ACT WORKING PAPER 2015-05

Spring 2014 ACT® Test Mode Comparability Study

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

April, 2015

Dongmei Li, is a Principle Psychometrician at ACT providing psychometric support for the ACT® test with specialties in educational measurement theories and practices, test equating and scaling, and growth modeling.

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

Deborah Harris, is a Chief Research Scientist at ACT, specializing in equating, linking, and the comparability of reported scores.

This report is a summary of work from many. The authors are in debt to David Woodruff, Tony Thompson, Yu Fang, Yang Lu, Wei Tao, Zhongmin Cui, Andrew Mroch, J.P. Kim, Tianli Li, Lu Wang, Qing Xie, and many others in the Measurement Research Department who participated in various components of the study. Their unique contributions and diligent work made this study and report possible. The authors would also like to thank all other ACT departments that were involved in the study and Creative Services for their editorial work on the report.

Executive Summary

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

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.

Data and Design

The four tests were administered online to approximately 3,000 examinees, with each examinee responding to one test. Students taking each subject 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. The three testing time limits and the four tests produced 12 different combinations with about 250 examinees in each condition.

Results

Student item and test level scores, item omit 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 Reading and Science would be more likely to be comparable to paper administration scores with an increase in testing time. Given the potential

confounding of motivation and familiarity with the online testing format in the timing study, the final decision was to tentatively increase testing time for the Reading and Science tests by five minutes and continue to evaluate the timing issue in the spring 2014 mode comparability study.

Spring 2014 Mode Comparability Study

To increase student motivation and to learn about administration issues, ACT conducted the 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, even if differences were found across mode in terms of item or raw score differences, a random groups design was implemented, allowing equating methodologies to be used to adjust for differences. The purposes of the mode comparability study were to (1) investigate the comparability of the ACT scores from the two testing modes; (2) adjust for the differences through equating methodology if there was evidence of incomparability of scores; (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.

Design and Data

A random equivalent groups design was used for the spring 2014 mode comparability study for the ACT. Students participating in the study could choose to register for the ACT with Writing or without Writing. 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 ACT forms (two online and one paper) that were administered in the study. (Note, the assignment was similar to distributing spiraled paper booklets.) The study took place in an operational testing environment on one of the ACT national test dates. 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. Proctor comments, phone logs, irregularity reports, and other documents were examined to help identify records whose scores may not be reflecting students' typical performance or standardized conditions so that these records could be excluded from further analyses. Test centers with large discrepancies in form counts across modes were deleted from analyses to better maintain group equivalency.

Results

Analyses were conducted to investigate construct equivalency and establish score comparability between the paper and online versions of the ACT. 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 multiple choice tests, the similarity of item score distributions, such as the average percent correct or item *p*-values, item response distributions and item omit rates were compared. ACT Writing scores were examined conditioning on examinees' English scores. In addition, measurement precision (reliability and conditional standard error of measurement) was compared between modes, and the item latency information for the online test items was also examined. Construct equivalency was examined by comparing the dimensionality and factor loadings and by examining differential item functioning (DIF) between paper and online 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 omit rates tended to be lower for the online group than for the paper group, especially for the Reading and the Science tests. Equating methodology was used 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, no matter under which mode they took the ACT.

The spring 2014 mode study revisited the time limit issue of the online administration of the ACT test as investigated in the fall 2013 Timing Study. In addition to analyses evaluating mode effect as mentioned above, item latency information of the two online forms and student survey results regarding whether they thought they had enough time to finish each test were also examined. Taking into account results from all these analyses and the improvements that had occurred to the online administration platform between the fall 2013 timing study and the spring 2014 mode comparability study and those expected for upcoming online administrations, it was concluded that, going forward, online and paper scores of the ACT would likely be more comparable without the extra five minutes for online Reading and Science. To best ensure the comparability of college reportable scores between online and paper administrations, planning and recruiting for another study in spring 2015 is already under way to further examine the comparability of scores with online and paper administration time being the same for all tests. Results from the spring 2015 study will be reported when they are available.

Spring 2014 ACT Mode Comparability Study

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

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

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 test. That is, to aid the online version development process, studies were conducted to evaluate the comparability of scores from paper and online delivery of the ACT under given conditions of the online delivery system and design and to inform decisions about revisions of the online version to be evaluated in further studies. Due to the intended high-stakes uses of the ACT test scores, if a study involves operational score reporting, scores are adjusted for students participating in the study using equating methodologies when mode effect exists.

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 tests were administered to approximately 3,000 examinees, 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 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. 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 mean and standard deviation (SD) 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 on 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 omit 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 percent of students omitting zero to three or more items under the three timing conditions (i.e., current, plus 5, or plus 10). It shows that 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.

Table 2

Percentage of Students Omitting Zero to Three or More Items

Test	# of Omits	Timing		
		Current	Plus 5	Plus 10
English	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	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	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	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 1 presents the item *p*-values, that is, the percentage of correct answers for each item, under each of the timing condition 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 are 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 are observed in tests with additional time, especially for items near the end of the test.

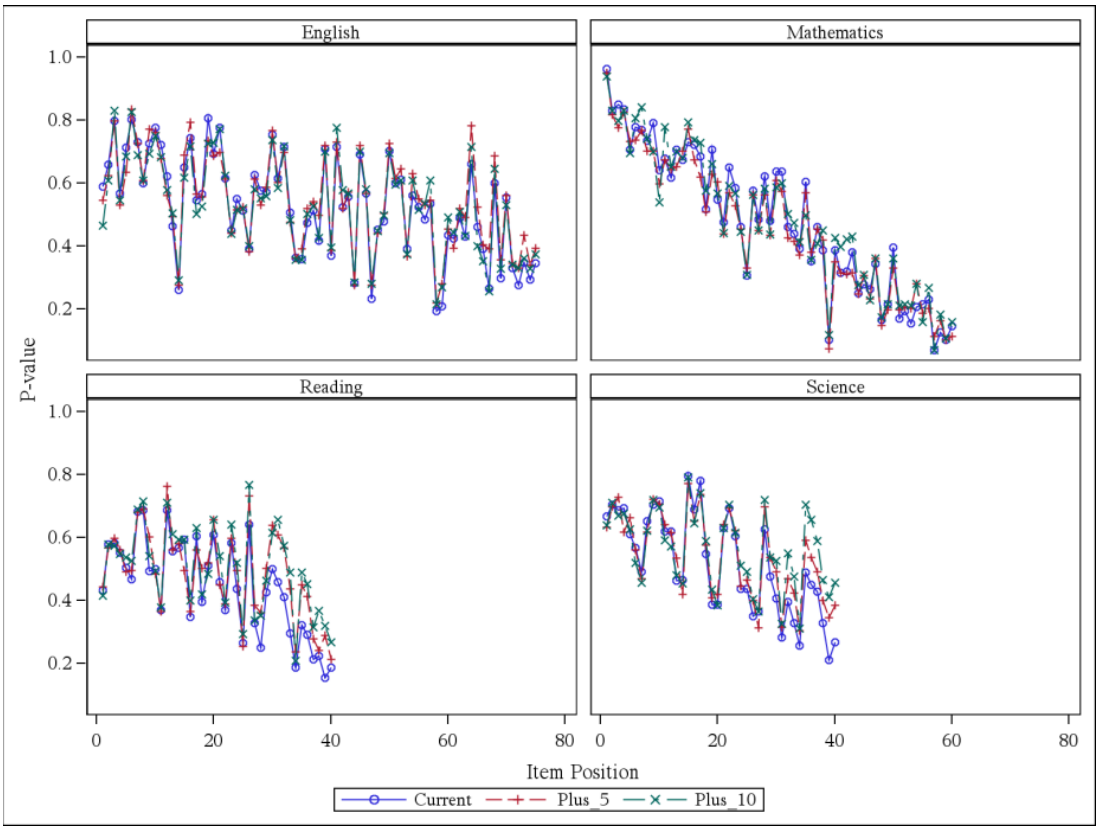

Figure 1. Item *p*-values by item position for the three timing conditions.

