# 2010 Nebraska State Accountability (NeSA)

# Paper and Pencil versus Computer Administered Assessment

# Comparability Study for Reading

**Prepared by**

**Computerized Assessments and Learning**

**October 2010**

# 2010 Nebraska State Accountability (NeSA):

# Paper and Pencil versus Computer Administered Comparability Studies for Reading

## Introduction

Comparability analysis is an important issue when two different modes of testing (e.g. traditional paper and pencil and computer-based) are utilized for the same test administration. Evaluation of the Nebraska reading assessment data from the Spring 2010 test administration will provide item performance statistics as a comparison between the two testing modes used during the NeSA administration: computer-based, vs. paper and pencil (P&P) administrations. Comparability analyses are appropriate in order to assess the performance of the test form modalities on each test item using differential item function analysis.

Differential item functioning (DIF) analysis will provide the state of Nebraska with information regarding how each test item performs between the two matched group modalities. This type of statistical analysis is appropriate as it detects potential item biases between the groups of examinees (Holland & Thayer, 1988) having comparable total test scores. The Mantel-Haenszel (MH) chi-square procedure with one degree of freedom is a method of detecting DIF in a way that is more powerful than other chi-square procedures (Holland & Thayer, 1988).

Comparability analyses using DIF will indicate if an item is favoring one group

over the other (i.e. computer-based versus paper and pencil). For example, an item that exhibits a large level of DIF in favor of the computer-based group indicates that these examinees responded with a correct answer more often than P&P examinees …more often than chance would expect. Analyses would show that this particular item was more difficult for P&P examinees and will be recommended for professional review for potential problems, indeed, unfairness due to the assessment modality.

In this report is a description of the data and the sampling technique employed for analyses, followed by a description of the procedure conducted and results obtained. Conclusions based on the results from data analyses are also provided.

## Data

Spring 2010 Nebraska reading assessment item-level response data for grades 3 through 8 and high school (grade 11) were analyzed for this comparability study. In all, 144,935 students were assessed. This occasion was the first operational administration of each grade level reading examination. The reading assessment for each grade consisted of 45 scoreable items for grades 3 and 4, 48 scoreable items for grades 5, 6 and 7, and 50 scoreable items for grades 8 and 11. Students were assessed either using a computer-based test delivery engine or a parallel paper and pencil modality. In all applications, which modality used was a district, building or teacher decision. In general, between 15 to 18% of tested students at each grade took their grade level reading assessment via

paper and pencil versus on a computer.

To begin analyses, data files were split by testing modality for each grade level. Random samples of students who tested using the computer-based testing modality were created from the computer-based groups of students to match identically the number of students who used the P&P modality at each grade to form equivalent grade samples sizes (see Table 1 for a cross tabulation of sample sizes per grade). Large sample sizes for each grade will yield more accurate and powerful results, as larger sample sizes are more representative of the population of examinees, and also as larger sample size are more stable.

**Table 1. Population and Resultant Sample Sizes by Grade by Modality**

| Grade | Online | Paper/Pencil | Total Students |
|---|---|---|---|
| 3 | 17,645 | 3,851 | 21,496 |
| 4 | 17,207 | 3,947 | 21,154 |
| 5 | 17,079 | 3,643 | 20,722 |
| 6 | 17,090 | 3,355 | 20,445 |
| 7 | 17,243 | 3,091 | 20,334 |
| 8 | 17,216 | 3,148 | 20,364 |
| 11 | 16,937 | 3,483 | 20,420 |

## Procedure

Comparability analyses were conducted to detect the presence of differential item functioning (DIF) amongst reading items on each test form (i.e. grade) between

computer-based and P&P tested groups. It should be noted at the outset that on the computerized assessments at the different grades, different test forms were prepared (different placement positions for items is possible) whereas only one arrangement of items was available in the P&P mode. Thus the computer placements, as as they are "random," are the more preferred arrangements for assessment. For this report and the ensuing analyses, the sequence of items has been arranged and studied using the ordering sequence from the P&P booklets. Continuing then, the performance on a specific item was compared between groups for differences in how examinees responded to test items across overall ability levels. The Mantel-Haenszel (MH) statistical procedure conducts comparability analyses between equivalent score groups. DIFAS 5.0 (Penfield, 2007) was used to calculate DIF using the MH nonparametric method for dichotomous models. The computer-based modality was designated as the Reference group and the P&P modality was designated as the Focal group in the matched group MH analyses. The use of equal and large groups for each grade level yields more powerful item statistics using the MH procedure. Items in each grade level were flagged according to the Educational Testing Service's (ETS) categorization scheme.

DIFAS 5.0 categorizes an item's level of DIF using the now common and standard ETS classification system of 'A' (negligible or nonsignificant DIF), 'B' (moderate DIF), and 'C' (large DIF) (Zieky, 1993). According to Zieky (1993), a classification of 'A' requires that the Mantel-Haenszel delta difference (MH D-DIF) statistic is not significantly different from zero, or has an absolute value less than 1.0. A classification of 'B' requires that the MH D-DIF statistic is significantly different from zero ($p < .05$)

and the absolute value is at least 1.0, and either less than 1.5 or not significantly greater than 1.0, thus evidence of some difference. Finally, a classification of 'C' requires that the MH D-DIF statistic is significantly greater than 1.0 and has an absolute value of 1.5 or more, thus a potentially meaningful difference between the groups being compared on the item.

