in the spring 2015 study by examining student survey data and online test item response time. In spring 2015, 490 students responded to the survey question of whether they felt they had enough time to finish each of the tests, with about 68% of the students taking the tests online. Table 17 contains the survey results for this question. Larger percentages of students felt they had enough time to finish the online tests than those who took the paper tests. This was true for all five subjects.

Figure 22 presents the average time spent on each item for all four multiple-choice tests. No evidence of speededness was found for the online tests. The patterns of time spent on items were quite similar between the spring 2014 study and the spring 2015 study. The comparability of scores between online and paper with removing the five extra minutes in online administration of

the reading and science tests was further evaluated in spring 2015, which confirmed the decision that the same administration time should be used for online and paper versions of all tests.

Table 17

Percentage of Responses to the Timing Related Question for Spring 2015

I felt I had enough time to finish the…test	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Either agree or strongly agree	Either disagree or strongly disagree
Online							
English	35	38	9	12	6	73	18
Mathematics	15	34	13	23	15	49	38
Reading	14	23	11	34	18	37	52
Science	12	17	18	31	22	29	53
Writing	39	33	15	8	5	72	13
Paper							
English	18	35	11	24	12	53	36
Mathematics	7	30	12	32	19	37	52
Reading	5	18	10	43	25	22	68
Science	4	23	19	28	26	27	54
Writing	28	31	20	13	8	59	21

Phase II mode comparability analyses and results for multiple-choice tests. Similar Phase II mode comparability analyses were conducted for the spring 2015 study. Figure 23 contains plots of the test characteristic curves (TCCs) across modes for each subject. Consistent with the patterns observed in Figure 17 for the raw and scale score relative cumulative frequency distributions, the between-mode TCC difference was the smallest for both the mathematics and science tests. There were some differences in TCC for the English and reading tests. Scatter plots of item parameter estimates from online and paper are presented in Figure 24. The *b*-parameter comparison showed that the online items tended to be easier than the paper items, especially for the reading test, while online items had higher *c*-values.

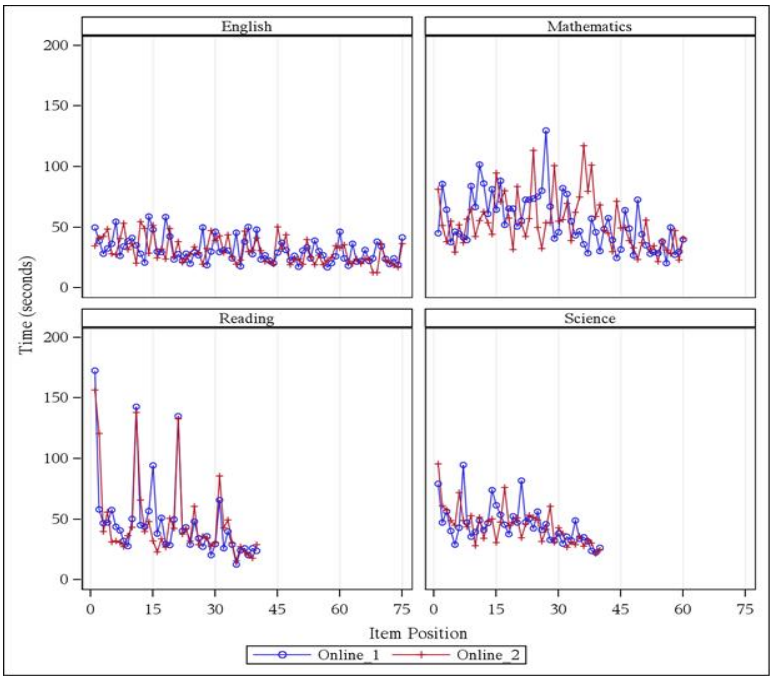


Figure 22. Average time spent on each item for spring 2015.