Figure 2 presents the percentage of omits for each item for 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 are much higher for the Reading and Science tests than those for the English and Mathematics tests. This is especially true for the Reading test, where the omit rate reached 40% for the last item under the current paper timing limits.

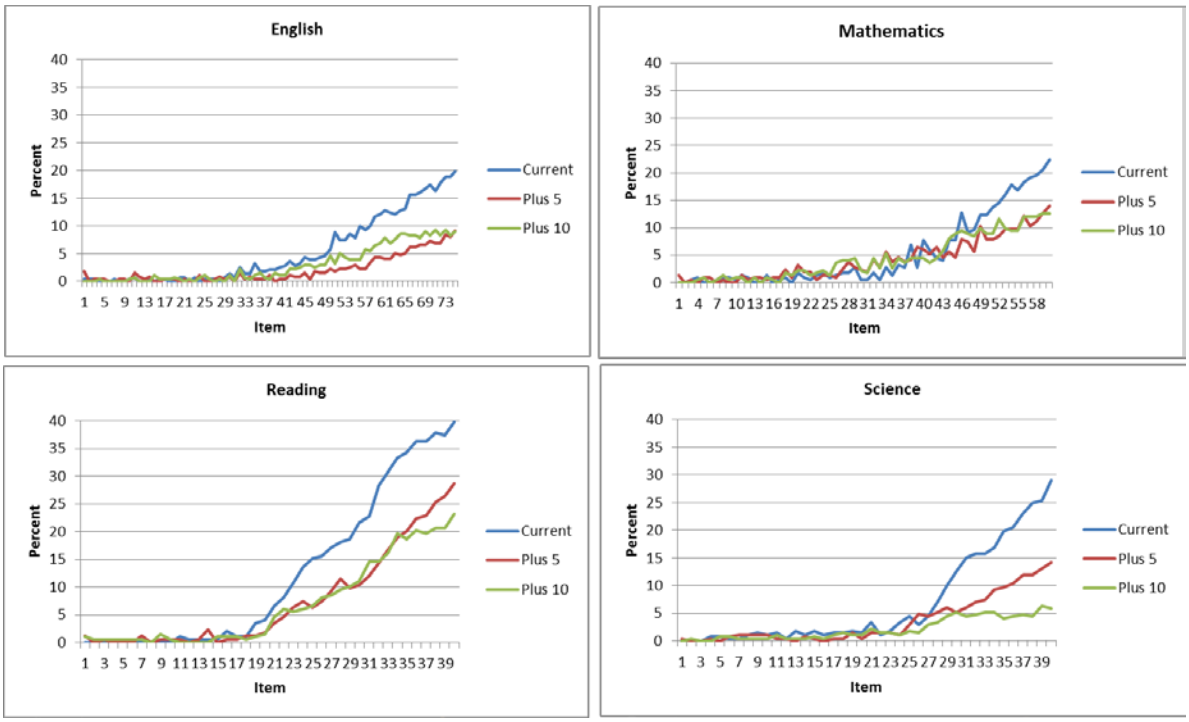

Figure 2. Percent of students omitting items under the three timing conditions.

Figure 3 shows the percentage of students who responded 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 and each test. 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 disagreement on the statement that they had enough time to finish the test. As testing time increased, this disagreement percentage was reduced. However, it was still higher than that for the English and Mathematics tests under the same timing condition.

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 is confounded with issues such as motivation

and familiarity with the online testing format. For example, while the Reading and Science tests show the most speededness in these analyses, it is also true that fewer examinees in those subjects reported watching the orientation videos about how the online testing worked based on the responses to a survey question (See Table 3).

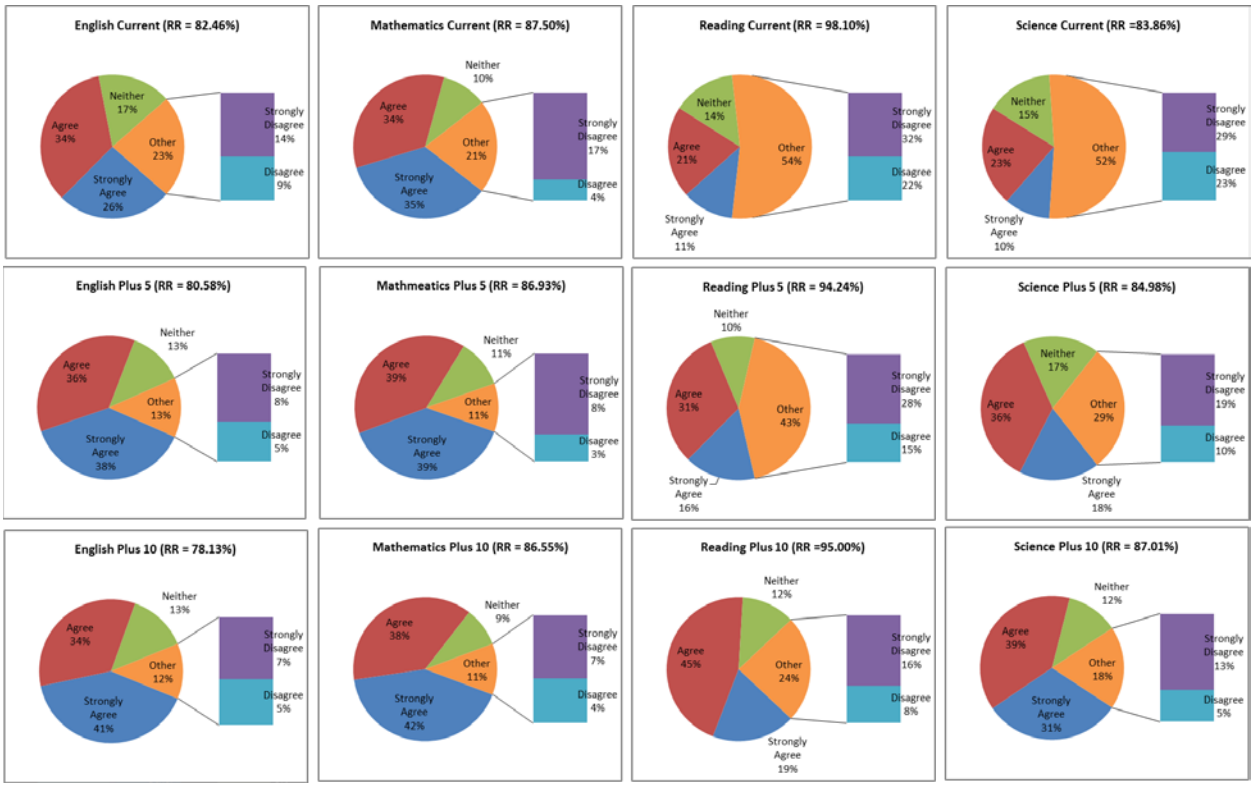

Figure 3. Student responses to survey question about if they had enough time under the three timing conditions.

