



**Comparability Analyses of the Online and Paper-and-Pencil MEAP**

**Final Report**

**Pearson Educational Measurement**

**July 27, 2006**

**Comparability Analyses of the Online and Paper-and-Pencil MEAP  
Final Report  
Pearson Educational Measurement  
July 27, 2006**

**Executive Summary**

As part of the fall 2005 MEAP administration, the Office of Educational Assessment and Accountability (OEAA) undertook a study of tests delivered by computer (referred to in this document as “online testing”). The purpose of the study was to pilot online testing for the grade 6 ELA and Social Studies tests administered as part of the MEAP program, and to compare performance on online and paper administrations of the test. The study drew primarily from schools that have been provided with laptop computers as part of the Freedom to Learn initiative managed by Ferris State University. For the participating schools, online tests were administered to grade 6 students using the laptop computers that were assigned to them. In addition, a small number of schools participated by administering the online MEAP tests through computer facilities maintained in classrooms or computer labs.

There were 2,403 online tests administered to students in 19 Michigan school districts. Participating schools tested online during the same administration window as the paper-and-pencil MEAP. Provisional (raw) scores on the online tests were provided within 48 hours of the online administration. These included multiple-choice total scores and scores on grade-level expectations. Preliminary scores on constructed response (CR) items for the online tests were generated using Pearson Knowledge Technologies’ (PKT) automated essay scoring engine, referred to as the Intelligent Essay Assessor (IEA). However, CR items were also scored along with responses based on the paper test administration using traditional essay scoring procedures prior to final score reporting.

The primary results of the study are highlighted below:

- Analyses of student scores indicated that the online and paper versions of each test were sufficiently comparable for the same score conversion tables to be used in both modes.
- Analyses of the CR items in reading (one six-point item) and writing (one six-point item and one four-point item) indicated lower performance for students testing online versus students testing by paper. Assigning online CR scores based on the higher of automated and human ratings reduced the performance differences so that the students participating in the pilot were not disadvantaged.
- Comparisons between automated and human scores of the online CR items indicated that the agreement of the automated and human ratings can be increased significantly by improving the representativeness of the CR responses used to train the automated scoring engine. In a follow up study where the automated scoring engine was trained using a representative sample of the operational online essays, agreement rates with human scores increased by 7 percent to 25 percent across the four CR items.
- Overall, student reactions to online testing were overwhelmingly positive. Over 90 percent of the online students agreed or strongly agreed with the statement, “Overall, I am happy that I took this test on computer rather than on paper” and over 90 percent indicated that they would recommend or strongly recommend taking the test by computer to other students.
- Survey results indicated that some students were not comfortable with writing essays online. Only 30 percent of ELA respondents agreed or strongly agreed with the statement, “It is easier for me to write an essay on paper than on the computer.” When asked, “How often do you use a computer for writing papers or essays?”, 26 percent of ELA students indicated either “less than once a month” or “never”.

In general, the results of the online MEAP comparability study indicated that the pilot test was successful and provided support for continued online testing efforts in Michigan. However, the performance

differences on the CR items and the evidence from the surveys of some student discomfort with composing essays online has implications for future online testing efforts in Michigan. One implication is that any students taking the MEAP online in subjects that include essay-type items should be given practice with online essay composition prior to testing, perhaps through the use of required tutorials. In addition, no student should be required to take the MEAP online in subjects that include essay-type items if they are not comfortable with composing essays online. Teachers should confirm that their students have had sufficient exposure to and comfort with writing online before assigning them to take the MEAP online. A second implication is that the introduction of online MEAP should be sensitive to the instructional practices within schools with respect to online writing. That is, if schools are not yet consistently instructing students in writing online, it may be premature to offer MEAP tests with extended constructed response items online. Of course, the opposite is true as well. When schools are to the point where writing instruction and student writing is almost entirely online, it will no longer make sense to offer paper-and-pencil MEAP tests with extended constructed response items. This is perhaps a natural aspect of the inevitable evolution of technology within both instruction and assessment in Michigan, yet one that needs to be delicately considered.

**Comparability Analyses of the Online and Paper-and-Pencil MEAP  
Final Report  
Pearson Educational Measurement  
July 27, 2006**

**Introduction**

The Michigan Educational Assessment Program (MEAP) is administered in mathematics, English language arts (ELA), which includes reading, writing and listening), science, and social studies to students at the elementary, middle, and high school levels. The program's purpose is to provide information on the status and progress of Michigan education in specified content areas to the Michigan students, parents, teachers, and other Michigan citizens, so that individual students are helped to achieve the skills that they have missed and educators can use the results to review and make improvements to the school's instructional program across grade levels.

As part of the fall 2005 MEAP administration, the Office of Educational Assessment and Accountability (OEAA) undertook a study of tests delivered by computer (referred to in this document as "online testing"). The purpose of the study was to pilot online testing for the grade 6 ELA and Social Studies tests administered as part of the MEAP program, and to compare performance on online and paper administrations of the test. The study drew primarily from schools that have been provided with laptop computers as part of the Freedom to Learn initiative managed by Ferris State University. For the participating schools, online tests were administered to grade 6 students using the laptop computers that were assigned to them. In addition, a small number of schools participated by administering the online MEAP tests through computer facilities maintained in classrooms or computer labs.

Pearson Educational Measurement (PEM), the MEAP testing vendor, administered the online MEAP pilot and analyzed the data to compare online and paper student performance. This document provides results of analyses to study the comparability of the online and paper assessments. These analyses included an assessment of comparability between students testing online compared with students testing by paper, comparisons of automated scoring and human scoring of the constructed-response items, comparisons of item-level performance on the online and paper test versions, and summary of surveys administered as part of the online test to obtain students' reactions to testing by computer.

**Research Methodology**

**Comparability Sample**

There were 2,403 online tests administered to students in 19 Michigan school districts. Table 1 summarizes the numbers of students testing online by test district. Note that in many cases the same students tested online in both social studies and ELA.

Insert Table 1

Participating schools tested online during the same administration window as the paper-and-pencil MEAP. Provisional (raw) scores on the online tests were provided within 48 hours of the online administration. These included multiple-choice total scores and scores on grade-level expectations. Preliminary scores on constructed response (CR) items for the online tests were generated using Pearson Knowledge Technologies' (PKT) automated essay scoring engine, referred to as the Intelligent Essay Assessor (IEA). However, CR items were also scored along with responses based on the paper test administration using traditional essay scoring procedures prior to final score reporting.

## Comparability Design

The comparability data were analyzed using a “matched groups” design. In this design, test scores from the spring 2005 administration of the grade 5 social studies test and additional demographic information were used as matching variables. Specifically, for each grade 6 student testing by computer in a given subject, a grade 6 student testing by paper with the same profile of matching variables was selected at random and assigned to the comparison group. Computer and paper performance was compared between the two matched groups, and the sampling and comparisons were replicated 100 times. Since grade 5 students did not take ELA in winter 2005, the grade 5 social studies scores were as a matching variable for both the social studies and ELA test comparisons. Separate analyses were done for students for the reading and writing portions of the ELA test, as these areas are scaled and reported separately. (Listening is not assessed in grade 6.) For the CR questions, the scores used for the comparability analyses were the final reported scores. As directed by the OEAA, the reported CR scores for online students were based on the *higher of the automated and human ratings*.

For both the online and paper conditions, the students selected for the comparability study were restricted to those students with spring 2005 social studies scale scores. The matching rates were 92.4 percent for social studies (113,824 of 123,128 possible matches) and 92.5 percent for ELA (113,548 of 122,795 possible matches). For ELA, the paper sample was limited to those students who were administered the same operational items as the students testing online (i.e., students taking forms 1 to 8).

PEM matched students in the online group to student in the paper group using on three variables: spring grade 5 social studies scale score, gender, and ethnicity. To eliminate empty cells, PEM sorted the spring grade 5 scale scores into 30 equal-sized groups and used these groups in the matching. This resulting in a 30 (previous score groups) by 8 (ethnic groups) by 2 (gender) “matching matrix” with 240 total cells.

Because scores were reported for students testing online, we carried out the comparability analyses using an approach outlined by Dorans and Lawrence (1990), which is based on evaluating differences in score conversions in the context of the standard error of equating. Essentially, we evaluated a hypothesis that the same score conversions would be appropriate for the online and paper forms. To assist in evaluating the equivalence of the online and paper conversions, we utilized the following suggestion from Dorans and Lawrence (1990): “To assess equivalence, it is convenient to compute the difference between the equating function and the identity transformation, and to divide this difference by the standard error of equating. If the resultant ratio falls within a bandwidth of plus or minus two, then the equating function is deemed to be within sampling error of the identity function” (p. 247). Our analyses calculated standard errors of the linkings based on 100 replications of the matched samples analyses.

We repeated these analyses twice for each content area. In one analysis, the CR scores for online students were based on the higher of the automated and human scores. This analysis was used to evaluate whether the paper-and-pencil score conversion tables would be appropriate for reporting scale scores for the online students. In the second analysis, only the human CR scores were used for the online students. This second analysis was used to more generally assess the impact of administration mode on student scores.

The matched sample analyses involved calibrations and linkings utilizing the WINSTEPS estimation program (Linacre, 2003) to estimate the parameters of the partial credit Rasch model. To obtain standard errors of the linkings, bootstrap sampling procedures over the 100 replications were used (c.f., Kolen & Brennan, 2004, p. 232-235). The advantage the bootstrap approach is that it incorporates the error in drawing matched samples of paper students for the comparability comparisons. The specific steps in the comparability analyses (repeated 100 times) were as follows:

Step 1: Draw a random sample of students with replacement from the online student data that is the same size as the overall online student sample.

Step 2: Draw a stratified sample of students with replacement from the paper student data that *exactly matches* the frequencies in the matrix of score group by ethnicity by gender frequencies observed for the online group in step 1.

Step 3: Use the paper test raw score-to-ability conversion table to assign ability scores for each student in the paper group sample drawn in step 2. Calculate the mean and standard deviation of the ability values.

Step 4: Use WINSTEPS to calibrate the online group data. Transform the estimated abilities from this calibration so that the mean and standard deviation is equal to the mean and standard deviation of ability values for the paper sample calculated in step 3.<sup>1</sup>

Step 5: Re-calibrate the online group data, anchoring the WINSTEPS run using the transformed ability values calculated in step 4. The WINSTEPS online raw score-to-ability conversions are obtained from this calibration. Use the paper form scaling constants to transform the WINSTEP ability values for each raw score to scale scores

These steps were repeated 100 times. The mean of the ability conversions and transformed scale scores over replications at each possible raw score were compared with the paper form ability conversion and transformed scale scores. The standard deviations of the ability conversions and scale scores over replications at each possible raw score represented the bootstrap standard errors and were used to interpret the differences between the paper and online abilities and scale scores.

### Item Level Performance Comparisons

For the online and paper samples selected at each bootstrap replication, the response data were analyzed to calculate average item scores (i.e., percent correct or p-values for multiple-choice items and mean scores for CR items). These values were averaged over the 100 replications and overall bootstrap means and standard deviations (or bootstrap standard errors) were calculated. In addition, a z-difference statistic was calculated as follows:

$$Z_{dif} = \frac{\bar{X}_{online} - \bar{X}_{paper}}{\sqrt{SE_{online}^2 + SE_{paper}^2}}$$

where  $\bar{X}_{online}$  and  $\bar{X}_{paper}$  are the mean of the online and paper p-values over the 100 replications, and  $SE_{online}$  and  $SE_{paper}$  are the bootstrap standard errors over the 100 replications

For these analyses, the CR scores based on human scoring were used rather than those based on the higher of the automated and human scores.

### Ethnic and Gender Performance Comparisons

For the online and paper samples selected at each bootstrap replication, mean scores were calculated by ethnic group and gender. (Only the American Indian/Native American, Black, Hispanic, and White groups were sufficiently large to be included in the ethnic comparisons.) These mean scores were also averaged over the 100 replications and overall bootstrap means and standard deviations (or bootstrap standard errors) were calculated. As with the item level statistics, a z-difference statistic was calculated for each subgroup as follows:

---

<sup>1</sup> This procedure was modified for the reading and writing measures in two ways: 1) two “dummy” records with perfect scores on the CR items were added to the online data and the paper sample frequencies to ensure that the calibration theta-to-score tables would extend to the entire score scales; and 2) WINSTEP calibrations were run using a two-step process as was done for the operational ELA calibrations.

$$Z_{dif} = \frac{\bar{X}_{online} - \bar{X}_{paper}}{\sqrt{SE_{online}^2 + SE_{paper}^2}}$$

where  $\bar{X}_{online}$  and  $\bar{X}_{paper}$  are the grand means of the online and paper means over the 100 replications, and  $SE_{online}$  and  $SE_{paper}$  are the bootstrap standard errors over the 100 replications. It should be noted that, although the procedures for selecting students within each bootstrap iteration ensured that the same number of online and paper students were selected for each subgroup, the numbers of students in the subgroups could differ slightly across iterations. The grand means and bootstrap standard errors for each subgroup weighted each iteration equally regardless of differences in sample sizes across iterations.

### Comparison of Automated and Human Constructed Response Scores

Automated and human scores were compared for four CR items: a social studies CR item worth a maximum of three points, a reading CR item worth a maximum of six points, and two writing CR items, one worth a maximum of six points and the other worth a maximum of four points. Only papers for students in the matched comparability analysis file with nonzero scores were included in this analysis. Two-way contingency tables of the automated and human scores were calculated that included the marginal score level distributions. In addition, exact agreement and adjacent agreement rates were calculated between the automated and human scores based on the CR items with nonzero scores. The agreement rates were calculated for papers with nonzero scores only because the thresholds for defining “unscorable” in the automated scoring were set liberally so that human scoring would take precedence for responses that were detected by the automated scoring as unusual. This resulted in a number of responses that were scored relatively high by human raters but that would have defaulted to a score of zero based on an unscorable automated rating.

### Survey Analyses

Each online student was administered a 14-item survey upon completing their test. Slightly different versions of the survey were administered for the social studies and ELA tests. In addition, each student was given the opportunity to enter an open-ended comment on their testing experience. Responses to the survey questions were summarized, but the open-ended comments were not formally analyzed. PEM will be happy to provide a file containing these comments to OEAA upon request.

## Results

### Summary Statistics

Table 2 presents summary statistics (mean, standard deviation, minimum and maximum scores) for the social studies, reading, and writing scores for the grade 6 students included in the comparability study. There were 1,095 online students and 112,729 paper students in the social studies comparability samples. For ELA, there were 1,133 online students and 42,872 students in the paper sample. The ELA statistics in Table 2 are presented separately for reading and writing, as these two measures are separately scaled on the operational test.

#### Insert Table 2

There are two entries of summary statistics for the online tests. One entry is based on using only the human scores for the CR items. The second entry is based on using the higher of the automated and human scores for the CR items. As would be expected, mean raw scores are higher when the CR scores are based on the higher of the automated and human scores.

It can be seen in Table 2 that for all three measures (social studies, reading, and writing), the mean raw score was slightly *higher* for the paper sample versus the online sample when the CR scores were human-based. When the CR item scored using the higher of the automated and human scores, the mean writing score for the online sample was higher than the mean score for the paper sample. However, the means of the spring 2005 grade 5 social studies scale scores were slightly *lower* for the paper samples versus the online samples. The last column of Table 2 presents the correlations between fall 2006 raw scores and spring 2005 social studies scale scores. As might be expected, these correlations were highest for social studies and lowest for writing. For social studies, correlations with the previous spring's scores were slightly higher for the online sample than for the paper sample. For reading and writing, the paper samples had a slightly higher correlation with the previous spring's scores than the online samples. Not shown in Table 2 are the correlations between the reading and writing measures, which were higher in the paper sample than in the online sample (0.64 vs. 0.59).