## Results

Table 2 displays the number of items on each grade level test form that were classified as 'A,' 'B,' and 'C' according to the ETS categorization scheme. Note there were no 'C' classifications for items in grades 3, 4, 6, and 11, indicating there are no items on these assessment forms that revealed notable levels of DIF. 'A' classifications indicate the computer-based and P&P modalities are equivalent for those particular test items. Few items are categorized into the 'B' and 'C' classifications for each grade; these item characteristics are discussed below.

**Table 2. Item Classification by Grade**

| Grade | Number of ‘A’ Items | Number of ‘B’ Items | Number of ‘C’ Items | Total Items |
|---|---|---|---|---|
| 3 | 40 | 5 | 0 | 45 |
| 4 | 44 | 1 | 0 | 45 |
| 5 | 44 | 1 | 3 | 48 |
| 6 | 45 | 3 | 0 | 48 |
| 7 | 41 | 4 | 3 | 48 |
| 8 | 47 | 2 | 1 | 50 |
| 11 | 47 | 3 | 0 | 50 |

Table 3 displays the actual flagged items exhibiting moderate to large DIF (ETS classification of 'B' and 'C'), including the Mantel-Haenszel chi-square value (MH CHI) for each item. The MH chi-square statistic determines whether there is a relationship between group membership and performance on an item, after taking into account student performance on the overall reading test, that is, the students total score. Specifically, analyses show if the probability for success of the focal group members (i.e., P&P examinees) is statistically significantly different than the probability for success of the reference group members (i.e. computer-based examinees). The last column of the tables indicates which group each item is favoring according to the Mantel-Haenszel common log-odds ratio values (i.e. focal vs. reference). Favoring the focal group indicates the P&P students more often responded with a correct answer than the computer-based testing students of similar proficiency on that particular item. Favoring the reference group indicates the computer-based testing students responded more often with a correct answer on the item than comparably scoring (based on total scores) P&P students.

**Table 3. Items Classified as 'B' or 'C' within Grade**

| Grade | Flagged Item | Classification | MH $CHI^2$ | Group Favored |
|---|---|---|---|---|
| 3 | 3 | B | 79.5241 | CBT |
| | 5 | B | 40.1232 | CBT |
| | 20 | B | 117.4507 | P&P |
| | 38 | B | 95.1671 | CBT |
| | 41 | B | 107.9685 | P&P |
| 4 | 5 | B | 39.1692 | P&P |
| 5 | 3 | C | 51.1050 | P&P |
| | 27 | C | 121.0628 | P&P |
| | 28 | C | 76.1440 | CBT |
| | 36 | B | 38.3112 | P&P |
| 6 | 15 | B | 70.7647 | CBT |
| | 33 | B | 48.9802 | P&P |
| | 38 | B | 98.0666 | CBT |
| 7 | 7 | B | 65.0195 | P&P |
| | 10 | B | 46.5281 | CBT |
| | 19 | C | 147.1604 | P&P |
| | 20 | C | 54.7936 | P&P |
| | 26 | B | 72.6113 | CBT |
| | 38 | C | 40.8782 | CBT |
| | 48 | B | 15.7819 | CBT |
| 8 | 25 | C | 132.5038 | CBT |
| | 26 | B | 41.1090 | CBT |
| | 45 | B | 47.6584 | P&P |
| 11 | 12 | B | 76.2890 | P&P |
| | 20 | B | 40.8298 | CBT |
| | 36 | B | 58.4195 | CBT |

## Findings and Conclusions

Twenty-six (26) of the 334 scoreable NeSA 2010 Reading Assessment items, or 7.8 percent, were flagged for potential DIF when evaluated for difference between computer versus paper and pencil performance difference. That means 92.2% of the items showed no mode effect. As will be discussed in the paragraphs that follow, we must not lose sight of this overall impressive finding. Of the 26 flagged items, 12 items signaled greater statistical likelihood favoring paper and pencil test takers, whereas 14 items yielded higher performance by examinees on computer. While the 12-14 division may suggest a "split" of the difference, thus a "wash" that would be an improper conclusion. Since 26 items present potentially differences in performance due to test mode effects, it is important that these items be examined by the test developers and content experts, and NDE staff.

The comparability of paper and pencil and computer-based testing modalities on the Nebraska state assessments sheds light on how these two groups of examinees respond or react to test items. The analyses show that the flagged items (7.8% of total items) need to be reviewed for how each item is worded and presented, as computer-based examinees are responding differently than paper and pencil examinees. The NDE will want to be sure that the two forms of testing are measuring the same construct of reading, which is the expected and intended goal of utilizing the two modes of testing. If construct equivalence is supported for the DIF items, then working to equalize performance through equating based on modality (i.e., CBT and P&P) must be considered and adopted as necessary. Finally, although the items in

Table 3 are flagged for review, there is a 5% chance that many of these items were flagged by chance alone. Item review may indicate that some items do not need to be changed or eliminated on test forms, as the items are truly equivalent between the test forms. However, a Type I error can occur (by chance alone, flagging an item as DIF when in fact no difference between administration modes truly exists) when items that have been flagged for DIF are overlooked in a review. Large equivalent samples were used for each grade level to control for error rates and increase the power of the statistical analyses used. The review of the flagged items by testing and content experts, and by NDE advisors will assist in controlling for Type I error and eliminate item biases in test forms.

