Students Who Take ACT Test Sections in a Different Order Earn Similar Scores

Krista Mattern, PhD, and Jeffrey T. Steedle, PhD

ACT Research Report 2020-1

ACT

ABOUT THE AUTHORS

Krista Mattern, PhD

Krista Mattern is a senior director of Applied Research and Services, which comprises validity and efficacy research, workforce research, and navigation research. Her research focuses on predicting education and workplace success through evaluating the validity and fairness of cognitive and non-cognitive measures. Krista is also known for work in evaluating the efficacy of learning products to help improve intended learner outcomes.

Jeffrey T. Steedle, PhD

Jeffrey Steedle is a director in Assessment Transformation leading the team responsible for statistical analyses for the ACT test and guiding research studies related to maintaining measurement quality while making changes to the assessment program. assessment programs.

ACKNOWLEDGEMENTS

We thank Benjamin Andrews for his contributions to this study, including data analysis and the initial write-up of findings. We also thank Wayne Camara and Melinda Taylor for their feedback and comments on earlier versions of the manuscript.

Conclusions

This report investigates the effects of changes to the standard administration of the ACT® test through three studies: a paper order study, an online paper study, and an online modular study. Overall, the results suggest that students perform the same regardless of the testing order.

So What?

Starting in 2021, ACT will offer section retesting on the ACT test. The results of these studies provide initial evidence of score comparability for this new test option. As such, higher education institutions can rest assured that ACT scores obtained via section retesting are scores that they can continue to trust and use in the college admissions process.

Now What?

While these studies provide initial evidence of score comparability for section retesting, ACT is committed to monitoring the degree to which the current findings generalize to an operational setting once section retesting becomes available.

Contents

Introduction

The purpose of this set of three studies was to gather information about the effects of changes to the standard administration of the ACT® test. Currently, all students completing the ACT are required to take the full ACT test in a fixed order—English, math, reading, and science. However, starting in 2021, students will have the opportunity to retest in a single section (i.e., section retesting). In particular, students who have completed the full test can sign up to take up to three subject areas through section retesting and choose the order in which those sections are administered. Section retesting should provide greater access to retesting among students who may have difficulty committing several hours to test on a Saturday. The three studies in this paper investigated the effects of changing the ordering of the section tests within the full ACT test and the feasibility of modular administrations in which students take only one section test at a time.

The first study, referred to as the paper order study, took place in the fall of 2015. Each student in the study took the full ACT test, with some students taking the test in an order different from the standard order of English, mathematics, reading, science. There were four different orders in total, including the standard order. Each of the four orders had a different test in the first position. When a particular test is taken in the first position, it could be considered comparable to taking that test individually in a modular format. A similar study in which students took the ACT online was conducted in the spring of 2016. The ACT was administered in four different orders in the online order study, as was done in the paper order study. The third study, which was completed in the fall of 2016, involved an actual modular administration of the ACT in which students took one section test from the full ACT test per day over four days. The focus of the report was to examine the extent to which the order of the tests within the full test affected test scores. This was evaluated at the item level and the overall test level.

Methodology

Both test- and item-level analyses were conducted across each study to evaluate the effects of test order on scores. The test-level analyses included descriptive statistics, statistical tests, and item response theory (IRT) analyses. At the item level, the differences in item difficulty (*p*-values—or proportion correct), differential item functioning (DIF) analyses, and item latency information (for the online order study only) were all considered.

The test-level analyses in this report were conducted on the 1–36 scale scores generated from the original raw-to-scale score conversion tables (based on the standard test order administration). It was possible to examine the effects of test order by comparing scale scores (from the four test order conditions) that were all based on the original conversion.

IRT Methods

The following is a brief description of the processes used to evaluate the tests taken in different orders using IRT. First, three-parameter logistic (3PL) item parameters for each subject were estimated separately for all four orders using BILOG-MG (Zimowski et al., 2003). Because randomly equivalent groups were used, the item parameters were considered to be on the same scale, and no additional scale transformations were applied.

Methods described by Kolen, Zeng, and Hanson (1996) were used throughout the study. These methods use IRT to estimate properties of scores such as conditional standard errors of measurement (CSEMs) and reliability.

Conditional on each θ value (i.e., student ability), the distribution of raw scores can be calculated using the recursion formula provided by Lord and Wingersky (1984). The mean of this conditional distribution is the expected raw score, and the standard deviation is the raw score CSEM. This distribution can also be used to calculate the expected scale score and the scale score CSEM. The expected scale score can be expressed as

$$\xi(\theta) = \sum_{i=0}^{k} S(i) \Pr(X = i \mid \theta) \qquad \textbf{1}$$

where k is the total number of items, $\Pr(X = i \mid \theta)$ is the proportion of students at score point i, and $S(i)$ is the scale score that corresponds to score i. The CSEM is written as

$$\sigma^2\left[S(X) \mid \theta\right] = \sum_{i=0}^{k} \left[S(i) - \xi(\theta)\right]^2 \Pr(X = i \mid \theta). \qquad \textbf{2}$$

This model was also used to calculate the scale score reliability. The average error variance is written as

$$\sigma^2(E_S) = \int_{\theta} \sigma^2\left[S(X) \mid \theta\right] \psi(\theta) d\theta \qquad \textbf{3}$$

where $\psi(\theta)$ is the proportion of examinees in the population with that θ value. The population raw score distribution is written as

$$\Pr(X = i) = \int_{\theta} \Pr(X = i \mid \theta) \psi(\theta) d\theta \qquad \textbf{4}$$

This distribution can then be converted to scale scores. The variance of that scale score distribution is used in the formula for reliability. The scale score reliability is expressed as

$$rel_S = 1 - \frac{\sigma^2(E_S)}{\sigma^2\left[S(X)\right]}. \qquad \textbf{5}$$

Differential Item Functioning

DIF analyses were conducted to determine whether average scores on items differed significantly due to test order. Specifically, the Mantel-Haenszel procedure (Holland & Thayer, 1988) was used to investigate whether there were differences in item performance between the standard order test and the alternative orders. For the DIF analyses, examinees were grouped together based on their reported scale score from equating. The students who took the standard order test were the reference group, and those who took the alternative orders were the focal groups.

The Mantel-Haenszel test statistic is written as

$$\hat{\alpha}_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j} \qquad \textbf{6}$$

where T_j is the total number of examinees, A_j is the number of examinees in the reference group who got the item correct, B_j is the number of examinees in the reference group who got the item incorrect, and C_j and D_j are the numbers of examinees in the focal group who got the item correct and incorrect, respectively. The index for ability level (scale score) is represented by j. This statistic is then transformed to the delta scale,

$$\text{MH D-DIF} = -2.35 \ln(\hat{\alpha}_{MH}) \qquad \textbf{7}$$

which has a standard error of

$$SE_{\text{MH-D-DIF}} = 2.35\left[1/(2U^2)\sum_j T_j^{-2}\left(A_j D_j + \hat{\alpha}_{MH} B_j C_j\right)\left(A_j + D_j + \hat{\alpha}_{MH}(B_j + C_j)\right)\right]^{1/2} \qquad \textbf{8}$$

where $U = \sum_j A_j D_j / T_j$

Positive values of the MH D-DIF statistic indicate DIF favoring the focal group (i.e., unusually high performance in the alternative test order group compared to the standard test order group), and negative values indicate DIF favoring the reference group (i.e., unusually high performance in the standard test order group compared to the alternative test order group).

The Mantel-Haenszel chi-squared statistic is used to test the statistical significance. It is expressed as

$$\frac{\left(\left|\sum_j A_j - \sum_j (A_j + B_j)(A_j + C_j)/T_j\right| - 1/2\right)^2}{Var(A_j)} \qquad \textbf{9}$$

where $Var(A_j) = \frac{(A_j + B_j)(A_j + C_j)(C_j + D_j)(B_j + D_j)}{T_j^2 (T_j - 1)}$. This follows a chi-squared distribution with one degree of freedom under the assumption of a constant odds ratio.

The ETS classification system was used to classify items into three different categories (Dorans & Holland, 1993). Category A items are considered to have minimal or nonsignificant DIF. Items are classified as Category A items if they have MH D-DIF less than an absolute value of one or if the Mantel-Haenszel chi-squared statistic is not statistically significant. Category B items are considered to have moderate DIF. An item is classified as Category B if MH D-DIF has an absolute value greater than one and a Mantel-Haenszel chi-squared statistic greater than 3.84. Category C items have MH D-DIF values significantly greater than one (tested by (|MH D-DIF|-1)/ $SE_{\text{MH-D-DIF}}$)>1.645) and have an absolute value greater than 1.5. Different testing programs have different procedures for handling each item type: Category C items are often not administered again, and Category B items are often investigated more thoroughly to identify potential sources of the DIF.

1. Paper Order Study

The paper order study was conducted in the fall of 2015. Schools were recruited to participate in the study, and students from those schools were able to take the test for half the usual price. Students who participated in the study received college-reportable scores, meaning that the scores could be sent to colleges and were treated as though they were from any standard administration of the ACT.

The study differed from a standard ACT administration in a few ways. First, the order of the sections in the test differed. Students tested in one of four different orders: standard order (English, mathematics, reading, science), mathematics first (mathematics, reading, science, English), reading first (reading, science, English, math) and science first (science, English, reading, mathematics). Students were randomly assigned to the test order conditions, which created randomly equivalent groups. Students were not told the order in which they would be taking the sections ahead of time to avoid influencing whether they actually showed up for the test.

The second, yet not substantive, difference was from an administration standpoint. The requirements for the number of rooms was different for the paper order study compared to standard ACT administrations. Ordinarily, there are no requirements concerning the number of rooms in which students test. The paper order study, however, required that all students in a particular room take the same order. This was due to small changes in the verbal directions depending on the order of the sections in the test and the timing of the different sections. Some of the tests have different time limits, so if multiple test orders were administered in the same room, there would be the potential for disruptions from students changing tests and going on breaks at different times. All students, regardless of the order in which they took the tests, took a break between the second and third tests, which is what is done during standard administrations. Schools were required to have rooms in multiples of four (e.g., four rooms, eight rooms, etc.) so that the different orders could have roughly the same number of students.

The content was the same for all four orders. That is, the items on the English test were the same if the test was taken first or in any other position, and the same was true for math, reading, and science. Students used the standard test booklets and answer sheets. Students assigned to an alternative order condition were given verbal instructions that were slightly different from the standard instructions to help ensure

that students filled in their responses in the proper place on the answer sheet. Aside from the different orders and instructions, the testing experience was the same as a standard administration.

A total of 8,312 students from 82 schools registered to participate in the study. Some schools, however, did not have enough rooms to test roughly the same number of students for each of the four orders. For instance, a school may have had enough students registered to fill 14 rooms. In that case, 12 of the rooms were used for the order study, and students in the other two rooms were given the standard order and were excluded from the analyses. Some students did not test in the room to which they were assigned and were subsequently removed from the analyses. A small number of schools also had to cancel the administration because of inclement weather. The final sample size included 5,861 students. Refer to the Appendix for more information about the sample characteristics and how the sample compared to the population of 2017 ACT-tested high school graduates.

Test-Level Results

Descriptive Statistics

For each test, raw scores (i.e., number correct) were converted to scale scores on a 1–36 scale. The differences in scale scores using the standard order conversion are shown in Table 1. For English, the greatest mean difference between order conditions was observed when English was administered second versus last; however, the magnitude of the difference was small (0.40 points). The largest difference for math occurred when administered in the third position versus the first position (0.09 points), with the lowest mean associated with the first position. For the reading test, the mean scale scores for the order conditions were all within a quarter of a scale score point. The highest means were associated with the first and third position (both with a mean of 22.28). For science, the largest difference occurred when it was administered first versus last—a difference of 0.62 scale score points.

It is important to note that, even though there were small differences in mean scale scores for the four order conditions, the results do not suggest that students perform systematically worse when a section appears later in the test (i.e., highest mean in the first position, second highest mean in the second position, and so forth).

Table 1. Paper Order Study Scale Score Descriptive Statistics Using the Standard Order Conversion

Test	Order	Mean	SD	Skew	Kurtosis	SEM	Reliability
English	EMRS	21.25	5.98	0.23	2.67	1.61	0.93
	MRSE	21.13	5.89	0.27	2.62	1.64	0.92
	RSEM	21.39	5.95	0.22	2.62	1.62	0.93
	SEMR	21.53	5.90	0.25	2.61	1.63	0.92
Mathematics	EMRS	21.21	4.87	0.47	2.52	1.41	0.92
	MRSE	21.15	4.77	0.47	2.49	1.40	0.91
	RSEM	21.24	4.92	0.40	2.38	1.41	0.92
	SEMR	21.42	4.92	0.39	2.29	1.42	0.92
Reading	EMRS	22.28	5.94	0.16	2.42	2.10	0.88
	MRSE	22.05	5.84	0.19	2.51	2.10	0.87
	RSEM	22.28	5.61	0.19	2.63	2.07	0.86
	SEMR	22.12	6.08	0.17	2.45	2.11	0.88
Science	EMRS	21.52	4.81	0.14	3.13	1.87	0.85
	MRSE	21.67	4.77	0.38	3.22	1.87	0.85
	RSEM	21.84	4.68	0.27	3.31	1.86	0.84
	SEMR	22.14	4.69	0.34	3.25	1.86	0.84

t-Tests

A series of *t*-Tests was used to evaluate the statistical significance of the mean differences between the scale scores from the standard order and the three alternative orders. The raw scores were converted to scale scores using the conversion for the standard order form, which allowed for comparisons among the orders on the 1–36 scale score metric. The results are shown in Table 2. The only statistically significant difference in average scale scores occurred when the science test was administered in the first position rather than the standard position (fourth). When science was administered in the second position, the average scale score difference was nearly statistically significant (p = .072). These results suggest that there could have been a small order effect for science. The effect sizes corresponding to these differences were -0.131 and -0.067 standard deviations, respectively, which would be considered small to very small effects (Cohen, 1992). All other effect sizes were less than 0.05 in magnitude.

Table 2. Paper Study *t*-tests Comparing the Scale Scores on the Standard Order with the Other Three Orders

Subject	Order	Mean	Effect Size	*t*	*df*	*p*-value
English	MRSE	0.1151	0.019	0.52	2910	0.601
	RSEM	-0.1405	-0.024	-0.63	2873	0.528
	SEMR	-0.2799	-0.047	-1.26	2864	0.208
Mathematics	MRSE	0.0647	0.013	0.36	2910	0.717
	RSEM	-0.0228	-0.005	-0.12	2873	0.901
	SEMR	-0.2113	-0.043	-1.16	2864	0.248
Reading	MRSE	0.2239	0.038	1.03	2910	0.305
	RSEM	-0.0036	-0.001	-0.02	2873	0.987
	SEMR	0.1561	0.026	0.70	2864	0.487
Science	MRSE	-0.1488	-0.031	-0.84	2910	0.402
	RSEM	-0.3185	-0.067	-1.80	2873	0.072
	SEMR	-0.6217	-0.131	-3.50	2864	0.001

Note. Mean differences were calculated by subtracting the alternative order mean from the standard order mean.

Kolmogorov-Smirnov Tests

Kolmogorov-Smirnov tests (Connover, 1999) were conducted to test the hypothesis that the standard order scale scores and the alternative order scale scores (generated using the standard order conversion table) were drawn from the same distribution. Even if the means were very similar (as indicated by previous analyses), test order could have affected the scale score distributions in other ways. The Kolmogorov-Smirnov test is a nonparametric test that compares the cumulative distribution functions of two variables. The test statistic, *D*, is the maximum difference between two cumulative distribution functions. The results of these tests for all four subjects are shown in Table 3. The only statistically significant difference ($p < .05$) was for the science test when it was in the first position. These tests provide evidence that differences in scale score distributions were minimal for English, mathematics, and reading for all orders and for science as long as it was not in the first position.