Table 3 presents the sample frequencies by ethnicity for the online and paper groups. For both social studies and ELA, there was a higher proportion of White and American Indian/Alaskan Native students and a lower proportion of Black and Hispanic students in the online samples compared to the paper samples.

Insert Table 3

Table 4 presents the sample frequencies by gender. For social studies, the proportion of students by gender were virtually identical in the online and paper samples. For ELA, the proportion of females in the online sample was slightly higher than it was in the paper sample.

Insert Table 4

Table 5 presents the sample frequencies of CR item scores. The social studies CR item frequencies are presented in the upper left-hand side, the reading CR item frequencies are presented in the upper right-hand side, and the writing two CR item frequencies are presented in the lower portion of the table. For the online samples, the statistics based on the higher of automated and human scoring are listed in parentheses following the statistics based on human scoring only. For all of the CR items, the performance of the online group was lower than the performance of the paper group under human scoring. However, for three of the four CR items, the performance of the online group was higher than the performance of the paper group when the higher of the automated and human scores was assigned.

Insert Table 5

### Equating Comparability Analyses

Tables 6 to 8 present the results of the bootstrap equating analyses over the 100 replications. The columns of these tables are defined as follows:

| <b>Statistic</b> | <b>Definition</b>   |
|------------------|---|
| RS               | Raw Score   |
| Paper_Th         | WINSTEPS theta value for the operational paper test                             |
| Online_Th        | Mean WINSTEPS theta value for the online test over 100 replications             |
| Diff             | Difference between paper minus online theta                                     |
| S.E.             | Bootstrap standard error of differences between paper minus online thetas       |
| Paper_SS         | Operational paper scale score   |
| Online_SS        | Unrounded mean scale score for the online test over 100 replications            |
| Diff             | Difference between paper minus online scale score                               |
| S.E.             | Bootstrap standard error of differences between paper minus online scale scores |

These analyses utilized CR scores for the online group that were based on the higher of the automated and human ratings, as did the operational scoring for the online students.

The trends in Tables 6 to 8 are consistent with the data previously summarized: for social studies and reading, the thetas and resulting scale scores at each raw score point are slightly higher for the online group than for the paper group. This indicates that the online versions of the tests were slightly more difficult than the paper versions. For writing, however, the thetas and scale scores at each raw score point are lower for the online group than for the paper group, indicating that the online version of the writing test was easier than the paper version.

#### Insert Tables 6 to 8

Figure 1 presents differences between the online and paper score conversions along with the intervals defined by plus and minus two bootstrap standard errors of equating suggested by Dorans and Lawrence (1990).<sup>2</sup> It can be seen that for social studies and reading, the differences are close to or exceed the +2 SE interval over most of the score range. For writing, the differences are close to or exceed the -2 SE interval over most of the score range.

#### Insert Figure 1

Based on the bootstrap equating results presented in Tables 6 to 8 and Figure 1, PEM recommended and OEAA approved using the paper raw to scale score conversion table for social studies, reading, and writing. The rationale for this decision was as follows:

- 1) the “met standard” cut scores for the online and paper conversions corresponded to the same raw score for all three tests
- 2) The online vs. paper scale score differences were mostly within the  $\pm 2$  SE interval, especially in the vicinity of the “met standard” cuts
- 3) Since mode differences for reading and writing tests will tend to cancel each other out when they are combined to create the English language composite score

Although not used for score reporting decisions, Tables 9 to 11 present bootstrap equating analyses utilizing CR scores for the online group that were based only on human ratings. In these tables, for virtually all score points across all three measures the online theta and resulting scale score at each raw score point is higher than the paper theta and resulting scale score at that raw score point. For most raw score points, the scale score differences exceed the +2 SE interval. In particular, the scale score differences for writing indicate that the online version of this test is more difficult than the paper version when the two CR items are scored using only the human ratings.

#### Insert Tables 9 to 11

### **Item Level Comparability Analyses**

Tables 12 to 14 summarize the item level comparability analyses. These tables list mean p-values over 100 bootstrap replications for the online and matched paper samples, bootstrap standard errors of these p-values, and z-statistics used to flag items that differed significantly in difficulty ( $|z\_dif| > 2$ ). These tables reflect the overall tendencies for the online version of the tests to be slightly more difficult than the paper versions. For social studies (Table 12), 11 of the 47 items were flagged. All of these items were more difficult online than they were on paper. It should be noted that the social studies CR item was *not* flagged using the  $z\_dif$  statistic and the flagging criterion of  $\pm 2$  SE. For reading (Table 13), five of 38 items were

---

<sup>2</sup> It should be noted that the scale score differences in Tables 6 to 8 are in comparison to *rounded* paper test scale scores, which explains the unsmoothed pattern in the differences across scale scores. In addition, the artificial assignment of paper test scale scores of 600 reduces the scale score differences at this score level.

flagged; all of these items were more difficult in the online version than in the paper version. There were particularly large differences between the online and paper versions of the Reading CR item. For writing (Table 13), four of the seven items were flagged; for one of these (item 3) online performance was *higher* than paper performance. Both of the writing CR items were flagged using the  $z_{dif}$  statistic and indicated relatively large performance differences between online and paper performance.

#### Insert Tables 12 to 14

Inspecting the content of the flagged items was beyond the scope of this study. However, it is recommended that content experts review the items that were flagged as they were presented online and consider hypotheses that might explain why the online and paper performance differed for these items.

### **Ethnic and Gender Performance Comparisons**

Table 15 presents comparisons of test raw scores for gender and those ethnic groups with sufficient samples for analysis. As with the item level analyses, Table 15 lists subgroup mean raw scores over 100 bootstrap replications for the online and matched paper samples, bootstrap standard errors of the raw scores, and z-statistics used to flag groups for which online and paper raw scores were significantly different ( $|z_{dif}| > 2$ ). The power of these analyses to detect significant differences were clearly related to the group sample sizes. For all three tests, the only ethnic group for which online and paper test raw scores differences were significant was whites. For social studies, online and paper test raw scores were not significantly different for either males or females, and for writing, online and paper test raw scores were significantly different for both males and females. For reading, online and paper test raw scores were significantly for females but not for males. This result may be deserving of further investigation.

Figure 2 presents the bootstrap means and  $\pm 2$  SE intervals around the means for all subgroups. These plots basically display graphically the same trends that are seen in Table 15.

### **Comparison of Automated and Human Constructed Response Scores**

Table 16 presents cross-tabulations of automated and human CR scores along with rater agreement statistics. These tables indicate varying levels of agreement across the four CR items, with the highest agreement occurring for the social studies and six-point writing CR items (60 and 64 percent, respectively), and lower agreement occurring for the reading and four-point writing items (46 and 48 percent, respectively). Table 16 also indicates that the mean scores for the reading and four-point writing CR items were noticeably different when based on automated versus human ratings. In the case of reading CR1, the mean automated rating was much lower than the mean human rating (1.85 vs. 2.19); however, for writing CR2, the mean automated rating was much higher than the mean human rating (2.15 vs. 1.67).

#### Insert Table 16

Figure 3 presents the percentage frequencies of human versus automated scores for each of the four CR items. These plots reveal differences in the frequencies across the score points based on the human scoring versus the automated scoring.

The relatively low agreement rates between the human and automated scores in the online pilot were expected for several reasons. For example, the social-studies prompt and rubric changed between field-testing and operation use. In addition, for the reading and writing prompts, the distributions of scores for field-test CR responses used to train the automated scoring engine were not representative of the operational distributions. Finally, for the six-point reading and writing CR items, the training set contained few or no examples at the upper two score points (5 and 6). Since the scoring engine learns the appropriate score points from examples in the training set, it is very difficult to accurately grade score points with few or no examples.

To provide a more objective comparison between the human and automated scores based on the pilot data, Pearson Knowledge Technologies staff re-trained the Intelligent Essay Assessor (IEA) using a subset of the operational online essays. Two training sets were used, a set of 200 essays with the same version of IEA that was applied in the pilot, and a larger set of 500 essays with a recently improved version of IEA. Agreement rates between the automated and human scores based on these training samples is shown below, along with the exact agreement rates from the operational pilot as well as agreement rates between human scorers based on responses from the paper administration that were double-scored for reliability purposes.

| CR Prompt         | 200 IEA<br>Exact | 500 IEA+<br>Exact | Oper.<br>Exact | Human<br>Exact |
|-------------------|------------------|-------------------|----------------|----------------|
| Social Studies CR | 61.1             | 66.8              | 60.1           | 71             |
| Reading CR1       | 50.8             | 56.1              | 45.9           | 59             |
| Writing CR1       | 71.0             | 74.7              | 64.6           | 65             |
| Writing CR2       | 61.8             | 73.1              | 47.8           | 75             |

These data indicate that with the use of better and more representative training samples, agreements between IEA and human ratings can be improved to the point where they will approach or exceed the agreement rates obtained between human scorers.

Additional in-depth analyses related to the automated scoring of the online essays was provided by PKT staff and is included in Appendix A of this report.

## Survey Results

The online survey questions and a summary of student responses are summarized in Appendix B for social studies and Appendix C for ELA. Overall, student reactions to online testing were overwhelmingly positive. For example 90 percent of the online students taking social studies and 92 percent of the online students taking ELA agreed or strongly agreed with the statement, “Overall, I am happy that I took this test on computer rather than on paper.” Students seemed comfortable with the computer interface and with navigating through the test. However, the survey results did indicate that some students were not completely comfortable with writing essays online. Only 36 percent of social studies respondents and 30 percent of ELA respondents agreed or strongly agreed with the statement, “It is easier for me to write an essay on paper than on the computer.” When asked, “How often do you use a computer for writing papers or essays?”, more than one-fourth of the students indicated either “less than once a month” or “never” (28 percent for social studies and 26 percent for ELA). This finding is particularly noteworthy since most of the students participating in the online pilot were part of the Freedom to Learn initiative and were given laptop computers to use in their schooling as part of this program.

## Discussion and Implications for Future Online MEAP Efforts

Overall, the comparability analyses of the online and paper-and-pencil MEAP pilot indicated that the online and paper versions of each test were sufficiently comparable for the same score conversion tables to be used in both modes. In general, the online version of the tests seemed to be slightly more difficult than the paper versions. However, assigning the higher of human and automated ratings to the CR items in computing the reported scores mitigated these differences. In general, the comparability results for this pilot were similar to results PEM has found in similar studies with other clients, that is, online tests tend to be slightly more difficult on average than paper tests, typically around one-half of a raw score point or less. There is good reason to believe that these differences will shrink and disappear as online testing becomes routine for students.

The MEAP online pilot was unusual in that it included four constructed response items. The comparability data for the reading and writing tests indicated that differences in online and paper test

performance tended to be relatively large for the CR items. In addition, survey data suggested that some students are not yet comfortable with composing essays online. This has implications for future online testing efforts in Michigan. One implication is that any students taking the MEAP online should be given practice with online essay composition prior to testing, perhaps through the use of required tutorials. In addition, no student should be required to take the MEAP online if they are not comfortable with composing essays online. Teachers should confirm that their students have had sufficient exposure to and comfort with writing online before assigning them to take the MEAP online. A second implication is that the introduction of online MEAP should be sensitive to the instructional practices within schools with respect to online writing. That is, if schools are not yet consistently instructing students in writing online, it may be premature to offer MEAP tests with extended constructed response items online. Of course, the opposite is true as well. When schools are to the point where writing instruction and student writing is almost entirely online, it will no longer make sense to offer paper-and-pencil MEAP tests with extended constructed response items.

With respect to the automated scoring of the CR items, relatively low agreement rates with human scoring appeared to be due to the samples of field-test responses used to train the automated scoring engine. In subsequent analyses done by PKT where the automated scoring engine was trained using a subset of the operational online essays, agreement rates with human scoring improved to a level that was close to or better than agreement rates between human scorers on the paper-and-pencil test. For future online MEAP effort involving automated scoring, alternatives for providing more representative training responses for the automated scoring engine should be explored. This may involve trade-offs between valid training and providing immediate feedback on essay performance (e.g., 48 hour turnaround on automated scores). For example, it may be possible to use essays collected at the beginning of the testing window to train the automated scoring so that immediate score turnaround could occur beginning somewhat later in the testing window.

Finally, there are implications in the overwhelming positive reactions to online testing on the part of the students participating in the pilot. In general, the students seem to understand and embrace the inevitability of online testing. This enthusiastic acceptance by the penultimate stakeholder of the MEAP program is perhaps the best indicator of success for the online pilot.

## References

- Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test forms. *Applied Measurement in Education*, 3, 245-254.
- Yu, L., Livingston, S.A., Larkin, K.C., & Bonet, J. (2004). *Investigating differences in examinee performance between computer-based and handwritten essays* (ETS Research Report RR-04-18). Princeton, NJ: Educational Testing Service.
- Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: methods and practices* (2nd ed.). New York: Springer.
- Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33-38.