According to the ETS categorization scheme, items categorized as 'B' represent moderate DIF, having MH D-DIF statistics statistically different from zero and absolute values of at least 1, and either less than 1.5 or not significantly greater than 1 (Zieky, 1993, p. 342). The items listed in Table 3 classified as 'B' represent moderate levels of DIF, with a favor towards either the computer-based or P&P testing modalities. Although only 7.8% of the items, the item statistics do indicate a need for review of the items on the two testing modalities for potential biases towards testing modality group membership.

Items categorized as 'C' represent large DIF, having MH D-DIF statistics significantly greater than 1.0 and has absolute values of 1.5 or greater (Zieky, 1993, p. 342). The items listed in Table 3 classified as 'C' represent large levels of DIF. These

items show greater differences between the testing modalities, as either the computer-based or P&P testing modality groups are performing better on specific items. It will be important to review these items closely, as it might be that there is a problem with an item as displayed, worded, etc. on the computer versus on paper.

The current comparability analyses of traditional paper and pencil versus computer-based administrations in the Nebraska Spring 2010 reading assessments serve as evidence of a need for a review of 7.8% of the total number of items flagged as potentially lacking equivalency across the two testing modes on specific items. The items reported in Table 3 are recommended as the items for thoughtful review and study. Reading experts and item writers should be contacted and asked to inspect the items flagged in this report to identify potential sources of problems between the computer-based and P&P groups, with a consideration for editing, revising or removing the item(s) from the test forms if bias is evident and could be corrected. It is also very possible that an outcome coming from thorough inspection by reviewers would find no rational or reasonable disparity (thus a potential Type 1 error), in which case the item should be left intact and in place, and then studied during another administration when possible. The relatively very few items triggering the need for further study thus suggests convincingly that the construct being evaluated on the assessments by grade is not different based on test mode. Therefore score adjustments based on equating would be sufficient, if indeed necessary. Further psychometric study will define a fair and equitable course of action.

Finally, the likelihood of differences between performance in the modality deserves attention in one other respect. As a reading comprehension measure, the 2010 NeSA assessments are formed using "passage dependent" items. That is, students read a passage, then respond to a number of specific test questions, usually 5 to 7 items, directly associated with the passage. Thus, items within passages are certainly **not** independent. Reviewing flagged items for patterns within passages will be important as the passage-item dependence may be a source of difficulty upon review and study.

In conclusion, it must be said that for first time operational Reading assessments, given in two very different test administration modes (the computer mode unquestionably new and novel to Nebraska students), having only 7.8% of the total number of items flagged for review (and under a statistical criterion wherein 5% will be flagged by chance alone) based on the mode effect should be most heartening to Nebraska. The effort in the creation and development of these items and resulting assessments and in the used of a computerized assessment delivery engine that achieved equivalence between the test modes is exemplary.

## References

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

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

Penfield, R. D. (2007). *DIFAS 4.0 user's manual*. Retrieved August 31, 2010, from http://www.education.miami.edu/facultysites/penfield/index.html

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.). *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum Associates.

# Appendices

## Appendix 1. Grade 3 MANTEL-HAENZEL Statistics

| Item | MH $CHI^2$ | LOR SE | ETS |
|---|---|---|---|
| 1 | 0.7229 | 0.0627 | A |
| 2 | 0.7355 | 0.0648 | A |
| 3 | 79.5241 | 0.0632 | B |
| 4 | 1.7520 | 0.0505 | A |
| 5 | 40.1232 | 0.0674 | B |
| 6 | 0.5947 | 0.0642 | A |
| 7 | 54.3937 | 0.0576 | A |
| 8 | 50.9717 | 0.0584 | A |
| 9 | 2.9394 | 0.0633 | A |
| 10 | 16.2926 | 0.0518 | A |
| 11 | 7.3131 | 0.0529 | A |
| 12 | 6.9276 | 0.0653 | A |
| 13 | 1.1922 | 0.0557 | A |
| 14 | 6.1809 | 0.0567 | A |
| 15 | 3.9622 | 0.0493 | A |
| 16 | 0.0793 | 0.0592 | A |
| 17 | 2.9482 | 0.0545 | A |
| 18 | 10.6318 | 0.0512 | A |
| 19 | 3.2426 | 0.0652 | A |
| 20 | 117.4507 | 0.0555 | B |
| 21 | 1.5017 | 0.0674 | A |
| 22 | 9.6889 | 0.0552 | A |
| 23 | 4.6206 | 0.0591 | A |
| 24 | 6.2066 | 0.0601 | A |
| 25 | 7.6708 | 0.0700 | A |
| 26 | 11.6153 | 0.0938 | A |
| 27 | 3.7145 | 0.0666 | A |
| 28 | 0.0000 | 0.0667 | A |
| 29 | 6.7557 | 0.0485 | A |
| 30 | 18.8156 | 0.0507 | A |
| 31 | 0.6073 | 0.0659 | A |
| 32 | 0.5700 | 0.0573 | A |
| 33 | 1.8327 | 0.0718 | A |
| 34 | 0.1516 | 0.0591 | A |

| | | | |
|---|---|---|---|
| 35 | 10.9581 | 0.0543 | A |
| 36 | 1.7369 | 0.0576 | A |
| 37 | 5.2024 | 0.0524 | A |
| 38 | 95.1671 | 0.0588 | B |
| 39 | 1.2203 | 0.0564 | A |
| 40 | 27.4247 | 0.0556 | A |
| 41 | 107.9685 | 0.0528 | B |
| 42 | 0.5222 | 0.0645 | A |
| 43 | 27.4821 | 0.0751 | A |
| 44 | 8.2225 | 0.0595 | A |
| 45 | 16.6558 | 0.0518 | A |

Note. LOR SE is the standard error of the Mantel-Haenszel common log-odds ratio.