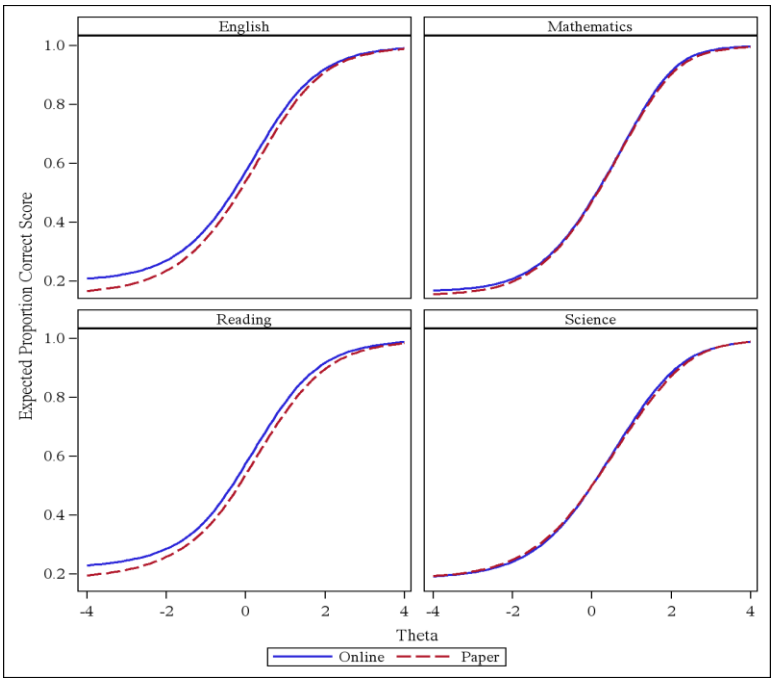
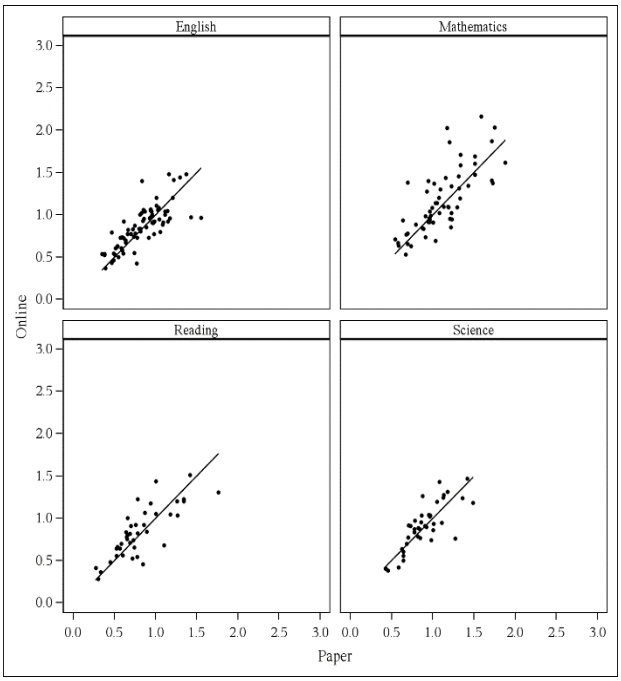
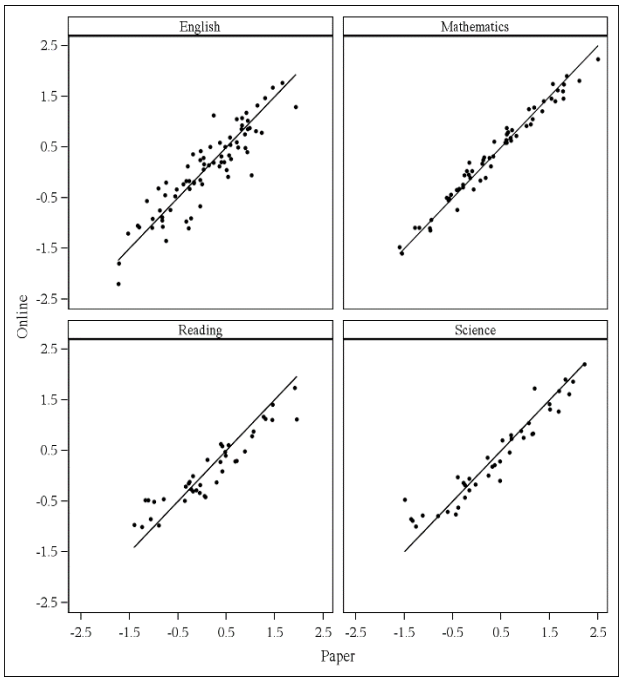


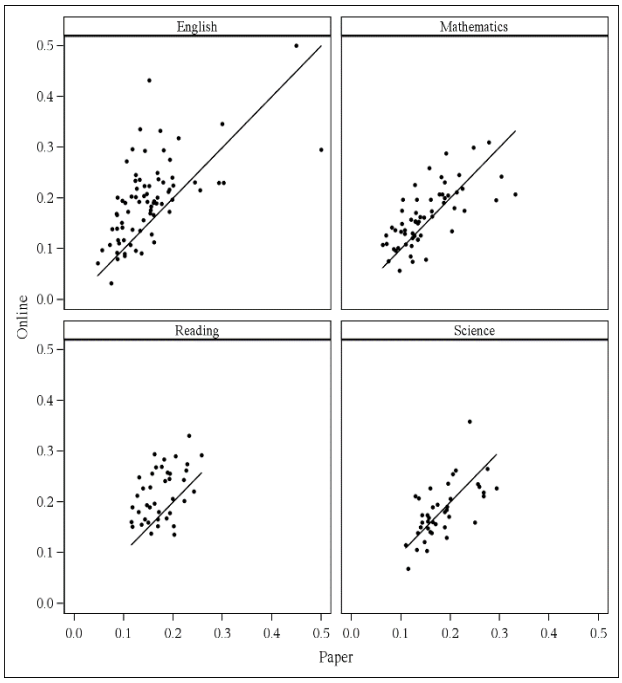
Figure 23. Test characteristic curves across modes for spring 2015.



a-parameter



b-parameter



c-parameter

Figure 24. IRT parameter comparison across modes for spring 2015.

Exploratory factor analysis was conducted to explore the dimensionality and construct equivalency of the online and paper tests. Eigenvalue scree plots for each test were examined across modes, as shown in Figure 25. The data were fit with both a one-factor and a two-factor model.

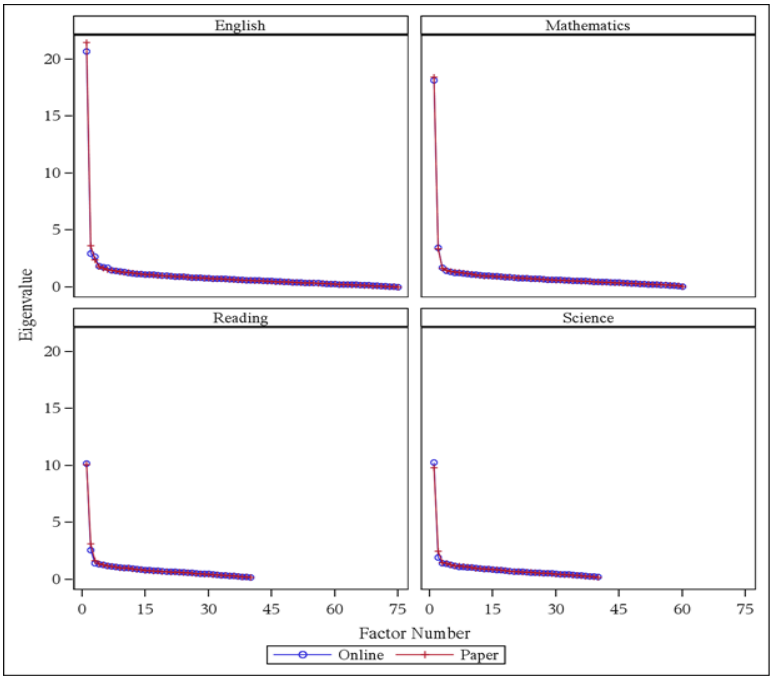


Figure 25. Eigenvalue scree plot for spring 2015.