Table 1: Participants in the MEAP Online Study by District

| <b>District</b>                 | <b>ELA</b>  | <b>Social Studies</b> |
|---------------------------------|-------------|-----------------------|
| BEAR LAKE SCHOOL DISTRICT       | 25          | 25                    |
| BENDLE PUBLIC SCHOOLS           | 78          | 79                    |
| BERRIEN SPRINGS PUBLIC SCHOOLS  | 82          | 82                    |
| BRANDYWINE COMMUNITY SCHOOLS    | 36          | 36                    |
| BRIDGMAN PUBLIC SCHOOLS         | 50          | 0                     |
| BRIMLEY AREA SCHOOLS            | 26          | 26                    |
| CASEVILLE PUBLIC SCHOOLS        | 26          | 26                    |
| CLARE PUBLIC SCHOOLS            | 105         | 105                   |
| COUNTRYSIDE CHARTER SCHOOL      | 24          | 26                    |
| DETOUR AREA SCHOOLS             | 13          | 13                    |
| GRAND RAPIDS PUBLIC SCHOOLS     | 43          | 0                     |
| LAKESHORE SCHOOL DISTRICT       | 179         | 179                   |
| NEW BUFFALO AREA SCHOOLS        | 41          | 41                    |
| NILES COMMUNITY SCHOOL DISTRICT | 209         | 208                   |
| ONEKAMA CONSOLIDATED SCHOOLS    | 33          | 33                    |
| RUDYARD AREA SCHOOLS            | 53          | 55                    |
| ST. IGNACE AREA SCHOOLS         | 42          | 43                    |
| WATERVLIET SCHOOL DISTRICT      | 70          | 70                    |
| WILLOW RUN COMMUNITY SCHOOLS    | 105         | 116                   |
| <b>Total</b>                    | <b>1240</b> | <b>1163</b>           |

Table 2: Fall 2005 Raw Scores and Spring 2005 Social Studies Scale Scores

| Data Set                   | Grade 6 Social Studies Raw Score |       |      |     |     | Grade 5 Social Studies Scale Score |       |     |     | Correlation<br>Gr 5-Gr 6 |
|----------------------------|----------------------------------|-------|------|-----|-----|------------------------------------|-------|-----|-----|--------------------------|
|                            | N                                | Mean  | Std  | Min | Max | Mean                               | Std   | Min | Max |                          |
| Online Sample <sup>1</sup> | 1095                             | 28.96 | 8.64 | 7   | 48  | 506.40                             | 37.42 | 399 | 620 | 0.79                     |
| Online Sample <sup>2</sup> | 1095                             | 29.16 | 8.57 | 8   | 48  | 506.40                             | 37.42 | 399 | 620 | 0.79                     |
| Paper Sample               | 112729                           | 29.60 | 9.35 | 1   | 49  | 505.96                             | 40.74 | 296 | 748 | 0.76                     |

| Data Set                    | Grade 6 Raw Score |       |      |     |     | Grade 5 Social Studies Scale Score |       |     |     | Correlation<br>Gr 5-Gr 6 |
|-----------------------------|-------------------|-------|------|-----|-----|------------------------------------|-------|-----|-----|--------------------------|
|                             | N                 | Mean  | Std  | Min | Max | Mean                               | Std   | Min | Max |                          |
| Online Reading <sup>1</sup> | 1133              | 26.12 | 6.80 | 5   | 40  | 507.19                             | 37.02 | 399 | 620 | 0.69                     |
| Online Reading <sup>2</sup> | 1133              | 26.29 | 6.80 | 6   | 40  | 507.19                             | 37.02 | 399 | 620 | 0.69                     |
| Online Writing <sup>1</sup> | 1133              | 6.72  | 2.02 | 0   | 14  | 507.19                             | 37.02 | 399 | 620 | 0.49                     |
| Online Writing <sup>2</sup> | 1133              | 7.45  | 2.09 | 0   | 14  | 507.19                             | 37.02 | 399 | 620 | 0.45                     |
| Paper Reading               | 42872             | 26.47 | 7.40 | 0   | 43  | 505.10                             | 41.11 | 296 | 748 | 0.71                     |
| Paper Writing               | 42872             | 7.11  | 1.97 | 0   | 15  | 505.10                             | 41.11 | 296 | 748 | 0.51                     |

<sup>1</sup>Based on human constructed response scores

<sup>2</sup>Based on the higher of human and automated constructed response scores

Table 3: Sample Frequencies by Ethnicity

| Social Studies<br>Ethnicity       | Online Sample |       | Paper Sample |       |
|-----------------------------------|---------------|-------|--------------|-------|
|                                   | N             | Pct.  | N            | Pct.  |
| American Indian or Alaskan Native | 54            | 4.93  | 1066         | 0.95  |
| Asian or Pacific Islander         | 10            | 0.91  | 2490         | 2.21  |
| Black, not of Hispanic Origin     | 118           | 10.78 | 21796        | 19.33 |
| Hispanic                          | 33            | 3.01  | 4346         | 3.86  |
| White, not of Hispanic Origin     | 878           | 80.18 | 81775        | 72.54 |
| Multiracial                       | 2             | 0.18  | 1014         | 0.90  |
| Other                             | 0             | 0.00  | 154          | 0.14  |
| Unknown                           | 0             | 0.00  | 88           | 0.08  |

| English Language Arts<br>Ethnicity | Online Sample |       | Paper Sample |       |
|------------------------------------|---------------|-------|--------------|-------|
|                                    | N             | Pct.  | N            | Pct.  |
| American Indian or Alaskan Native  | 52            | 4.59  | 322          | 0.75  |
| Asian or Pacific Islander          | 11            | 0.97  | 865          | 2.02  |
| Black, not of Hispanic Origin      | 112           | 9.89  | 8943         | 20.86 |
| Hispanic                           | 33            | 2.91  | 1729         | 4.03  |
| White, not of Hispanic Origin      | 923           | 81.47 | 30526        | 71.20 |
| Multiracial                        | 2             | 0.18  | 405          | 0.94  |
| Other                              | 0             | 0.00  | 68           | 0.16  |
| Unknown                            | 0             | 0.00  | 14           | 0.03  |

Table 4: Sample Frequencies by Gender

| Social Studies<br>Gender        | Online Sample |       | Paper Sample |       |
|---------------------------------|---------------|-------|--------------|-------|
|                                 | N             | Pct.  | N            | Pct.  |
| Female                          | 543           | 49.59 | 55955        | 49.64 |
| Male                            | 552           | 50.41 | 56774        | 50.36 |
|                                 |               |       |              |       |
| English Language Arts<br>Gender | Online Sample |       | Paper Sample |       |
|                                 | N             | Pct.  | N            | Pct.  |
| Female                          | 570           | 50.31 | 21298        | 49.68 |
| Male                            | 563           | 49.69 | 21574        | 50.32 |

Table 5: Sample Frequencies of Constructed Response Item Scores\*

| Soc. St. CR  | Online Sample |               | Paper Sample |       | Reading CR1 | Online Sample |               | Paper Sample |       |
|--|---------------|---------------|--------------|-------|-------------|---------------|---------------|--------------|-------|
|  | N             | Pct.          | N            | Pct.  |             | N             | Pct.          | N            | Pct.  |
| 0  | 72 (28)       | 6.58 (2.56)   | 8891         | 7.89  | 0           | 169 (101)     | 14.93 (8.92)  | 4252         | 9.92  |
| 1  | 570 (430)     | 52.05 (39.27) | 52520        | 46.59 | 1           | 258 (258)     | 22.79 (22.79) | 9029         | 21.06 |
| 2  | 345 (529)     | 31.51(48.31)  | 37098        | 32.91 | 2           | 467 (480)     | 41.25 (42.40) | 15204        | 35.46 |
| 3  | 108 (108)     | 9.86 (9.86)   | 14220        | 12.61 | 3           | 107 (157)     | 9.45 (13.87)  | 6843         | 15.96 |
|  |               |               |              |       | 4           | 127 (132)     | 11.22 (11.66) | 7120         | 16.61 |
|  |               |               |              |       | 5           | 4 (4)         | 0.35 (0.35)   | 406          | 0.95  |
|  |               |               |              |       | 6           | 0 (0)         | 0 (0)         | 18           | 0.04  |
| Mean Score   | 1.45 (1.65)   |               | 1.50         |       | Mean Score  | 1.80 (1.98)   |               | 2.11         |       |
| * Online entries outside of parentheses based on human scores; entries within parentheses based on the higher of human & automated scores. |               |               |              |       |             |               |               |              |       |
| Writing CR1  | Online Sample |               | Paper Sample |       | Writing CR2 | Online Sample |               | Paper Sample |       |
|  | N             | Pct.          | N            | Pct.  |             | N             | Pct.          | N            | Pct.  |
| 0  | 17 (14)       | 1.50 (1.24)   | 462          | 1.08  | 0           | 46 (27)       | 4.06 (2.39)   | 1245         | 2.90  |
| 1  | 127 (87)      | 11.22 (7.69)  | 2493         | 5.81  | 1           | 465 (175)     | 41.08 (15.46) | 13592        | 31.70 |
| 2  | 555 (440)     | 49.03 (38.87) | 18727        | 43.68 | 2           | 547 (659)     | 48.32 (58.22) | 23820        | 55.56 |
| 3  | 337 (430)     | 29.77 (37.99) | 16610        | 38.74 | 3           | 71 (260)      | 6.27 (22.97)  | 4049         | 9.44  |
| 4  | 77 (127)      | 6.80 (11.22)  | 3779         | 8.81  | 4           | 3 (11)        | 0.27 (0.97)   | 166          | 0.39  |
| 5  | 13 (26)       | 1.15 (2.30)   | 689          | 1.61  |             |               |               |              |       |
| 6  | 6 (8)         | 0.53 (0.71)   | 112          | 0.26  |             |               |               |              |       |
| Mean Score   | 2.35 (2.60)   |               | 2.54         |       | Mean Score  | 1.58 (2.05)   |               | 1.73         |       |

Table 6: Online to Paper Bootstrap Linking Results – Social Studies  
 Online Scores Based on the Higher of Human and Automated Ratings of CR Items

| RS | Paper_Th | Online_Th | Diff    | S.E.   | Paper_SS | Online_SS | Diff  | S.E.   |
|----|----------|-----------|---------|--------|----------|-----------|-------|--------|
| 0  | -5.3556  | -5.4624   | 0.1068  | 0.0677 | 479      | 476.58    | 2.42  | 1.6058 |
| 1  | -4.1261  | -4.2062   | 0.0801  | 0.0566 | 508      | 506.40    | 1.60  | 1.3442 |
| 2  | -3.3965  | -3.4469   | 0.0504  | 0.0463 | 526      | 524.42    | 1.58  | 1.0994 |
| 3  | -2.9546  | -2.9830   | 0.0284  | 0.0402 | 536      | 535.43    | 0.57  | 0.9551 |
| 4  | -2.6304  | -2.6422   | 0.0118  | 0.0366 | 544      | 543.52    | 0.48  | 0.8687 |
| 5  | -2.3707  | -2.3696   | -0.0011 | 0.0343 | 550      | 549.99    | 0.01  | 0.8142 |
| 6  | -2.1516  | -2.1404   | -0.0112 | 0.0328 | 555      | 555.43    | -0.43 | 0.7780 |
| 7  | -1.9605  | -1.9413   | -0.0192 | 0.0317 | 560      | 560.15    | -0.15 | 0.7531 |
| 8  | -1.7899  | -1.7640   | -0.0259 | 0.0309 | 564      | 564.36    | -0.36 | 0.7343 |
| 9  | -1.6347  | -1.6034   | -0.0313 | 0.0304 | 567      | 568.17    | -1.17 | 0.7203 |
| 10 | -1.4915  | -1.4558   | -0.0357 | 0.0299 | 571      | 571.68    | -0.68 | 0.7090 |
| 11 | -1.3581  | -1.3186   | -0.0395 | 0.0295 | 574      | 574.93    | -0.93 | 0.6997 |
| 12 | -1.2326  | -1.1899   | -0.0427 | 0.0292 | 577      | 577.99    | -0.99 | 0.6924 |
| 13 | -1.1135  | -1.0680   | -0.0455 | 0.0289 | 580      | 580.88    | -0.88 | 0.6856 |
| 14 | -0.9998  | -0.9520   | -0.0478 | 0.0287 | 582      | 583.63    | -1.63 | 0.6801 |
| 15 | -0.8907  | -0.8407   | -0.0500 | 0.0285 | 585      | 586.27    | -1.27 | 0.6752 |
| 16 | -0.7852  | -0.7336   | -0.0516 | 0.0283 | 588      | 588.82    | -0.82 | 0.6711 |
| 17 | -0.6830  | -0.6298   | -0.0532 | 0.0281 | 590      | 591.28    | -1.28 | 0.6672 |
| 18 | -0.5835  | -0.5288   | -0.0547 | 0.0280 | 592      | 593.68    | -1.68 | 0.6637 |
| 19 | -0.4862  | -0.4302   | -0.0560 | 0.0278 | 595      | 596.02    | -1.02 | 0.6609 |
| 20 | -0.3906  | -0.3336   | -0.0570 | 0.0277 | 597      | 598.31    | -1.31 | 0.6582 |
| 21 | -0.2965  | -0.2384   | -0.0581 | 0.0276 | 600      | 600.57    | -0.57 | 0.6556 |
| 22 | -0.2035  | -0.1445   | -0.0590 | 0.0275 | 601      | 602.80    | -1.80 | 0.6535 |
| 23 | -0.1113  | -0.0514   | -0.0599 | 0.0274 | 604      | 605.01    | -1.01 | 0.6515 |
| 24 | -0.0196  | 0.0411    | -0.0607 | 0.0274 | 606      | 607.20    | -1.20 | 0.6496 |
| 25 | 0.0718   | 0.1334    | -0.0616 | 0.0273 | 608      | 609.39    | -1.39 | 0.6479 |
| 26 | 0.1634   | 0.2257    | -0.0623 | 0.0272 | 610      | 611.58    | -1.58 | 0.6465 |
| 27 | 0.2552   | 0.3184    | -0.0632 | 0.0272 | 612      | 613.78    | -1.78 | 0.6449 |
| 28 | 0.3478   | 0.4117    | -0.0639 | 0.0271 | 614      | 616.00    | -2.00 | 0.6435 |
| 29 | 0.4413   | 0.5061    | -0.0648 | 0.0271 | 617      | 618.24    | -1.24 | 0.6421 |
| 30 | 0.5362   | 0.6018    | -0.0656 | 0.0270 | 619      | 620.51    | -1.51 | 0.6410 |
| 31 | 0.6327   | 0.6993    | -0.0666 | 0.0270 | 621      | 622.82    | -1.82 | 0.6400 |
| 32 | 0.7312   | 0.7989    | -0.0677 | 0.0269 | 624      | 625.19    | -1.19 | 0.6389 |
| 33 | 0.8323   | 0.9011    | -0.0688 | 0.0269 | 626      | 627.62    | -1.62 | 0.6378 |
| 34 | 0.9365   | 1.0066    | -0.0701 | 0.0268 | 628      | 630.12    | -2.12 | 0.6368 |
| 35 | 1.0442   | 1.1159    | -0.0717 | 0.0268 | 631      | 632.71    | -1.71 | 0.6359 |
| 36 | 1.1564   | 1.2299    | -0.0735 | 0.0267 | 634      | 635.42    | -1.42 | 0.6349 |
| 37 | 1.2738   | 1.3494    | -0.0756 | 0.0267 | 636      | 638.25    | -2.25 | 0.6342 |
| 38 | 1.3976   | 1.4756    | -0.0780 | 0.0267 | 639      | 641.25    | -2.25 | 0.6334 |
| 39 | 1.5292   | 1.6101    | -0.0809 | 0.0267 | 643      | 644.44    | -1.44 | 0.6327 |
| 40 | 1.6703   | 1.7547    | -0.0844 | 0.0266 | 646      | 647.87    | -1.87 | 0.6321 |
| 41 | 1.8234   | 1.9120    | -0.0886 | 0.0266 | 650      | 651.61    | -1.61 | 0.6321 |
| 42 | 1.9918   | 2.0856    | -0.0938 | 0.0267 | 654      | 655.73    | -1.73 | 0.6328 |
| 43 | 2.1806   | 2.2806    | -0.1000 | 0.0267 | 658      | 660.36    | -2.36 | 0.6347 |
| 44 | 2.3973   | 2.5051    | -0.1078 | 0.0269 | 663      | 665.68    | -2.68 | 0.6388 |
| 45 | 2.6545   | 2.7721    | -0.1176 | 0.0273 | 669      | 672.02    | -3.02 | 0.6470 |
| 46 | 2.9762   | 3.1060    | -0.1298 | 0.0279 | 677      | 679.95    | -2.95 | 0.6627 |
| 47 | 3.4156   | 3.5609    | -0.1453 | 0.0292 | 687      | 690.74    | -3.74 | 0.6919 |
| 48 | 4.1425   | 4.3078    | -0.1653 | 0.0314 | 705      | 708.47    | -3.47 | 0.7460 |
| 49 | 5.3702   | 5.5528    | -0.1826 | 0.0340 | 734      | 738.02    | -4.02 | 0.8072 |