## Appendix 2. Grade 3 Conditional Differences

| Lower Score Value | 3 | 8 | 12 | 16 | 21 | 25 | 29 | 33 | 38 | 42 |
|---|---|---|---|---|---|---|---|---|---|---|
| Upper Score Value | 7 | 11 | 15 | 20 | 24 | 28 | 32 | 37 | 41 | 45 |
| Item 3 – B | -0.03 | 0.04 | 0.07 | 0.10 | 0.17 | 0.15 | 0.04 | 0.09 | 0.06 | 0.02 |
| Item 5 – B | -0.40 | -0.04 | -0.01 | 0.08 | 0.14 | 0.13 | 0.07 | 0.02 | 0.03 | 0.00 |
| Item 20 – B | 0.17 | -0.06 | -0.01 | -0.05 | -0.12 | -0.15 | -0.13 | -0.12 | -0.10 | -0.07 |
| Item 38 – B | 0.00 | -0.03 | 0.09 | 0.04 | 0.10 | 0.14 | 0.17 | 0.12 | 0.06 | 0.03 |
| Item 41 – B | -0.20 | -0.02 | -0.13 | -0.11 | -0.07 | -0.16 | -0.15 | -0.14 | -0.09 | -0.03 |

Notes. Items classified by ETS as 'B' or 'C' only are reported and labeled next to respective items. Cell differences represent mean proportional differences between online and paper/pencil categories of assessment. For example, a value of 0.10 means there is a 10% difference at the score values between 16-20 favoring the paper and pencil examinees.

**Appendix 3. Grade 4 MANTEL-HAENZEL Statistics**

| Item | MH $CHI^2$ | LOR SE | ETS |
|---|---|---|---|
| 1 | 32.7840 | 0.0705 | A |
| 2 | 21.4605 | 0.0560 | A |
| 3 | 3.9048 | 0.0612 | A |
| 4 | 0.8996 | 0.1354 | A |
| 5 | 39.1692 | 0.0765 | B |
| 6 | 4.9944 | 0.0493 | A |
| 7 | 0.0161 | 0.0652 | A |
| 8 | 10.2541 | 0.0579 | A |
| 9 | 0.5403 | 0.0470 | A |
| 10 | 16.6233 | 0.0650 | A |
| 11 | 4.3777 | 0.0523 | A |
| 12 | 39.9016 | 0.0620 | A |
| 13 | 40.9307 | 0.0618 | A |
| 14 | 45.0811 | 0.0512 | A |
| 15 | 16.9712 | 0.0498 | A |
| 16 | 32.4398 | 0.0566 | A |
| 17 | 56.5433 | 0.0523 | A |
| 18 | 19.1755 | 0.0643 | A |
| 19 | 1.3602 | 0.0502 | A |
| 20 | 18.9745 | 0.0585 | A |
| 21 | 38.0238 | 0.0608 | A |
| 22 | 15.8048 | 0.0617 | A |
| 23 | 4.3048 | 0.0585 | A |
| 24 | 1.5220 | 0.0599 | A |
| 25 | 0.7035 | 0.0580 | A |
| 26 | 20.6036 | 0.0554 | A |
| 27 | 5.6330 | 0.0666 | A |
| 28 | 0.9127 | 0.0518 | A |
| 29 | 14.2849 | 0.0505 | A |
| 30 | 18.2595 | 0.0539 | A |
| 31 | 2.8754 | 0.0508 | A |
| 32 | 0.6717 | 0.0626 | A |
| 33 | 0.6502 | 0.0628 | A |
| 34 | 0.5038 | 0.0807 | A |
| 35 | 3.5525 | 0.0541 | A |
| 36 | 0.9315 | 0.0529 | A |

| | | | |
|---|---|---|---|
| 37 | 14.4309 | 0.0576 | A |
| 38 | 1.7992 | 0.0601 | A |
| 39 | 0.2390 | 0.0582 | A |
| 40 | 1.0503 | 0.0480 | A |
| 41 | 0.6169 | 0.0773 | A |
| 42 | 0.1289 | 0.0787 | A |
| 43 | 60.7867 | 0.0521 | A |
| 44 | 0.4065 | 0.0692 | A |
| 45 | 9.8943 | 0.0647 | A |

Note. LOR SE is the standard error of the Mantel-Haenszel common log-odds ratio.

## Appendix 4. Grade 4 Conditional Differences

| Lower Score Value | 5 | 9 | 13 | 17 | 21 | 25 | 29 | 33 | 37 | 41 |
|---|---|---|---|---|---|---|---|---|---|---|
| Upper Score Value | 9 | 13 | 17 | 21 | 25 | 29 | 33 | 37 | 41 | 45 |
| Item 5 – B | 0.21 | 0.08 | -0.06 | -0.03 | -0.17 | -0.14 | -0.02 | -0.03 | 0.00 | 0.00 |

Note. Items classified by ETS as 'B' or 'C' only are reported and labeled next to respective items. Cell differences represent mean proportional differences between online and paper/pencil categories of assessment. For example, a value of 0.21 means there is a 21% difference at the score values between 5-9 favoring the paper and pencil examinees.