Table 3

Survey Results for the Question Regarding Online Tutorial Video

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 tests by five minutes and to revisit the time limit issue in the spring 2014 mode comparability study. Because equating had been planned in case of sufficient evidence suggesting mode differences, comparable scores for the examinees in the mode study can be ensured regardless of possible changes in administration time in the future.

Spring 2014 Mode Comparability Study

In the 2014 mode comparability study, as in the fall 2013 timing study, the online versions of the ACT test forms were a strict computer delivery of the paper versions of the tests. That is, the content of the test questions between the online version and the paper version of a test form were intended to be exactly the same, even though some improvements had been made to the online test delivery system based on experiences and feedback from the timing study. The testing time for the paper and online administration were the same for the English and Mathematics tests, but they were different for the Reading and Science tests. As mentioned earlier, five additional minutes were added to the online versions of these two tests based on the recommendation from the fall 2013 timing study. The purposes of this mode comparability study were to (1) investigate the comparability of the ACT scores from the online and paper testing modes; (2) adjust for the differences through equating if there is evidence of incomparability of scores; (3) re-evaluate the timing decisions for the online administration of the ACT tests; and (4) gain insights into the online administration process

Design

A random equivalent groups design was used for the spring 2014 ACT mode comparability study. Students participating in the study were randomly assigned to take one of the three ACT forms that were administered in the 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.

The study took place in operational testing environment on one of the ACT national test dates. Schools with sufficient numbers of computers that met ACT requirements of online testing were recruited to participate in this study. 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 this study.

Only Grade 11 students from participating schools were eligible for the study. These students registered for testing via the normal ACT registration process, and they could register for either the ACT without Writing or the ACT with Writing. 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 that random equivalent groups could be ensured for the Writing 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. Students participating in the study took the ACT for free and received college-reportable scores. After the testing, survey questions were sent to students who participated in the study to ask for their comments and feedback on their testing experience.

Data

More than 7,000 students from about 80 schools across the country signed up for the mode comparability study. As expected, not everyone was able to show up on the day of testing due to various reasons. Computer issues, power outages, and other problems also prevented some

students from testing on that day. Those students were rescheduled to be tested on paper on another test date, and their scores were not included in the analyses. Proctor comments, phone logs, and irregularity reports were examined to help identify records whose scores may not be reflecting students' normal performance so they could be excluded from further analyses.

Group equivalency was to be ensured by the random assignment of students to test forms. After data cleaning, the frequency of students taking each test form was checked again for each test center. Centers with large discrepancies in form counts were deleted to better maintain group equivalency. All subsequent analyses were based on the final cleaned dataset, which contained more than 5,500 students—with at least 1,800 students for each form.

Procedures

Though the content of items on the online versions were intended to be exactly the same as the paper versions, differences in the presentations of these items, such as text font, page size, page layout, graphics, etc., may still exist. Before conducting statistical analyses, ACT first examined comparability of the paper and online versions of the tests through a qualitative comparison of the items in the paper booklets and those in the online version, and potential sources of differential effect were documented.

Mode comparability was examined at two levels: 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, the constructs reasonably can be assumed to be the same, especially with other supporting evidence such as that the test questions are the same on the two modes. However, score equivalency cannot ensure construct equivalency, so additional comparisons to determine whether the two modes measure the same construct is still needed. If score equivalency does not

hold, however, it is a strong indication that scores cannot be used interchangeably for high-stakes decision making such as college admission.

Analyses to evaluate mode comparability were carried out in two phases. Phase I analyses focused more on score equivalency. 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 multiple choice tests, the similarity of item score distributions, such as the average percent correct or item *p*-values, item response distributions and item omission rates were compared. The ACT Writing scores were examined conditioning on examinees' English scores.

Once sufficient evidence was gathered indicating un-equivalency of scores between modes under the random groups design, equating was conducted for the English, Mathematics, Reading, and Science tests to ensure the college reportable scores for students participating in this study from both modes are comparable. Timing decisions were then re-evaluated based on the new evidence gathered from this study.

After that, Phase II mode comparability analyses were conducted, focusing more on construct equivalency as well as some additional analyses, including item and test comparison based on item response theory (IRT), factor analysis, differential item functioning (DIF), generalizability analysis, and evaluation of measurement precision after mode effect was adjusted through equating. Results from Phase II analyses were not expected to have any significant impact on the decisions whether equating was necessary or how timing needed to be adjusted.

The following sections present results from these analyses, in the order of Phase I comparability analyses, equating, timing evaluation, and Phase II comparability analyses. Major findings are summarized in the end.

Phase I Mode Comparability Results for Multiple Choice Tests

Table 4 presents the samples size of each test form in the study as wells as the mean, standard deviation, minimum, and maximum of the observed total raw scores and scale scores of the two online forms (Online_1 and Online_2) and the one paper form (Paper_1). Online Form 1 and Paper Form 1 contained the same test questions, and were the focus of most analyses. They are referred to simply 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 all refer to scale scores obtained by applying the paper conversions regardless of testing mode.*** 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. Final reported scale scores for the online forms are based on conversions obtained through equating to be discussed later.

On average, the Online Form 1 scores tend to be higher than the Paper Form 1 scores for all tests. Though Online Form 2 has higher raw score means than Online Form 1, their scale score means are similar. Phase I comparability analyses for the multiple-choice tests included an examination of test and item level score distributions, test reliabilities, and item omit rates across modes.

Table 4

Descriptive Statistics of Raw and Scale Scores of all Test Forms in Mode Comparability Study

Form	Test	N	Raw Score				Scale Score			
			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-test of mean differences

Scores across modes can be compared either on raw scores or scale scores, with scale scores for the online forms obtained by applying the paper version conversions. 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 may have any practical impact because decisions are usually made based on scale scores.

Figure 4 is a graphical presentation of the raw and scale score mean differences across modes. Mean differences, effect sizes, and *p*-values of 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. For all tests, the online group tended to have higher mean scores than the paper group. Except for the Mathematics test, the mean differences were all statistically significant, for both raw scores and scale scores.