Table 7: Online to Paper Bootstrap Linking Results – Reading  
 Online Scores Based on the Higher of Human and Automated Ratings of CR Items

| RS | Paper_Th | Online_Th | Diff    | S.E.   | Paper_SS | Online_SS | Diff  | S.E.   |
|----|----------|-----------|---------|--------|----------|-----------|-------|--------|
| 0  | -5.4520  | -5.4240   | -0.0280 | 0.0370 | 477      | 478.10    | -1.10 | 0.8573 |
| 1  | -4.2183  | -4.1888   | -0.0295 | 0.0366 | 506      | 506.74    | -0.74 | 0.8485 |
| 2  | -3.4822  | -3.4505   | -0.0317 | 0.0361 | 523      | 523.86    | -0.86 | 0.8363 |
| 3  | -3.0329  | -2.9992   | -0.0337 | 0.0356 | 534      | 534.32    | -0.32 | 0.8250 |
| 4  | -2.7006  | -2.6648   | -0.0358 | 0.0351 | 541      | 542.07    | -1.07 | 0.8143 |
| 5  | -2.4319  | -2.3942   | -0.0377 | 0.0347 | 547      | 548.35    | -1.35 | 0.8039 |
| 6  | -2.2032  | -2.1634   | -0.0398 | 0.0342 | 553      | 553.70    | -0.70 | 0.7939 |
| 7  | -2.0016  | -1.9599   | -0.0417 | 0.0338 | 557      | 558.42    | -1.42 | 0.7844 |
| 8  | -1.8196  | -1.7760   | -0.0436 | 0.0334 | 562      | 562.68    | -0.68 | 0.7748 |
| 9  | -1.6523  | -1.6067   | -0.0456 | 0.0330 | 566      | 566.61    | -0.61 | 0.7656 |
| 10 | -1.4962  | -1.4486   | -0.0476 | 0.0326 | 569      | 570.27    | -1.27 | 0.7564 |
| 11 | -1.3489  | -1.2993   | -0.0496 | 0.0322 | 573      | 573.74    | -0.74 | 0.7477 |
| 12 | -1.2085  | -1.1569   | -0.0516 | 0.0319 | 576      | 577.04    | -1.04 | 0.7391 |
| 13 | -1.0737  | -1.0200   | -0.0537 | 0.0315 | 579      | 580.21    | -1.21 | 0.7305 |
| 14 | -0.9434  | -0.8874   | -0.0560 | 0.0311 | 582      | 583.29    | -1.29 | 0.7222 |
| 15 | -0.8164  | -0.7582   | -0.0582 | 0.0308 | 585      | 586.28    | -1.28 | 0.7140 |
| 16 | -0.6922  | -0.6317   | -0.0605 | 0.0305 | 588      | 589.21    | -1.21 | 0.7062 |
| 17 | -0.5699  | -0.5070   | -0.0629 | 0.0301 | 591      | 592.10    | -1.10 | 0.6988 |
| 18 | -0.4492  | -0.3837   | -0.0655 | 0.0298 | 593      | 594.96    | -1.96 | 0.6916 |
| 19 | -0.3292  | -0.2612   | -0.0680 | 0.0295 | 596      | 597.81    | -1.81 | 0.6848 |
| 20 | -0.2096  | -0.1389   | -0.0707 | 0.0292 | 600      | 600.64    | -0.64 | 0.6781 |
| 21 | -0.0900  | -0.0164   | -0.0736 | 0.0290 | 602      | 603.48    | -1.48 | 0.6720 |
| 22 | 0.0301   | 0.1067    | -0.0766 | 0.0287 | 605      | 606.33    | -1.33 | 0.6664 |
| 23 | 0.1512   | 0.2309    | -0.0797 | 0.0285 | 607      | 609.21    | -2.21 | 0.6613 |
| 24 | 0.2738   | 0.3567    | -0.0829 | 0.0283 | 610      | 612.13    | -2.13 | 0.6566 |
| 25 | 0.3982   | 0.4846    | -0.0864 | 0.0282 | 613      | 615.10    | -2.10 | 0.6528 |
| 26 | 0.5249   | 0.6151    | -0.0902 | 0.0280 | 616      | 618.12    | -2.12 | 0.6497 |
| 27 | 0.6545   | 0.7486    | -0.0941 | 0.0279 | 619      | 621.22    | -2.22 | 0.6473 |
| 28 | 0.7875   | 0.8860    | -0.0985 | 0.0279 | 622      | 624.40    | -2.40 | 0.6465 |
| 29 | 0.9248   | 1.0278    | -0.1030 | 0.0279 | 625      | 627.69    | -2.69 | 0.6472 |
| 30 | 1.0671   | 1.1749    | -0.1078 | 0.0280 | 629      | 631.10    | -2.10 | 0.6497 |
| 31 | 1.2156   | 1.3284    | -0.1128 | 0.0283 | 632      | 634.66    | -2.66 | 0.6553 |
| 32 | 1.3717   | 1.4898    | -0.1181 | 0.0286 | 636      | 638.40    | -2.40 | 0.6638 |
| 33 | 1.5376   | 1.6609    | -0.1233 | 0.0292 | 640      | 642.37    | -2.37 | 0.6762 |
| 34 | 1.7162   | 1.8444    | -0.1282 | 0.0299 | 644      | 646.63    | -2.63 | 0.6924 |
| 35 | 1.9118   | 2.0445    | -0.1327 | 0.0307 | 648      | 651.26    | -3.26 | 0.7126 |
| 36 | 2.1311   | 2.2676    | -0.1365 | 0.0317 | 653      | 656.44    | -3.44 | 0.7355 |
| 37 | 2.3850   | 2.5244    | -0.1394 | 0.0328 | 659      | 662.39    | -3.39 | 0.7610 |
| 38 | 2.6921   | 2.8338    | -0.1417 | 0.0342 | 666      | 669.56    | -3.56 | 0.7925 |
| 39 | 3.0872   | 3.2318    | -0.1446 | 0.0371 | 675      | 678.79    | -3.79 | 0.8613 |
| 40 | 3.6387   | 3.7848    | -0.1461 | 0.0487 | 688      | 691.62    | -3.62 | 1.1295 |
| 41 | 4.4472   | 4.5407    | -0.0935 | 0.0692 | 707      | 709.14    | -2.14 | 1.6047 |
| 42 | 5.5525   | 5.4494    | 0.1031  | 0.0675 | 733      | 730.21    | 2.79  | 1.5651 |
| 43 | 6.9298   | 6.5756    | 0.3542  | 0.1114 | 765      | 756.32    | 8.68  | 2.5829 |

Table 8: Online to Paper Bootstrap Linking Results – Writing  
 Online Scores Based on the Higher of Human and Automated Ratings of CR Items

| RS | Paper_Th | Online_Th | Diff   | S.E.   | Paper_SS | Online_SS | Diff | S.E.   |
|----|----------|-----------|--------|--------|----------|-----------|------|--------|
| 0  | -4.5650  | -4.9276   | 0.3626 | 0.2408 | 494      | 487.17    | 6.83 | 4.8579 |
| 1  | -3.2616  | -3.5454   | 0.2838 | 0.1865 | 521      | 515.06    | 5.94 | 3.7629 |
| 2  | -2.3918  | -2.5806   | 0.1888 | 0.1253 | 538      | 534.52    | 3.48 | 2.5268 |
| 3  | -1.7519  | -1.8807   | 0.1288 | 0.0859 | 551      | 548.64    | 2.36 | 1.7338 |
| 4  | -1.1613  | -1.2669   | 0.1056 | 0.0640 | 563      | 561.02    | 1.98 | 1.2914 |
| 5  | -0.5678  | -0.6769   | 0.1091 | 0.0523 | 575      | 572.92    | 2.08 | 1.0545 |
| 6  | 0.0406   | -0.0822   | 0.1228 | 0.0450 | 587      | 584.92    | 2.08 | 0.9087 |
| 7  | 0.6652   | 0.5245    | 0.1407 | 0.0408 | 600      | 597.16    | 2.84 | 0.8238 |
| 8  | 1.3047   | 1.1357    | 0.1690 | 0.0402 | 613      | 609.49    | 3.51 | 0.8117 |
| 9  | 1.9477   | 1.7423    | 0.2054 | 0.0430 | 626      | 621.73    | 4.27 | 0.8666 |
| 10 | 2.5692   | 2.3405    | 0.2286 | 0.0480 | 638      | 633.80    | 4.20 | 0.9684 |
| 11 | 3.1564   | 2.9313    | 0.2251 | 0.0586 | 650      | 645.71    | 4.29 | 1.1822 |
| 12 | 3.7370   | 3.5286    | 0.2085 | 0.0823 | 662      | 657.76    | 4.24 | 1.6603 |
| 13 | 4.3921   | 4.1986    | 0.1935 | 0.1237 | 675      | 671.28    | 3.72 | 2.4947 |
| 14 | 5.3366   | 5.1676    | 0.1690 | 0.1833 | 694      | 690.83    | 3.17 | 3.6979 |
| 15 | 6.7450   | 6.6208    | 0.1242 | 0.2390 | 723      | 720.15    | 2.85 | 4.8222 |

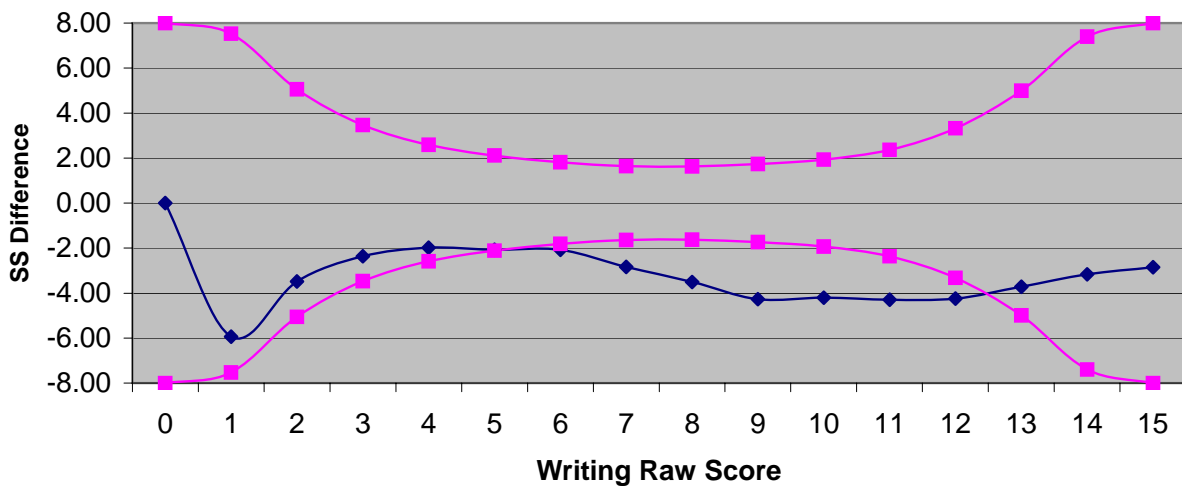
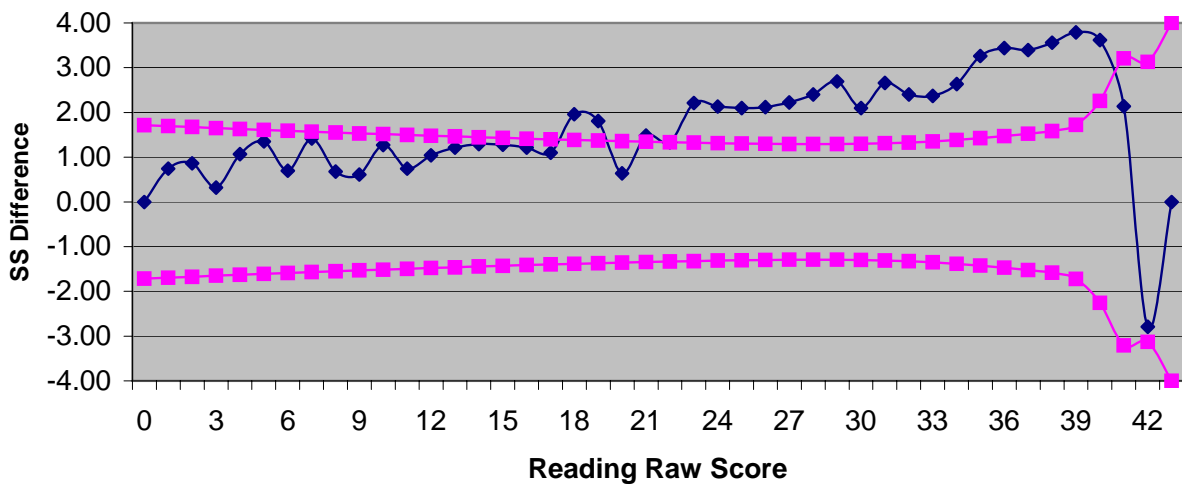
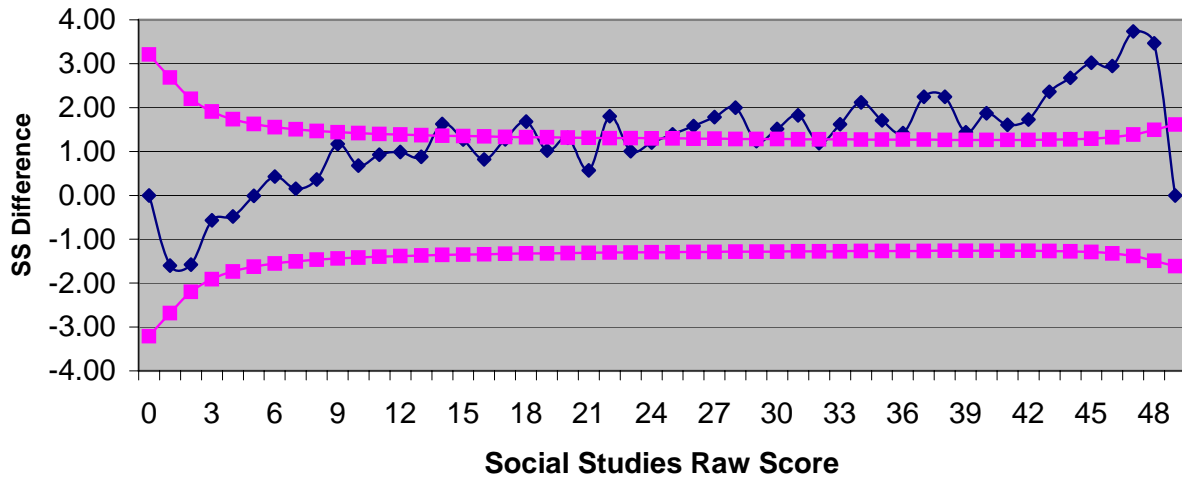


Figure 1: Online Minus Paper Scale Score Differences and  $\pm 2$  Equating Standard Errors (Online Scores based on Higher of Human and Automated CR Ratings of CR Items)