**Appendix 5. Grade 5 MANTEL-HAENZEL Statistics**

| Item | MH $CHI^2$ | LOR SE | ETS |
|---|---|---|---|
| 1 | 28.3919 | 0.0660 | A |
| 2 | 1.6580 | 0.0837 | A |
| 3 | 51.1050 | 0.1077 | C |
| 4 | 0.1178 | 0.0557 | A |
| 5 | 3.6962 | 0.0532 | A |
| 6 | 8.4829 | 0.0873 | A |
| 7 | 0.0060 | 0.0662 | A |
| 8 | 3.2157 | 0.0748 | A |
| 9 | 0.0004 | 0.0507 | A |
| 10 | 2.9682 | 0.0603 | A |
| 11 | 0.0546 | 0.0528 | A |
| 12 | 8.3303 | 0.0546 | A |
| 13 | 0.8843 | 0.0516 | A |
| 14 | 2.2772 | 0.0551 | A |
| 15 | 0.1562 | 0.0543 | A |
| 16 | 2.1361 | 0.0510 | A |
| 17 | 20.2466 | 0.0574 | A |
| 18 | 0.7625 | 0.0553 | A |
| 19 | 2.7714 | 0.0696 | A |
| 20 | 1.6995 | 0.0810 | A |
| 21 | 0.5397 | 0.0631 | A |
| 22 | 1.0316 | 0.0595 | A |
| 23 | 2.5188 | 0.0525 | A |
| 24 | 1.1198 | 0.0548 | A |
| 25 | 4.9157 | 0.0528 | A |
| 26 | 9.0402 | 0.0543 | A |
| 27 | 121.0628 | 0.0609 | C |
| 28 | 76.1440 | 0.0773 | C |
| 29 | 0.1113 | 0.0525 | A |
| 30 | 18.1274 | 0.0679 | A |
| 31 | 6.5880 | 0.0509 | A |
| 32 | 22.2550 | 0.0546 | A |
| 33 | 0.2406 | 0.0635 | A |
| 34 | 7.5968 | 0.0806 | A |
| 35 | 11.3801 | 0.0950 | A |
| 36 | 38.3112 | 0.0782 | B |

| | | | |
|---|---|---|---|
| 37 | 4.7338 | 0.1105 | A |
| 38 | 7.4471 | 0.0582 | A |
| 39 | 3.3351 | 0.0728 | A |
| 40 | 21.4886 | 0.0545 | A |
| 41 | 19.0308 | 0.0589 | A |
| 42 | 2.8235 | 0.0605 | A |
| 43 | 0.5061 | 0.0544 | A |
| 44 | 1.5227 | 0.0519 | A |
| 45 | 0.0220 | 0.0622 | A |
| 46 | 1.8125 | 0.0814 | A |
| 47 | 1.4726 | 0.0510 | A |
| 48 | 0.0191 | 0.0513 | A |

Note. LOR SE is the standard error of the Mantel-Haenszel common log-odds ratio.

## Appendix 6. Grade 5 Conditional Differences

| Lower Score Value | 7 | 12 | 16 | 20 | 24 | 29 | 33 | 37 | 41 | 45 |
|---|---|---|---|---|---|---|---|---|---|---|
| Upper Score Value | 11 | 15 | 19 | 23 | 28 | 32 | 36 | 40 | 44 | 48 |
| Item 3 – C | -0.05 | -0.13 | -0.17 | -0.11 | -0.06 | -0.05 | -0.04 | -0.02 | 0.00 | -0.01 |
| Item 27 – C | -0.15 | -0.05 | -0.13 | -0.15 | -0.19 | -0.18 | -0.14 | -0.08 | -0.04 | -0.03 |
| Item 28 – C | 0.07 | 0.18 | 0.00 | 0.14 | 0.15 | 0.11 | 0.05 | 0.05 | 0.02 | 0.01 |
| Item 36 – B | -0.10 | 0.01 | -0.14 | -0.09 | -0.08 | -0.08 | -0.06 | -0.02 | -0.01 | 0.00 |

Note. Items classified by ETS as 'B' or 'C' only are reported and labeled next to respective items. Cell differences represent mean proportional differences between online and paper/pencil categories of assessment. For example, a value of -0.18 means there is an 18% difference at the score values between 29-32 favoring the online examinees.