Table 3. Kolmogorov-Smirnov Statistics Comparing the Standard Order Scale Scores with the Scale Score Distributions from Alternative Orders

Test	Order	*D*	*p*-value
English	MRSE	0.023	0.821
	RSEM	0.023	0.845
	SEMR	0.024	0.803
Mathematics	MRSE	0.022	0.860
	RSEM	0.017	0.987
	SEMR	0.026	0.727
Reading	MRSE	0.033	0.423
	RSEM	0.026	0.731
	SEMR	0.026	0.708
Science	MRSE	0.022	0.857
	RSEM	0.034	0.376
	SEMR	0.060	0.012

Effective Weights

The relationships among the different subject test scores for alternative orders must be similar to those from the standard order if scores are to be considered comparable, providing evidence of construct equivalence (i.e., the ACT measures the same constructs regardless of the order in which the tests are administered). Effective weights are the statistical contributions of subject test scores to the variance of ACT Composite scores, and they should be similar for the different order conditions if construct equivalence holds. The effective weights are shown in Table 4 for the scale scores using the standard conversion. The results indicate that the effective weights are similar, which provides evidence of construct equivalence regardless of test order.

Table 4. Scale Score Effective Weights When Using the Standard Order Conversion

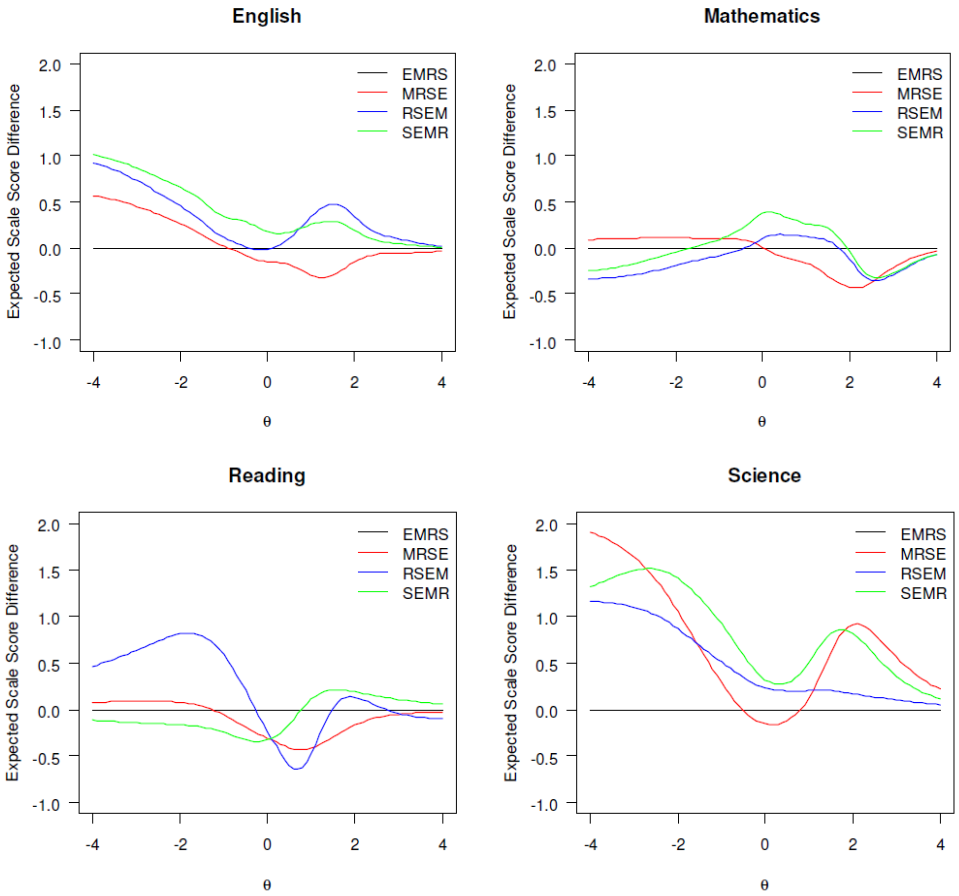
Order	English	Mathematics	Reading	Science
EMRS	0.286	0.216	0.276	0.222
MRSE	0.282	0.215	0.278	0.225
RSEM	0.289	0.226	0.265	0.221
SEMR	0.280	0.220	0.284	0.216

IRT Analyses

To gather additional information on the differences in scores from tests administered in different positions, IRT analyses were also conducted. The responses for each test in each order were fit with the unidimensional three parameter logistic model (3PL; Lord, 1980) using BIOLOG-MG. The item parameters were then used to estimate the expected raw scores in a range of θ values. Those raw scores were then converted to scale scores using the standard order conversion tables.

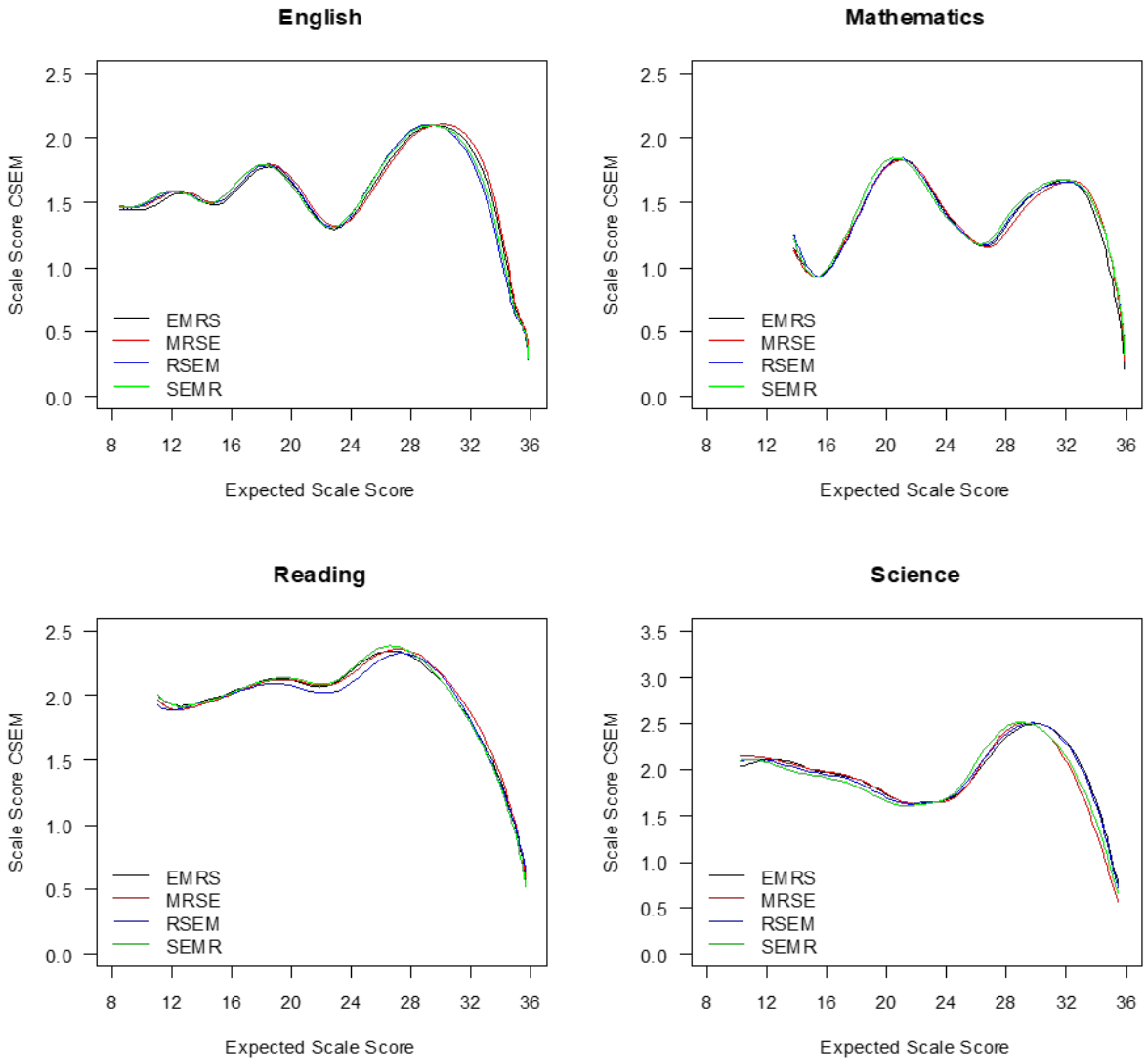
Figure 1 shows the differences in expected scale scores using the standard conversion. The y-axis is the expected scale score (calculated using equation [1] for each alternative order minus the expected scale score for the standard order). The majority of scale differences were less than half a point; the most notable exception was among low θ levels when English was second or third in the testing order where differences were about one point. For mathematics, the differences in scale scores were all within half a scale score point. For reading, the lower θ values had higher expected scale scores when it appeared in the first position compared to the other orders. The largest differences were for the science test. At low θ values, expected scale scores were one to two points higher when the test was taken in a position other than last (as it is in the standard order). Science also showed relatively large differences between scale scores for higher ability students. Specifically, expected scale scores were nearly one point higher when science was in either the first or the third position compared to the last position for students with θ around 2.0.

Figure 1. Differences in Expected Scale Scores Using the Standard Conversions



CSEM. In addition to the differences in expected scale scores, the CSEMs are also an important characteristic of a test. If the measurement error differs substantially from one order to the next, the scores could no longer be treated as comparable. The scale score CSEMs for the standard order conversion were calculated using equation (2) (Figure 2). The CSEMs using the standard order conversions are very similar among the four orders for each of the four tests. The variability of the CSEM estimates is comparable to what would be seen for several forms with different items that were all equated.

Figure 2. Scale Score Standard Errors of Measurement Using the Standard Conversions

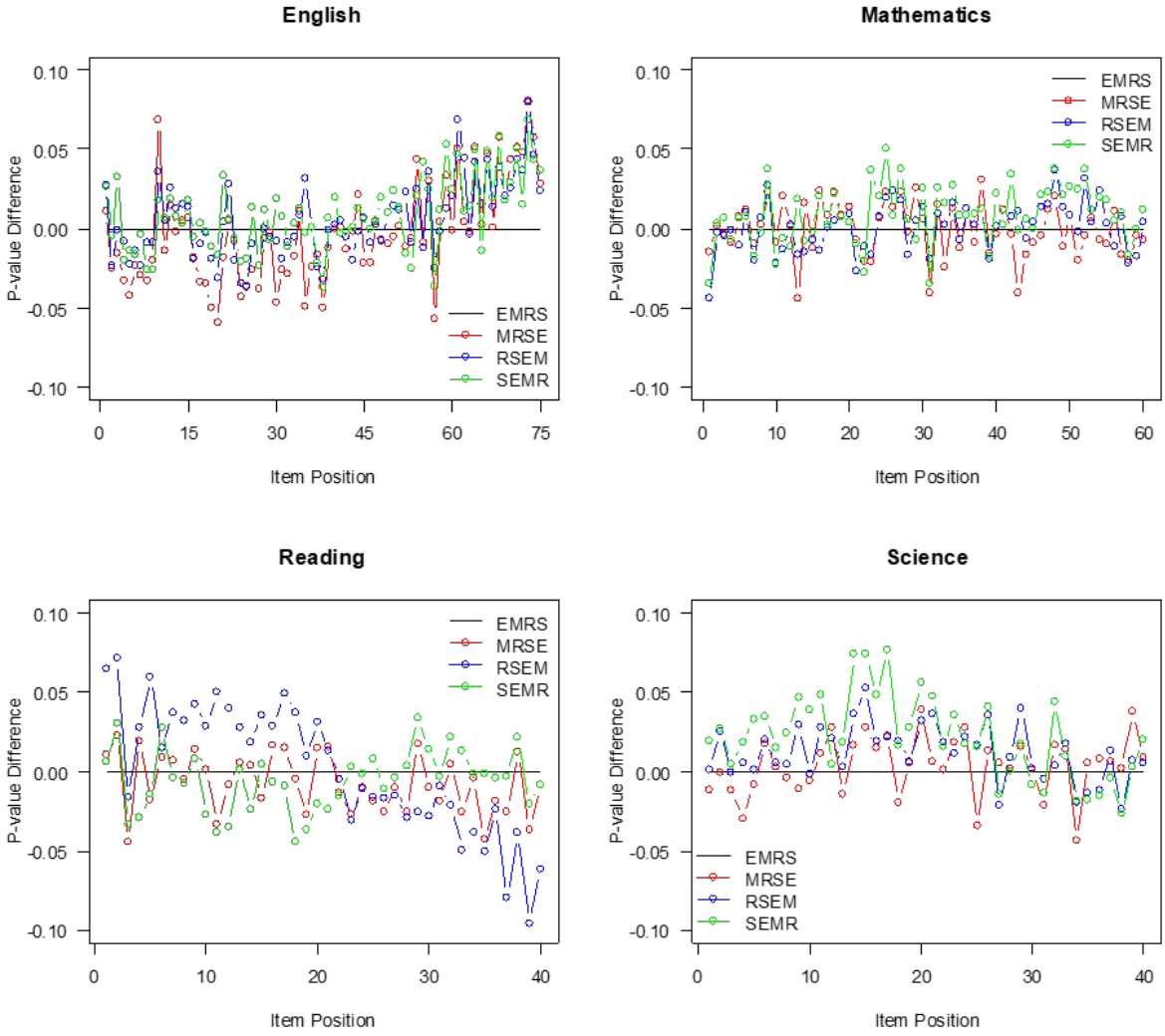


Item-Level Results

Item Difficulty

Figure 3 shows the differences in p-values (proportion correct) between items from each alternative order and the standard order. The y-axis is the p-value for the each order minus the p-value for the standard order. The x-axis is the item position within the test. For English, most items in the first half of the test had slightly lower p-values when the test appeared in the last position within the testing order. At the end of the test, the p-values for the other three orders were higher than the standard order. Analyses of mathematics p-values did not reveal any discernable trends. Most differences were small, and they fluctuated around zero. For reading, however, there were noticeable differences in p-values even though the magnitudes were still small. When reading was in the first position, p-values were highest for the first half of the test and lowest for the second half of the test. The pattern was not so clear when reading was in the other positions. As for science, the p-values were slightly higher when administered in the first and second position as compared to the standard condition for the first three-quarters of the test.

Figure 3. Differences in *P*-Values for the Four Different Orders



Differential Item Functioning

In addition to p-value differences, DIF methods were also used to evaluate differences in the item performances. The DIF analyses were conducted comparing the standard order items (the reference group) with those from the other three orders (the focal groups). The ETS classifications (Dorans & Holland, 1993) were used to evaluate the magnitude of DIF (Table 5). There were no C-DIF items for any subject, and the mathematics test had no items flagged for B-DIF. English had two B-DIF items for the order with English in the last position. The items were the 20th and 73rd items. The item in the 20th position favored the standard order, while the 73rd item favored the English last order. When English was in the third position, item 73 was also flagged as a B-DIF item and favored that condition. For the reading and science tests, the items that were flagged as B-DIF items were for the orders with those tests were in the first position. The reading test had only a single B-DIF item. When reading was in the first position, the second item on the test was classified as a B-DIF item, and it favored the reading first condition. Items 14 and 15 were flagged as B-DIF items for science, and they both favored the science first condition.

Since this study employed randomly equivalent groups (matched on ability), the DIF results should mirror the p-value analyses. Though there were small differences in p-values for some subjects, there were only six items that were identified as having B-DIF (moderate) based on the processes proposed by Dorans and Holland (1993).