Table 9: Online to Paper Bootstrap Linking Results – Social Studies  
 Online Scores Based on Human Ratings of CR Items

| RS | Paper_Th | Online_Th | Diff    | S.E.   | Paper_SS | Online_SS | Diff  | S.E. |
|----|----------|-----------|---------|--------|----------|-----------|-------|------|
| 0  | -5.3556  | -5.3175   | -0.0381 | 0.0340 | 479      | 480.02    | -1.02 | 0.81 |
| 1  | -4.1261  | -4.0839   | -0.0422 | 0.0332 | 508      | 509.30    | -1.30 | 0.79 |
| 2  | -3.3965  | -3.3492   | -0.0473 | 0.0324 | 526      | 526.74    | -0.74 | 0.77 |
| 3  | -2.9546  | -2.9029   | -0.0517 | 0.0317 | 536      | 537.33    | -1.33 | 0.75 |
| 4  | -2.6304  | -2.5749   | -0.0555 | 0.0312 | 544      | 545.12    | -1.12 | 0.74 |
| 5  | -2.3707  | -2.3118   | -0.0588 | 0.0309 | 550      | 551.36    | -1.36 | 0.73 |
| 6  | -2.1516  | -2.0899   | -0.0617 | 0.0306 | 555      | 556.63    | -1.63 | 0.73 |
| 7  | -1.9605  | -1.8963   | -0.0642 | 0.0303 | 560      | 561.22    | -1.22 | 0.72 |
| 8  | -1.7899  | -1.7233   | -0.0666 | 0.0301 | 564      | 565.33    | -1.33 | 0.71 |
| 9  | -1.6347  | -1.5660   | -0.0687 | 0.0299 | 567      | 569.06    | -2.06 | 0.71 |
| 10 | -1.4915  | -1.4210   | -0.0705 | 0.0298 | 571      | 572.50    | -1.50 | 0.71 |
| 11 | -1.3581  | -1.2859   | -0.0722 | 0.0296 | 574      | 575.71    | -1.71 | 0.70 |
| 12 | -1.2326  | -1.1588   | -0.0738 | 0.0295 | 577      | 578.73    | -1.73 | 0.70 |
| 13 | -1.1135  | -1.0383   | -0.0752 | 0.0294 | 580      | 581.59    | -1.59 | 0.70 |
| 14 | -0.9998  | -0.9232   | -0.0766 | 0.0293 | 582      | 584.32    | -2.32 | 0.70 |
| 15 | -0.8907  | -0.8128   | -0.0779 | 0.0292 | 585      | 586.94    | -1.94 | 0.69 |
| 16 | -0.7852  | -0.7062   | -0.0790 | 0.0292 | 588      | 589.47    | -1.47 | 0.69 |
| 17 | -0.6830  | -0.6029   | -0.0801 | 0.0291 | 590      | 591.92    | -1.92 | 0.69 |
| 18 | -0.5835  | -0.5023   | -0.0812 | 0.0290 | 592      | 594.31    | -2.31 | 0.69 |
| 19 | -0.4862  | -0.4039   | -0.0823 | 0.0290 | 595      | 596.64    | -1.64 | 0.69 |
| 20 | -0.3906  | -0.3074   | -0.0832 | 0.0289 | 597      | 598.93    | -1.93 | 0.69 |
| 21 | -0.2965  | -0.2124   | -0.0841 | 0.0289 | 600      | 601.19    | -1.19 | 0.69 |
| 22 | -0.2035  | -0.1185   | -0.0850 | 0.0288 | 601      | 603.42    | -2.42 | 0.68 |
| 23 | -0.1113  | -0.0254   | -0.0859 | 0.0288 | 604      | 605.63    | -1.63 | 0.68 |
| 24 | -0.0196  | 0.0672    | -0.0868 | 0.0287 | 606      | 607.82    | -1.82 | 0.68 |
| 25 | 0.0718   | 0.1596    | -0.0878 | 0.0287 | 608      | 610.02    | -2.02 | 0.68 |
| 26 | 0.1634   | 0.2520    | -0.0886 | 0.0286 | 610      | 612.21    | -2.21 | 0.68 |
| 27 | 0.2552   | 0.3448    | -0.0896 | 0.0286 | 612      | 614.41    | -2.41 | 0.68 |
| 28 | 0.3478   | 0.4383    | -0.0905 | 0.0285 | 614      | 616.63    | -2.63 | 0.68 |
| 29 | 0.4413   | 0.5327    | -0.0914 | 0.0285 | 617      | 618.87    | -1.87 | 0.68 |
| 30 | 0.5362   | 0.6285    | -0.0923 | 0.0285 | 619      | 621.15    | -2.15 | 0.68 |
| 31 | 0.6327   | 0.7260    | -0.0933 | 0.0284 | 621      | 623.46    | -2.46 | 0.67 |
| 32 | 0.7312   | 0.8256    | -0.0944 | 0.0283 | 624      | 625.82    | -1.82 | 0.67 |
| 33 | 0.8323   | 0.9278    | -0.0955 | 0.0283 | 626      | 628.25    | -2.25 | 0.67 |
| 34 | 0.9365   | 1.0331    | -0.0966 | 0.0282 | 628      | 630.75    | -2.75 | 0.67 |
| 35 | 1.0442   | 1.1421    | -0.0979 | 0.0282 | 631      | 633.34    | -2.34 | 0.67 |
| 36 | 1.1564   | 1.2556    | -0.0992 | 0.0281 | 634      | 636.03    | -2.03 | 0.67 |
| 37 | 1.2738   | 1.3745    | -0.1007 | 0.0280 | 636      | 638.85    | -2.85 | 0.67 |
| 38 | 1.3976   | 1.4999    | -0.1023 | 0.0280 | 639      | 641.83    | -2.83 | 0.66 |
| 39 | 1.5292   | 1.6331    | -0.1039 | 0.0279 | 643      | 644.99    | -1.99 | 0.66 |
| 40 | 1.6703   | 1.7761    | -0.1058 | 0.0279 | 646      | 648.38    | -2.38 | 0.66 |
| 41 | 1.8234   | 1.9313    | -0.1079 | 0.0279 | 650      | 652.07    | -2.07 | 0.66 |
| 42 | 1.9918   | 2.1021    | -0.1103 | 0.0279 | 654      | 656.12    | -2.12 | 0.66 |
| 43 | 2.1806   | 2.2934    | -0.1128 | 0.0280 | 658      | 660.66    | -2.66 | 0.66 |
| 44 | 2.3973   | 2.5130    | -0.1157 | 0.0282 | 663      | 665.87    | -2.87 | 0.67 |
| 45 | 2.6545   | 2.7736    | -0.1191 | 0.0285 | 669      | 672.06    | -3.06 | 0.68 |
| 46 | 2.9762   | 3.0990    | -0.1228 | 0.0291 | 677      | 679.78    | -2.78 | 0.69 |
| 47 | 3.4156   | 3.5425    | -0.1269 | 0.0300 | 687      | 690.31    | -3.31 | 0.71 |
| 48 | 4.1425   | 4.2744    | -0.1319 | 0.0315 | 705      | 707.68    | -2.68 | 0.75 |
| 49 | 5.3702   | 5.5058    | -0.1356 | 0.0330 | 734      | 736.90    | -2.90 | 0.78 |

Table 10: Online to Paper Bootstrap Linking Results – Reading  
 Online Scores Based on Human Ratings of CR Items

| RS | Paper_Th | Online_Th | Diff    | S.E.   | Paper_SS | Online_SS | Diff  | S.E. |
|----|----------|-----------|---------|--------|----------|-----------|-------|------|
| 0  | -5.4520  | -5.4020   | -0.0500 | 0.0442 | 477      | 478.61    | -1.61 | 1.02 |
| 1  | -4.2183  | -4.1660   | -0.0523 | 0.0435 | 506      | 507.27    | -1.27 | 1.01 |
| 2  | -3.4822  | -3.4265   | -0.0557 | 0.0427 | 523      | 524.41    | -1.41 | 0.99 |
| 3  | -3.0329  | -2.9739   | -0.0590 | 0.0419 | 534      | 534.91    | -0.91 | 0.97 |
| 4  | -2.7006  | -2.6384   | -0.0622 | 0.0412 | 541      | 542.69    | -1.69 | 0.95 |
| 5  | -2.4319  | -2.3667   | -0.0652 | 0.0405 | 547      | 548.99    | -1.99 | 0.94 |
| 6  | -2.2032  | -2.1349   | -0.0683 | 0.0398 | 553      | 554.36    | -1.36 | 0.92 |
| 7  | -2.0016  | -1.9305   | -0.0711 | 0.0392 | 557      | 559.10    | -2.10 | 0.91 |
| 8  | -1.8196  | -1.7457   | -0.0739 | 0.0386 | 562      | 563.38    | -1.38 | 0.89 |
| 9  | -1.6523  | -1.5757   | -0.0766 | 0.0380 | 566      | 567.33    | -1.33 | 0.88 |
| 10 | -1.4962  | -1.4172   | -0.0790 | 0.0375 | 569      | 571.00    | -2.00 | 0.87 |
| 11 | -1.3489  | -1.2675   | -0.0814 | 0.0369 | 573      | 574.47    | -1.47 | 0.86 |
| 12 | -1.2085  | -1.1249   | -0.0836 | 0.0364 | 576      | 577.78    | -1.78 | 0.84 |
| 13 | -1.0737  | -0.9879   | -0.0858 | 0.0360 | 579      | 580.95    | -1.95 | 0.83 |
| 14 | -0.9434  | -0.8555   | -0.0879 | 0.0355 | 582      | 584.02    | -2.02 | 0.82 |
| 15 | -0.8164  | -0.7267   | -0.0897 | 0.0351 | 585      | 587.01    | -2.01 | 0.81 |
| 16 | -0.6922  | -0.6006   | -0.0916 | 0.0347 | 588      | 589.94    | -1.94 | 0.80 |
| 17 | -0.5699  | -0.4766   | -0.0933 | 0.0343 | 591      | 592.81    | -1.81 | 0.79 |
| 18 | -0.4492  | -0.3540   | -0.0952 | 0.0339 | 593      | 595.65    | -2.65 | 0.79 |
| 19 | -0.3292  | -0.2324   | -0.0968 | 0.0336 | 596      | 598.47    | -2.47 | 0.78 |
| 20 | -0.2096  | -0.1110   | -0.0986 | 0.0332 | 600      | 601.29    | -1.29 | 0.77 |
| 21 | -0.0900  | 0.0105    | -0.1005 | 0.0329 | 602      | 604.10    | -2.10 | 0.76 |
| 22 | 0.0301   | 0.1326    | -0.1025 | 0.0326 | 605      | 606.94    | -1.94 | 0.76 |
| 23 | 0.1512   | 0.2559    | -0.1047 | 0.0323 | 607      | 609.79    | -2.79 | 0.75 |
| 24 | 0.2738   | 0.3809    | -0.1071 | 0.0320 | 610      | 612.69    | -2.69 | 0.74 |
| 25 | 0.3982   | 0.5080    | -0.1098 | 0.0318 | 613      | 615.64    | -2.64 | 0.74 |
| 26 | 0.5249   | 0.6378    | -0.1129 | 0.0315 | 616      | 618.65    | -2.65 | 0.73 |
| 27 | 0.6545   | 0.7709    | -0.1164 | 0.0313 | 619      | 621.74    | -2.74 | 0.73 |
| 28 | 0.7875   | 0.9079    | -0.1204 | 0.0311 | 622      | 624.91    | -2.91 | 0.72 |
| 29 | 0.9248   | 1.0495    | -0.1247 | 0.0310 | 625      | 628.19    | -3.19 | 0.72 |
| 30 | 1.0671   | 1.1965    | -0.1294 | 0.0309 | 629      | 631.60    | -2.60 | 0.72 |
| 31 | 1.2156   | 1.3498    | -0.1342 | 0.0308 | 632      | 635.16    | -3.16 | 0.72 |
| 32 | 1.3717   | 1.5107    | -0.1390 | 0.0309 | 636      | 638.89    | -2.89 | 0.72 |
| 33 | 1.5376   | 1.6808    | -0.1432 | 0.0311 | 640      | 642.83    | -2.83 | 0.72 |
| 34 | 1.7162   | 1.8627    | -0.1465 | 0.0314 | 644      | 647.05    | -3.05 | 0.73 |
| 35 | 1.9118   | 2.0601    | -0.1483 | 0.0319 | 648      | 651.63    | -3.63 | 0.74 |
| 36 | 2.1311   | 2.2793    | -0.1482 | 0.0325 | 653      | 656.71    | -3.71 | 0.75 |
| 37 | 2.3850   | 2.5308    | -0.1458 | 0.0331 | 659      | 662.54    | -3.54 | 0.77 |
| 38 | 2.6921   | 2.8338    | -0.1417 | 0.0342 | 666      | 669.56    | -3.56 | 0.79 |
| 39 | 3.0872   | 3.2252    | -0.1380 | 0.0376 | 675      | 678.64    | -3.64 | 0.87 |
| 40 | 3.6387   | 3.7764    | -0.1377 | 0.0523 | 688      | 691.42    | -3.42 | 1.21 |
| 41 | 4.4472   | 4.5450    | -0.0978 | 0.0775 | 707      | 709.24    | -2.24 | 1.80 |
| 42 | 5.5525   | 5.4679    | 0.0846  | 0.0698 | 733      | 730.64    | 2.36  | 1.62 |
| 43 | 6.9298   | 6.5981    | 0.3317  | 0.1276 | 765      | 756.84    | 8.16  | 2.96 |

Table 11: Online to Paper Bootstrap Linking Results – Writing  
 Online Scores Based on Human Ratings of CR Items

| RS | Paper_Th | Online_Th | Diff    | S.E.   | Paper_SS | Online_SS | Diff  | S.E. |
|----|----------|-----------|---------|--------|----------|-----------|-------|------|
| 0  | -4.5650  | -4.8891   | 0.3241  | 0.2238 | 494      | 487.95    | 6.05  | 4.51 |
| 1  | -3.2616  | -3.4600   | 0.1984  | 0.1784 | 521      | 516.78    | 4.22  | 3.60 |
| 2  | -2.3918  | -2.4167   | 0.0249  | 0.1227 | 538      | 537.83    | 0.17  | 2.48 |
| 3  | -1.7519  | -1.6346   | -0.1173 | 0.0876 | 551      | 553.61    | -2.61 | 1.77 |
| 4  | -1.1613  | -0.9438   | -0.2175 | 0.0694 | 563      | 567.54    | -4.54 | 1.40 |
| 5  | -0.5678  | -0.2884   | -0.2794 | 0.0599 | 575      | 580.76    | -5.76 | 1.21 |
| 6  | 0.0406   | 0.3523    | -0.3117 | 0.0544 | 587      | 593.69    | -6.69 | 1.10 |
| 7  | 0.6652   | 0.9838    | -0.3186 | 0.0513 | 600      | 606.43    | -6.43 | 1.03 |
| 8  | 1.3047   | 1.6068    | -0.3021 | 0.0506 | 613      | 619.00    | -6.00 | 1.02 |
| 9  | 1.9477   | 2.2166    | -0.2689 | 0.0555 | 626      | 631.30    | -5.30 | 1.12 |
| 10 | 2.5692   | 2.7935    | -0.2243 | 0.0704 | 638      | 642.93    | -4.93 | 1.42 |
| 11 | 3.1564   | 3.3217    | -0.1653 | 0.0980 | 650      | 653.59    | -3.59 | 1.98 |
| 12 | 3.7370   | 3.8305    | -0.0935 | 0.1372 | 662      | 663.86    | -1.86 | 2.77 |
| 13 | 4.3921   | 4.4072    | -0.0151 | 0.1898 | 675      | 675.49    | -0.49 | 3.83 |
| 14 | 5.3366   | 5.2787    | 0.0579  | 0.2644 | 694      | 693.07    | 0.93  | 5.33 |
| 15 | 6.7450   | 6.6301    | 0.1149  | 0.3090 | 723      | 720.33    | 2.67  | 6.23 |