**Appendix 7. Grade 6 MANTEL-HAENZEL Statistics**

| Item | MH $CHI^2$ | LOR SE | ETS |
|---|---|---|---|
| 1 | 5.1136 | 0.1172 | A |
| 2 | 4.4563 | 0.0720 | A |
| 3 | 25.3656 | 0.0651 | A |
| 4 | 4.8268 | 0.1031 | A |
| 5 | 7.4537 | 0.0517 | A |
| 6 | 15.1966 | 0.1002 | A |
| 7 | 7.7513 | 0.0550 | A |
| 8 | 0.3911 | 0.0673 | A |
| 9 | 0.3067 | 0.0659 | A |
| 10 | 2.1835 | 0.0511 | A |
| 11 | 23.0771 | 0.0665 | A |
| 12 | 0.2032 | 0.0669 | A |
| 13 | 4.9803 | 0.0589 | A |
| 14 | 1.4592 | 0.0729 | A |
| 15 | 70.7647 | 0.0632 | B |
| 16 | 5.3438 | 0.0560 | A |
| 17 | 21.2390 | 0.0573 | A |
| 18 | 22.4903 | 0.0577 | A |
| 19 | 0.6101 | 0.1146 | A |
| 20 | 29.3498 | 0.0695 | A |
| 21 | 26.8434 | 0.0754 | A |
| 22 | 9.7063 | 0.0629 | A |
| 23 | 1.0563 | 0.0707 | A |
| 24 | 3.0244 | 0.0553 | A |
| 25 | 0.9628 | 0.0645 | A |
| 26 | 2.9764 | 0.0590 | A |
| 27 | 14.7220 | 0.0786 | A |
| 28 | 0.2404 | 0.0562 | A |
| 29 | 0.0120 | 0.0532 | A |
| 30 | 8.2581 | 0.0546 | A |
| 31 | 56.1617 | 0.0560 | A |
| 32 | 10.8868 | 0.0586 | A |
| 33 | 48.9802 | 0.0622 | B |
| 34 | 31.7145 | 0.0552 | A |
| 35 | 0.9042 | 0.0714 | A |
| 36 | 12.9521 | 0.0600 | A |

| | | | |
|---|---|---|---|
| 37 | 0.0227 | 0.0712 | A |
| 38 | 98.0666 | 0.0627 | B |
| 39 | 7.9660 | 0.0598 | A |
| 40 | 0.9275 | 0.0797 | A |
| 41 | 0.1643 | 0.0603 | A |
| 42 | 2.3199 | 0.0587 | A |
| 43 | 0.0027 | 0.0612 | A |
| 44 | 10.6674 | 0.0610 | A |
| 45 | 6.0060 | 0.0764 | A |
| 46 | 0.1523 | 0.0809 | A |
| 47 | 1.6550 | 0.0531 | A |
| 48 | 9.8131 | 0.0571 | A |

Note. LOR SE is the standard error of the Mantel-Haenszel common log-odds ratio.

## Appendix 8. Grade 6 Conditional Differences

| Lower Score Value | 5 | 10 | 15 | 19 | 23 | 28 | 32 | 36 | 40 | 45 |
|---|---|---|---|---|---|---|---|---|---|---|
| Upper Score Value | 9 | 14 | 18 | 22 | 27 | 31 | 35 | 39 | 44 | 48 |
| Item 15 – B | 0.39 | 0.09 | 0.07 | 0.10 | 0.15 | 0.10 | 0.15 | 0.07 | 0.04 | 0.04 |
| Item 33 – B | -0.11 | -0.01 | 0.05 | -0.08 | -0.09 | -0.13 | -0.11 | -0.13 | 0.00 | -0.02 |
| Item 38 – B | -0.11 | -0.01 | 0.13 | 0.14 | 0.14 | 0.18 | 0.13 | 0.13 | 0.05 | 0.00 |

Note. Items classified by ETS as 'B' or 'C' only are reported and labeled next to respective items. Cell differences represent mean proportional differences between online and paper/pencil categories of assessment. For example, a value of 0.39 means there is a 39% difference at the score values between 5-9 favoring the paper and pencil examinees.

**Appendix 9. Grade 7 MANTEL-HAENZEL Statistics**

| Item | MH $CHI^2$ | LOR SE | ETS |
|---|---|---|---|
| 1 | 12.7848 | 0.0566 | A |
| 2 | 2.3389 | 0.0673 | A |
| 3 | 0.3202 | 0.0569 | A |
| 4 | 22.2778 | 0.0745 | A |
| 5 | 0.2876 | 0.0676 | A |
| 6 | 11.3900 | 0.0837 | A |
| 7 | 65.0195 | 0.0576 | B |
| 8 | 0.8540 | 0.0562 | A |
| 9 | 12.9583 | 0.0740 | A |
| 10 | 46.5281 | 0.0675 | B |
| 11 | 9.4151 | 0.0626 | A |
| 12 | 1.8327 | 0.0575 | A |
| 13 | 0.2902 | 0.0584 | A |
| 14 | 11.6721 | 0.0597 | A |
| 15 | 21.1871 | 0.0738 | A |
| 16 | 6.0092 | 0.0607 | A |
| 17 | 1.7218 | 0.0763 | A |
| 18 | 0.0005 | 0.0609 | A |
| 19 | 147.1604 | 0.0635 | C |
| 20 | 54.7936 | 0.0949 | C |
| 21 | 1.1352 | 0.0543 | A |
| 22 | 20.1132 | 0.0679 | A |
| 23 | 0.5063 | 0.0549 | A |
| 24 | 6.7073 | 0.0579 | A |
| 25 | 1.0941 | 0.0554 | A |
| 26 | 72.6113 | 0.0530 | B |
| 27 | 2.3551 | 0.0656 | A |
| 28 | 7.3652 | 0.0619 | A |
| 29 | 10.1692 | 0.0520 | A |
| 30 | 30.5267 | 0.0548 | A |
| 31 | 0.1766 | 0.0682 | A |
| 32 | 12.8191 | 0.0580 | A |
| 33 | 0.0001 | 0.0629 | A |
| 34 | 10.6502 | 0.0696 | A |
| 35 | 14.3028 | 0.0663 | A |
| 36 | 5.2952 | 0.0566 | A |

| | | | |
|---|---|---|---|
| 37 | 11.8599 | 0.0695 | A |
| 38 | 40.8782 | 0.1034 | C |
| 39 | 22.1554 | 0.0642 | A |
| 40 | 0.2187 | 0.0825 | A |
| 41 | 0.5512 | 0.1034 | A |
| 42 | 1.0274 | 0.0681 | A |
| 43 | 0.2069 | 0.0589 | A |
| 44 | 23.4454 | 0.0546 | A |
| 45 | 8.8259 | 0.0563 | A |
| 46 | 0.3396 | 0.0622 | A |
| 47 | 0.5528 | 0.0539 | A |
| 48 | 15.7819 | 0.1150 | B |

Note. LOR SE is the standard error of the Mantel-Haenszel common log-odds ratio.