Table 5. DIF Classifications for the Paper Order Study

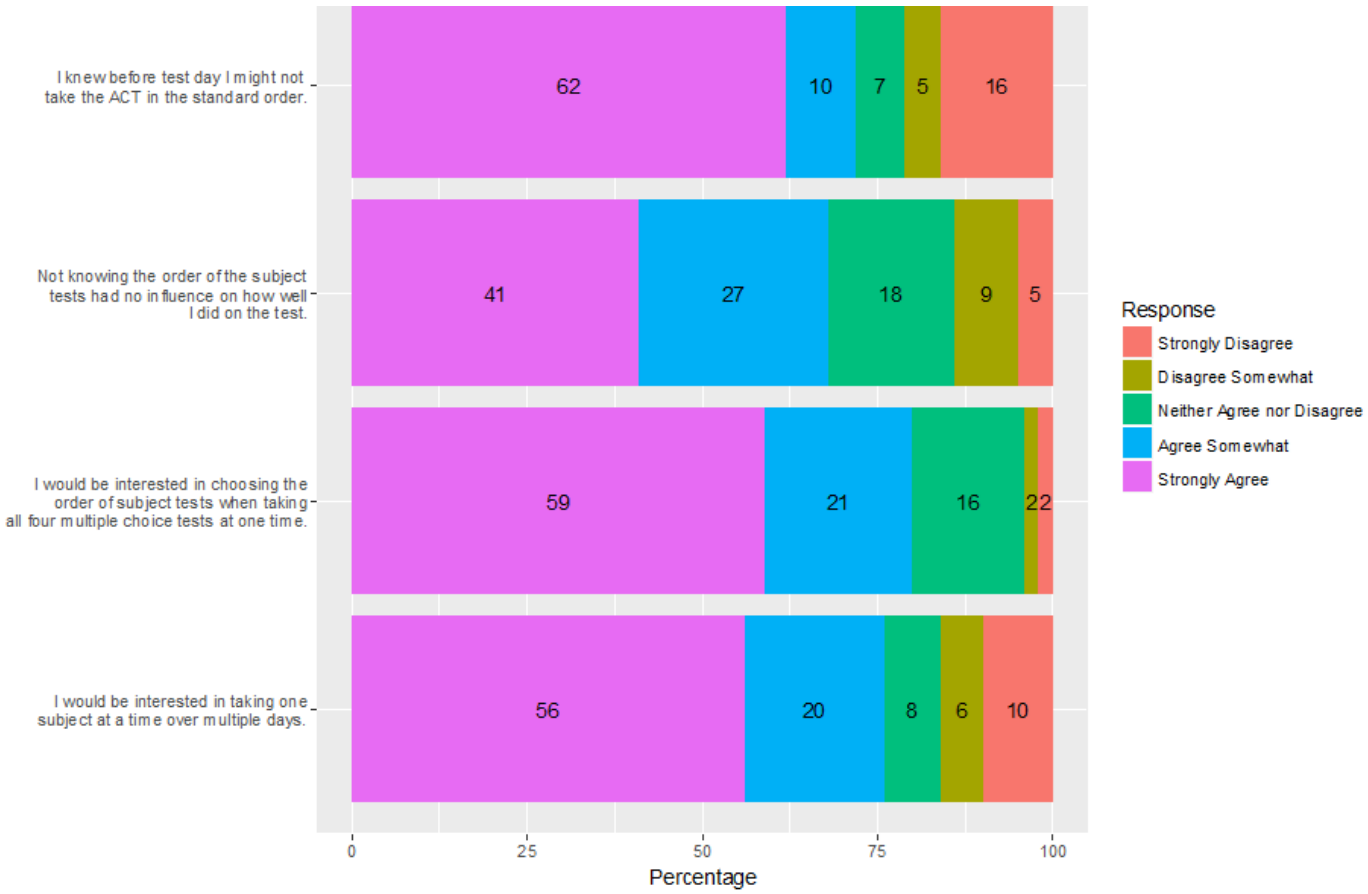
Test	Order	# of A Items	# of B Items	# of C Items
English	MRSE	73	2	0
	RSEM	74	1	0
	SEMR	75	0	0
Mathematics	MRSE	60	0	0
	RSEM	60	0	0
	SEMR	60	0	0
Reading	MRSE	40	0	0
	RSEM	39	1	0
	SEMR	40	0	0
Science	MRSE	40	0	0
	RSEM	40	0	0
	SEMR	38	2	0

Student Survey Results

This section presents results from a digital survey that was sent to students after they participated in the study. A total of 1,271 student responded to the survey (response rate of 22.4%). Among those who responded, about 22 percent had taken the ACT once before, 24 percent had taken the ACT twice before, and about 54 percent had never taken the ACT before.

The percentages of responses to each question are shown in Figure 4. The first question asked students to rate their level of agreement with the statement “I knew before test day I might not take the ACT in the standard order.” Most students (about 78 percent) agreed somewhat or strongly agreed with this statement while around 21 percent disagreed somewhat or strongly disagreed. If students disagreed somewhat or strongly disagreed to a series of statements, they were given the option to provide additional written feedback. Most students didn’t provide additional feedback, but a small number indicated that they learned of the possibility of taking it in a different order the morning of the test. Some students also indicated that their parents completed the registration for them so they did not know about the potential for a different order.

Figure 4. Response to Survey Questions for the Paper Order Study



The responses to the statement “Not knowing the order of the subject tests had no influence on how well I did on the test” are shown in the second line of Figure 4. Around 68% of students said they agreed somewhat or strongly agreed with this statement, and around 14% disagreed somewhat or strongly disagreed. Some students indicated that they experienced a little more stress for having not known the order ahead of time.

The responses to the statement “I would be interested in choosing the order of subject tests when taking all four multiple choice tests at one time” are shown the third line of Figure 4. Around 80% selected either Agree Somewhat or Strongly Agree, and 4% selected Disagree Somewhat or Strongly Disagree. The comments from those students reflected a few different ways of thinking. Some students seemed to think it made no difference so they did not want to choose. Others did not believe they knew the best order for them so they were willing to go with any order. Similarly, they thought choosing the order would add more anxiety and complication to the whole situation.

About 76% of students said they agreed or strongly agreed with the statement “I would be interested in taking one subject at a time over multiple days.” The results are shown in the fourth line of Figure 4. About 16% said they disagreed somewhat or strongly disagreed with that statement. Of those that disagreed with the statement, most of the students expressed the preference to complete the entire battery in one day. Some students indicated it would add too much stress to take the test over multiple days. Some also thought that it would be more difficult to fit into their schedules if it were over multiple days.

Students were given the opportunity to provide additional thoughts about their experience. The most common comment was that students felt that there was not enough time to complete the tests. Most of these comments mentioned the reading test in particular. Of the comments related to the different orders of the study, most were related to which tests they thought would be better to take in which positions. Some students believed that taking the tests they felt most comfortable with earlier in the testing order was preferable, while others wanted to take those at the end. Some students thought they would do better on the tests at the beginning of the order because they were more alert and fresh while others thought they needed time to warm up before they were performing their best. Some students liked taking math first, others really disliked it. Others said reading at the end made it harder to focus. Some students said they liked taking science in a position other than last because they felt they were better able to interpret the content compared to when it appears in the last position. There were also some students who expressed disappointment that they were assigned the standard order. No discernable pattern was found for these comments.

Summary – Paper Order Study

The results from the paper order study showed minimal differences at the test level. There were small differences in mean scale scores for the four order conditions, but the results do not suggest that students perform systematically worse when the subject test appears later in the test order (i.e., highest mean in the first position, second highest mean in the second position, and so forth). For example, English scale scores were highest when administered second, followed by third, first, and finally fourth. For mathematics, scale scores were highest when administered third, followed by fourth,

second, and finally first. For reading, scales scores were highest in the first and third position (both 22.28), followed by fourth, and finally second. Only for science do we see that scale scores decrease as test position increases. At the item level, minimal differences in item difficulty were observed. According to DIF analyses, very few differences were statistically significant. In general, results suggest that students earn similar ACT subject test scores when taking a test section first rather than the typical test position, providing initial evidence in support of section retakes.

2. Online Order Study

If modular administrations of the ACT are going to be feasible, computer administrations will likely be required. Students could then simply log in to take whatever test they needed, and the test center or school would not need to worry about managing a large number of test booklets and answer sheets. This is why a second order study was conducted to see if the differences among the orders seen in a paper administration were similar when the test was taken on computer.

The online order study took place in April of 2016. Schools were again recruited to participate in the study. A total of 4,106 students from 54 different schools registered for the online study. Similar to the paper order study, students who participated in the study could register using a voucher to take the test for half the normal price. The final sample included 3,587 students; refer to the Appendix for more detailed information.

The online study was very similar to the paper study. Students received college-reportable scores and were required to test in rooms within their school that were comprised of students taking only a single order. For this study, the orders were different. The standard order was again administered along with three other orders, each of them starting with a different subject. The mathematics first condition was mathematics, science, English, and reading. The reading first condition was reading, English, science, and mathematics. The science first condition was science, reading, mathematics, and English. The TestNav platform was used to deliver the ACT.[1]

Because of the unusual circumstances of this administration, the way students logged in and moved between tests differed slightly from normal administrations on the computer. Students had to log in to each test separately instead of seamlessly moving from one to the next. The proctors also had to make some adjustments beyond their usual responsibilities so the tests could be administered in the proper order. ACT staff worked with the schools ahead of time to ensure that the schools had the resources to administer the ACT online. Training sessions were also conducted to ensure that any questions the schools had about the administration were answered before the test date.

The next section describes results from the online order study. All analyses conducted for the paper order study were also conducted for the online order study. In addition, because the test was taken on computer, item latency information was available.

[1]Detailed description of minimum hardware requirements for the TestNav platform is located here: https://www.act.org/content/dam/act/unsecured/documents/TechnicalGuidefortheACTTakenOnline.pdf

Test-Level Results

Descriptive Statistics

Scale Score Descriptive Statistics. The descriptive statistics using the standard conversions are shown in Table 6. The scale scores for the mathematics test showed small differences among the four orders. For the English and science tests, students performed about half a scale score point higher when the test appeared in the first position compared to the last position. For reading, the mean was highest when it was in the standard order position (third) and lowest when it was in the last position (difference of 0.56). On the math test, the largest mean difference was 0.33 (between math in the third and fourth positions). In general, there was no discernible pattern between test order and scale scores (e.g., highest mean in the first position, second highest mean in the second position, third highest mean in the third position, and lowest mean in the last position).

Table 6. Online Order Study Scale Score Descriptive Statistics Using the Standard Order Conversion

Test	Order	Mean	SD	Skew	Kurtosis	SEM	Reliability
English	EMRS	19.85	5.73	0.34	2.86	1.66	0.92
	MSER	19.57	5.53	0.28	2.76	1.66	0.91
	RESM	19.45	5.60	0.29	3.02	1.65	0.91
	SRME	19.39	5.45	0.33	3.02	1.67	0.91
Mathematics	EMRS	20.00	4.88	0.51	2.68	1.52	0.90
	MSER	20.04	4.84	0.39	2.45	1.54	0.90
	RESM	19.79	4.82	0.47	2.49	1.52	0.90
	SRME	20.11	4.94	0.52	2.62	1.53	0.91
Reading	EMRS	20.66	5.35	0.19	2.78	2.03	0.86
	MSER	20.10	5.52	0.27	2.57	2.02	0.87
	RESM	20.36	5.09	0.23	3.04	1.97	0.85
	SRME	20.57	5.19	0.27	3.12	2.02	0.85
Science	EMRS	20.42	4.34	0.41	3.89	1.87	0.82
	MSER	20.54	4.13	0.40	3.64	1.89	0.80
	RESM	20.65	4.22	0.41	4.04	1.86	0.81
	SRME	20.93	4.01	0.21	4.09	1.83	0.79

t-Tests

A series of *t*-tests was conducted to investigate whether mean scale score differences between the standard order and the other three orders were statistically significant (Table 7). The effect sizes were all small (≥ 0.12 in magnitude). Only two of the *t*-tests yielded statistically significant results. The reading scale scores were significantly lower for those who took it in the final position compared to when it appeared in the standard position (third), and the science scale scores were significantly higher when it was taken in the first position compared to when it was taken in the last position. Despite being statistically

significant, the effect sizes (0.081 for reading and -0.121 for science) are far below the threshold of 0.20, which is the conventional threshold for a small effect (Cohen, 1992).

Table 7. *T*-test Results Comparing the Scale Scores on the Standard Order with the Alternative Orders Using the Standard Order Conversion

Subject	Order	Mean Difference	Effect Size	*t*	*df*	*p*-value
English	MSER	0.2717	0.048	1.030	1824	0.303
	RESM	0.3996	0.071	1.500	1808	0.134
	SRME	0.4519	0.081	1.710	1789	0.088
Mathematics	MSER	-0.0364	-0.008	-0.160	1824	0.873
	RESM	0.2135	0.044	0.940	1808	0.350
	SRME	-0.1056	-0.022	-0.450	1789	0.649
Reading	MSER	0.5594	0.103	2.200	1824	0.028
	RESM	0.2970	0.057	1.210	1808	0.227
	SRME	0.0837	0.016	0.340	1789	0.737
Science	MSER	-0.1213	-0.029	-0.610	1824	0.541
	RESM	-0.2299	-0.054	-1.140	1808	0.254
	SRME	-0.5071	-0.121	-2.560	1789	0.010

Note: The differences were calculated by subtracting each of the orders from the standard order.

Kolmogorov-Smirnov Tests

The differences in the scale score distributions were evaluated using Kolmogorov-Smirnov tests (Connover, 1999). For each subject, the scale score distributions using the standard conversions were compared between the standard order group and the three alternative orders. The results are shown in Table 8. The *D* statistic was significant only for the science test when it was taken in the first position compared to the standard position (fourth).

Table 8. Kolmogorov-Smirnov Statistics Comparing the Standard Order Scale Scores with the Scale Score Distributions from the Other Orders When Using the Standard Order Conversion

Test	Order	*D* Statistic	*p*-value
English	MSER	0.027	0.883
	RESM	0.041	0.444
	SRME	0.037	0.565
Mathematics	MSER	0.030	0.803
	RESM	0.030	0.814
	SRME	0.025	0.948
Reading	MSER	0.063	0.052
	RESM	0.039	0.495
	SRME	0.048	0.259
Science	MSER	0.035	0.616
	RESM	0.035	0.627
	SRME	0.070	0.026

Effective Weights

The effective weights for scale scores are shown in Table 9 for the standard order conversions. The largest differences in effective weights when using the standard conversion was for reading where the difference between the largest and the smallest effective weight was about 0.02. The other three subjects had even smaller differences.