Table 12: Social Studies Online and Paper Bootstrap P-Values, Standard Errors, and Z-Statistics for Differences

| Item | cpval | ppval | cstd  | pstd  | z_dif |    |
|------|-------|-------|-------|-------|-------|----|
| 1    | 0.83  | 0.85  | 0.011 | 0.012 | -1.39 |    |
| 2    | 0.51  | 0.54  | 0.014 | 0.014 | -1.45 |    |
| 3    | 0.49  | 0.52  | 0.011 | 0.017 | -1.45 |    |
| 4    | 0.58  | 0.57  | 0.014 | 0.015 | 0.75  |    |
| 5    | 0.56  | 0.58  | 0.014 | 0.015 | -1.17 |    |
| 6    | 0.74  | 0.75  | 0.013 | 0.012 | -0.70 |    |
| 7    | 0.76  | 0.78  | 0.013 | 0.013 | -0.86 |    |
| 8    | 0.56  | 0.58  | 0.016 | 0.014 | -0.96 |    |
| 9    | 0.54  | 0.55  | 0.015 | 0.013 | -0.54 |    |
| 10   | 0.47  | 0.47  | 0.013 | 0.015 | -0.18 |    |
| 11   | 0.72  | 0.73  | 0.013 | 0.014 | -0.14 |    |
| 12   | 0.69  | 0.71  | 0.014 | 0.013 | -1.06 |    |
| 13   | 0.66  | 0.64  | 0.015 | 0.015 | 0.67  |    |
| 14   | 0.60  | 0.60  | 0.014 | 0.015 | 0.29  |    |
| 15   | 0.48  | 0.54  | 0.014 | 0.014 | -2.93 | ** |
| 16   | 0.58  | 0.62  | 0.015 | 0.012 | -2.00 | ** |
| 17   | 0.54  | 0.61  | 0.017 | 0.014 | -3.04 | ** |
| 18   | 0.57  | 0.56  | 0.016 | 0.012 | 0.59  |    |
| 19   | 0.57  | 0.59  | 0.014 | 0.015 | -0.51 |    |
| 20   | 0.56  | 0.52  | 0.014 | 0.015 | 1.62  |    |
| 21   | 0.51  | 0.51  | 0.014 | 0.015 | -0.34 |    |
| 22*  | 1.45  | 1.50  | 0.020 | 0.026 | -1.71 |    |
| 23   | 0.69  | 0.70  | 0.014 | 0.013 | -0.44 |    |
| 24   | 0.74  | 0.73  | 0.011 | 0.012 | 0.66  |    |
| 25   | 0.59  | 0.60  | 0.014 | 0.016 | -0.61 |    |
| 26   | 0.53  | 0.53  | 0.015 | 0.014 | -0.16 |    |
| 27   | 0.62  | 0.62  | 0.014 | 0.015 | 0.11  |    |
| 28   | 0.84  | 0.87  | 0.009 | 0.009 | -2.27 | ** |
| 29   | 0.74  | 0.79  | 0.012 | 0.011 | -2.95 | ** |
| 30   | 0.59  | 0.69  | 0.015 | 0.014 | -4.57 | ** |
| 31   | 0.57  | 0.61  | 0.016 | 0.016 | -1.82 |    |
| 32   | 0.48  | 0.53  | 0.015 | 0.016 | -2.31 | ** |
| 33   | 0.70  | 0.70  | 0.014 | 0.013 | 0.04  |    |
| 34   | 0.62  | 0.61  | 0.015 | 0.014 | 0.08  |    |
| 35   | 0.48  | 0.48  | 0.014 | 0.016 | 0.00  |    |
| 36   | 0.44  | 0.45  | 0.014 | 0.013 | -0.63 |    |
| 37   | 0.25  | 0.30  | 0.012 | 0.015 | -2.29 | ** |
| 38   | 0.62  | 0.66  | 0.015 | 0.015 | -1.55 |    |
| 39   | 0.72  | 0.78  | 0.013 | 0.013 | -2.94 | ** |
| 40   | 0.83  | 0.83  | 0.012 | 0.012 | -0.08 |    |
| 41   | 0.82  | 0.80  | 0.012 | 0.012 | 1.41  |    |
| 42   | 0.47  | 0.46  | 0.015 | 0.014 | 0.32  |    |
| 43   | 0.58  | 0.67  | 0.012 | 0.015 | -4.48 | ** |
| 44   | 0.52  | 0.55  | 0.012 | 0.013 | -1.58 |    |
| 45   | 0.58  | 0.57  | 0.013 | 0.015 | 0.45  |    |
| 46   | 0.53  | 0.53  | 0.013 | 0.014 | -0.27 |    |
| 47   | 0.40  | 0.46  | 0.015 | 0.015 | -2.46 | ** |

\* Online CR scores based on human ratings only

\*\* Denotes absolute z-difference > 2.0

Table 13: Reading Online and Paper Bootstrap P-Values,  
Standard Errors, and Z-Statistics for Differences

| Obs | cpval | ppval | cstd  | pstd  | z_dif |    |
|-----|-------|-------|-------|-------|-------|----|
| 1   | 0.89  | 0.87  | 0.010 | 0.011 | 1.12  |    |
| 2   | 0.83  | 0.83  | 0.011 | 0.012 | -0.08 |    |
| 3   | 0.53  | 0.59  | 0.014 | 0.014 | -2.73 | ** |
| 4   | 0.84  | 0.84  | 0.010 | 0.012 | -0.57 |    |
| 5   | 0.90  | 0.90  | 0.009 | 0.007 | 0.07  |    |
| 6   | 0.87  | 0.89  | 0.010 | 0.010 | -1.39 |    |
| 7   | 0.68  | 0.71  | 0.012 | 0.013 | -1.60 |    |
| 8   | 0.83  | 0.84  | 0.011 | 0.012 | -0.48 |    |
| 9   | 0.46  | 0.48  | 0.016 | 0.016 | -0.73 |    |
| 10  | 0.74  | 0.78  | 0.013 | 0.012 | -2.62 | ** |
| 11  | 0.88  | 0.88  | 0.009 | 0.009 | -0.24 |    |
| 12  | 0.77  | 0.79  | 0.014 | 0.013 | -1.26 |    |
| 13  | 0.62  | 0.61  | 0.013 | 0.015 | 0.42  |    |
| 14  | 0.61  | 0.63  | 0.014 | 0.015 | -1.19 |    |
| 15  | 0.32  | 0.29  | 0.015 | 0.013 | 1.65  |    |
| 16  | 0.85  | 0.86  | 0.011 | 0.010 | -0.66 |    |
| 17  | 0.51  | 0.55  | 0.017 | 0.014 | -1.96 |    |
| 18  | 0.55  | 0.57  | 0.014 | 0.013 | -1.17 |    |
| 19  | 0.74  | 0.73  | 0.014 | 0.012 | 0.49  |    |
| 20  | 0.38  | 0.42  | 0.015 | 0.016 | -1.76 |    |
| 21  | 0.54  | 0.53  | 0.016 | 0.015 | 0.72  |    |
| 22* | 1.81  | 2.15  | 0.034 | 0.035 | -6.88 | ** |
| 23  | 0.76  | 0.78  | 0.012 | 0.013 | -1.22 |    |
| 24  | 0.50  | 0.53  | 0.014 | 0.014 | -1.58 |    |
| 25  | 0.78  | 0.80  | 0.011 | 0.012 | -1.11 |    |
| 26  | 0.51  | 0.52  | 0.012 | 0.014 | -0.40 |    |
| 27  | 0.64  | 0.65  | 0.017 | 0.014 | -0.72 |    |
| 28  | 0.54  | 0.53  | 0.015 | 0.014 | 0.63  |    |
| 29  | 0.31  | 0.35  | 0.014 | 0.015 | -2.01 | ** |
| 30  | 0.54  | 0.53  | 0.013 | 0.014 | 0.41  |    |
| 31  | 0.81  | 0.82  | 0.011 | 0.011 | -0.86 |    |
| 32  | 0.71  | 0.77  | 0.015 | 0.012 | -2.94 | ** |
| 33  | 0.43  | 0.47  | 0.015 | 0.015 | -1.89 |    |
| 34  | 0.36  | 0.34  | 0.015 | 0.013 | 0.93  |    |
| 35  | 0.79  | 0.81  | 0.013 | 0.012 | -1.19 |    |
| 36  | 0.80  | 0.77  | 0.011 | 0.013 | 1.62  |    |
| 37  | 0.75  | 0.75  | 0.011 | 0.015 | 0.31  |    |
| 38  | 0.78  | 0.79  | 0.012 | 0.013 | -0.26 |    |

\* Online CR scores based on human ratings only

\*\* Denotes absolute z-difference > 2.0

Table 14: Writing Online and Paper Bootstrap P-Values,  
Standard Errors, and Z-Statistics for Differences

| Item | cpval | ppval | cstd  | pstd  | z_dif |    |
|------|-------|-------|-------|-------|-------|----|
| 1*   | 2.34  | 2.56  | 0.026 | 0.021 | -6.49 | ** |
| 2    | 0.62  | 0.66  | 0.014 | 0.015 | -1.79 |    |
| 3    | 0.23  | 0.19  | 0.013 | 0.013 | 2.29  | ** |
| 4    | 0.33  | 0.39  | 0.013 | 0.015 | -2.85 | ** |
| 5    | 0.79  | 0.79  | 0.011 | 0.010 | -0.10 |    |
| 6    | 0.83  | 0.86  | 0.011 | 0.011 | -1.87 |    |
| 7*   | 1.58  | 1.75  | 0.022 | 0.019 | -5.92 | ** |

\* Online CR scores based on human ratings only

\*\* Denotes absolute z-difference > 2.0

Table 15: Online and Paper Bootstrap Means, Standard Errors, and Z-Difference Statistics for Online minus Paper Differences by Ethnicity and Gender

| Social Studies Subgroup           | Ave. N | Online Mean | Paper Mean | Online SE | Paper SE | Z-Dif   |
|-----------------------------------|--------|-------------|------------|-----------|----------|---------|
| American Indian or Alaskan Native | 53.0   | 25.41       | 26.29      | 1.18      | 1.07     | -0.55   |
| Black, not of Hispanic Origin     | 116.7  | 24.81       | 24.80      | 0.75      | 0.78     | 0.01    |
| Hispanic                          | 33.4   | 25.66       | 25.72      | 1.31      | 1.40     | -0.03   |
| White, not of Hispanic Origin     | 879.8  | 29.82       | 30.86      | 0.28      | 0.29     | -2.56** |
| Male                              | 555.0  | 29.37       | 30.28      | 0.37      | 0.38     | -1.69   |
| Female                            | 540.0  | 28.54       | 29.41      | 0.36      | 0.38     | -1.68   |
| Reading Subgroup                  | Ave. N | Online Mean | Paper Mean | Online SE | Paper SE | Z-Dif   |
| American Indian or Alaskan Native | 51.7   | 24.59       | 24.31      | 1.00      | 0.92     | 0.21    |
| Black, not of Hispanic Origin     | 112.2  | 23.58       | 23.94      | 0.61      | 0.71     | -0.38   |
| Hispanic                          | 33.1   | 23.92       | 24.22      | 1.25      | 1.20     | -0.18   |
| White, not of Hispanic Origin     | 924.3  | 26.63       | 27.56      | 0.21      | 0.25     | -2.84** |
| Male                              | 565.0  | 25.69       | 26.33      | 0.31      | 0.33     | -1.42   |
| Female                            | 569.0  | 26.62       | 27.59      | 0.24      | 0.29     | -2.56** |
| Writing Subgroup                  | Ave. N | Online Mean | Paper Mean | Online SE | Paper SE | Z-Dif   |
| American Indian or Alaskan Native | 51.4   | 6.80        | 6.63       | 0.31      | 0.22     | 0.45    |
| Black, not of Hispanic Origin     | 112.7  | 6.05        | 6.60       | 0.22      | 0.20     | -1.88   |
| Hispanic                          | 32.7   | 6.28        | 6.76       | 0.28      | 0.32     | -1.12   |
| White, not of Hispanic Origin     | 924.4  | 6.83        | 7.30       | 0.06      | 0.06     | -5.27** |
| Male                              | 566.2  | 6.50        | 6.88       | 0.08      | 0.07     | -3.55** |
| Female                            | 567.8  | 6.98        | 7.51       | 0.08      | 0.08     | -4.50** |

\*\* Denotes absolute z-difference > 2.0

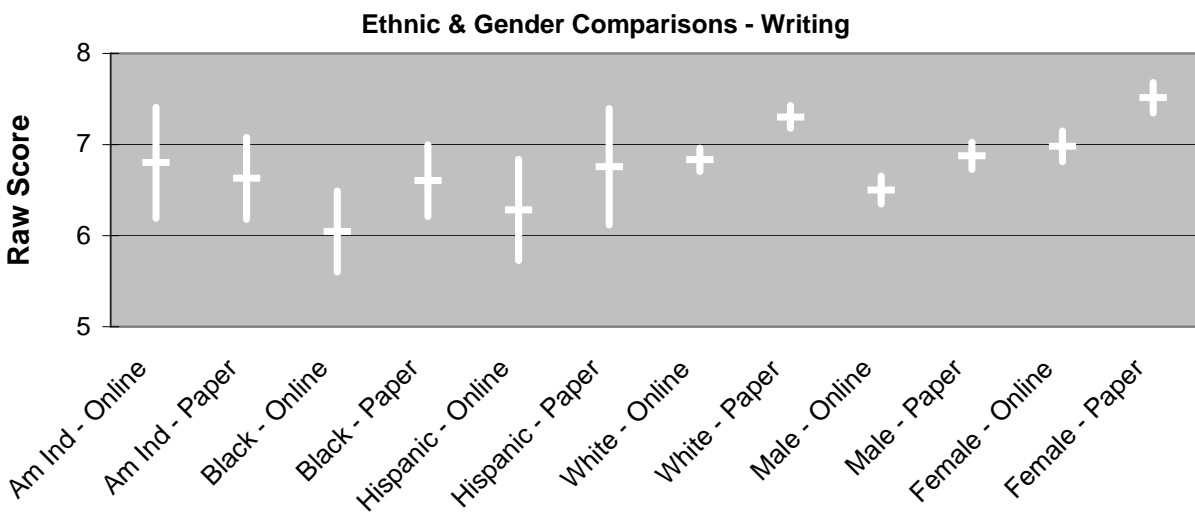
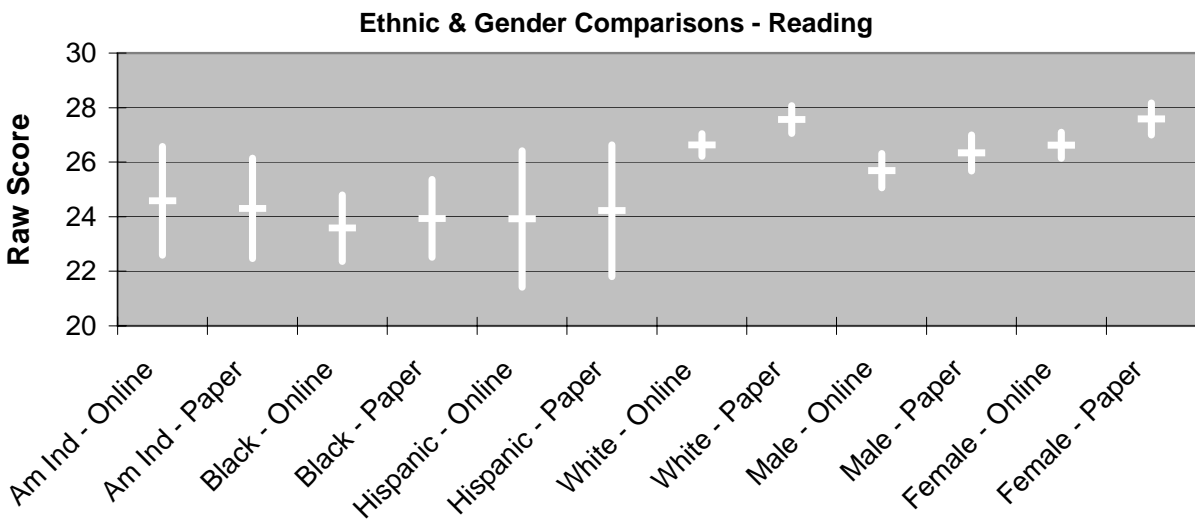
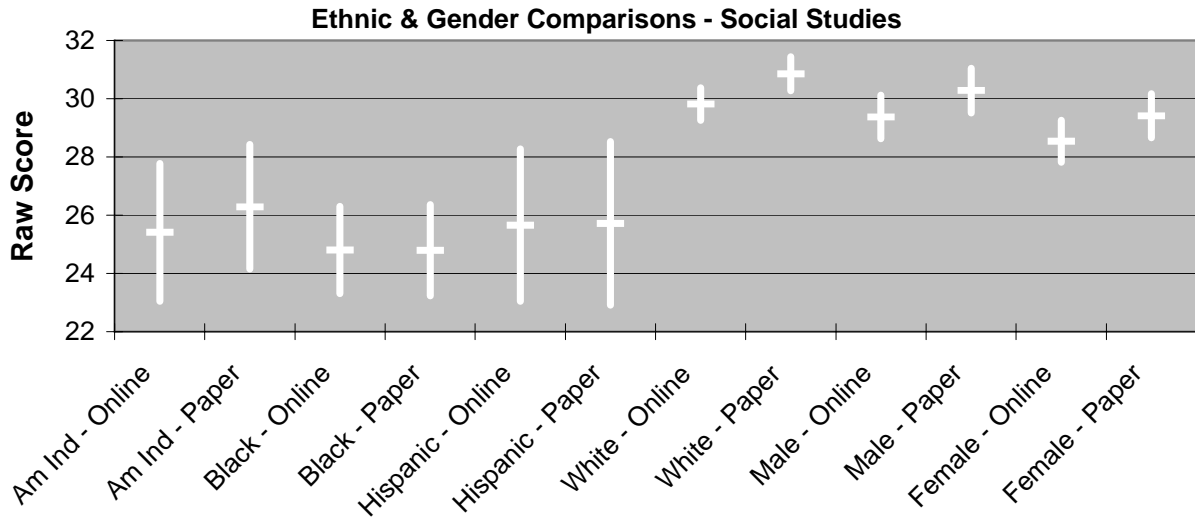