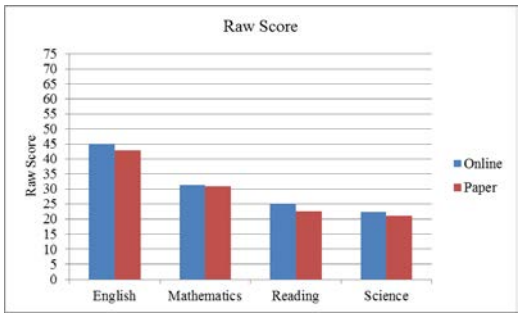

Figure 4(a)

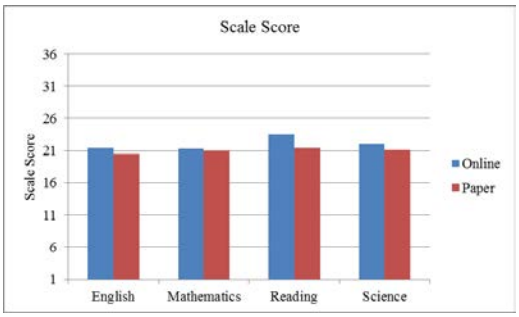

Figure 4(b)

Figure 4. Raw and scale score mean comparison across modes.

Table 5

Raw and Scale Score Mean Differences between Modes (Online minus Paper)

	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 and scale score frequency distributions and cumulative frequency distributions were also compared between modes. The plots of the relative cumulative frequency distributions of raw proportion correct scores and scale scores are shown in Figure 5. Again, scores tend to be higher for the online group than for the paper group for all tests except the Mathematics test.

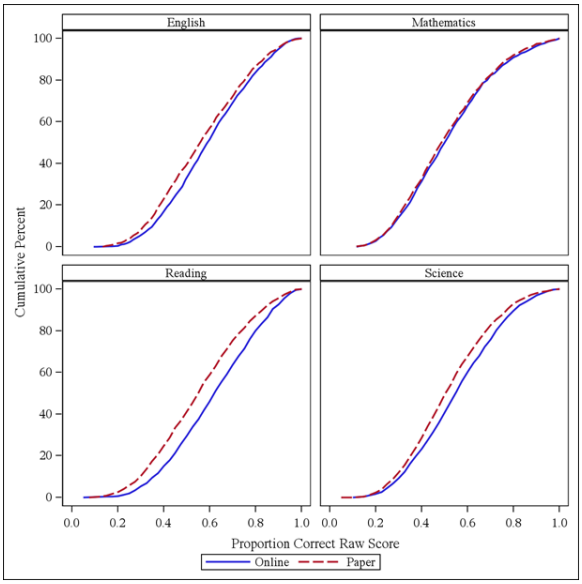

Figure 5(a)

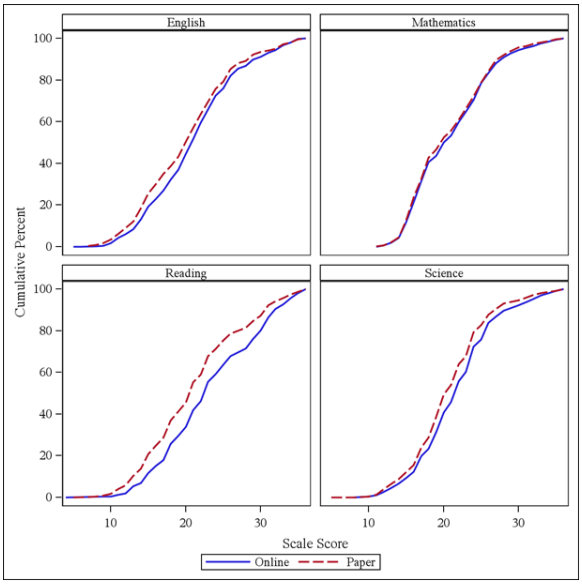

Figure 5(b)

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

The Kolmogorov-Smirnov (KS) test of equivalency of distributions was conducted for the raw and scale scores for each test. Similar with results of the t-tests of mean differences, the KS tests showed that the between-mode raw and scale score distributions were statistically significant for all subject tests ($p < .001$) except for the Mathematics test ($p = .40$).

Correlations, effective weights, and Cronbach alpha

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

Table 6

Scale Score Correlations, Effective Weights, and Cronbach Alpha

		Online Form				Paper Form			
		English	Mathematics	Reading	Science	English	Mathematics	Reading	Science
Correlations	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 Weights		.26	.22	.28	.23	.28	.22	.28	.22
Cronbach Alpha		.93	.92	.87	.86	.93	.92	.87	.86

P-values, omit rates, and option analysis

Item difficulty was compared across modes. Figure 6(a) presents the proportion of correct responses for each item (item *p*-values) by item position across modes, with smaller values indicating harder items. Figure 6(b) presents the *p*-value differences between modes, with positive difference indicating that the item was easier for the online administration. Figure 6 shows that whereas 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 later in the test.

Consistent with the effect size differences observed in Table 5, the *p*-value differences are smallest for the Mathematics test, but largest for the Reading test. 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 regarding item position. 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.

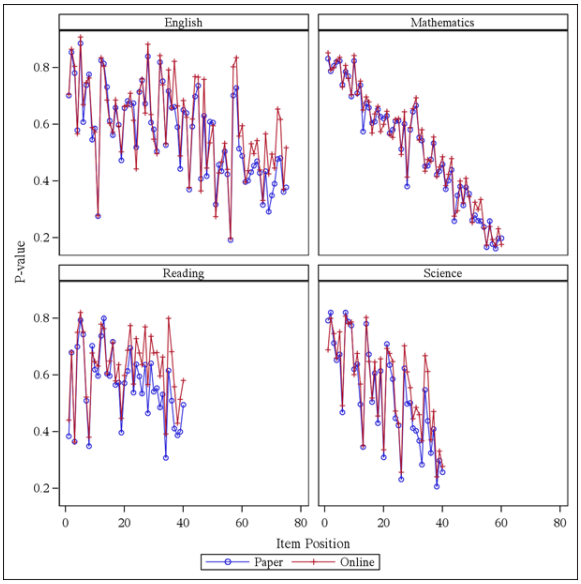

Figure 6(a)

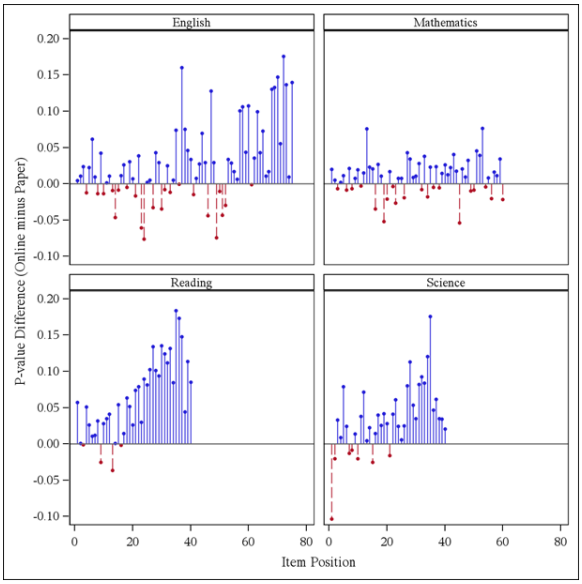

Figure 6(b)

Figure 6. Scatter plots of item *p*-values and needle plots of *p*-value differences between modes.

Omit rates, that is, 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 omit 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.