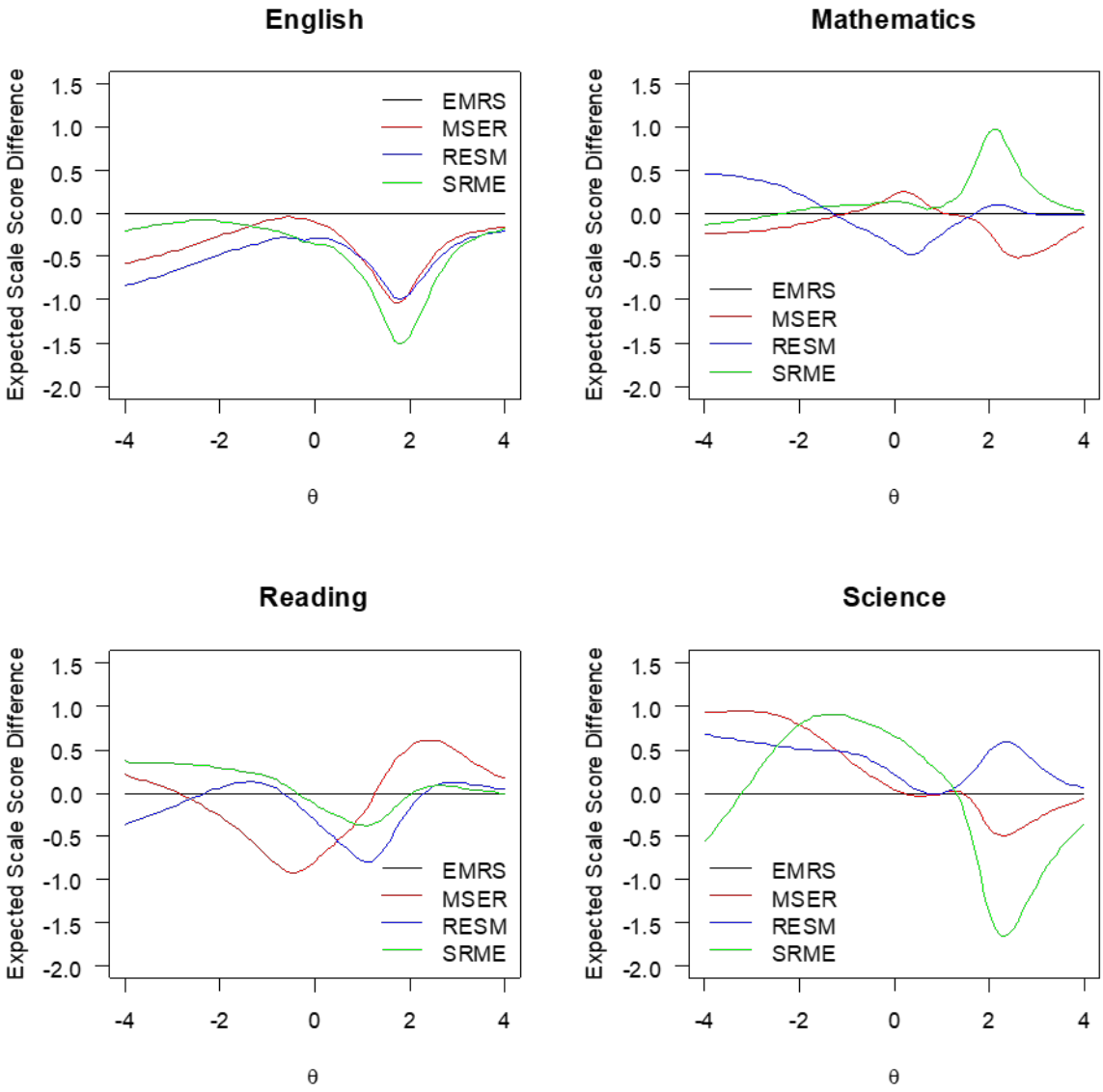
Table 9. Scale Score Effective Weights When Using the Standard Order Conversion

Order	English	Mathematics	Reading	Science
EMRS	0.289	0.234	0.265	0.212
MSER	0.282	0.235	0.281	0.202
RESM	0.293	0.233	0.261	0.213
SRME	0.285	0.244	0.269	0.202

IRT Analyses

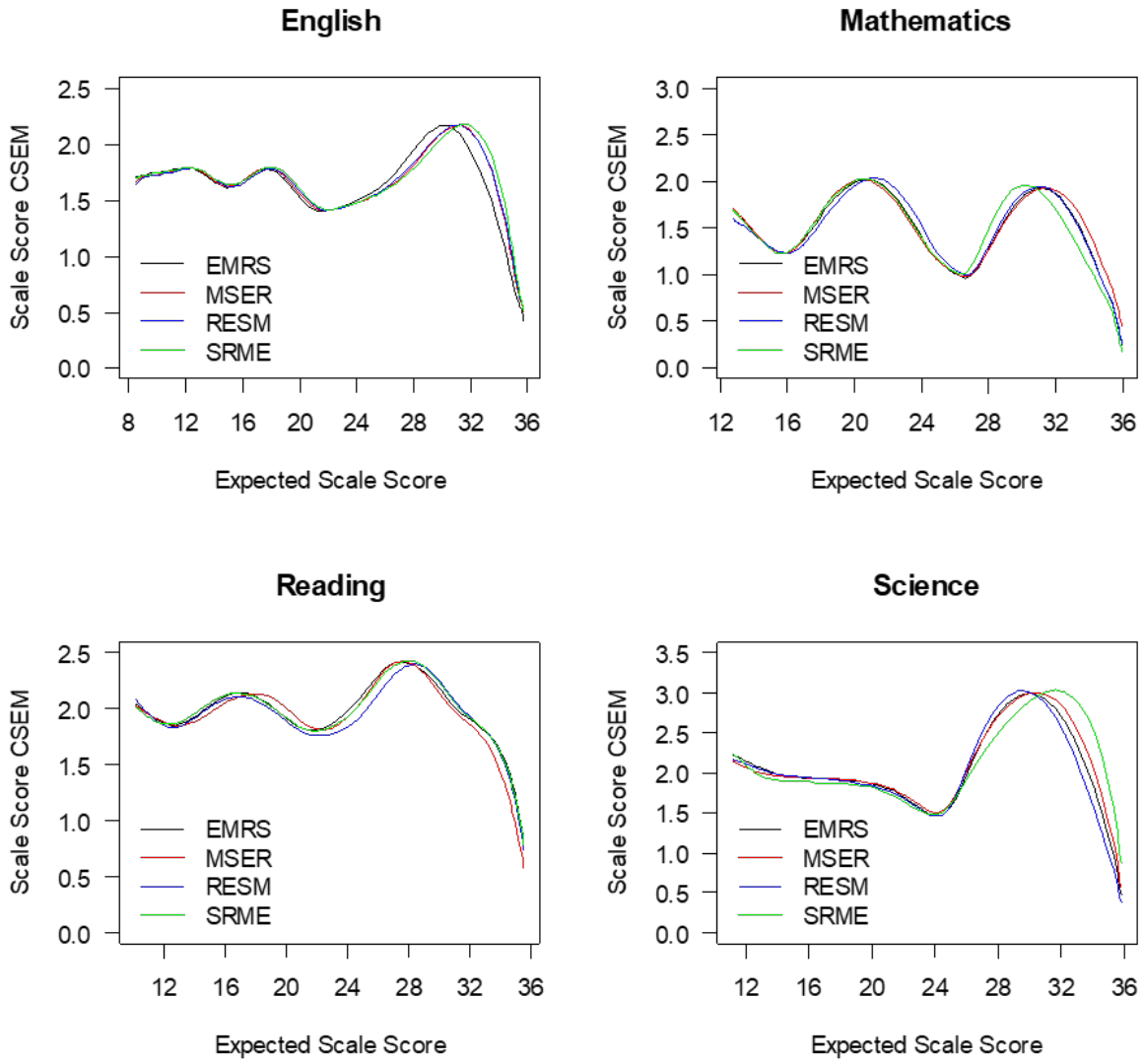
The magnitudes of the differences in expected scale scores on the 1–36 scale using the standard order conversion are shown in Figure 5. The differences are the expected scale scores minus the standard order expected scale scores. The largest differences for English were when θ was around 2.0, where the differences between the standard order (English first) and the form with English last were about 1.5 scale score points. Similarly, the largest difference for science were when θ was around 2.0, where the expected scale scores were about 1.5 points lower for the science first condition compared to the standard order (fourth). The differences for mathematics and reading were relatively small, all within a point of the expected scale score for the standard order.

Figure 5. Differences in Expected Scale Scores Using the Standard Conversions



CSEM. The CSEM plots using the standard order conversion are shown in Figure 6. Similar to the paper order study results, the CSEMs for the four orders were quite similar when using the standard order conversion.

Figure 6. Scale Score Standard Errors of Measurement Using the Standard Conversions



Item-Level Results

Latency

The average item latencies are shown in Figure 7. The differences between the average latencies of the standard order and the alternative orders are shown in Figure 8. The y-axis is the average latency on an alternative order minus the average latency on the standard order. The medians for each order and the difference plots for the medians are shown in Figure 9 and Figure 10, respectively. The biggest differences in time spent on items occurred at the beginning of the test. For English, students tended to spend less time on earlier items the later the test appeared in the order. For mathematics, there did not seem to be much of a pattern in the differences between the standard order and the alternatives. For reading, students tended to spend more time on earlier items when the test was administered earlier in the testing order. Specifically, when reading appeared first, students spent more time on the items at the beginning and less time on the items at the end. A similar pattern was observed for science.