Table 18 contains several fit indices resulting from fitting the one- and two-factor models for the four multiple-choice tests across modes. The bolded numbers are values that did not meet the criteria presented in Table 8, showing that a couple of the fit statistics for the reading and science online tests were not optimal, and the two-factor model seemed to improve the mode fit for those tests. However, based on the principle of parsimony, the one-factor model was considered to be adequate and the factor loadings of each test for the one-factor model were compared across modes. Table 19 presents the descriptive statistics of the factor loadings of each mode and the correlations of the factor loadings between the two modes.

Table 18

Fit Statistics of One- and Two-Factor Models for Spring 2015

Test	Fit Statistic	Online			Paper		
		One Factor	Two Factors	DIFF	One Factor	Two Factors	DIFF
English	CFI	0.96	0.98	0.02	0.96	0.98	0.03
	TLI	0.96	0.98	0.02	0.96	0.98	0.03
	RMSEA	0.03	0.03	0.01	0.04	0.03	0.01
	SRMR	0.06	0.05	0.01	0.07	0.05	0.01
Mathematics	CFI	0.97	0.99	0.02	0.97	0.99	0.02
	TLI	0.97	0.99	0.03	0.97	0.99	0.02
	RMSEA	0.04	0.02	0.02	0.03	0.02	0.02
	SRMR	0.06	0.05	0.02	0.06	0.05	0.02
Reading	CFI	**0.94**	0.98	0.05	**0.90**	0.98	0.07
	TLI	**0.93**	0.98	0.05	**0.90**	0.97	0.07
	RMSEA	0.04	0.02	0.02	0.05	0.03	0.03
	SRMR	0.07	0.05	0.02	0.08	0.05	0.03
Science	CFI	0.97	0.99	0.02	**0.94**	0.98	0.04
	TLI	0.97	0.99	0.02	**0.94**	0.98	0.04
	RMSEA	0.03	0.02	0.01	0.04	0.02	0.01
	SRMR	0.05	0.05	0.01	0.07	0.05	0.02

Table 19

Descriptive Statistics and Correlation of Factor Loadings across Modes for Spring 2015

Test	Form	Mean	SD	Minimum	Maximum	Correlation
English	Online	0.51	0.12	0.24	0.72	.85
	Paper	0.52	0.12	0.22	0.76	
Mathematics	Online	0.53	0.12	0.19	0.74	.84
	Paper	0.54	0.11	0.28	0.69	
Reading	Online	0.48	0.12	0.19	0.74	.87
	Paper	0.47	0.13	0.19	0.74	
Science	Online	0.47	0.13	0.19	0.71	.88
	Paper	0.46	0.12	0.15	0.70	

Generalizability coefficients ($E\rho^2$) and dependability indices or phi coefficients (Φ) are reported in Table 20. Similar to the alpha results, reliability indices from the generalizability analyses showed barely any differences across modes.

Table 20

Raw Score Generalizability Coefficient, Phi Coefficient, and Alpha for Spring 2015

	Online				Paper			
	English	Mathematics	Reading	Science	English	Mathematics	Reading	Science
$E\rho^2$	0.93	0.92	0.87	0.86	0.93	0.93	0.87	0.86
Φ	0.92	0.91	0.86	0.84	0.92	0.91	0.86	0.83
Alpha	0.93	0.92	0.86	0.86	0.93	0.92	0.86	0.85

The Mantel-Haenszel procedure was used to flag DIF items that needed to be further reviewed. Three English items were flagged for DIF based on raw score, and a few more items were flagged when controlling for scale scores after using the equating methodology. One reading and one science item was flagged based on scale score. No concrete sources of DIF for those items were identified. Table 21 presents the scale score moments, SEM, and reliability of each form. Figure 26 contains plots of the conditional SEM of each true scale score point for all three forms.

Table 21

Scale Score Moments, Standard Error of Measurement (SEM), and Reliability for Spring 2015

Test		Mean	SD	Skewness	Kurtosis	SEM	Reliability
English	Online_1	19.79	6.06	0.32	2.71	1.75	0.92
	Paper_1	19.79	6.03	0.27	2.69	1.71	0.92
	Online_2	19.76	6.11	0.31	2.72	1.67	0.93
Mathematics	Online_1	20.65	5.18	0.44	2.44	1.61	0.90
	Paper_1	20.58	5.16	0.44	2.43	1.59	0.91
	Online_2	20.60	5.12	0.40	2.33	1.63	0.90
Reading	Online_1	20.93	6.09	0.27	2.44	2.26	0.86
	Paper_1	20.91	6.07	0.29	2.54	2.28	0.86
	Online_2	21.01	6.18	0.31	2.47	2.16	0.88
Science	Online_1	20.82	5.06	0.23	3.27	2.14	0.82
	Paper_1	20.80	4.96	0.25	3.13	2.11	0.82
	Online_2	20.84	5.05	0.24	3.17	2.17	0.82