Equating

Due to the differences observed between the paper and online scores as discussed above, equating was conducted for the multiple-choice tests so that the college reportable scores are comparable regardless under which mode and time limits the student took the tests. Consistent with the methodology used for equating paper forms of the ACT, equipercentile equating with post smoothing was used to equate the online test forms to the paper form, based on the random equivalent groups design.

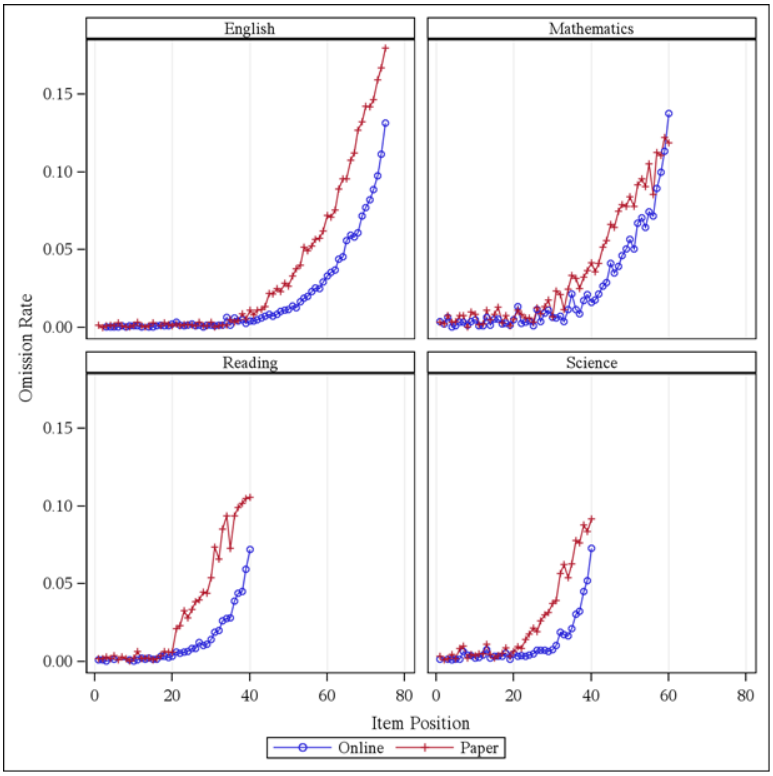

Figure 7. Item omit rates by item position.

Equating 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. Figure 8(a) plotted the raw to scale score

conversions for the two online forms together with their counterpart paper conversions, and Figure 8(b) plotted 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 is converted to a lower scale score in the online conversion than in the paper conversion. 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.

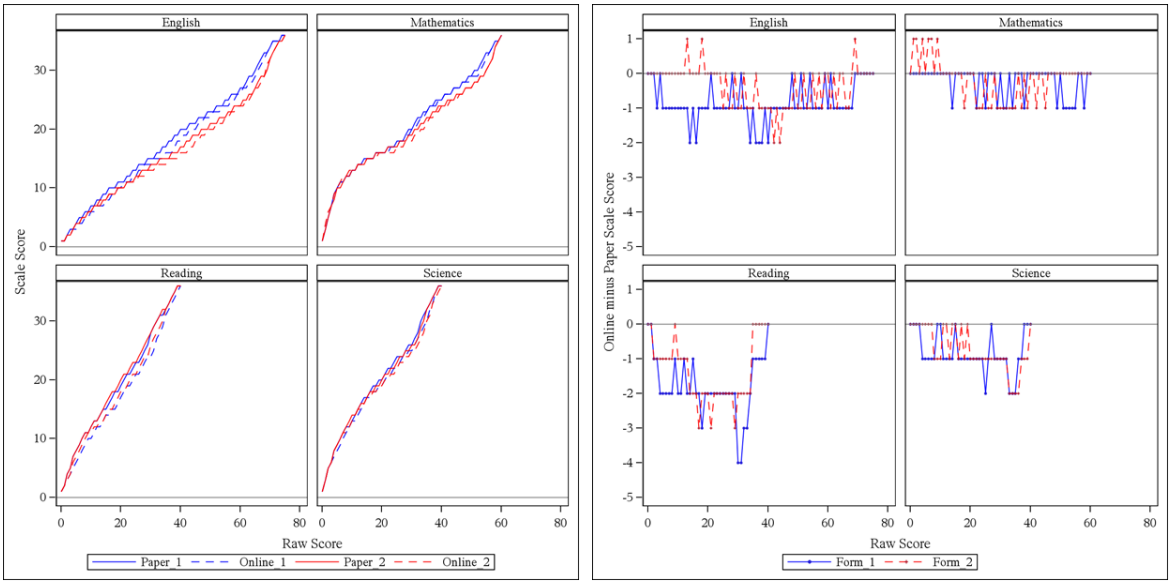

Figure 8(a) Figure 8(b)

Figure 8. Conversion tables before and after adjusting for mode effect.

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 of scale scores based on the different conversions were also calculated for each student in the dataset as the scale score under the online conversions minus the scale score under the paper conversions. The magnitude of the difference scores indicates the amount of adjustment for mode effect after equating. Figure 9 presents the distribution of these difference scores. For the English, Mathematics, and Science tests, the adjustment was within one score point for the majority of

students. For Reading, however, the adjustment was two scale score points for the majority of students.

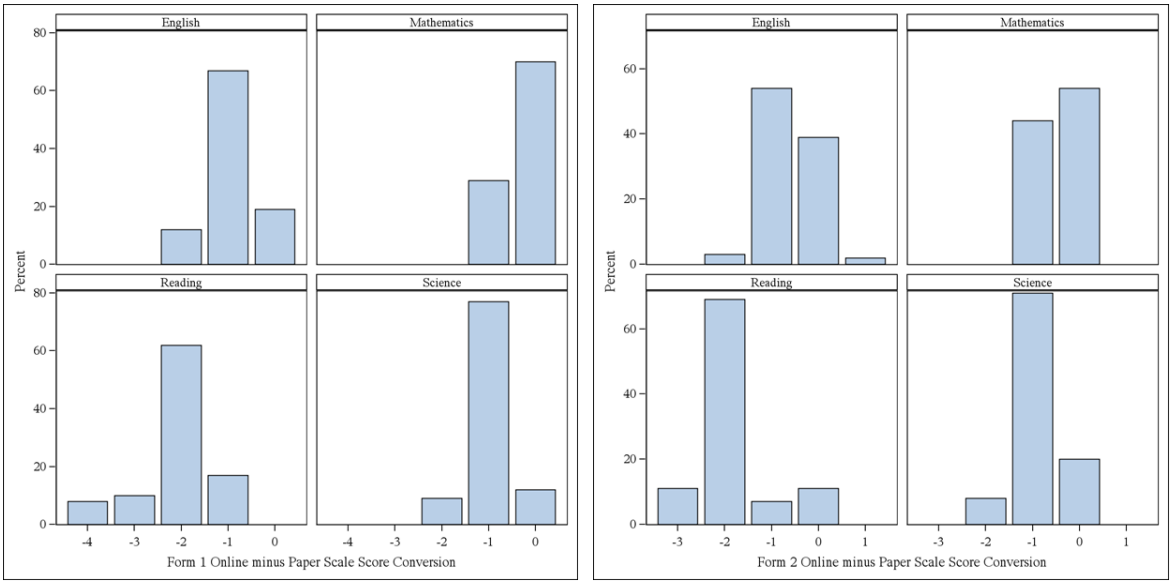

Figure 9. Distributions of scale score adjustment for the two online forms.