Figure 7. Average Item Latencies for the Online Order Study

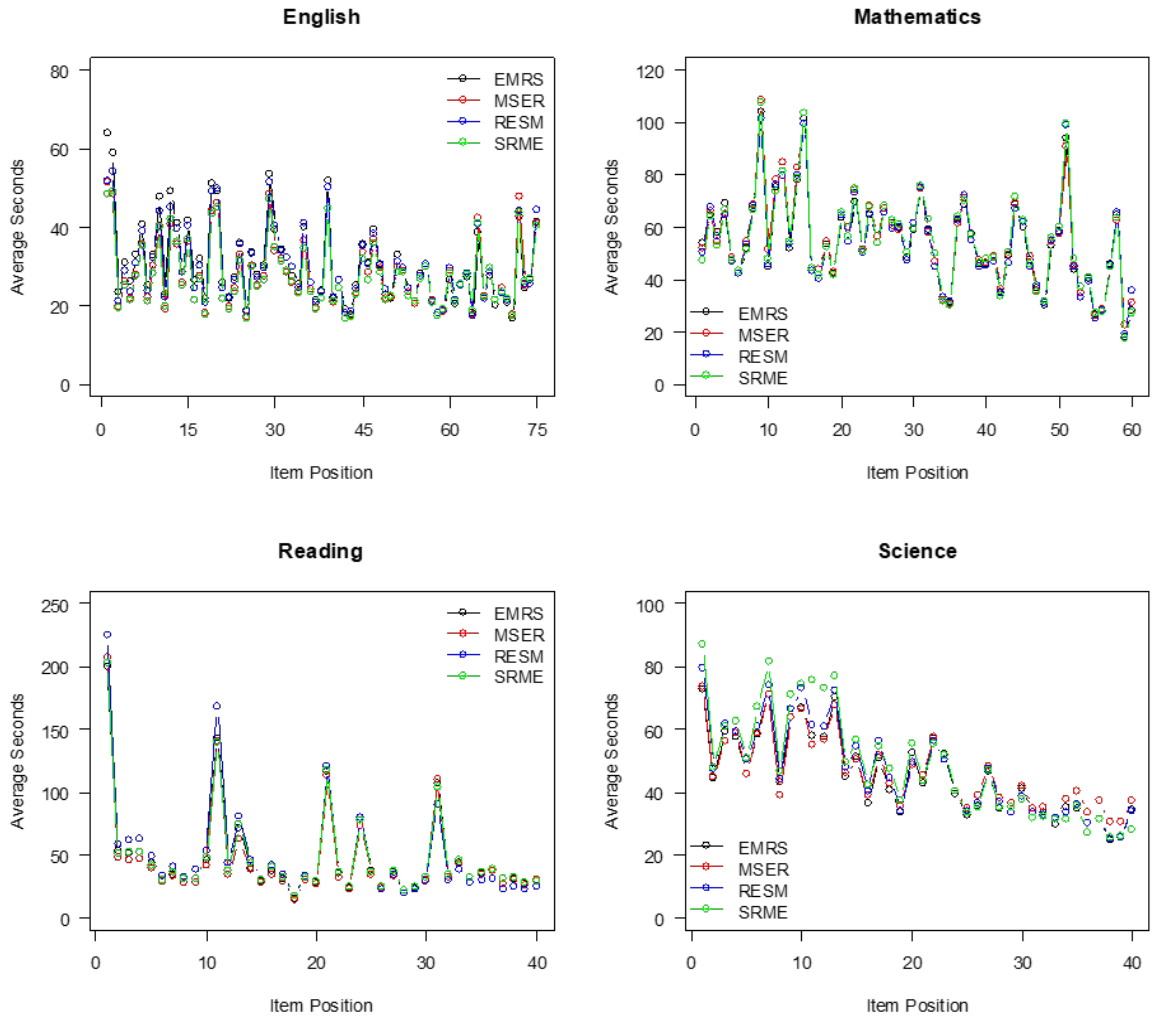


Figure 8. Differences in Average Item Latencies for the Online Order Study

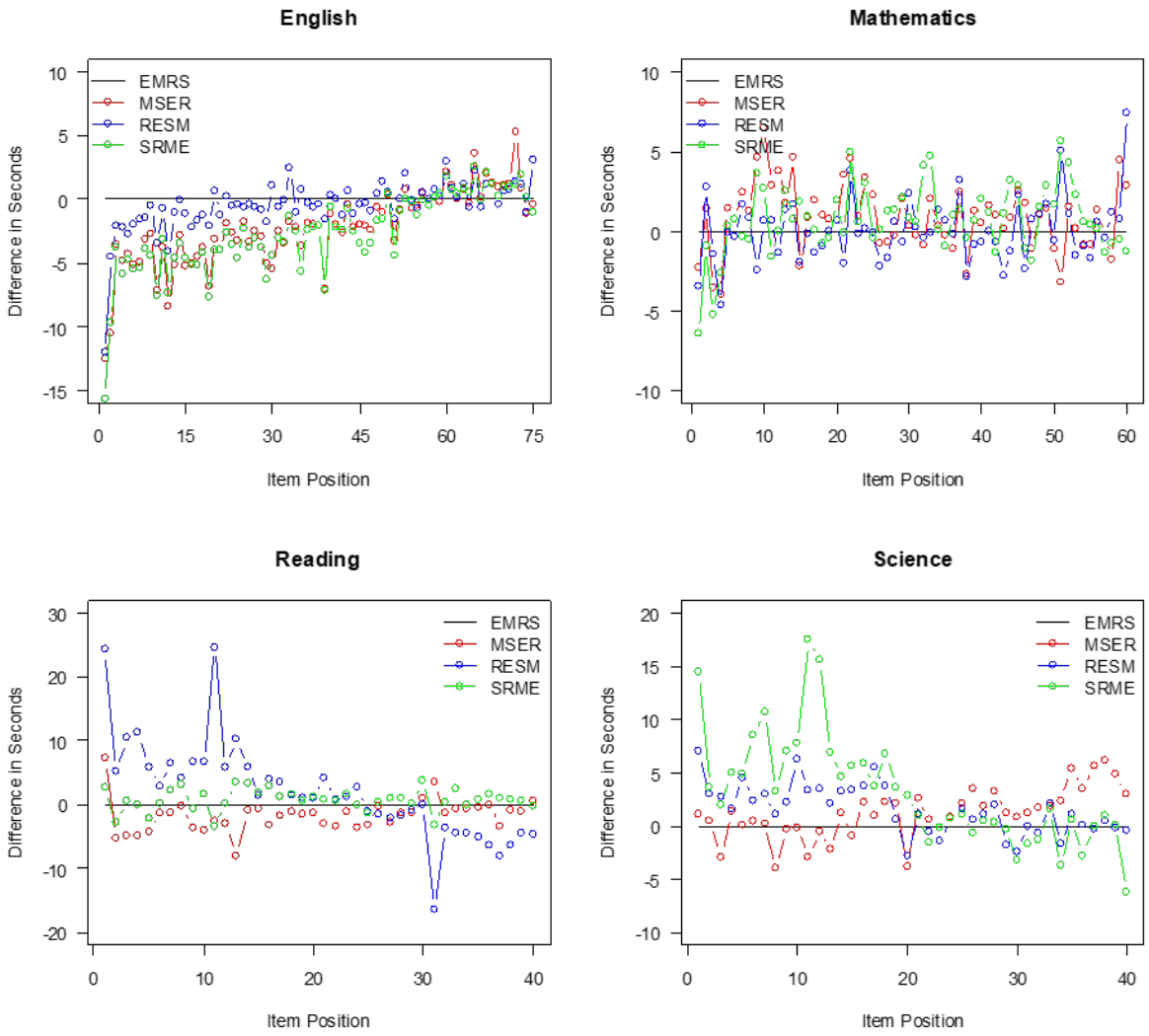


Figure 9. Median Item Latencies for the Online Order Study

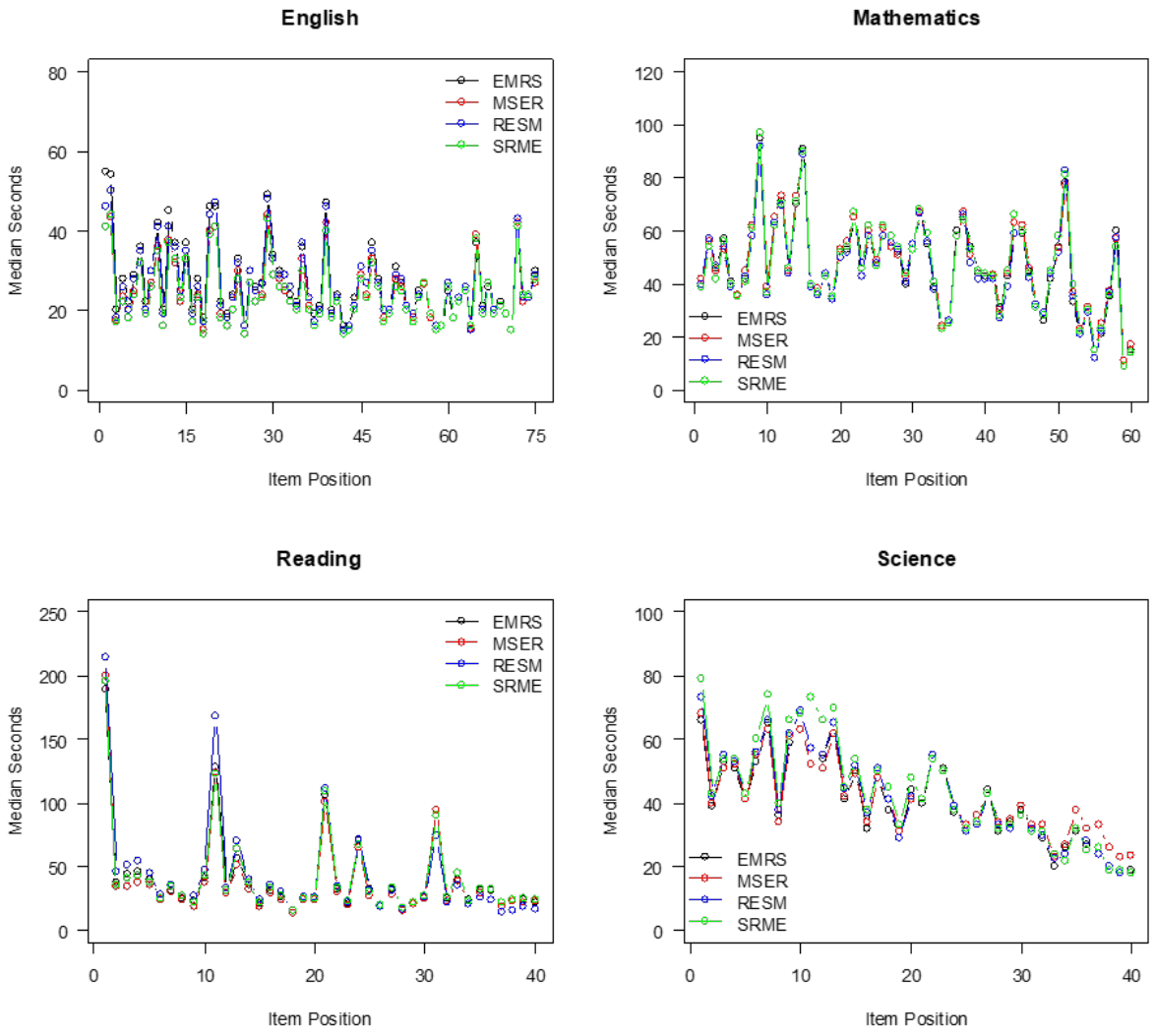
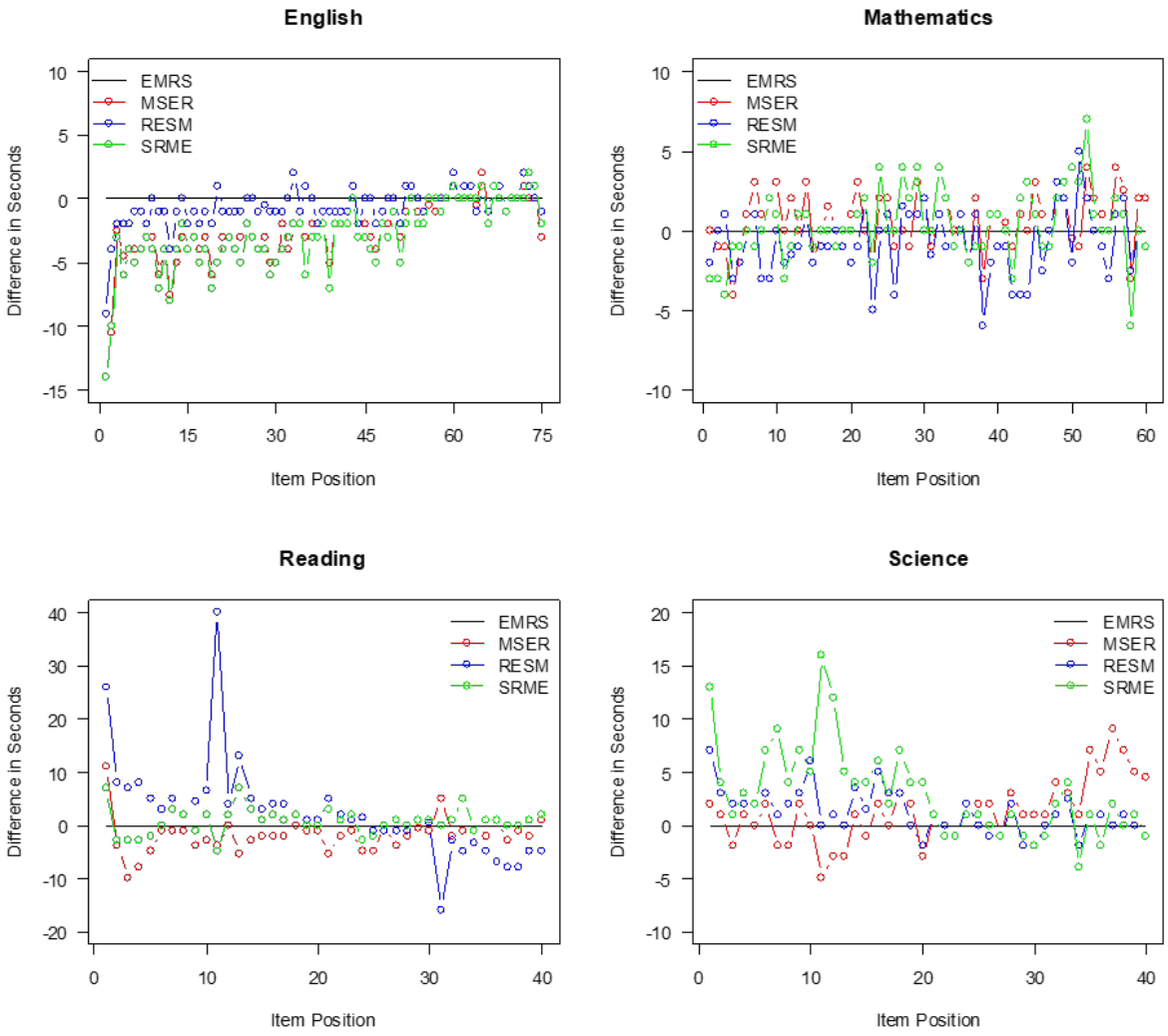


Figure 10. Differences in Median Item Latencies for the Online Order Study



Overall latency statistics are shown in Table 10. The mean, median, and standard deviations for total time spent on each test are presented. The mean time spent on the English, reading, and science tests tended to decrease when the test was moved later in the order. This pattern did not hold for mathematics, however, where the means and medians for the four different orders were similar. For all four subjects, when the test was administered later in the order, the standard deviations of the overall latencies tended to increase. The increases were especially large for the reading and science tests.

Table 10. Test-Level Latency Descriptive Statistics in Seconds

Test	Order	Mean	Median	SD
English	EMRS	2278	2373	302
	MSER	2111	2151	354
	RESM	2240	2319	309
	SRME	2080	2106	344
Mathematics	EMRS	3260	3421	399
	MSER	3313	3423	315
	RESM	3263	3411	370
	SRME	3304	3433	329
Reading	EMRS	1874	1951	209
	MSER	1804	1893	262
	RESM	1955	1998	137
	SRME	1904	1970	178
Science	EMRS	1812	1929	291
	MSER	1858	1949	234
	RESM	1873	1966	241
	SRME	1937	1989	167

The extent to which the order of the tests within the full ACT test affected the latencies of high-performing students and low-performing students was also investigated. Students were separated into two different groups based on the scale score they received from the study. The top 10% in terms of scale scores was the high-performing group, and the bottom 10% was the low-performing group. Figure 11 shows the median latencies for the low-performing group for each order and test. Figure 12 shows the median difference between alternative orders and the standard order for the low-performing group. On the English test, the low-performing group tended to spend slightly more time on items at the beginning of the test and slightly less time on items toward the end of the test the later it appeared in the testing order. There were a few items near the beginning of the mathematics test with much higher latencies when mathematics appeared first in the testing order. Students who tested in the standard order also spent much more time on one item near the end of the test compared to the other three orders. For reading, the overall group spent a relatively large amount of time on the first item of each passage. This is time mostly devoted to reading the passage. The peaks were much less pronounced for this low-performing group. Similar to the full group, students spent more time on the items at the beginning of the reading test the earlier the test appeared in the order. For science, low-performing students tended to spend more time on early items and less time on the later items compared to the whole group. The students spent substantially more time on the early items when they took science in the first position compared to the standard order (last position). However, differences among the four orders at the end of the test were very small.

Figure 11. Median Item Latencies for the Bottom 10% in Terms of Scale Score

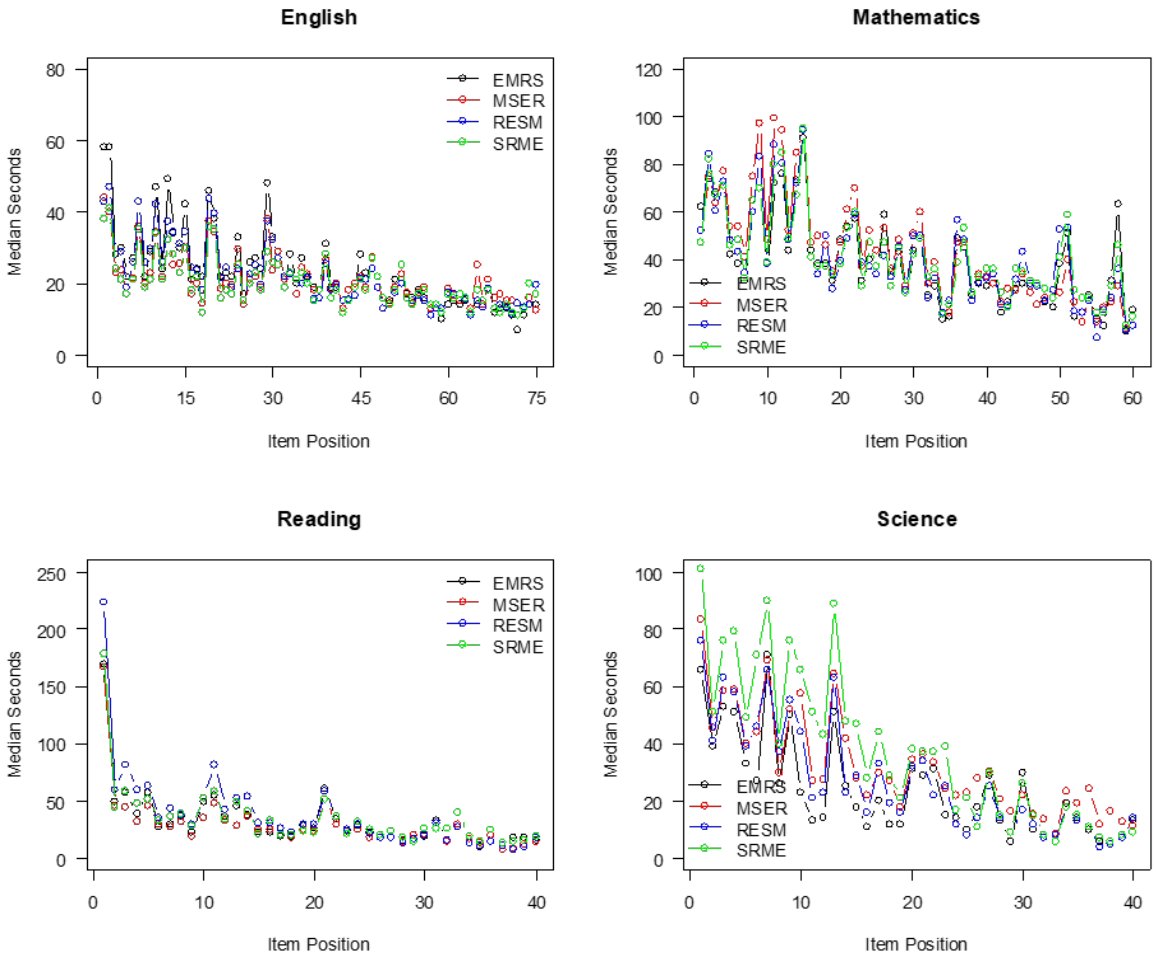
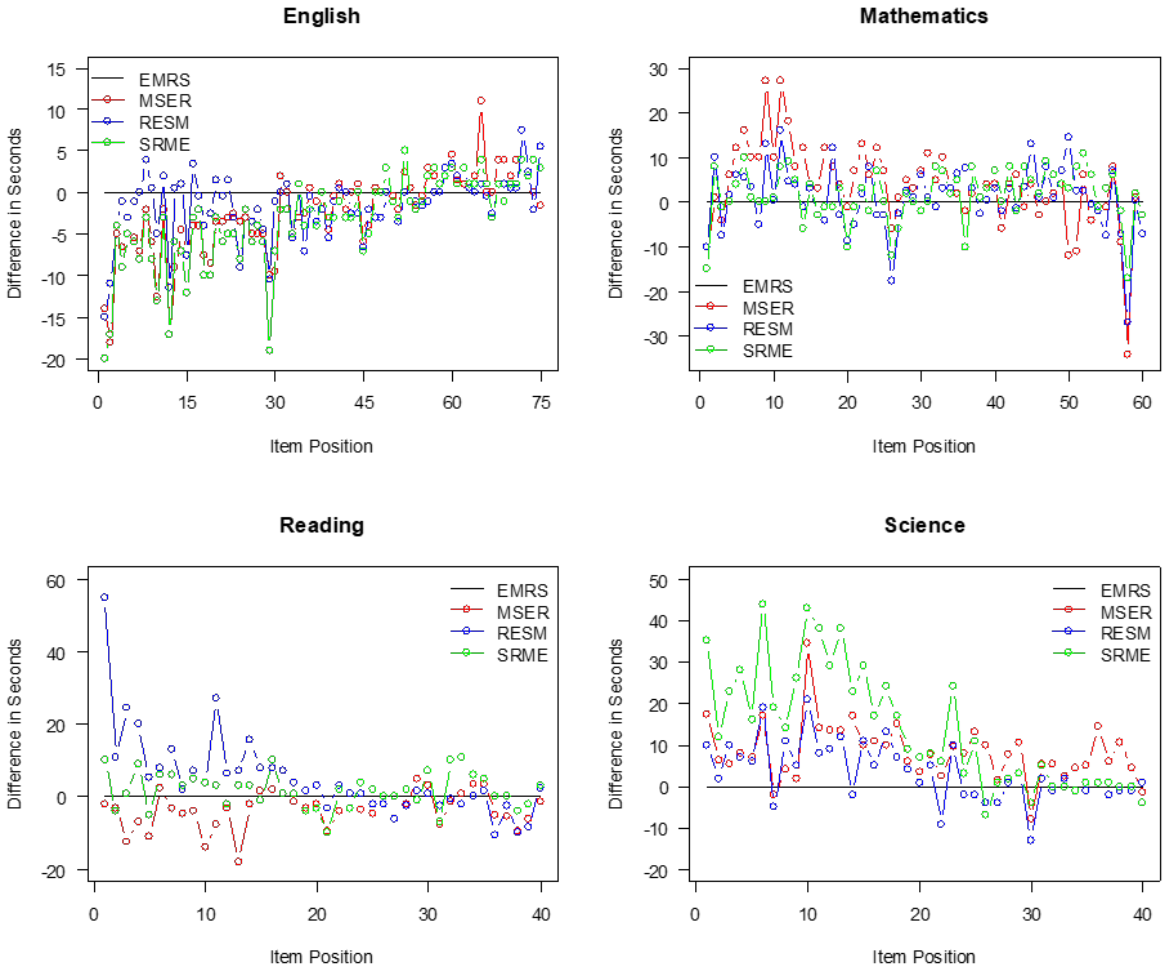


Figure 12. Differences in Median Item Latencies for the Bottom 10% in Terms of Scale Score



The same information for a group consisting only of students in the top 10% of scale scores is presented in Figures 13 and 14. Students tended to spend about as much time per item for English and reading as the overall group. For this group of high-performing students, there were minimal differences overall for all subjects among the four orders. The order of the tests within the full ACT test did not appear to have much of an influence on the amount of time the high-performing students spent on each item.