Figure 2: Bootstrap Raw Score Means and Standard Errors by Ethnicity and Gender for Online and Paper Tests

Table 16: Summary of Comparisons Between Automated Scores and Human Scores

| Automated Scores | Human Scores - Social Studies CR |     |    | Total |
|------------------|----------------------------------|-----|----|-------|
|                  | 1                                | 2   | 3  |       |
| 1                | 277                              | 64  | 4  | 345   |
| 2                | 178                              | 191 | 66 | 435   |
| 3                | 0                                | 1   | 4  | 5     |
| Total            | 455                              | 256 | 74 | 785   |

Exact Agreement (scorable) 60.1%  
 Adjacent Agreement (scorable) 99.5%  
 Mean Score Automated Rating 1.57 Mean Score Human Rating 1.51

| Automated Scores | Human Scores - Reading CR1 |     |     |     |   | Total |
|------------------|----------------------------|-----|-----|-----|---|-------|
|                  | 1                          | 2   | 3   | 4   | 5 |       |
| 1                | 119                        | 149 | 6   | 17  | 0 | 291   |
| 2                | 61                         | 213 | 46  | 57  | 0 | 377   |
| 3                | 10                         | 47  | 45  | 44  | 1 | 147   |
| 4                | 0                          | 2   | 4   | 2   | 2 | 10    |
| 5                | 0                          | 0   | 0   | 0   | 0 | 0     |
| Total            | 190                        | 411 | 101 | 120 | 3 | 825   |

Exact Agreement (scorable) 45.9%  
 Adjacent Agreement (scorable) 94.3%  
 Mean Score Automated Rating 1.85 Mean Score Human Rating 2.19

| Automated Scores | Human Scores - Writing CR1 |     |     |    |    |   | Total |
|------------------|----------------------------|-----|-----|----|----|---|-------|
|                  | 1                          | 2   | 3   | 4  | 5  | 6 |       |
| 1                | 59                         | 25  | 0   | 0  | 0  | 0 | 84    |
| 2                | 43                         | 312 | 31  | 0  | 0  | 0 | 386   |
| 3                | 0                          | 162 | 227 | 13 | 0  | 0 | 402   |
| 4                | 0                          | 5   | 57  | 50 | 5  | 0 | 117   |
| 5                | 0                          | 0   | 3   | 12 | 6  | 1 | 22    |
| 6                | 0                          | 0   | 0   | 1  | 1  | 1 | 3     |
| Total            | 102                        | 504 | 318 | 76 | 12 | 2 | 1014  |

Exact Agreement (scorable) 64.6%  
 Adjacent Agreement (scorable) 99.1%  
 Mean Score Automated Rating 2.62 Mean Score Human Rating 2.41

| Automated Scores | Human Scores - Writing CR2 |     |    |   | Total |
|------------------|----------------------------|-----|----|---|-------|
|                  | 1                          | 2   | 3  | 4 |       |
| 1                | 88                         | 27  | 0  | 0 | 115   |
| 2                | 262                        | 332 | 16 | 0 | 610   |
| 3                | 46                         | 153 | 47 | 1 | 247   |
| 4                | 0                          | 3   | 5  | 2 | 10    |
| Total            | 396                        | 515 | 68 | 3 | 982   |

Exact Agreement (scorable) 47.8%  
 Adjacent Agreement (scorable) 99.6%  
 Mean Score Automated Rating 2.15 Mean Score Human Rating 1.67

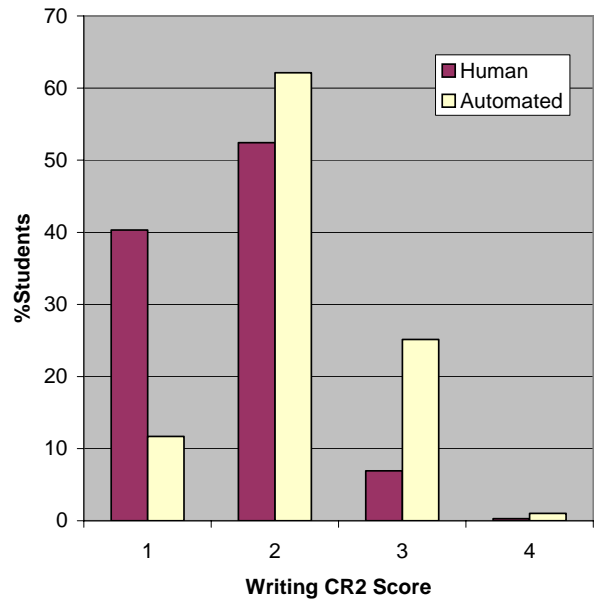
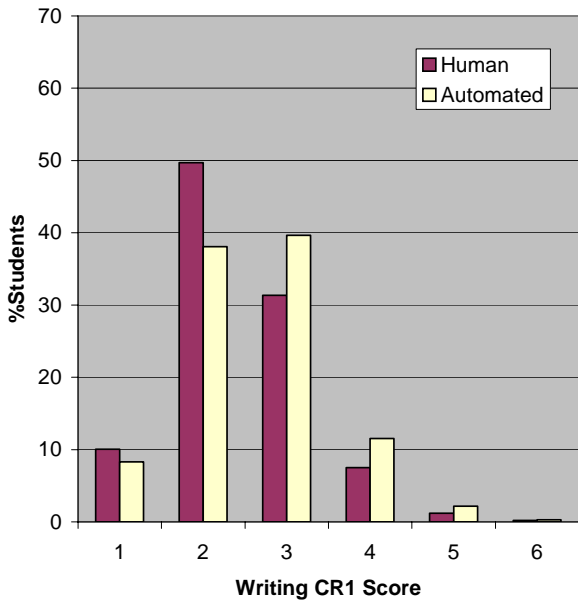
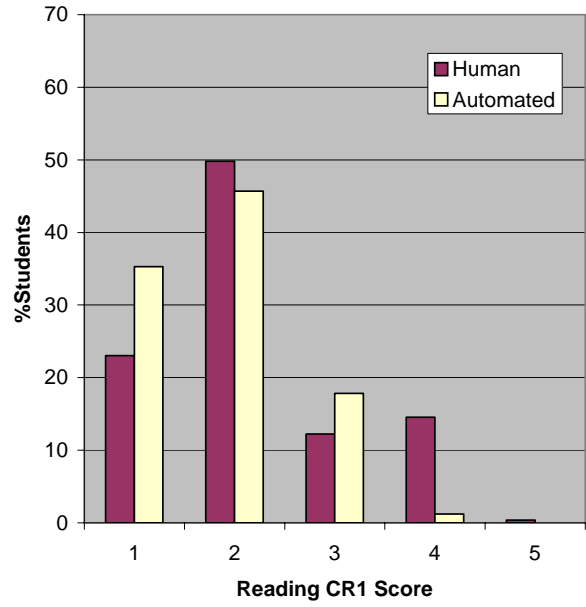
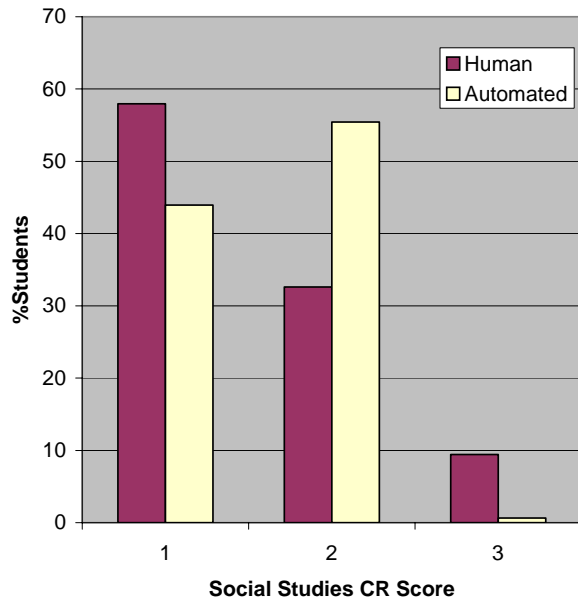


Figure 3: Percentage Frequencies of Human Versus Automated Constructed Response Scores

## **Appendix A: Further Analysis of Automated Scoring for the Online Constructed Response Items Pearson Knowledge Technologies**

The PKT automated scoring engine (IEA) is based on statistical machine learning techniques. This means that it uses a set of human scored training responses to build a scoring model that is then used to score operational responses. The quality of the scoring is limited by the accuracy and representativeness of the training set with respect to operational responses. In the case of this pilot, there were a priori reasons to believe the training set would not be representative. (a) In the case of the social studies CR item (mich-305MI06SS\_27), the prompt and rubric were changed between the administration of the training responses and the operational responses. (b) The training set was derived from transcripts of handwritten responses while the test set was answered on-line. Differences between answers and scoring in the two modes can decrease the representativeness between training and operational responses and may decrease IEA-Human agreement. If practical, it is best to hold mode of entry for training and test constant. (c) Since the scoring engine learns the appropriate score points from examples in the training set, it is very difficult to accurately grade score points with few or no examples. For the six point responses, the training set contained few or no examples at the upper two score points (5 and 6).

Since the scoring engine is statistically based, it is possible for it to determine when a response is outside of the bounds it can accurately grade and so the response needs to be passed to a human grader. As operational responses diverge from the responses of the training set, the number of these “unscorable” responses increases. In this pilot, the result of lack of upper score points can be seen in Writing CR1 (mich-605MI06WR\_51) where of the six responses that humans gave a score of six to, four of them were rated “unscorable” by IEA indicating that the answers should be scored by a human (not that they were bad answers).

Comparison of the training responses (i.e., responses collected during field-testing) and operational responses as described below indicates this concern was warranted. In this appendix, we examine the distribution of scores and differences in language features for training and operational responses.

Figure 1 shows comparisons of the human score point distributions for each of the 4 prompts for training and operations. The distributions are clearly quite different for the two writing prompts and the reading prompt. For example with writing CR2 (mich-405MI06WR\_57) there are very noticeable differences between training and operational percentages of essays at score points 1 and 3. Similar imbalances occur with Reading CR1 (mich-605MI06RD\_22) and writing CR1 (mich-605MI06WR\_51).

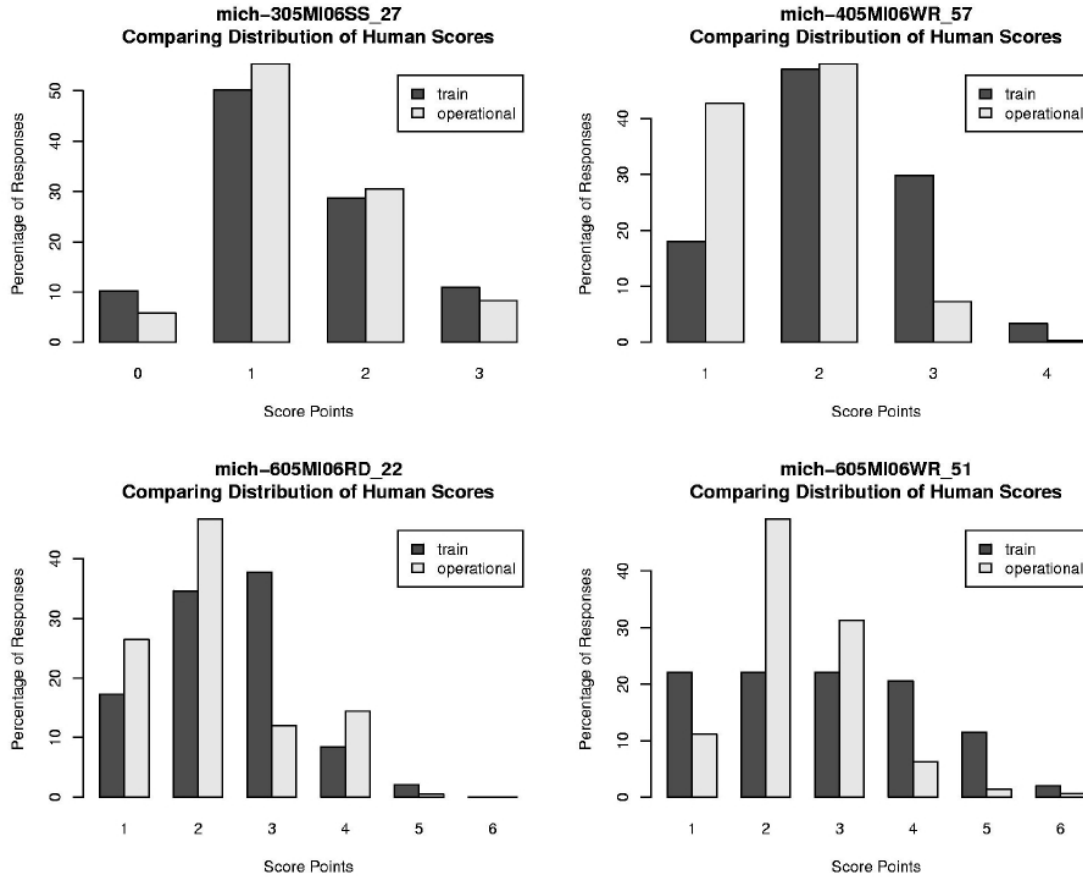


Figure 1: Human Score Distributions for Field Test (train) and Operational Online Papers

To look beyond the differences in the grade distribution, statistical language theory has developed many measures of the features of text passages. These measures can be as simple as the number of words in the passage to more sophisticated measures of how close the word choice is to a typical English passage or how coherent is the sentence structure of the passage. Table 1 presents the average value for a number of text measures for the training and operational sets by prompt, as well as t-tests to compare the means. These measures are one way of comparing the training and operational sets and show that they do vary in important ways. Note that many of these features, for example number of words, are not used in our scoring, but are given here to aid understanding of the comparative properties of the responses.