ACT Writing Test

More than half of the students in the final data set took the ACT with Writing. The mode effect for the Writing test was examined after equating was done for the multiple choice tests, by comparing the online and paper Writing mean scores and by comparing the conditional Writing scores after controlling for the English scale scores.

Table 7 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 paper and online 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 Writing versus the paper Writing. The effect size of between-mode group difference for English was small, and

the t-test of mean difference was not 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.

Table 7

Between Mode Comparison for Students Taking the ACT Writing

	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 of Writing between 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 10(a) is a scatter plot of the online and paper Writing scores against students' English scale scores, and Figure 10(b) presents the conditional mean Writing scores for each mode. 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 to the ACT Writing scores.

Computer-Based Testing (CBT) Timing Re-evaluation

As pointed out earlier, the online administration in the mode comparability study added five minutes to the current paper administration time for the Reading and Science tests based on

recommendations from the fall 2013 timing study. However, the timing study had limitations that made it necessary to continue to gather information to inform the timing decisions, which was indeed one of the purposes of this mode comparability study. Results from the spring 2014 mode comparability study indicated that, between randomly equivalent groups of students, the online scores tended to be higher than the paper scores, especially for the Reading and Science tests.

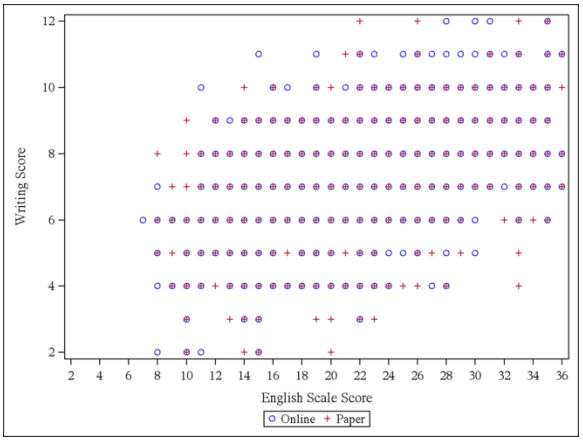

Figure 10(a)

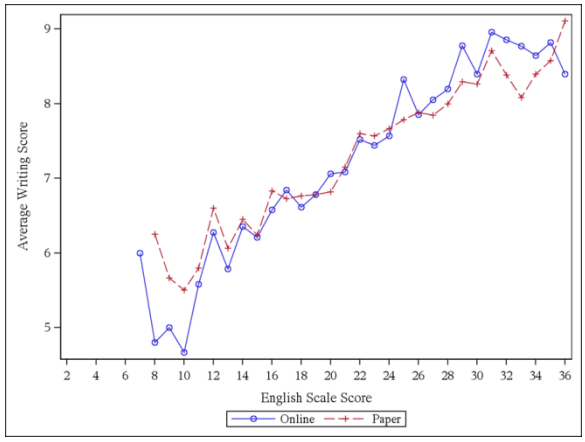

Figure 10(b)

Figure 10. Writing scores conditioning on English scale scores.

Since the mode comparability study was conducted in an operational testing environment with a paper control group, this study provided information for timing decisions that were less confounded than the earlier timing study. Results from analyses presented in previous sections were considered together with the following additional information from student survey and online item latency information to re-evaluate the online timing decisions.

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 completed the survey, among which about two thirds of the students took the online versions of the tests.

Table 8 presents the 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 than for the paper administration.

Table 8

Student Survey Results for the Timing Related Question

I felt I had enough time to finish the…	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Either agree or strongly agree	Either disagree or strongly disagree	Viewed item but did not respond	Mean	SD	N
Online											
English	40	39	7	9	4	79	12	1	4.05	1.08	1,031
Mathematics	21	32	13	24	9	53	33	2	3.33	1.30	1,027
Reading	21	34	12	21	10	55	31	2	3.35	1.30	1,022
Science	15	32	17	23	11	47	34	2	3.18	1.27	1,024
Writing	19	28	16	18	14	48	32	4	3.22	1.35	635
Paper											
English	27	36	11	18	7	62	25	1	3.58	1.26	447
Mathematics	14	34	12	25	14	49	38	2	3.11	1.31	446
Reading	11	26	12	30	18	37	48	2	2.81	1.31	442
Science	9	29	15	29	17	38	45	2	2.84	1.27	445
Writing	29	42	11	9	7	71	17	1	3.77	1.19	264

Online form item response time

Item latency information was examined for the two online forms. Figure 11 presents the average time spent on each item for all four subject tests. If the time spent on the last few items of each test was significantly less than on the other items, then the test may be speeded. However, no such evidence was found for the online tests. Note that the peaks in the tests are usually the first item associated with a passage, which included the time spent reading the passage.

Online timing decision

Based on the results from comparability analyses, equating, student survey, and item latency information, it was decided that the extra five minutes for the online administration of Reading and Science be removed, resulting in testing time being the same for all tests administered, whether paper or online. The comparability of scores between paper and online with this change in online administration time will be further evaluated in future studies.

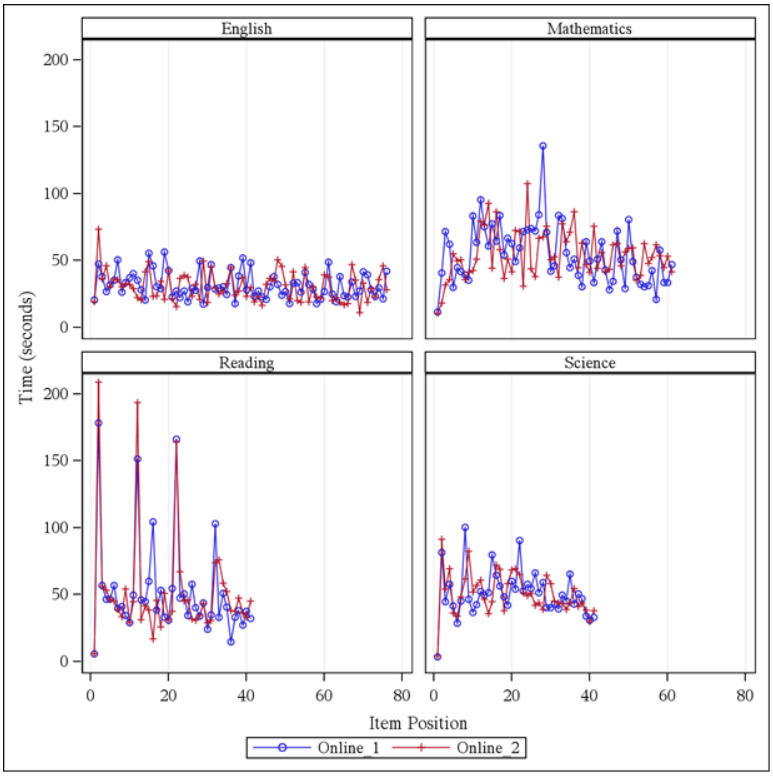

Figure 11. Average time spent on each item.