Figure 13. Median Item Latencies for the Top 10% in Terms of Scale Score

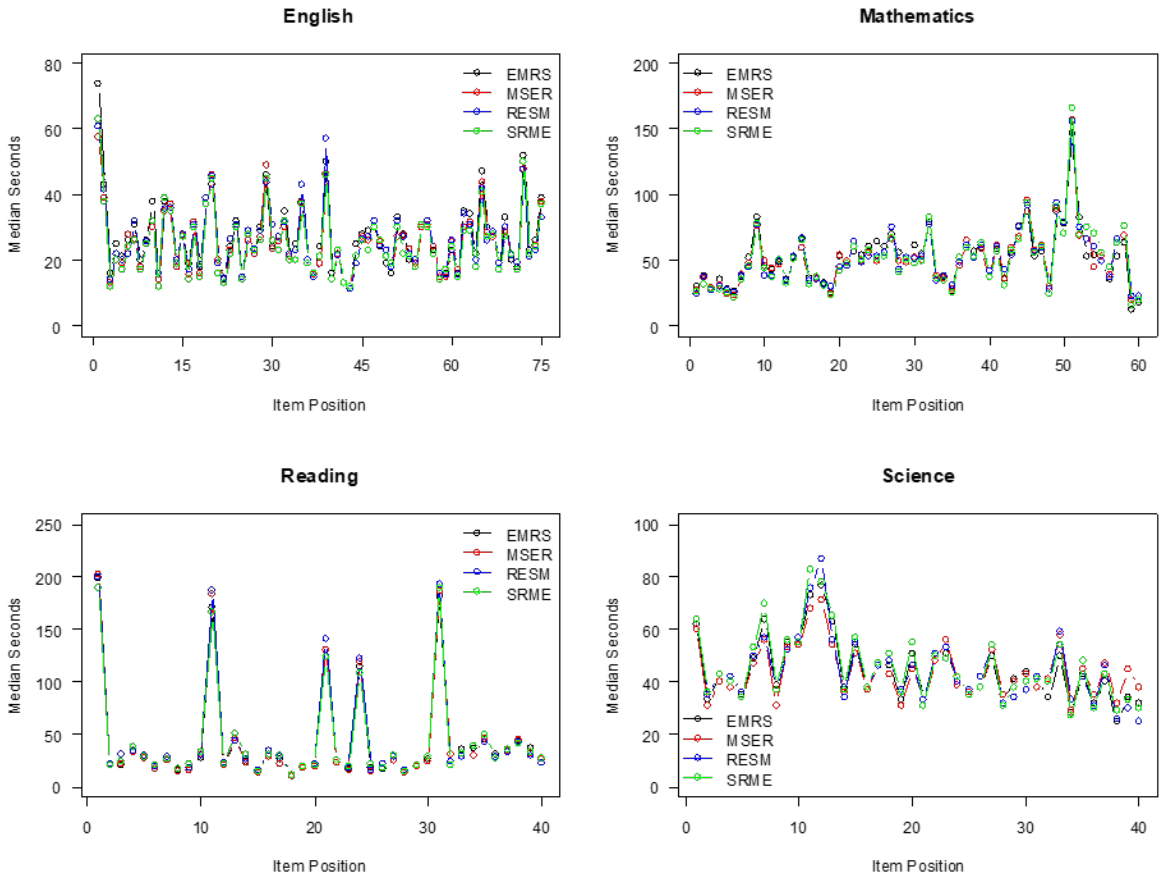
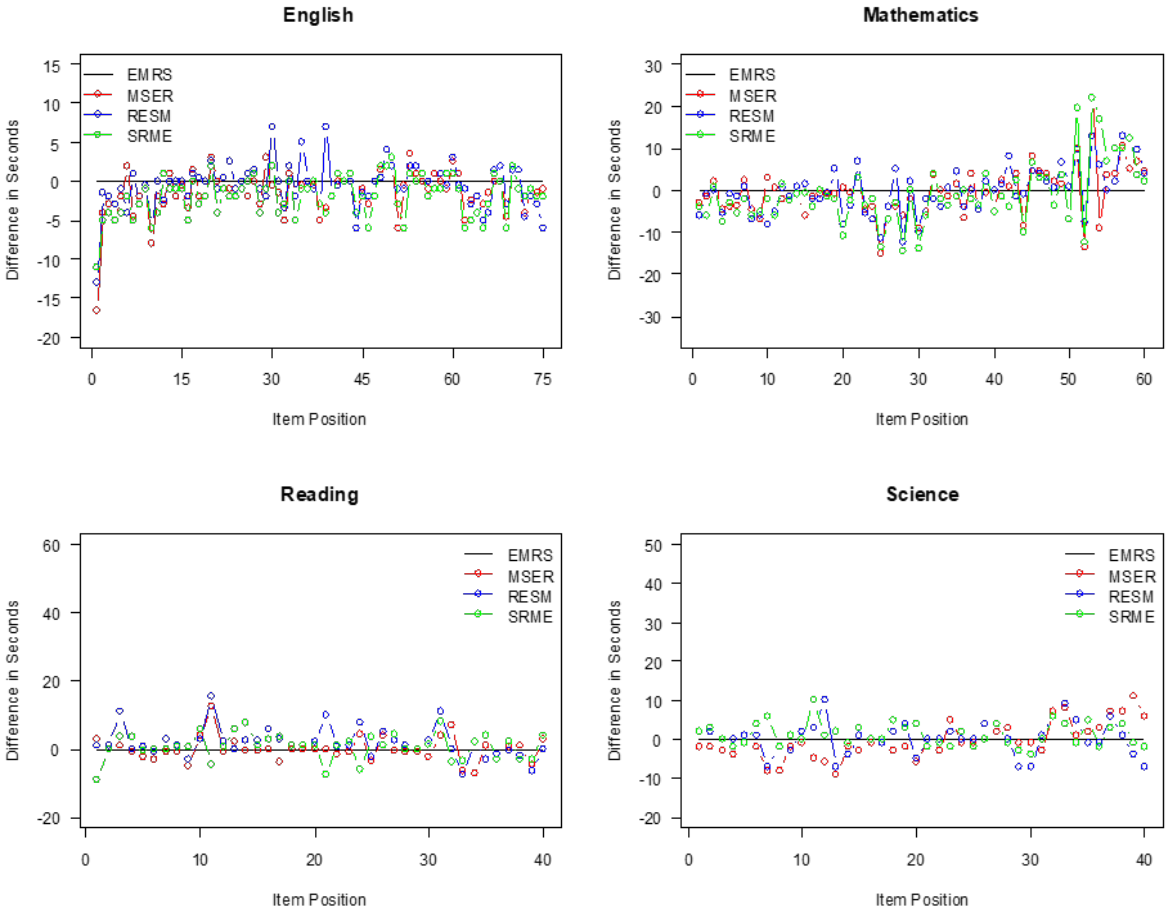


Figure 14. Differences in Median Item Latencies for the Top 10% in Terms of Scale Score

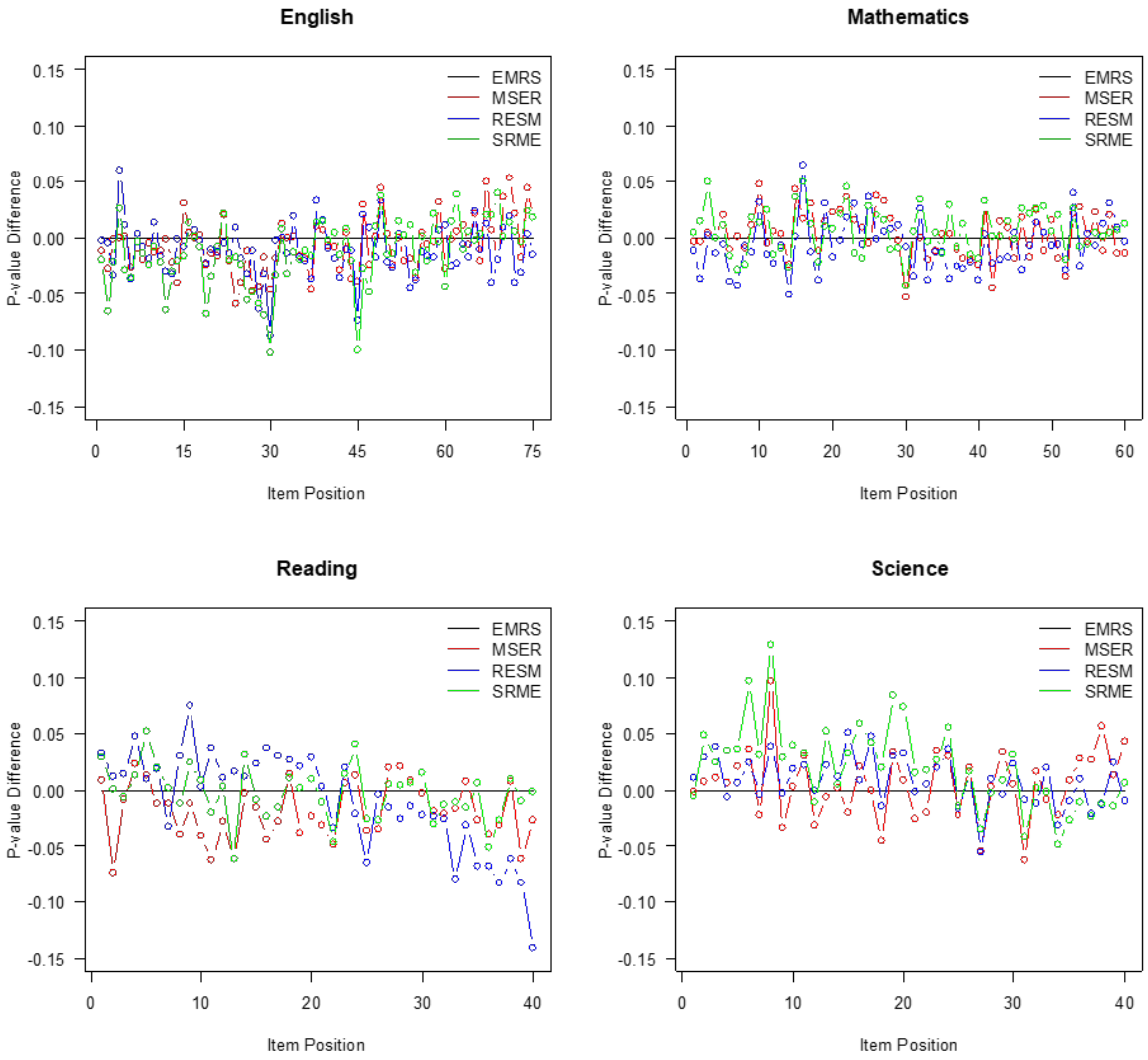


It is difficult to ascertain the reasons for these overall differences in latencies among the four orders. One potential explanation is that students spent too much time at the beginning of the first test and then had to rush to finish. They then moved more quickly throughout subsequent tests. Perhaps running short on time in the first tests prompts students to work faster on later tests. Another possibility is that students need time to “warm up” or “settle in” the testing environment. That is, it takes them a little longer to do some tasks at the beginning of the first test. In later tests, tasks become easier for students to do quickly. Additional studies would be needed to provide more information on what affects how much time students spend on each item and how the amount of time spent is related to their performance on those items and the overall test. In summary, it is clear that section location had some impact on item latency for online tests; however, the effect is small. Item latency information was not available for the previous study because the test was administered on paper; therefore, we are unable to compare these results with the previous study. With that in mind, these findings may not generalize to other modes.

Item Difficulty

Figure 15 shows the differences in p-values between the standard order group and the alternative orders for each item for all four subjects. For the English test, the items that appeared at the beginning of the test tended to have slightly higher p-values for the standard condition than the condition where English was in the last position. There were minimal differences for the rest of the test. The other two orders seemed to be similar to the standard order for English. For the mathematics test, there were small differences between the four orders in terms of p-values. There were some bigger differences for the reading test, however. The earlier the reading test appeared in the testing order, the higher the p-values tended to be for the first half of the test. The differences in p-values were larger at the end of the test when reading was in the first position. The items at the end had lower p-values when the reading test appeared in the first position compared to the other positions. One potential reason for this is related to how much time students spent on those items. The latency data showed that when reading was in first position, students spent more time on the items at the beginning of the test compared to the end of the test. These differences were not nearly as large for the other orders. Similarly, for science, the items at the beginning of the test tended to have slightly higher p-values when the science test appeared first compared to when it appeared last in the standard testing order. This could be related to students spending more time on those items when science was first compared to when it was last.

Figure 15. Differences in *P*-Values for the Four Different Orders



Differential Item Functioning

In these analyses, the reference group was the group that took the ACT test in the standard order. The focal group took the test in one of the other three orders. The numbers of items classified into each of the three ETS classifications (Dorans & Holland, 1993) are shown in Table 11. There were four B-DIF items when the English test was last in the order. These items were toward the middle of the test (positions 19, 29, 30, and 45) and all favored the standard order. There was one item flagged for B-DIF (moderate) when English was in the second position, and there were no flagged items when English appeared in the third position. No mathematics items were flagged for DIF. There were three B-DIF items for reading when it appeared in the first position. These were three out of the last four items, and performance on the standard order was better. For science, there were four B-DIF items and one C-DIF (strong) item when the science test was in the first position. These items, all in the first half of the test, favored those that took science in the first position. The latencies for these items were also a little higher when the science test was taken first compared to when it is taken last in the standard order. The difference in latencies is likely one factor influencing the functioning of those items.

Table 11. DIF Classifications for the Online Order Study

Test	Order	# of A Items	# of B Items	# of C Items
English	MSER	75	0	0
	RESM	74	1	0
	SRME	71	4	0
Mathematics	MSER	60	0	0
	RESM	60	0	0
	SRME	60	0	0
Reading	MSER	40	0	0
	RESM	37	3	0
	SRME	40	0	0
Science	MSER	39	1	0
	RESM	40	0	0
	SRME	35	4	1

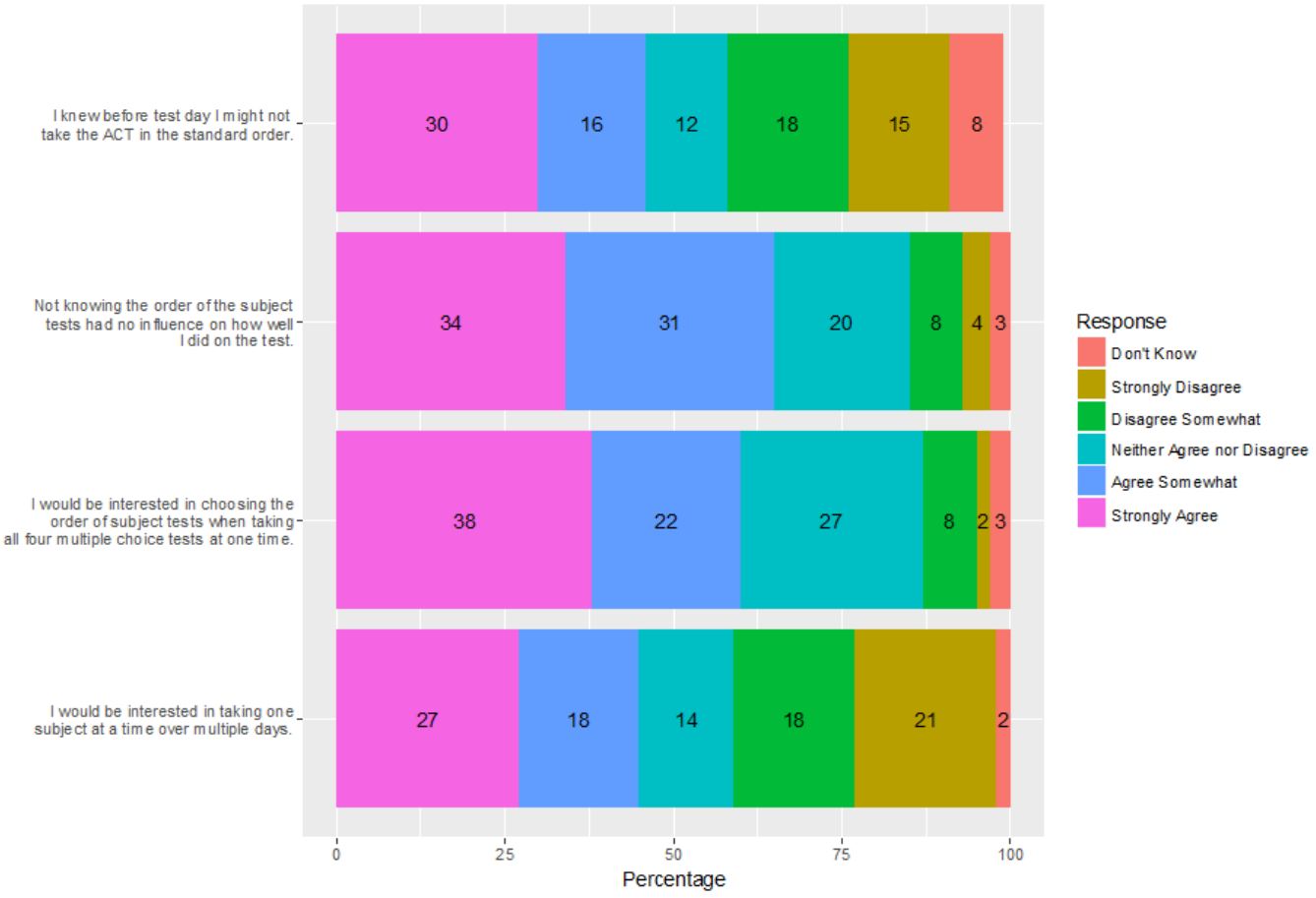
Student Survey Results

After the test, students were emailed a survey to provide feedback on taking the test online and in a potentially different order. ACT received 1,443 responses (response rate of 40.2%). Of the students who responded to the survey, roughly 67% had never taken the ACT prior to the online order study, 21% had taken it one previous time, and 12% had taken it two or more times previously.