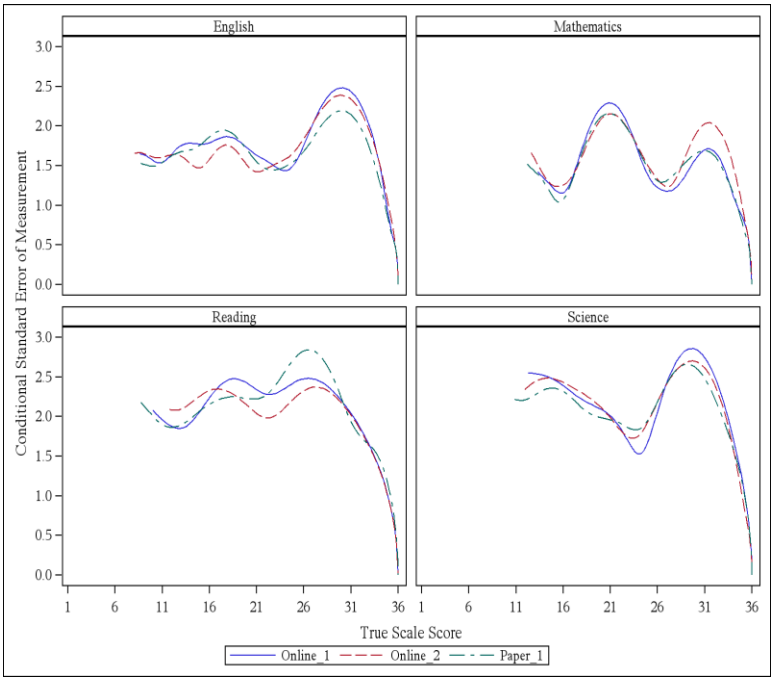


Figure 26. Conditional standard errors of measurement for spring 2015.

ACT writing test. Two writing prompts were administered along with the two online and one paper forms of the ACT in the spring 2015 ACT mode comparability study. The online

versions of the two writing prompts were spiraled within the examinees taking the two online ACT forms, and the paper versions of the two prompts were spiraled within examinees taking the one paper form. This design resulted in randomly equivalent groups of students taking each mode and form combination of the writing prompts. Since there was only one paper ACT form, the number of students taking the paper prompts was about half of the students taking the online prompts.

Before examining potential mode effect for the writing prompts, the groups taking the online and paper versions of each prompt were examined to verify the equivalency of the groups. First, form distributions within each test center were checked so centers with large form count differences could be excluded from further analyses. Then, the distributions of student demographics (e.g., gender and ethnicity) and students' ACT English scale scores were compared across the groups of students taking each of the writing mode and form combinations. The groups were found to be similar in terms of demographics and English test scores. In addition, rater ID distributions by prompt were examined. It was found that basically the same sets of raters rated the online and paper versions of each prompt, which eliminated to an extent the potential concern about the confounding of rater group and mode effect. Table 22 includes the demographic information. Figure 27 displays the English scale score distributions for all the online and paper versions of the writing test.

Before conducting mode comparability analyses for the writing test, the latency information from the online test was also examined. Table 23 contains the descriptive statistics of the latency information of the two online prompts. As shown in Table 23, the two online prompts had similar latency information. The average time students spent on each prompt was around 30 minutes with a standard deviation of around nine minutes.

Table 22

Demographic Distribution of the Online and Paper Examinees for Spring 2015

Prompt	Demographic		Online		Paper	
			N	Percent	N	Percent
Form_1	Gender	Male	475	45	237	43
		Female	575	55	317	57
Form_2		Male	448	43	211	41
		Female	600	57	307	59
Form_1	Race/Ethnicity	African American	128	12	58	10
		American Indian/Alaska Native	14	1		
		White	484	46	266	48
		Hispanic/Latino	248	24	138	25
		Asian	84	8	44	8
		Two or more races	60	6	31	6
		Prefer not to respond	30	3	17	3
Form_2		African American	119	11	63	12
		American Indian/Alaska Native	9	1	3	1
		White	511	49	251	48
		Hispanic/Latino	273	26	119	23
		Asian	81	8	42	8
		Two or more races	25	2	21	4
		Prefer not to respond	30	3	19	4

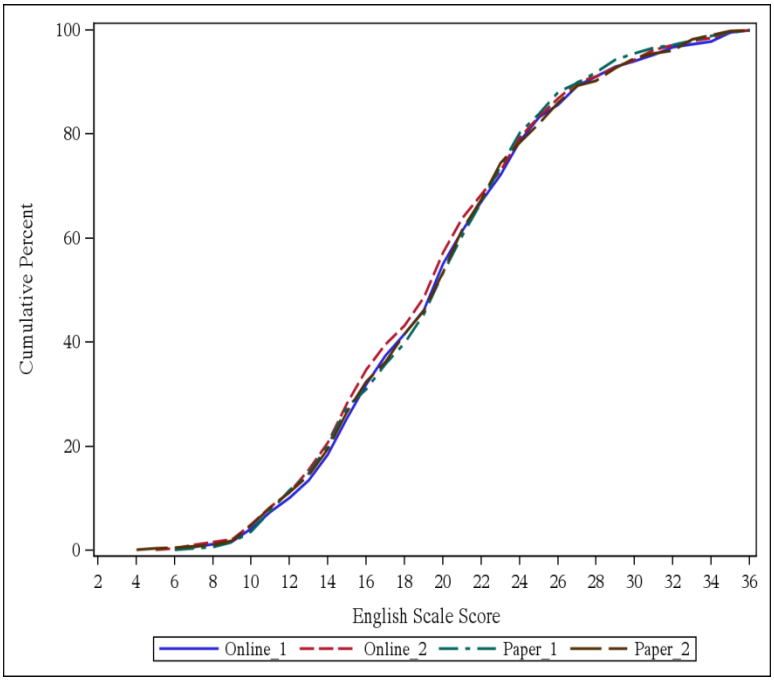


Figure 27. English scale score distribution of the students who took writing test for spring 2015.

Table 23

Descriptive Statistics of Online Latency Information of Writing Test for Spring 2015

Form	N	Mean	SD	Min	Max
Online_1	1086	29.55	8.67	0	39.88
Online_2	1079	29.68	8.69	0	39.90

Assuming the groups are indeed equivalent across modes, for each prompt, the following comparisons were made to evaluate the mode effect of the writing test: (1) means and standard deviations of the writing domain scores, rounded average domain raw scores, and scale scores; (2) mean score differences across modes, effect sizes of mean differences, and t-test *p*-values of mean differences; (3) correlations among the domain scores, rounded average domain raw scores, raw and scale scores, and with English scale scores; (4) plots of the relative cumulative frequency distributions of raw and scale scores; (5) scatter plots of English and writing raw and scale scores; and (6) plots of mean writing scores conditional on English scale scores. As was done for the multiple-choice tests, the scale scores for the online writing prompts used at this stage of comparison were obtained by applying the corresponding paper version raw to scale score conversions.