Phase II Mode Comparability Analyses and Results

IRT analysis

Even though item response theory (IRT) has not been used for the current operational scoring of the ACT tests for college reporting, it may be used for other purposes. Mode effects were also examined under IRT at the test and item level by comparing the test characteristics curves and item parameters across modes, using the three-parameter logistic model.

Figure 12 contains plots of the test characteristic curve (TCC) comparison for each subject. Consistent with the patterns observed in Figure 5 for the raw and scale score relative cumulative frequency distributions, the between-mode TCC difference is smallest for the Mathematics test, but largest for Reading.

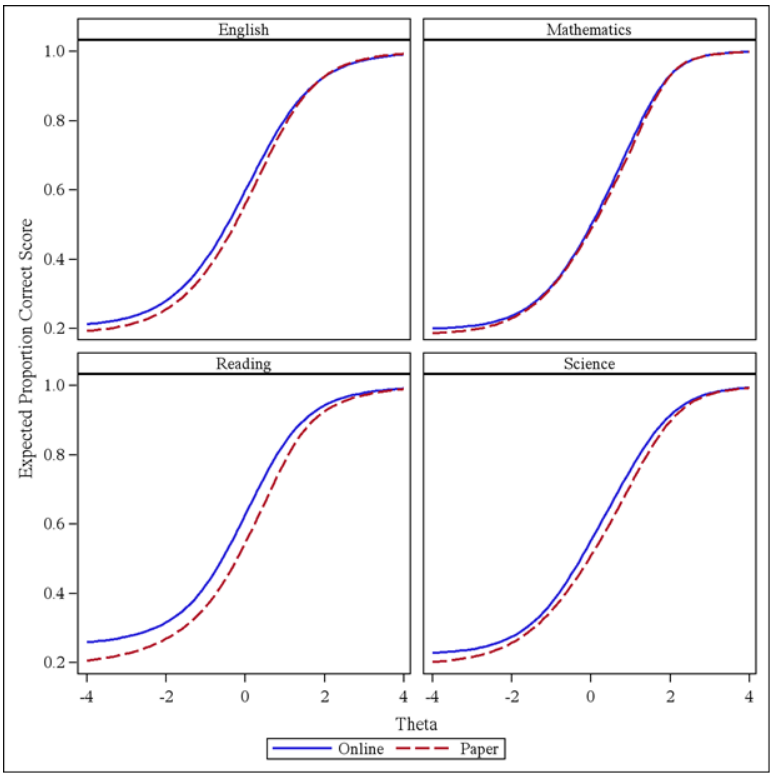

Figure 12. Test characteristic curves across modes.

Scatter plots of item parameter estimates from paper and online are presented in Figure 13, with parameters of *a*, *b*, and *c* plotted in Figure 13(a), Figure 13(b), and Figure 13(c), respectively. Also consistent with the comparison of item *p*-values, 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 also tended to be higher for the online items, which may indicate that low-performing students had a higher chance of answering the online items correctly than the paper items.

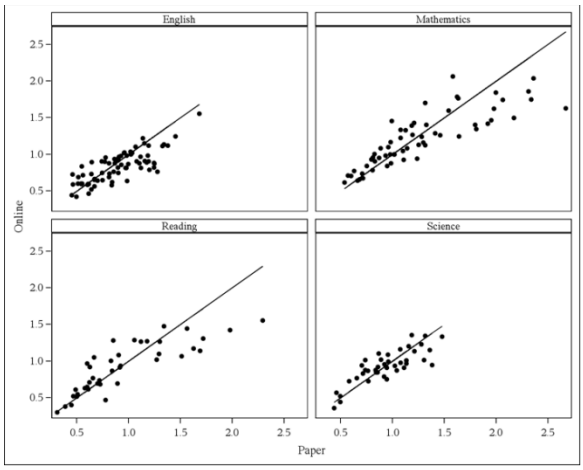

Figure 13(a) *a*-parameter

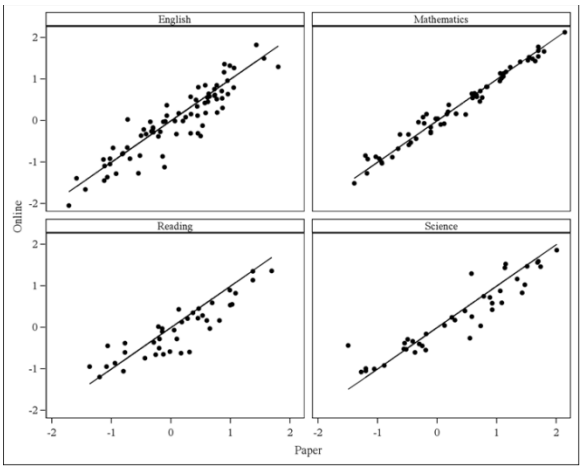

Figure 13(b) *b*-parameter

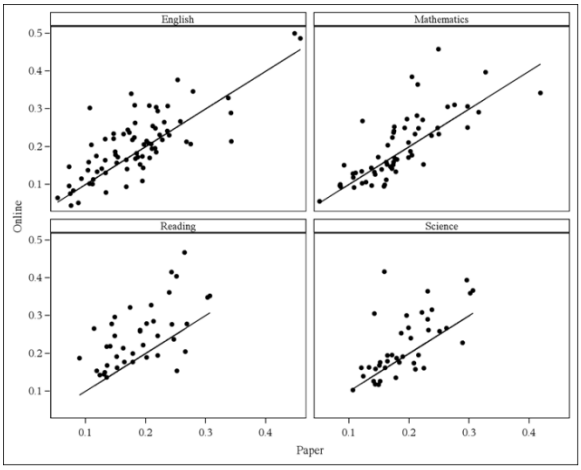

Figure 13(c) *c*-parameter

Figure 13. Between mode IRT parameter comparison.

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 subject test were examined across modes and they all showed that both online and paper tests measured an essentially unidimensional construct. In addition, the data were fit with both a one-factor and a two-factor model. Table 9 presents the criteria used for evaluating model fit and Table 10 contains several fit indices resulting from fitting the one- and two-factor models for the four tests

across modes. All statistics indicated good model fit for the one-factor model expect 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. Based on the principle of parsimony, the one factor model was considered to be adequate and the factor loadings of each test on the one factor were compared across modes. Table 11 presents the descriptive statistics of the factor loadings of each mode and the correlations of the factor loadings between the two modes.

Table 9

Criteria for Good Model Fit

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

Generalizability analysis

Raw score reliability was further examined using generalizability theory based on the content specifications of the tests from both a univariate and a multivariate perspective, with content considered as a facet within which items are nested or with the different content categories considered as different variables. Analyses from these two perspectives produced identical results for the generalizability coefficients ($E\rho^2$) and dependability indices or phi coefficients (Φ). These coefficients are reported in Table 12, together with the Cronbach alpha reliability already reported in Table 6 to facilitate comparison. Whereas the phi coefficients are

slightly lower than the generalizability coefficients, as expected, these values are also very similar to the alpha reliability. Similar with alpha, reliability indices from the generalizability analyses showed barely any differences across modes.