## Appendix 10. Grade 7 Conditional Differences

| Lower Score Value | 7 | 12 | 16 | 20 | 24 | 29 | 33 | 37 | 41 | 45 |
|---|---|---|---|---|---|---|---|---|---|---|
| Upper Score Value | 11 | 15 | 19 | 23 | 28 | 32 | 36 | 40 | 44 | 48 |
| Item 7 – B | -0.05 | -0.07 | -0.11 | -0.12 | -0.09 | -0.13 | -0.06 | -0.10 | -0.11 | -0.02 |
| Item 10 – B | 0.01 | -0.01 | 0.10 | 0.09 | 0.14 | 0.07 | 0.10 | 0.06 | 0.03 | -0.01 |
| Item 19 – C | -0.03 | -0.04 | 0.00 | -0.09 | -0.14 | -0.14 | -0.24 | -0.13 | -0.13 | -0.04 |
| Item 20 – C | -0.16 | -0.06 | -0.12 | -0.16 | -0.11 | -0.11 | -0.03 | 0.00 | 0.00 | 0.00 |
| Item 26 – B | 0.01 | 0.10 | 0.04 | -0.02 | 0.11 | 0.11 | 0.10 | 0.15 | 0.15 | 0.07 |
| Item 38 – C | 0.06 | 0.13 | 0.15 | 0.15 | 0.05 | 0.02 | 0.04 | 0.02 | 0.02 | 0.00 |
| Item 48 – B | 0.12 | 0.22 | 0.12 | 0.03 | 0.06 | 0.04 | -0.01 | -0.01 | 0.00 | 0.01 |

Note. Items classified by ETS as 'B' or 'C' only are reported and labeled next to respective items. Cell differences represent mean proportional differences between online and paper/pencil categories of assessment. For example, a value of -0.12 means there is a 12% difference at the score values between 20-23 favoring the online examinees.

**Appendix 11. Grade 8 MANTEL-HAENZEL Statistics**

| Item | MH $CHI^2$ | LOR SE | ETS |
|---|---|---|---|
| 1 | 2.1593 | 0.0566 | A |
| 2 | 2.1858 | 0.0555 | A |
| 3 | 2.2570 | 0.1136 | A |
| 4 | 0.0035 | 0.0691 | A |
| 5 | 8.1420 | 0.0626 | A |
| 6 | 0.2481 | 0.1089 | A |
| 7 | 0.2246 | 0.0603 | A |
| 8 | 2.2270 | 0.0735 | A |
| 9 | 0.1413 | 0.0979 | A |
| 10 | 1.6762 | 0.0544 | A |
| 11 | 6.1067 | 0.0752 | A |
| 12 | 8.2516 | 0.0668 | A |
| 13 | 12.2163 | 0.0583 | A |
| 14 | 31.3411 | 0.0618 | A |
| 15 | 5.4557 | 0.0810 | A |
| 16 | 28.2951 | 0.0610 | A |
| 17 | 0.6198 | 0.0600 | A |
| 18 | 0.8844 | 0.0698 | A |
| 19 | 23.5661 | 0.0578 | A |
| 20 | 11.6362 | 0.0623 | A |
| 21 | 26.5352 | 0.0604 | A |
| 22 | 4.5221 | 0.0715 | A |
| 23 | 27.1582 | 0.0556 | A |
| 24 | 35.5420 | 0.0694 | A |
| 25 | 132.5038 | 0.0665 | C |
| 26 | 41.1090 | 0.0767 | B |
| 27 | 17.0964 | 0.0534 | A |
| 28 | 1.0801 | 0.0578 | A |
| 29 | 50.0212 | 0.0562 | A |
| 30 | 0.4038 | 0.0714 | A |
| 31 | 6.0963 | 0.0585 | A |
| 32 | 4.5769 | 0.0825 | A |
| 33 | 10.4273 | 0.0661 | A |
| 34 | 10.3516 | 0.0612 | A |
| 35 | 15.7986 | 0.0542 | A |
| 36 | 1.0819 | 0.0566 | A |

| | | | |
|---|---|---|---|
| 37 | 0.2661 | 0.0627 | A |
| 38 | 0.6660 | 0.0893 | A |
| 39 | 1.4638 | 0.0634 | A |
| 40 | 0.4847 | 0.0618 | A |
| 41 | 2.0849 | 0.0548 | A |
| 42 | 0.3002 | 0.0531 | A |
| 43 | 0.0043 | 0.0546 | A |
| 44 | 9.4766 | 0.0576 | A |
| 45 | 47.6584 | 0.0670 | B |
| 46 | 1.8936 | 0.0648 | A |
| 47 | 0.0382 | 0.0788 | A |
| 48 | 1.8448 | 0.0624 | A |
| 49 | 9.4953 | 0.0568 | A |
| 50 | 4.2389 | 0.0554 | A |

Note. LOR SE is the standard error of the Mantel-Haenszel common log-odds ratio.