Results for the above analyses are presented in the following tables and figures. The comparison of means and plots of the relative cumulative frequencies both indicate that the two prompts appear to have differential mode effects. Online students scored higher on one prompt than paper students, but not on the other prompt. In addition, the online writing scores were slightly more correlated with the English scores for both prompts. However, the higher correlation might be an artifact of sample size differences.

Table 24 has the descriptive statistics, mean differences of online and paper prompt scores, effect sizes, and t-test *p*-values. Table 25 shows the correlations among writing scores and with

English scale scores. Figures 28 to 30 display the writing score distributions, scatter plots of English scale score and writing scores, and average writing scores conditioning on English scale scores.

Table 24

Descriptive Statistics, Effect Sizes, and t-test p-values of English and Writing Scores across Modes for Spring 2015

	Online			Paper			Online – Paper	Effect Size	t-test *p*
	N	Mean	SD	N	Mean	SD			
Form_1									
Domain 1	1050	6.60	1.86	554	6.04	1.83	0.56	0.30	<.0001
Domain 2	1050	6.40	1.79	554	5.86	1.74	0.55	0.31	<.0001
Domain 3	1050	6.58	1.82	554	6.10	1.77	0.47	0.26	<.0001
Domain 4	1050	7.13	1.78	554	6.57	1.75	0.55	0.31	<.0001
Average Domain	1050	6.76	1.79	554	6.23	1.77	0.52	0.29	<.0001
Raw Score	1050	26.70	7.06	554	24.57	6.92	2.13	0.30	<.0001
Scale Score	1050	19.16	6.59	554	17.17	6.44	2.00	0.31	<.0001
English	1050	20.02	6.02	554	19.90	5.85	0.11	0.02	0.7187
Form_2									
Domain 1	1048	5.98	1.95	518	5.94	1.83	0.03	0.02	0.7404
Domain 2	1048	5.87	1.93	518	5.81	1.79	0.06	0.03	0.5761
Domain 3	1048	6.15	1.94	518	6.21	1.81	-0.06	-0.03	0.5578
Domain 4	1048	6.26	1.93	518	6.41	1.71	-0.14	-0.08	0.1339
Average Domain	1048	6.14	1.93	518	6.18	1.78	-0.04	-0.02	0.6832
Raw Score	1048	24.26	7.60	518	24.37	6.98	-0.11	-0.02	0.7698
Scale Score	1048	17.50	7.65	518	17.66	7.06	-0.15	-0.02	0.6919
English	1048	19.68	6.07	518	19.95	6.08	-0.26	-0.04	0.4223

Table 25

Correlation among Writing Scores and with English Scores for Spring 2015

			Domain 1	Domain 2	Domain 3	Domain 4	Average Domain	Raw Score	Scale Score	English
Form_1	Online	Domain 1	1.00	.95	.96	.91	.97	.98	.98	.53
		Domain 2	.95	1.00	.96	.90	.96	.98	.98	.52
		Domain 3	.96	.96	1.00	.92	.97	.98	.98	.53
		Domain 4	.91	.90	.92	1.00	.95	.96	.96	.56
		Average Domain	.97	.96	.97	.95	1.00	.99	.99	.54
		Raw Score	.98	.98	.98	.96	.99	1.00	1.00	.55
		Scale Score	.98	.98	.98	.96	.99	1.00	1.00	.55
		English	.53	.52	.53	.56	.54	.55	.55	1.00
	Paper	Domain 1	1.00	.95	.94	.92	.97	.98	.98	.46
		Domain 2	.95	1.00	.96	.91	.96	.98	.97	.45
		Domain 3	.94	.96	1.00	.93	.97	.98	.98	.45
		Domain 4	.92	.91	.93	1.00	.96	.97	.97	.48
		Average Domain	.97	.96	.97	.96	1.00	.99	.99	.46
		Raw Score	.98	.98	.98	.97	.99	1.00	1.00	.47
		Scale Score	.98	.97	.98	.97	.99	1.00	1.00	.47
		English	.46	.45	.45	.48	.46	.47	.47	1.00
Form_2	Online	Domain 1	1.00	.98	.96	.93	.98	.99	.98	.59
		Domain 2	.98	1.00	.95	.92	.97	.98	.97	.57
		Domain 3	.96	.95	1.00	.95	.98	.98	.98	.60
		Domain 4	.93	.92	.95	1.00	.96	.97	.97	.64
		Average Domain	.98	.97	.98	.96	1.00	.99	.99	.60
		Raw Score	.99	.98	.98	.97	.99	1.00	1.00	.61
		Scale Score	.98	.97	.98	.97	.99	1.00	1.00	.61
		English	.59	.57	.60	.64	.60	.61	.61	1.00
	Paper	Domain 1	1.00	.97	.95	.93	.97	.98	.98	.44
		Domain 2	.97	1.00	.93	.90	.95	.97	.96	.42
		Domain 3	.95	.93	1.00	.96	.98	.98	.98	.45
		Domain 4	.93	.90	.96	1.00	.96	.97	.96	.47
		Average Domain	.97	.95	.98	.96	1.00	.99	.99	.45
		Raw Score	.98	.97	.98	.97	.99	1.00	.99	.45
		Scale Score	.98	.96	.98	.96	.99	.99	1.00	.45
		English	.44	.42	.45	.47	.45	.45	.45	1.00