Table 10

Fit Statistics of One- and Two-Factor Model

		Online_1			Paper_1		
Test	Fit Statistic	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 11

Descriptive Statistics and Correlation of Factor Loadings between Two Modes

Test	Form	Mean	SD	Minimum	Maximum	Correlation
English	Online_1	0.51	0.11	0.25	0.70	
	Paper_1	0.52	0.11	0.26	0.75	.88
Mathematics	Online_1	0.53	0.10	0.25	0.72	
	Paper_1	0.52	0.11	0.26	0.70	.90
Reading	Online_1	0.50	0.12	0.22	0.73	
	Paper_1	0.49	0.11	0.26	0.76	.87
Science	Online_1	0.48	0.13	0.23	0.71	
	Paper_1	0.47	0.11	0.26	0.69	.87

Table 12

Raw Score Generalizability Coefficient, Phi Coefficient, and Alpha

	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

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 analyses results were also compared across modes. No noteworthy differences were found except that correlations among the content categories for the online tests tended to be higher than the paper versions for Reading and Science.

Differential item functioning

The purpose of conducting differential item functioning (DIF) analysis was to examine whether there were some items that function significantly 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 were used as a basis for judging the practical significance of the statistically identified items.

The qualitative comparison documented differences between modes that may or may not 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 does not have any effect on students' performance. Other differences may or may not affect performances. For example, the paper version may have the entire passage or entire set of tables and figures visible on a single page whereas online may need scrolling, and the online version used highlighting but paper used underlying or reference to line numbers. Items that were potentially affected by these differences were identified.

Note also that the random equivalent groups design of this study ensured group level equivalency of overall proficiency level, and thus the item *p*-value differences as well as omission rates comparison across modes as presented in Figure 6 and Figure 7 all contribute to the understanding of item DIF. In addition, DIF was examined by comparing the transformed item index or the delta plot method and the Mantel-Haenszel procedure (Camilli & Shepard, 1994; Mantel & Haenszel, 1959).

The transformed item difficulty index converts the *p*-values to normalized z-scores corresponding to the $(1\text{-}p)^{th}$ percentile to remove curvilinearity in the relationship between the

two sets of *p*-values so that very easy and very hard items do not always tend to produce the smallest differences due to floor or ceiling effect.

The Mantel-Haenszel procedure calculates the odds ratios of the weighted average of item proportion correct scores conditioning on each level of the overall ability. In this study, items with odd ratio values smaller than 0.5 or larger than 2 were flagged for further review. In addition, different scores were available as indicators of the overall ability—test scores before and scores after equating that adjusted for the mode effect, with scores after equating being better indicators of student ability. When controlling for raw or scale scores before equating, two English items, one Reading, and one Science items were flagged, and a few more were flagged when controlling for scale scores after equating. These items were those with the largest *p*-value or delta 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 equating

Scale score properties after equating 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 (1963) four parameter beta compound binomial model for raw scores. Table 13 presents the scale score moments, SEM, and reliability of each form, and Figure 14 contains plots of the conditional SEM of each true scale score point for all three forms.

Table 13

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

Test		Mean	SD	Skewness	Kurtosis	SEM	Reliability
English	Paper_1	20.47	6.12	0.29	2.62	1.71	0.92
	Online_1	20.47	6.13	0.32	2.63	1.76	0.92
	Online_2	20.43	6.12	0.26	2.61	1.72	0.92
Mathematics	Paper_1	21.02	5.15	0.56	2.59	1.58	0.91
	Online_1	21.01	5.21	0.58	2.60	1.57	0.91
	Online_2	21.02	5.18	0.57	2.57	1.45	0.92
Reading	Paper_1	21.47	6.43	0.32	2.37	2.29	0.87
	Online_1	21.48	6.46	0.30	2.35	2.35	0.87
	Online_2	21.56	6.49	0.32	2.36	2.41	0.86
Science	Paper_1	21.14	5.03	0.40	3.29	2.11	0.82
	Online_1	21.15	5.12	0.41	3.28	2.05	0.84
	Online_2	21.07	5.07	0.40	3.24	1.94	0.85

Conclusions and Discussion

The spring 2014 ACT mode comparability study was the first time that the ACT was administered online for operational purposes. Administration time for the online versions had five more minutes for the Reading and Science tests than the paper versions, a decision based on the results from the fall 2013 timing study that showed evidence of speededness for the Reading and the Science tests.

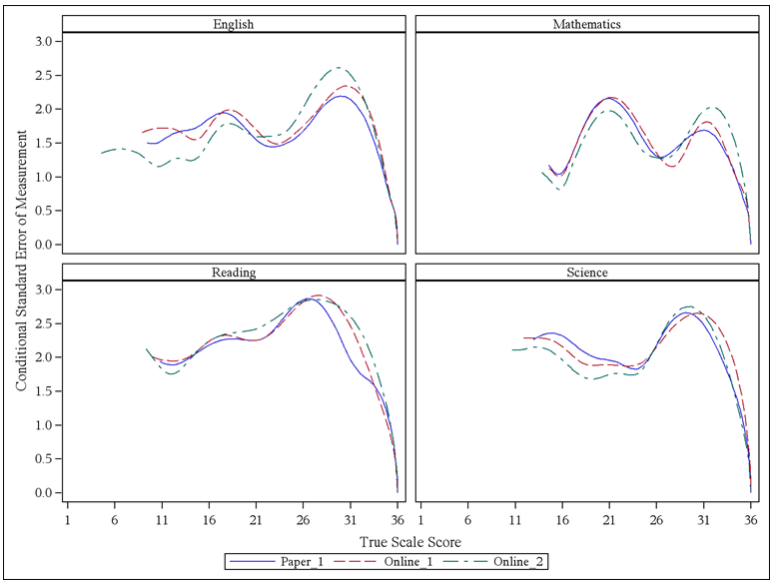

Figure 14. Conditional standard error of measurement.

The study examined both item and test level differences across paper and online versions of the tests. Results showed that although no difference was found between the two modes in terms of test reliability, correlations among tests, effective weights, factor structure, etc., item scores and test scores tended to be higher for the online group than for the paper group. Equating was conducted 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 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, online and paper scores of the ACT would probably be more comparable without the extra five minutes for online Reading and

Science. To best ensure the comparability of college reportable scores between online and paper administrations, planning and recruiting for another study in spring 2015 is already under way to further examine the comparability of scores with online and paper administration time being the same for all tests. Results from the spring 2015 study will be reported when they are available.

In addition to supplying the data for the analyses in this report, the mode comparability study and the earlier timing study also provided ACT with valuable experience in online administration of the ACT. During both studies, feedback from students and test administrators were collected. Besides questions on the sufficiency of testing time, students were also asked various other questions, including their preparation for the online testing, computer experience and typing skills, easiness of navigation and use of various features of the online test, use of scratch papers, and their preference of the testing mode, etc. They were also asked to provide any additional comments that they had regarding their testing experience. Though some students experienced difficulty during the online testing mainly due to technology issues, a larger proportion (53%) 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%). 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.

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

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.

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://www.jtla.org

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