Students were asked to select their level of agreement with various statements about their experience taking the test in a different order. These results are shown in Figure 16. Students who disagreed or strongly disagreed with the statements were given the

opportunity to provide further details. In response to the statement, “I knew before test day I might not take the ACT in the standard order,” 46% selected Agree Somewhat or Strongly Agree while 23% selected Disagree Somewhat or Strongly Disagree. In response to this statement, some students expressed that they had little experience with the ACT and were unaware of what the standard order was. Others said they did not find out until they arrived in the testing room and were surprised. Some wished they had known ahead of time, while others said they did not mind and enjoyed taking it in a different order.

Figure 16. Online Order Study Survey Questions About Taking the ACT test in Different Orders



Students were asked to respond to the statement “Not knowing the order of the subject tests had no influence on how well I did on the test.” Around 65% of students said they either agreed somewhat or strongly agreed, while 12% disagreed somewhat or strongly disagreed. Of those that disagreed or strongly disagreed, many seemed to convey in their written comments that they wished they had known the order prior. Some said that not knowing the order added additional stress.

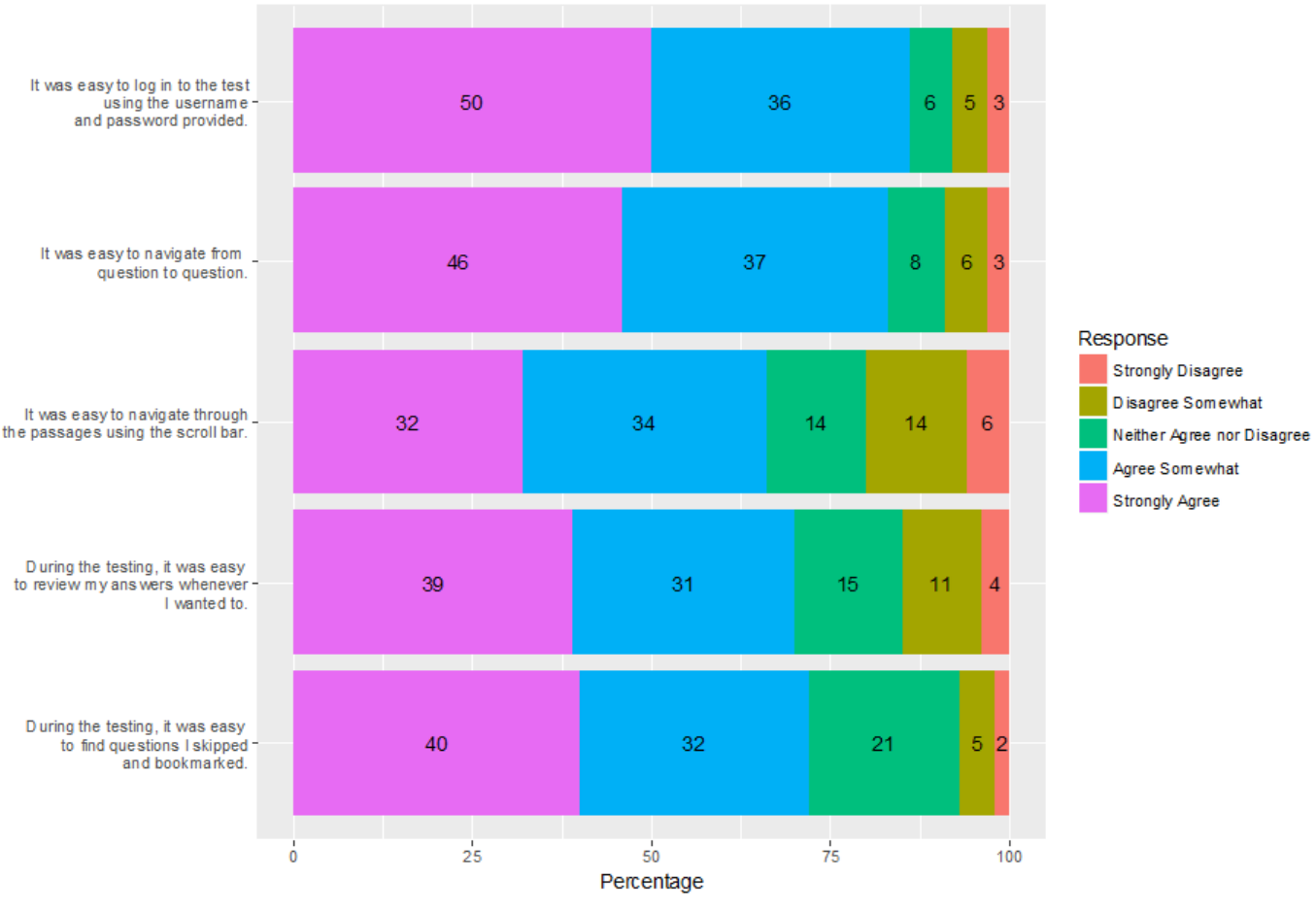
Most students (around 60%) said they either agreed somewhat or strongly agreed with the statement “I would be interested in choosing the order of subject tests when taking all four multiple choice tests at one time.” About 27% selected “neither agree

nor disagree." Around 10% selected either disagree or strongly disagree. Of those students, many said they did so because they did not think that the order mattered and consequently had little interest in selecting an order. A small number also thought it was an unnecessary complication of the whole process.

The responses to the statement "I would be interested in taking one subject test at a time, over multiple days" was more evenly split. Forty five percent said they either strongly agreed or agreed somewhat, while 39% said they disagreed or strongly disagreed. Most of the written responses from students who disagreed somewhat or strongly disagreed seemed to convey that those students preferred to just complete all testing in one sitting. Some said they didn't like the idea of carrying over stress and anxiety over multiple days. A few also expressed that it could be potentially difficult to schedule tests on multiple days with all the other commitments they have.

There were also some questions regarding the students' experience with the testing platform. First, students were asked "Did you experience any technical problems with the online test delivery system (TestNav)?" Twenty eight percent of students indicated they had issues, while 72% indicated they had no issues with TestNav. Students were also asked to indicate their level of agreement with several statements concerning their interactions with the testing platform. These results are summarized in Figure 17. The majority of students agreed somewhat or strongly agreed with the statements about it being easy to log in to the system (86%) and easy to navigate from question to question (83%). Generally, the responses concerning several characteristics of the TestNav platform were positive.

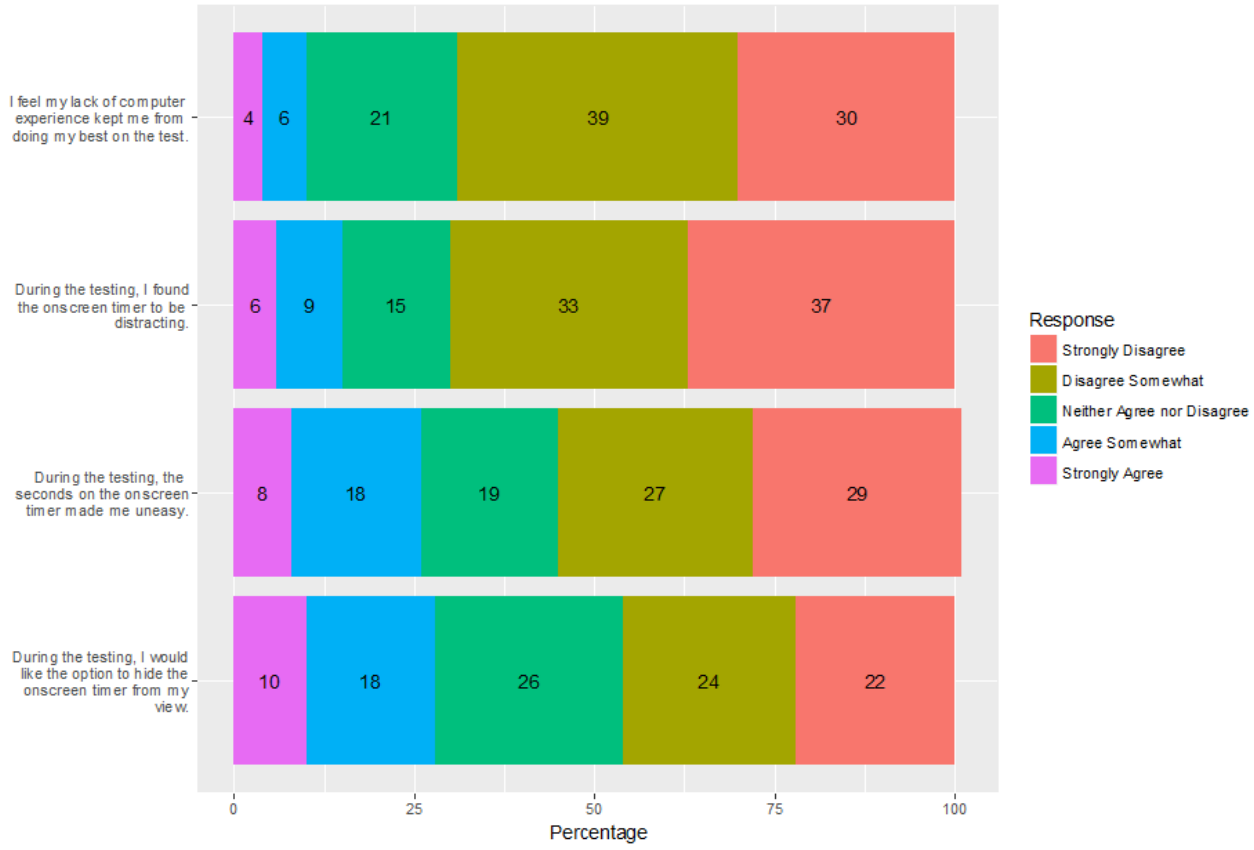
Figure 17. Online Order Study Questions Related to Test Delivery



Another set of questions addressed some of students' experience taking the test on a computer. These results are shown in Figure 18. Overall, the responses to testing on a computer were fairly positive. Only a very small proportion of students felt that their lack of computer experience kept them from doing their best (10%). Students did not seem to have substantial objections to the way the timer was presented on the screen.

A few additional questions were asked concerning students' experiences and preferences regarding technology. Students were also asked "How would you describe your level of comfort using computers (e.g., laptop computers, desktop computers, tablets)?" Over half of the 805 responses were "Very comfortable" and 36 percent said, "Moderately comfortable." Only 3% said they were "not at all comfortable." They were then asked to select the option that best described their preferences for taking the ACT online or on paper. Of the 796 responses, 48% said they would definitely or somewhat prefer to take the ACT online, 20% had no preference, and 31% said they somewhat preferred or definitely preferred taking the ACT on paper.

Figure 18. Online Order Study Questions Related to Taking the Test on a Computer



Students were also given the opportunity to provide any additional thoughts they had about their experience in the online order study. The most common complaint was concerning the lag between questions. Many students complained that it would take anywhere between 2 and 5 seconds (there was some variability in the estimates provided by the students) for the next question to appear. There were also quite a few comments about how the scrolling and navigation, particularly for science and reading, were problematic at times. There were some general complaints about online testing,

but there were also those who expressed that they really liked being able to take it online. Some students said they would feel even more positively towards online testing if it meant they could get scores back sooner. Some even expressed an interest in having scores immediately upon completion. There were also quite a few comments about the timer. Of those, most seemed to say that they thought it helped their time management.

Summary – Online Order Study

The results from the online order study were similar to the results from the paper order study. Even though we do find small differences in mean scale scores for the four order conditions, the results do not suggest that students perform worse when the subject test appears later in the testing order (i.e., highest mean in the first position, second highest mean in the second position, and so forth). The online order study was also able to provide more information on test-taker behaviors using the latency data. The placement of a test within the order had a small effect on the amount of time spent on particular items within that test, particularly for the reading and science tests. These differences were more pronounced for low-performing students. Overall, the results suggest that students earn similar scores regardless the order in which the section tests are administered.

3. Online Modular Study

The third study was a modular administration of the ACT conducted in September of 2016. Two different groups participated in the study. One group (n = 94) came from a single school that agreed to administer the ACT during the school day, one test at a time throughout the course of a week. It is likely that if modular administrations were an option for schools, many would prefer to do it during the school day, so it was important to conduct the study during the day to see if there were any issues associated with testing during normal school periods. For instance, internet bandwidth could be an issue when many students are using the internet during the school day as opposed to during a Saturday morning when there are fewer students online. The goal was to simulate the conditions of an operational administration with modular testing so that potential obstacles or difficulties could be identified and hopefully avoided when modular testing becomes an option. The second group (n = 8) that participated in the study took a different test on each of four consecutive mornings at the ACT campus. These students were tested on their personal laptops. Students received college-reportable scores as part of the modular study. Final sample size included 101 students; see the Appendix for detailed information on the study sample and how it compared to the population of ACT-tested students.

The administration was as similar as possible to the standard administration of the ACT. The directions read aloud to the students were as similar as possible. There was some additional language regarding how to log in and select the correct test, but those were the only differences. The main difference between this study and the previous two was that one test was administered per day over the course of a week instead of during a single sitting. Students took one test Monday through Thursday, with Friday serving as a makeup time for students who missed a test earlier in the week.

The test form used in this study was the same one used in the online order study. In addition to using the standard conversion to convert raw scores to scale scores, conversions derived from the online order study (e.g., each section test administered first) were used in the modular study. The conversion from when each test appeared first in the order was used because those were similar conditions to taking it modularly.

Descriptive Statistics

The descriptive statistics for the modular study are shown in Table 12. Because there was only a single condition in this study, there is no comparable group with which to make comparisons. These results do show, however, the magnitude of the adjustment needed to make the modular scores comparable to those from a standard ACT test. English comes first in the standard order so no adjustment was needed for those scores. For mathematics, scale scores for this group went down around 0.1 when using the equated conversion. The mean scale scores for the reading test were about 0.4 higher for the equated conversion than they would have been if the standard conversion had been used. Scale scores were 0.3 of a point lower when using the equated conversion compared to the standard conversion for science. In general, the average scale scores obtained via the standard order conversion were similar to the average score scores using the order-specific conversion.