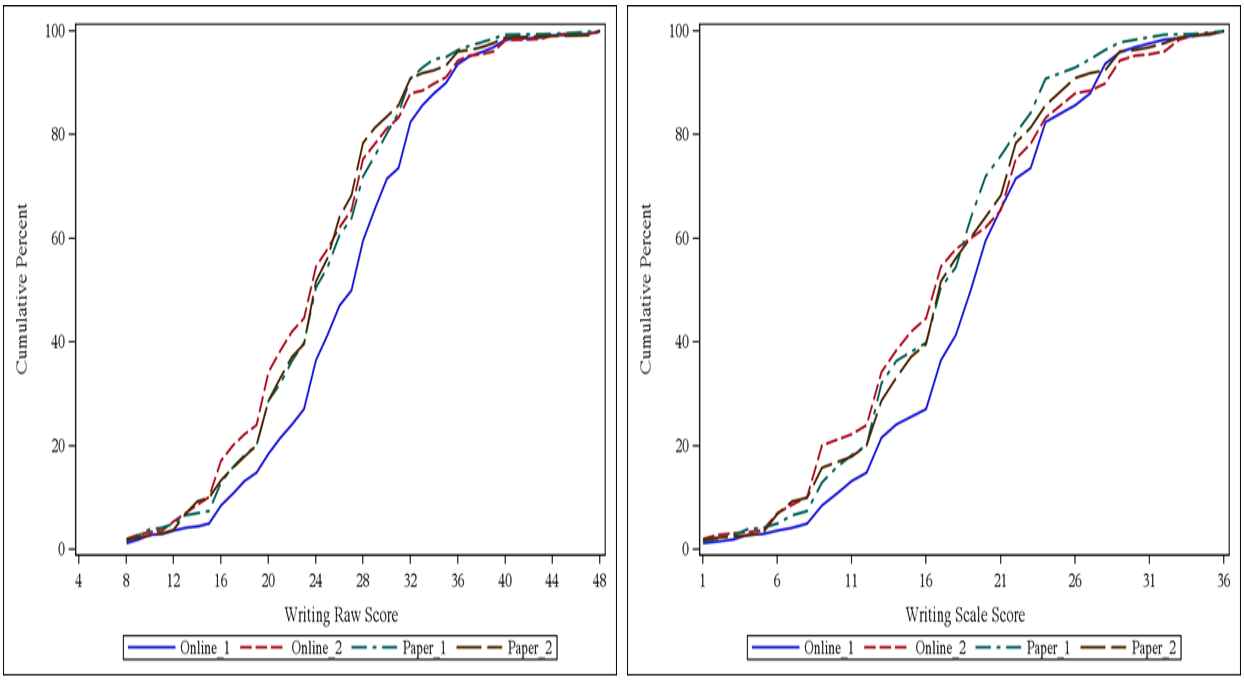


Figure 28. Writing raw score and scale score distribution across prompts for spring 2015.

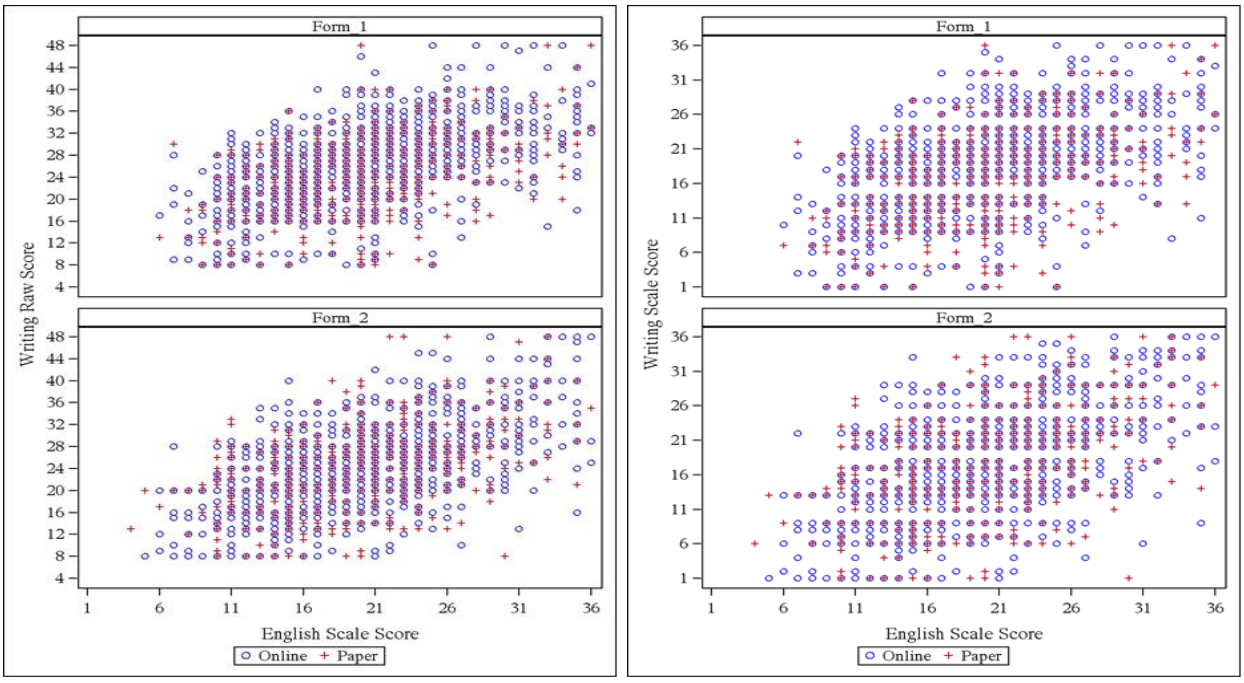


Figure 29. Scatter plots of English scale score and writing scores for spring 2015.