## Appendix 12. Grade 8 Conditional Differences

| Lower Score Value | 3 | 9 | 13 | 18 | 23 | 28 | 32 | 37 | 42 | 46 |
|---|---|---|---|---|---|---|---|---|---|---|
| Upper Score Value | 8 | 12 | 17 | 22 | 27 | 31 | 36 | 41 | 45 | 50 |
| Item 25 – C | 0.00 | 0.10 | 0.04 | 0.23 | 0.17 | 0.21 | 0.15 | 0.11 | 0.06 | 0.02 |
| Item 26 – B | 0.00 | -0.04 | 0.04 | 0.07 | 0.10 | 0.12 | 0.09 | 0.03 | 0.03 | -0.01 |
| Item 45 – B | 0.00 | -0.12 | -0.09 | -0.02 | -0.10 | -0.09 | -0.09 | -0.07 | -0.03 | -0.01 |

Note. Items classified by ETS as 'B' or 'C' only are reported and labeled next to respective items. Cell differences represent mean proportional differences between online and paper/pencil categories of assessment. For example, a value of 0.10 means there is a 10% difference at the score values between 9-12 favoring the paper and pencil examinees.

**Appendix 13. Grade 11 MANTEL-HAENZEL Statistics**

| Item | MH $CHI^2$ | LOR SE | ETS |
|---|---|---|---|
| 1 | 0.1753 | 0.052 | A |
| 2 | 0.0522 | 0.0545 | A |
| 3 | 14.1088 | 0.0749 | A |
| 4 | 2.1366 | 0.0762 | A |
| 5 | 13.0986 | 0.0518 | A |
| 6 | 2.001 | 0.0727 | A |
| 7 | 5.1492 | 0.0498 | A |
| 8 | 4.7591 | 0.06 | A |
| 9 | 0.9466 | 0.0718 | A |
| 10 | 16.4062 | 0.0913 | A |
| 11 | 3.8415 | 0.063 | A |
| 12 | 76.289 | 0.0563 | B |
| 13 | 14.0296 | 0.0897 | A |
| 14 | 0.4987 | 0.0526 | A |
| 15 | 3.6807 | 0.0575 | A |
| 16 | 31.4004 | 0.0512 | A |
| 17 | 1.0072 | 0.056 | A |
| 18 | 1.2272 | 0.0552 | A |
| 19 | 1.3717 | 0.0502 | A |
| 20 | 40.8298 | 0.0762 | B |
| 21 | 4.0772 | 0.0519 | A |
| 22 | 2.678 | 0.0502 | A |
| 23 | 22.6206 | 0.059 | A |
| 24 | 2.46 | 0.0641 | A |
| 25 | 3.9983 | 0.0582 | A |
| 26 | 25.0708 | 0.061 | A |
| 27 | 8.1239 | 0.0525 | A |
| 28 | 12.0675 | 0.0659 | A |
| 29 | 0.8474 | 0.061 | A |
| 30 | 3.7984 | 0.0695 | A |
| 31 | 5.4505 | 0.0833 | A |
| 32 | 2.6385 | 0.1063 | A |
| 33 | 0.0371 | 0.068 | A |
| 34 | 18.7351 | 0.0492 | A |
| 35 | 33.8048 | 0.0575 | A |
| 36 | 58.4195 | 0.0659 | B |

| | | | |
|---|---|---|---|
| 37 | 28.0323 | 0.0789 | A |
| 38 | 0.5346 | 0.0531 | A |
| 39 | 1.9181 | 0.0545 | A |
| 40 | 6.2472 | 0.0614 | A |
| 41 | 25.8784 | 0.0551 | A |
| 42 | 31.2416 | 0.0611 | A |
| 43 | 0.1312 | 0.1041 | A |
| 44 | 4.7674 | 0.0576 | A |
| 45 | 5.4484 | 0.0678 | A |
| 46 | 10.9344 | 0.0701 | A |
| 47 | 12.0477 | 0.0605 | A |
| 48 | 5.8027 | 0.0565 | A |
| 49 | 9.0911 | 0.053 | A |
| 50 | 0.1711 | 0.0564 | A |

Note. LOR SE is the standard error of the Mantel-Haenszel common log-odds ratio.

## Appendix 14. Grade 11 Conditional Differences

| Lower Score Value | 7 | 12 | 17 | 21 | 25 | 30 | 34 | 38 | 42 | 47 |
|---|---|---|---|---|---|---|---|---|---|---|
| Upper Score Value | 11 | 16 | 20 | 24 | 29 | 33 | 37 | 41 | 46 | 50 |
| Item 12 – B | -0.21 | -0.09 | -0.02 | -0.07 | -0.10 | -0.08 | -0.16 | -0.10 | -0.06 | -0.04 |
| Item 20 – B | -0.08 | 0.18 | 0.06 | 0.18 | 0.06 | 0.10 | 0.04 | 0.02 | 0.01 | 0.00 |
| Item 36 – B | 0.00 | 0.16 | 0.05 | 0.18 | 0.13 | 0.08 | 0.08 | 0.04 | 0.02 | -0.01 |

Note. Items classified by ETS as 'B' or 'C' only are reported and labeled next to respective items. Cell differences represent mean proportional differences between online and paper/pencil categories of assessment. For example, a value of -0.21 means there is a 21% difference at the score values between 7-11 favoring the online examinees.