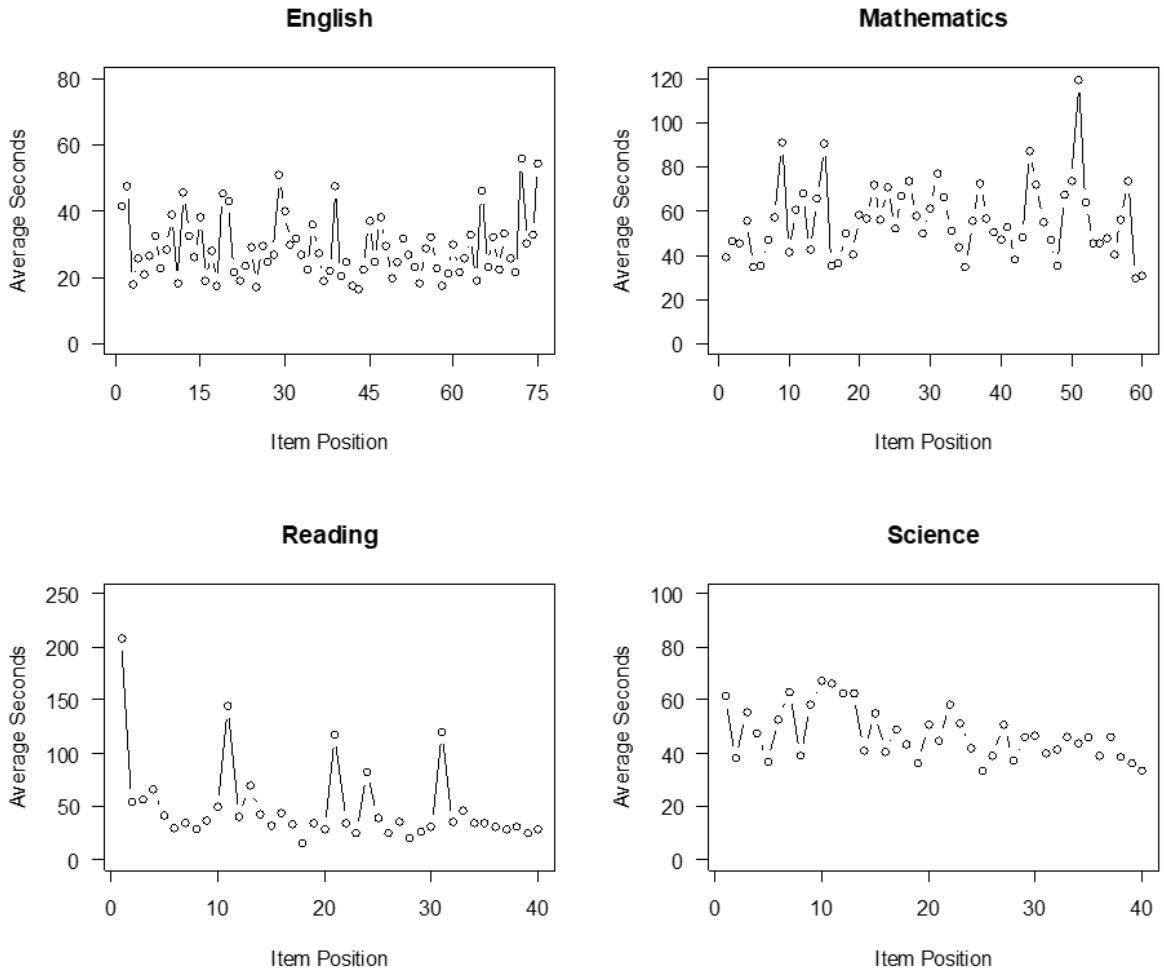
Table 12. Descriptive Statistics for the Modular Study for Raw and Scale Scores

	Test	Mean	SD	Skew	Kurtosis
Scale Scores Using the Standard Conversion	English	22.02	5.54	0.39	2.89
	Mathematics	21.84	4.31	0.08	1.90
	Reading	22.12	4.92	0.40	3.11
	Science	22.14	3.80	0.63	4.50
Scale Scores Using the Order-Specific Conversion	English	22.02	5.54	0.39	2.89
	Mathematics	21.73	4.30	0.17	2.19
	Reading	22.53	5.18	0.28	2.75
	Science	21.82	4.27	0.67	4.41

Latency

The latencies for the modular study are shown in Figure 19. The average number of seconds per item is on the y-axis, and the item position is on the x-axis. The general pattern of the latencies is similar to that for the online order study. This is a different sample of test-takers so the patterns might not be expected to be identical, but the similarity shows that there was not a drastically different latency for this group.

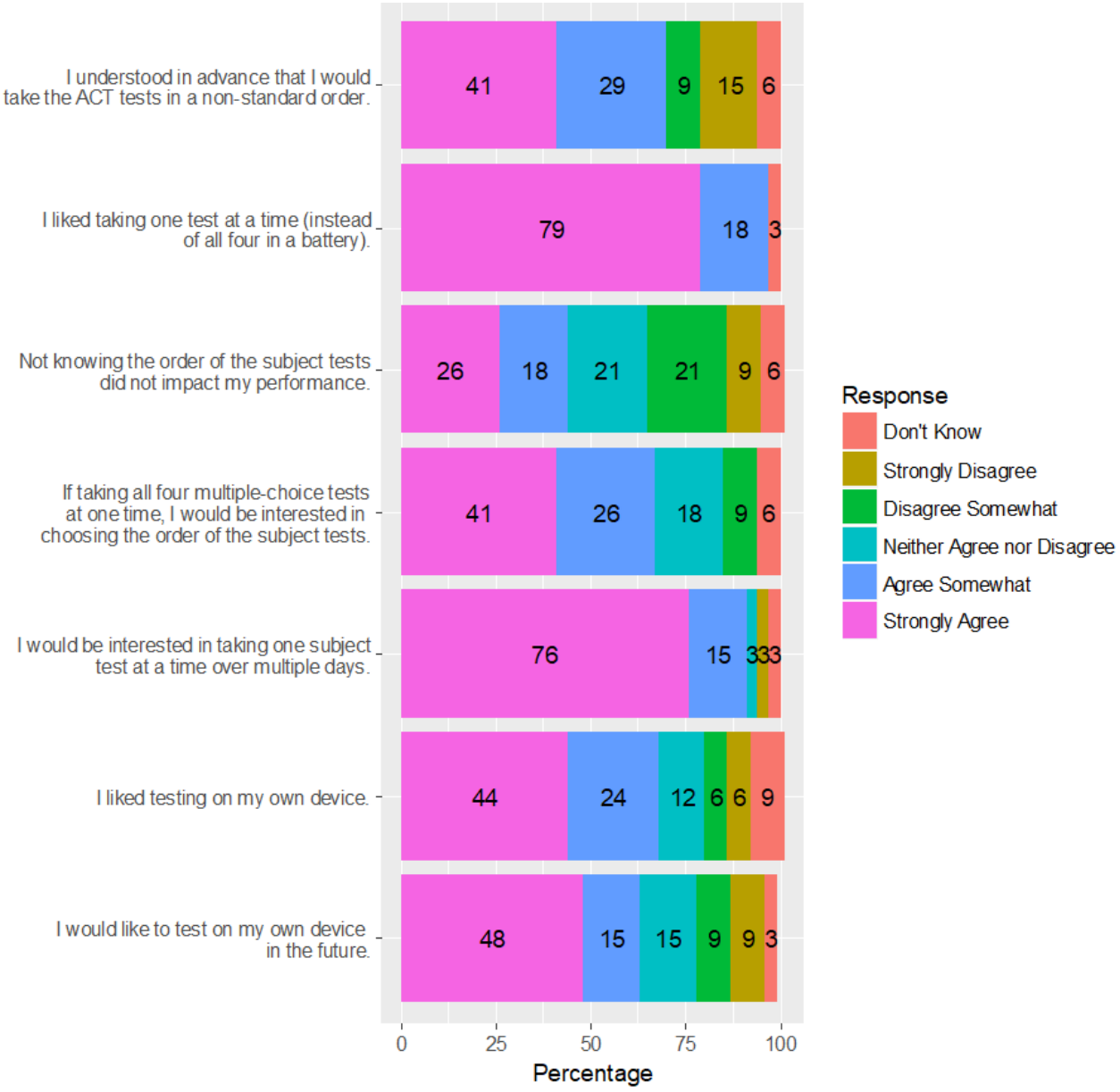
Figure 19. Average Latency for the Modular Study



Survey Results

At the end of the testing window, students were emailed a survey that asked them about their experiences during the modular study. About a third of students responded to the survey. A summary of student responses is shown in Figure 20. Around 97% of students selected either Strongly Agree or Agree Somewhat for the statement “I liked taking one test at a time (instead of all four in a battery).” No students selected Strongly Disagree or Disagree Somewhat. The responses for the statement “Not knowing the order of the subject tests did not impact my performance” were much more varied. About 44% of students strongly agreed or agreed somewhat, while 21% neither agreed nor disagreed, and 21% disagreed somewhat. About 76% strongly agreed with the statement “I would be interested in taking one subject at a time or over multiple days in the future.” This is a much larger percentage than those in the other studies. The majority of respondents expressed a positive experience in using their own laptops and a preference to test on their own device in the future.

Figure 20. Survey Responses for the Modular Study



Students were given the option to provide any additional thoughts they had regarding their experience in the modular study. About half of the students who responded to the survey provided a response. There were several students who expressed that they didn't like testing on computer. The most common issue was with the navigation and scrolling for the reading test. Some also said internet problems caused them to not be able to complete the test as quickly as they would have had it been on paper. Only a few students commented on the modular administration aspect of the study. Two of them said they felt less stress taking it one at a time and enjoyed feeling more rested for each test. One person said that he or she wished there was an option to take a second test in a day.

Summary – Online Modular Study

The modular study provided useful information regarding the feasibility of administering the ACT modularly. In this study, we found that the average scale scores obtained via the standard order conversion were similar to the average score scores using the order-specific conversion and that the item latencies results were similar to that for the online order study. Receiving feedback from students about their experiences with the modular administration was also important and will also help guide future directions.

Discussion

The three studies investigated the differences in test scores and item performance when a test is administered under conditions that differ from the historic test administration model. In the paper and online order studies, the majority of comparisons were not statistically significant (Table 13); when significant differences were observed, the effect sizes were small or very small. In particular, the paper order study detected one significant difference when science was administered first as compared to the traditional order (last); the online study detected two significant differences when reading was administered last as compared to the traditional order (third) and when science was administered first as compared to the traditional order (last). However, in all three cases, the effect size was far below the threshold of 0.20, which is commonly used for classifying an effect as small. Table 13 provides a summary of the three studies including an overview of the design, conditions, and general findings in terms of the impact of different test administration models on ACT scores.

Table 13. Summary of Three Studies

Study	Design	Conditions	Groups	Results
Paper Order Study	Random Assignment, Full Battery Testing, Test Sections Administered in Different Orders	4	EMRS MRSE RSEM SEMR	Of the 12 score comparisons, 1 significant finding (EMRS vs. SEMR for science)
Online Paper Study	Random Assignment, Full Battery Testing, Test Sections Administered in Different Orders	4	EMRS MSER RESM SRME	Of the 12 score comparisons, 2 significant findings (EMRS vs. MSER for reading; EMRS vs. SRME for science)
Online Modular Study	Single Section Testing, One test per Day, Monday through Thursday	1	EMRS	No control group; scores were largely identical after applying the order-specific conversion

It should be noted that no single type of evidence or statistical test can demonstrate that scores from modular administrations with no adjustment from equating can be used interchangeably with scores from standard administrations. The analyses conducted for this study were meant to evaluate a wide range of characteristics that could be affected. Many of the analyses showed no significant differences among the orders.

Some caveats should also be addressed. There were only two forms used in these studies. The content for the paper order study came from a single form, and the content for the online order study and the modular study came from another form. There could be variability in the degree to which different orders affect the performance on different forms. There was not always consistency in the direction of the differences between the results from the paper order study and the online order study. Also, the behaviors of the test-takers could change over time as they become more familiar with different types of administrations. The differences that were observed in these studies could become larger or smaller as students become more familiar with taking the ACT in a format other than on paper and in the standard order. Consequently, when section retesting becomes an available testing option, the data should be continuously monitored to evaluate the degree to which the current findings generalize to an operational setting.

Overall, these three studies demonstrated the validity and feasibility of administering the ACT in different orders or modularly. The magnitudes of the differences among the orders tended to be fairly small.

References

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.

Connover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York, NY: John Wiley.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 36-66). Hillsdale, NJ: Lawrence Erlbaum.

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

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*(2), 129-140.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed score “equatings.” *Applied Psychological Measurement*, *8*(4), 452–461.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG (Version 3.0) [Computer software]. Chicago, IL: Scientific Software International.

Appendix

Table A1. Demographic Characteristics of Three Study Samples and the 2017 ACT-Tested Graduating Class

	Paper Order Study (Study 1)		Online Order Study (Study 2)		Online Modular Study (Study 3)		2017 ACT-tested high school graduates	
	N	%	N	%	N	%	N	%
All students	5,681		3,587		101		2,030,038	
Test date	Oct-15		Apr-16		Sep-16		Sept-14 to Jun-17	
Grade level								
Missing	7	0.1	0	0.0	0	0.0	861	0.0
Sophomore or earlier	26	0.5	1,065	29.7	9	8.9	14	0.0
Junior	2,529	44.5	2,165	60.4	52	51.5	1,056,282	52.0
Senior	3,119	54.9	357	10.0	40	39.6	972,881	47.9
Gender								
Missing	7	0.1	0	0.0	0	0.0	43,138	2.1
Female	3,379	59.5	2,059	57.4	62	61.4	1,047,170	51.6
Male	2,295	40.4	1,528	42.6	39	38.6	939,730	46.3
Race/ethnicity								
African American	556	9.8	294	8.2	0	0.0	256,756	12.7
American Indian	20	0.4	23	0.6	0	0.0	16,135	0.8
White	3,491	61.5	2,482	69.2	95	94.1	1,062,439	52.3
Hispanic	937	16.5	439	12.2	2	2.0	347,906	17.1
Asian	266	4.7	72	2.0	1	1.0	96,097	4.7
Native Hawaiian/ Pacific Islander	13	0.2	6	0.2	0	0.0	6,503	0.3
Multiracial	200	3.5	157	4.4	1	1.0	86,119	4.2
Missing	198	3.5	114	3.2	2	2.0	158,083	7.8
Parents' ed level								
Missing	900	15.8	534	14.9	4	4.0	429,006	21.1
No college	799	14.1	623	17.4	1	1.0	357,527	17.6
Some college	1,169	20.6	943	26.3	16	15.8	417,107	20.6
Bach degree	1,526	26.9	874	24.4	42	41.6	455,622	22.4
Beyond Bach	1,287	22.7	613	17.1	38	37.6	370,776	18.3
Income level								
Less than $36,000	844	14.9	668	18.6	2	2.0	426,337	21.0
$36,000 to $80,000	1,364	24.0	935	26.1	24	23.8	453,604	22.3
More than $80,000	1,904	33.5	1,155	32.2	62	61.4	538,964	26.6
Missing	1,569	27.6	829	23.1	13	12.9	611,133	30.1

Table A2. Academic Characteristics of Three Study Samples and the 2017 ACT-Tested Graduating Class

Academic Indicator	Paper Order Study (Study 1)		Online Order Study (Study 2)		Online Modular Study (Study 3)		2017 ACT-tested high school graduates	
	N	Mean	N	Mean	N	Mean	N	Mean
HSGPA	5,021	3.46	3,196	3.42	98	3.58	1,646,388	3.27
ACT Composite score	5,681	21.70	3,587	20.30	101	22.20	2,030,038	21.00
ACT English score	5,681	21.30	3,587	19.80	101	22.20	2,030,038	20.30
ACT math score	5,681	21.20	3,587	20.00	101	21.70	2,030,038	20.70
ACT reading score	5,681	22.30	3,587	20.60	101	22.70	2,030,038	21.40
ACT science score	5,681	21.50	3,587	20.40	101	21.90	2,030,038	21.00