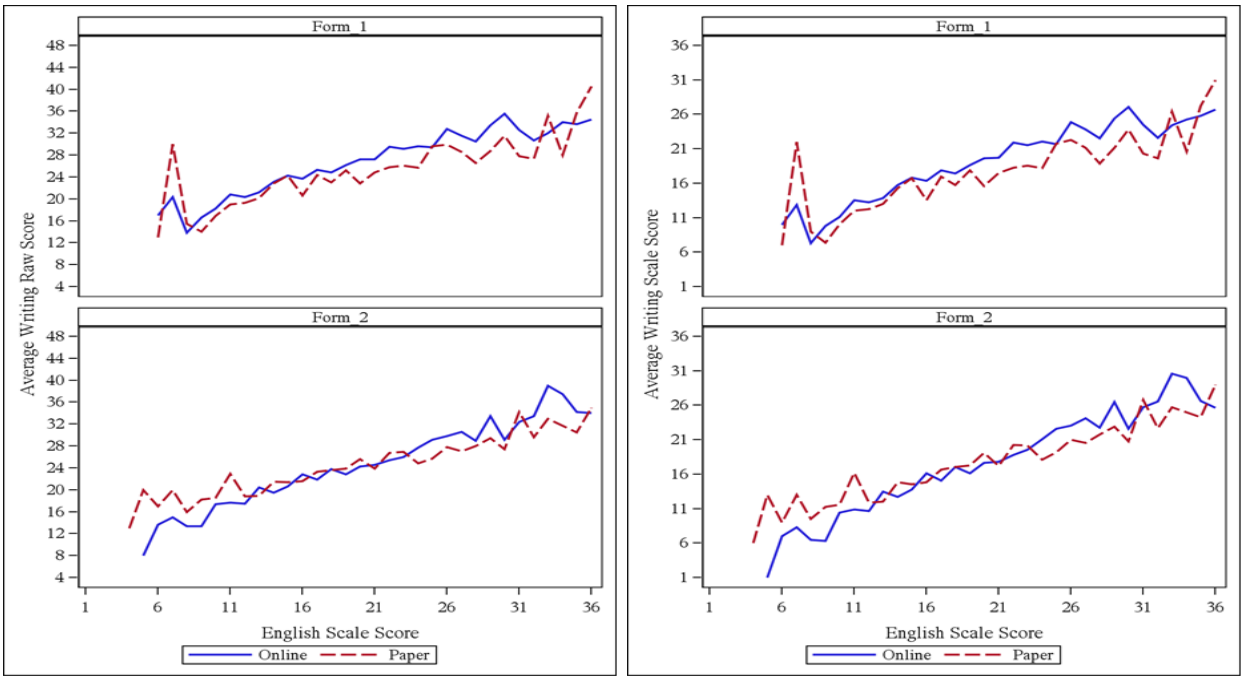


Figure 30. Average writing raw scores (left) and scale scores (right) conditioning on English scale score for spring 2015.

Equating methodology was applied to remove any mode differences between the writing scores. Table 26 contains the descriptive statistics of the writing scale score after using the equating methodology. Note that except for Table 26, all writing scale scores included in the tables and figures in this section were from applying the paper conversions to the online prompts.

Table 26

Descriptive Statistics, Effect Sizes, and t-test p-values of Writing Scale Scores after Applying Equating Methodology for Spring 2015

	Online			Paper					
Scale Score	N	Mean	SD	N	Mean	SD	Online – Paper	Effect Size	t-test *p*
Form_1	1050	17.08	6.47	554	17.17	6.44	-0.08	-0.01	0.8064
Form_2	1048	17.10	6.51	518	17.66	7.06	-0.56	-0.08	0.1183

Spring 2015 ACT Online Administration

Shortly after the second mode comparability study in spring 2015, the ACT was administered online in another administration. More than 4,000 students participated in this ACT testing within a multi-day testing window. On one of the testing days, the two online forms used in the spring 2015 mode comparability study were randomly assigned to more than 1,800 students. The sample sizes of the two forms were 890 and 922, respectively. This provided additional data for validation of the mode comparability online "equating" results.

With randomly equivalent groups of examinees taking each form, the distributions of scale scores across forms are expected to be very similar. The equivalency of the score distributions of the two online forms from these data was examined. In addition, the distributions of the online form scale scores were also compared with their distributions in the spring 2015 mode comparability study.

Figure 31 presents plots of the relative cumulative frequency distributions of the scale scores of the two online forms. It shows that the scale score distributions were similar across the two forms for all subjects. Figure 32 has the distributions of the two forms in the spring 2015 mode comparability study added as a comparison, which shows obvious differences between the samples of students in these two administrations. Even though examinee proficiencies were different in the mode comparability study and this subsequent administration, the distributions of scale scores of the two forms were similar within each administration (i.e., the mode comparability study and the online administration). These results provided further support for the stability of the online form conversions obtained from the mode comparability study when equating methodology was applied.

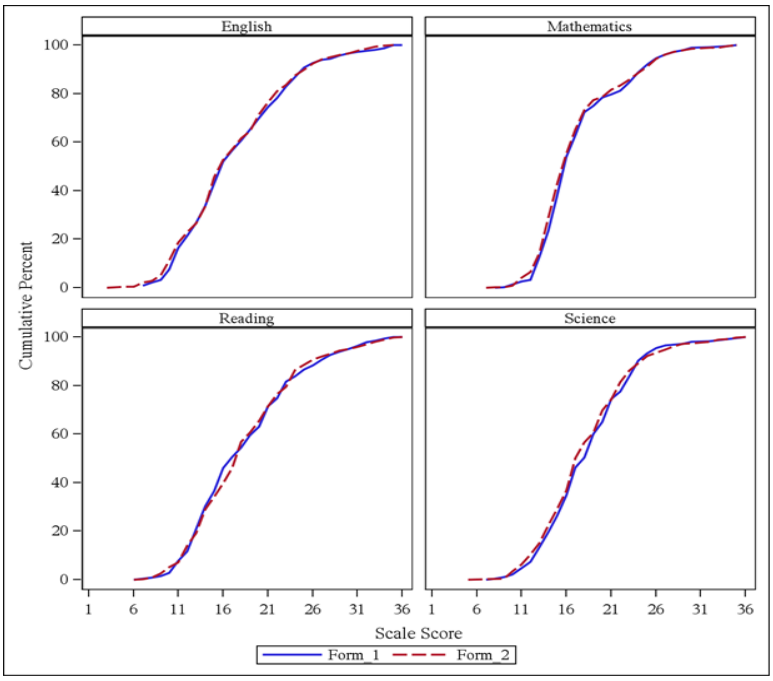


Figure 31. Relative cumulative frequency distribution of scale scores for online testing.