Some features immediately catch the eye. In 3 of the 4 prompts, the mean word count very significantly differs between the training and operational sets—operational higher in one case, lower in the other two. Word count is known to influence human scores (which is one reason why IEA scoring avoids using it - to reduce “gaming”).

In 3 of the 4 prompts, the readability of the answers by popular measures (again in different directions) differs between the training and operational responses. In half of the prompts the use of English is different in training and operational answers, as evidenced by the significant t-statistics for English Word Distribution or Typical Word Ordering in Table 1.

Table A.1: Mean and SD of Various Text Measures for the Training and Operational Prompts

| mich-305MI06SS_27            |               |              |          |         |        |                 |         |
|------------------------------|---------------|--------------|----------|---------|--------|-----------------|---------|
| Feature                      | mean<br>train | mean<br>oper | sd train | sd oper | t-stat | degrees<br>free | p value |
| Word Count                   | 91.46         | 121.37       | 52.72    | 70.72   | -7.81  | 690.70          | <.001   |
| Sentence Coherence           | 0.27          | 0.32         | 0.14     | 0.14    | -4.45  | 432.90          | <.001   |
| unique words/total<br>words  | 0.66          | 0.61         | 0.14     | 0.13    | 4.99   | 489.40          | <.001   |
| English Word<br>Distribution | 0.83          | 0.86         | 0.11     | 0.08    | -4.12  | 413.30          | <.001   |
| Typical Word<br>Ordering     | 8.44          | 8.19         | 1.02     | 0.97    | 3.78   | 493.40          | <.001   |
| Readability Measure          | 6.18          | 6.57         | 1.85     | 1.83    | -3.19  | 513.20          | <.001   |
| mich-605MI06RD_22            |               |              |          |         |        |                 |         |
| Feature                      | mean<br>train | mean<br>oper | sd train | sd oper | t-stat | degrees<br>free | p value |
| Word Count                   | 166.59        | 123.39       | 78.95    | 95.88   | 6.71   | 303.00          | <.001   |
| Sentence Coherence           | 0.27          | 0.26         | 0.09     | 0.12    | 1.23   | 322.80          | 0.218   |
| unique words/total<br>words  | 0.59          | 0.65         | 0.10     | 0.13    | -7.13  | 334.00          | <.001   |
| English Word<br>Distribution | 0.88          | 0.84         | 0.05     | 0.10    | 8.28   | 508.60          | <.001   |
| Typical Word<br>Ordering     | 8.36          | 8.41         | 0.77     | 0.99    | -0.65  | 316.30          | 0.519   |
| Readability Measure          | 7.05          | 7.34         | 1.60     | 2.46    | -2.11  | 379.00          | 0.036   |
| mich-605MI06WR_51            |               |              |          |         |        |                 |         |
| Feature                      | mean<br>train | mean<br>oper | sd train | sd oper | t-stat | degrees<br>free | p value |
| Word Count                   | 200.84        | 160.94       | 137.24   | 114.85  | 3.89   | 248.00          | <.001   |
| Sentence Coherence           | 0.35          | 0.39         | 0.14     | 0.14    | -4.06  | 253.90          | <.001   |
| unique words/total<br>words  | 0.55          | 0.55         | 0.13     | 0.12    | -0.30  | 259.70          | 0.768   |
| English Word<br>Distribution | 0.88          | 0.88         | 0.08     | 0.07    | -0.52  | 251.30          | 0.602   |
| Typical Word<br>Ordering     | 7.20          | 7.04         | 0.96     | 0.90    | 2.22   | 260.60          | 0.028   |
| Readability Measure          | 4.28          | 3.87         | 1.68     | 1.51    | 3.26   | 255.70          | 0.001   |
| mich-405MI06WR_57            |               |              |          |         |        |                 |         |
| Feature                      | mean<br>train | mean<br>oper | sd train | sd oper | t-stat | degrees<br>free | p value |
| Word Count                   | 71.20         | 69.90        | 50.56    | 50.40   | 0.34   | 280.10          | 0.734   |
| Sentence Coherence           | 0.40          | 0.32         | 0.23     | 0.20    | 3.94   | 211.40          | <.001   |
| unique words/total<br>words  | 0.70          | 0.70         | 0.13     | 0.13    | -0.07  | 279.10          | 0.945   |
| English Word<br>Distribution | 0.79          | 0.80         | 0.12     | 0.12    | -1.20  | 277.00          | 0.233   |
| Typical Word<br>Ordering     | 8.34          | 8.18         | 1.13     | 1.13    | 1.84   | 280.60          | 0.067   |
| Readability Measure          | 6.10          | 5.50         | 1.78     | 2.26    | 4.26   | 331.30          | <.001   |

Given the differences between the training and operational data, we decided to see if IEA performance would improve by using a randomly selected subset of the operational data as a more representative training set, and measuring our performance on the remainder of the operational data (those responses we

didn't train on). For this experiment, we excluded responses that received non-numeric human grades. IEA was trained on 200 operational items, since 200 is roughly comparable to the number of responses in original training set, (social studies CR– 303 responses, reading CR1 – 191 responses, writing CR1 – 200 responses, and writing CR2 – 205 responses). The results are shown in Table 2 in the columns labeled 200 IEA exact and 200 IEA adjacent. Also included in the first column of Table 2 is the percentages of exact agreement among the human scorers based on the papers that were read twice for reliability purposes.

Table A.2: IEA Agreement with Human Ratings Under Different Training Scenarios

| CR Prompt         | Human Exact | Oper. Exact | 200 IEA Exact | 500 IEA+ Exact | Oper. Adj. | 200 IEA Adj | 500 IEA+ Adj |
|-------------------|-------------|-------------|---------------|----------------|------------|-------------|--------------|
| Social Studies CR | 71          | 60.1        | 61.1          | 66.8           | 98.3       | 97.3        | 98.8         |
| Reading CR1       | 59          | 45.9        | 50.8          | 56.1           | 88.1       | 91.7        | 88.6         |
| Writing CR1       | 65          | 64.6        | 71.0          | 74.7           | 99.1       | 99.7        | 99.8         |
| Writing CR2       | 75          | 47.8        | 61.8          | 73.1           | 99.6       | 99.4        | 99.8         |

Comparing the first two columns of agreement rates, it can be seen that agreement rates between IEA and human scoring for the operational online responses in the pilot were lower than the paper test agreement rates. However, there was substantial improvement of IEA agreement when training with 200 items from the operational online data as compared to the performance using the original training set. This improvement again suggests that the original training items were not representative of the operational data. As the data in Table A.2 indicate, if IEA had been trained on data more representative of the operational data its performance would have been substantially better.

We have also provided in Table A.2 results using a very recently improved version of IEA. This version, shown in columns labeled 500 IEA+ exact and 500 IEA+ adjacent, requires a slightly larger training set -- an example with 500 items is shown here -- and indicates that we are continuing to improve our automated scoring technology. We believe the agreements for IEA+ are more representative of how well we would score the October 2005 pilot data at this date.

**Appendix B: Results of Student Survey for Social Studies**

| Question   | Omit        | Strongly Agree | Agree          | Neutral       | Disagree          | Strongly Disagree |                   |
|--|-------------|----------------|----------------|---------------|-------------------|-------------------|-------------------|
| 1 The online tutorial was helpful to me.   | N 97<br>% . | 369<br>34      | 616<br>57      | 70<br>6       | 30<br>3           | .<br>.            |                   |
| 2 I understood the special directions for taking this test on the computer.  | N 98<br>% . | 464<br>43      | 553<br>51      | 53<br>5       | 14<br>1           | .<br>.            |                   |
| 3 The test questions were easy for me to read on the computer screen.  | N 97<br>% . | 548<br>51      | 439<br>40      | 80<br>7       | 18<br>2           | .<br>.            |                   |
| 4 It was easy for me to read the reading passage, maps, charts, or tables on the computer screen to answer the questions on the test   | N 97<br>% . | 423<br>39      | 493<br>45      | 130<br>12     | 39<br>4           | .<br>.            |                   |
| 5 It was confusing for me to move between the reading passages, maps, charts, or tables and the test questions on the computer screen. | N 97<br>% . | 99<br>9        | 196<br>18      | 426<br>39     | 364<br>34         | .<br>.            |                   |
| 6 Which of the following online tools did you use while taking the test? Check all that apply.   |             | Highlighter    |                | Eraser        | Answer Eliminator |                   | None used         |
|  | N<br>%      | 467<br>40      | 363<br>31      | 516<br>44     | 423<br>36         |                   |                   |
| 7 Overall, I am happy that I took this test on computer rather than on paper.  | N 97<br>% . | 750<br>69      | 233<br>21      | 60<br>6       | 42<br>4           | .<br>.            |                   |
|  |             | Omit           | Expert         | Above Average | Average           | Below Average     | Beginner          |
| 8 I would rate my computer skills as (please choose only one)  | N 97<br>% . | 187<br>17      | 464<br>43      | 385<br>35     | 25<br>2           | 24<br>2           |                   |
|  |             | Omit           | Strongly Agree | Agree         | Neutral           | Disagree          | Strongly Disagree |
| 9 It is easier for me to write an essay on paper than on the computer  | N 97<br>% . | 234<br>22      | 153<br>14      | 282<br>26     | 416<br>38         | .<br>.            |                   |

Appendix B: Results of Student Survey for Social Studies (Continued)

|    |  |   | Every day | A few times<br>each week | A few<br>times a<br>month | < once a<br>month  | Never                          |     |
|----|--|---|-----------|--------------------------|---------------------------|--------------------|--------------------------------|-----|
| 10 | How often do you use a computer for writing papers or essays?                                    | N | 97        | 111                      | 342                       | 333                | 194                            | 105 |
|    |  | % | .         | 10                       | 32                        | 31                 | 18                             | 10  |
| 11 | How often do you use a computer for your schoolwork or homework?                                 | N | 97        | 142                      | 338                       | 258                | 162                            | 185 |
|    |  | % | .         | 13                       | 31                        | 24                 | 15                             | 17  |
| 12 | How often do you use a computer for personal use at home or outside school?                      | N | 97        | 377                      | 336                       | 130                | 89                             | 153 |
|    |  | % | .         | 35                       | 31                        | 12                 | 8                              | 14  |
| 13 | How often do you use email or internet chat or discussion sessions at home or outside of school? | N | 97        | 271                      | 217                       | 111                | 99                             | 387 |
|    |  | % | .         | 25                       | 20                        | 10                 | 9                              | 36  |
|    |  |   | Omit      | Highly<br>recommend      | Recommend                 | Don't<br>recommend | Strongly<br>don't<br>recommend |     |
| 14 | Would you recommend taking the test on computer to other students?                               | N | 100       | 640                      | 346                       | 61                 | 35                             | .   |
|    |  | % | .         | 59                       | 32                        | 6                  | 3                              | .   |

**Appendix C: Results of Student Survey for English Language Arts**

| Question  |   | Strongly Agree |                |               |                   |               | Strongly Disagree |
|---|---|----------------|----------------|---------------|-------------------|---------------|-------------------|
|   |   | Omit           | Agree          | Neutral       | Disagree          |               |                   |
| 1 The online tutorial was helpful to me.  | N | 63             | 349            | 702           | 67                | 24            | .                 |
|   | % | .              | 31             | 61            | 6                 | 2             | .                 |
| 2 I understood the special directions for taking this test on the computer.                                 | N | 59             | 509            | 588           | 40                | 9             | .                 |
|   | % | .              | 44             | 51            | 3                 | 1             | .                 |
| 3 The test questions were easy for me to read on the computer screen.                                       | N | 60             | 588            | 466           | 78                | 13            | .                 |
|   | % | .              | 51             | 41            | 7                 | 1             | .                 |
| 4 It was easy for me to read the passages on the computer screen  | N | 58             | 552            | 500           | 78                | 17            | .                 |
|   | % | .              | 48             | 44            | 7                 | 1             | .                 |
| 5 It was confusing for me to move between the reading passage and the test questions on the computer screen | N | 59             | 94             | 210           | 485               | 357           | .                 |
|   | % | .              | 8              | 18            | 42                | 31            | .                 |
|   |   |                | Highlighter    | Eraser        | Answer Eliminator | None used     |                   |
| 6 Which of the following online tools did you use while taking the test? Check all that apply.              | N |                | 614            | 464           | 580               | 367           |                   |
|   | % |                | 51             | 39            | 48                | 30            |                   |
| 7 Overall, I am happy that I took this test on computer rather than on paper.                               | N | 60             | 812            | 246           | 54                | 33            | .                 |
|   | % | .              | 71             | 21            | 5                 | 3             | .                 |
|   |   | Omit           | Expert         | Above Average | Average           | Below Average | Beginner          |
| 8 I would rate my computer skills as (please choose only one)   | N | 61             | 193            | 484           | 412               | 29            | 26                |
|   | % | .              | 17             | 42            | 36                | 3             | 2                 |
|   |   | Omit           | Strongly Agree | Agree         | Neutral           | Disagree      | Strongly Disagree |
| 9 It is easier for me to write an essay on paper than on the computer                                       | N | 60             | 179            | 155           | 323               | 488           | .                 |
|   | % | .              | 16             | 14            | 28                | 43            | .                 |

Appendix C: Results of Student Survey for English Language Arts (Continued)

|    |  |   | Every day | A few times<br>each week | A few<br>times a<br>month | < once a<br>month  | Never                          |     |
|----|--|---|-----------|--------------------------|---------------------------|--------------------|--------------------------------|-----|
| 10 | How often do you use a computer for writing papers or essays?                                    | N | 60        | 112                      | 350                       | 385                | 173                            | 125 |
|    |  | % | .         | 10                       | 31                        | 34                 | 15                             | 11  |
| 11 | How often do you use a computer for your schoolwork or homework?                                 | N | 61        | 175                      | 355                       | 284                | 162                            | 168 |
|    |  | % | .         | 15                       | 31                        | 25                 | 14                             | 15  |
| 12 | How often do you use a computer for personal use at home or outside school?                      | N | 61        | 430                      | 327                       | 168                | 81                             | 138 |
|    |  | % | .         | 38                       | 29                        | 15                 | 7                              | 12  |
| 13 | How often do you use email or internet chat or discussion sessions at home or outside of school? | N | 61        | 266                      | 258                       | 121                | 108                            | 391 |
|    |  | % | .         | 23                       | 23                        | 11                 | 9                              | 34  |
|    |  |   | Omit      | Highly<br>recommend      | Recommend                 | Don't<br>recommend | Strongly<br>don't<br>recommend |     |
| 14 | Would you recommend taking the test on computer to other students?                               | N | 62        | 644                      | 410                       | 60                 | 29                             | .   |
|    |  | % | .         | 56                       | 36                        | 5                  | 3                              | .   |