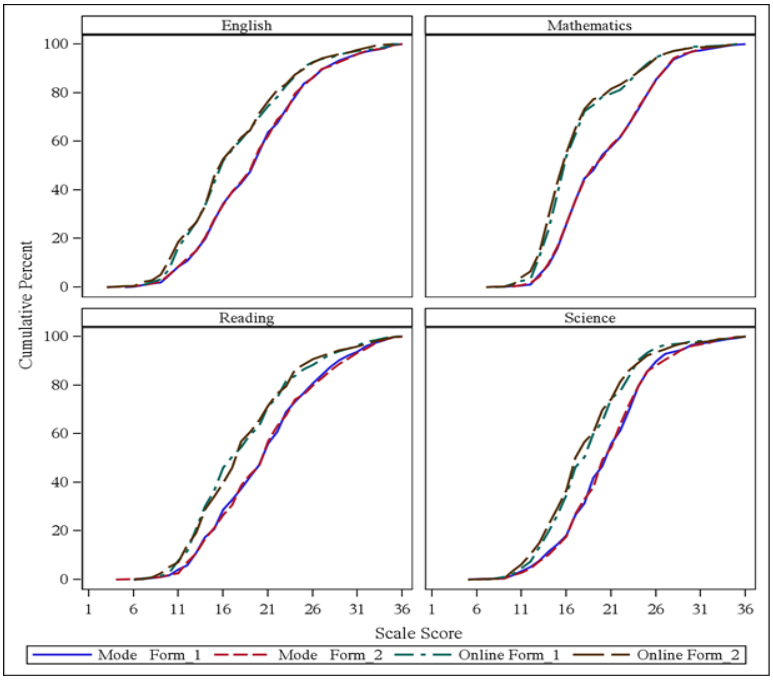


Figure 32. Relative cumulative frequency distribution of scale scores of online testing compared with their distributions in the spring 2015 mode comparability study.

Conclusion and Discussion

The spring 2014 ACT mode comparability study was the first time that the ACT was administered online for operational purposes. The online versions of the reading and science tests had an extra five minutes compared with their paper versions, a decision made based on the results that showed evidence of speededness for those two tests from the fall 2013 timing study. The spring 2014 study examined both item and test level differences across online and paper versions of the multiple-choice tests. Results showed that very small differences were found between the two modes in terms of test reliability, correlations among tests, effective weights, and factor structure. However, item level and test level scores tended to be higher for the online group than for the paper group. Equating methodology was used to adjust for the differences so that scale scores from the two administration mode versions were comparable.

To minimize the potential between-mode differences in forthcoming online administrations, the online timing issue was revisited based on the spring 2014 comparability analyses results, an examination of the item latency information of the two online forms, and the student survey results regarding whether they thought they had enough time to finish each test. Taking into account results from all these analyses and the changes that had occurred to the online administration platform between the fall 2013 timing study and the spring 2014 mode comparability study, it was concluded that, going forward, the extra five minutes for the online reading and science tests should be eliminated.

Another mode comparability study was conducted in spring 2015 in which the online and paper administration time were the same for all tests. Results from the spring 2015 study showed that the mathematics and science test scores were relatively more comparable between online and paper administration modes, while the English and reading scores tended to be higher for the online

versions. The ACT writing test with the enhanced design was administered along with the multiple-choice tests in spring 2015. Online scores on one prompt tended to be higher than the paper scores, but similar to the paper scores on the other prompt. Equating methodology was applied to link the online forms to their paper counterparts for both the multiple-choice tests and the writing test in spring 2015.

In addition to supplying the data for the analyses in this report, the mode comparability studies and the earlier timing study also provided ACT with valuable experience in online administration of the ACT. During the studies, feedback from students and test administrators were collected. Besides questions on the sufficiency of testing time, students were also asked various other questions concerning their preparation for the online testing, computer experience and typing skills, easiness of navigation and use of various features of the online test, their use of scratch papers, their preference of the testing mode, and others. They were also asked to provide any additional comments that they might have regarding their testing experience. Though some students experienced difficulty during the online testing mainly due to technology issues, a larger proportion (53% in spring 2014 and 45% in spring 2015) of the students who took the online tests expressed preference of online testing over paper testing than those who expressed preference of paper over online (33% in spring 2014 and 26% in spring 2015). Analyses of the feedback, together with experiences gained in dealing with the various issues encountered, are valuable resources that ACT can utilize in creating optimal online testing experiences for examinees while maintaining the comparability of scores to the paper versions for future online administrations.

References

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. SAGE Publications.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29(4),* 285-307.

Leeson, H. V. (2006). The mode effect: a literature review of human and technological issues in computerized testing. *International Journal of Testing, 6(1),* 1-24.

Lord, F. M. (1965). A strong true score theory, with applications. *Psychometrika, 30,* 239-270.

Lottridge, S, Nicewander, A., Schulz, M., & Mitzel, H. (2008). *Comparability of paper-based and computer-based tests: A review of the methodology.* Paper submitted to the CCSSO Technical Issues in Large Scale Assessment Comparability Research Group. Available from https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxwYXBlcnZlcnN1c3NjcmVlbnxneDoxNTU0NmE0NDY0NTQ4MzA4

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Mutler, P. (1996). Interface design and optimization of reading of continuous text. In H. van Oostendorp & S. de Mul (Eds.), *Cognitive aspects of electronic next processing* (pp. 161-180). Norwood, NJ: Ablex.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer-Verlag.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment,* 2(6). Available from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1666/1508

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71(5),* 849-869.