Report of the Expert Panel on the Impact of the ISTEP+ Scoring Issue

Edward Roeber, Derek Briggs and Wes Bruce

December 21, 2015

Summary

An expert panel's independent review of data provided by CTB finds no evidence that students were erroneously given a lower score on the Spring 2015 ISTEP+ Writing tests, the first part of the two-part E/LA assessments. The State Board of Education (SBOE) independent expert panel was comprised of Derek Briggs, Professor, University of Colorado; Wes Bruce, Consultant; and Edward Roeber, Assessment Director, Michigan Assessment Consortium.

The panel analyzed the anonymous allegation that a software glitch caused CTB to erroneously assign a lower score to the Writing assessments. In response to this allegation, CTB asserted that the software glitch in question was extremely difficult to reproduce, had no effect on student scores, was addressed through a procedural change when it was brought to the attention of scoring manager, and then the software was updated to avoid the possibility of the glitch happening again eight days after being brought to CTB's attention.

The panel's analyses included, among other things, a comparison of: (1) the percentage of students receiving high scores (e.g., 5s and 6s) scores on the Writing portion of the ISTEP+ E/LA assessment before and after the software problem was fixed, and (2) the percentage of students receiving identical scores on both parts of the ISTEP+ Writing tests before and after the software problem was fixed. These were the areas where a glitch, if it occurred, would have had the most pronounced impact. The panel found no evidence of changes in student scores on the writing section of the ISTEP+. Based on these analyses, the expert panel also believes that this issue did not have an impact on the scores in the remaining parts of the ISTEP+ assessments.

Introduction

On Sunday, December 13, 2015, the Indianapolis Star newspaper ran a story based on an anonymous letter it received alleging that a glitch that occurred during the scoring of the 2015 ISTEP+ assessments had resulted in erroneous scores being given to students on both the ISTEP+ mathematics and English language arts assessments. This newspaper story resulted in additional efforts to investigate this situation to determine whether or not such a glitch did occur and if it did, how many students' scores were affected. The purpose of this report is to summarize the steps taken to investigate this situation, the data examined, and the conclusions drawn.

The anonymous letter writer indicated that scorers were encouraged to enter the two scores using number keypads attached to the computers, since this would be more efficient than using pull-down menus. Scoring software was set so that the scorer would enter the first score, then the computer would move on to show where the second score was to be entered, and the scorer would then enter the second score. The computer would then move on to the next student to be scored.

The anonymous letter also indicated that the glitch was first reported on April 22, 2015, eight days after scoring students' responses began. According to the letter, CTB was apprised of the issue and scorers were told not to use the number keypad but instead, to use the mouse and dropdown menus for score entry. The letter indicated that a meeting with scoring supervisors was held on April 30, 2015.

CTB was asked, according to the anonymous letter writer, about re-scoring all of the student responses already scored during these eight days. CTB staff indicated that the responses would not be rescored.

ISTEP+ Scoring

A number of written-response mathematics and English language arts ISTEP+ assessment items require scorers to score each of the items on two dimensions. For example, in ELA the writing prompt at each grade is first scored using a 6-point writing rubric and then on a 4-point grammar rubric; and in mathematics the problem solving items at each grade are scored using a 2- or 3-point content rubric and a corresponding 2- or 3-point process rubric. Samples of these item types can be found at http://www.doe.in.gov/assessment/istep-grades-3-8.

During scoring, two checks were made on the reliability and validity of scoring respectively. First, a random sample of 5% of student responses was read by a second reader. These so-called "read-behinds" are performed by scoring leaders. This serves as a check to make sure that scorers are rating the responses consistently with one another. The second check gives scorers expert-pre-scored responses at pre-determined intervals, such as after each scorer has scored 25 responses. These checksets are embedded among the items being scored and appear like all other responses. Failure to score a checkset correctly creates an immediate flag for the scoring supervisor, who may review with the scorer the rubric being used, re-train the scorer if this re-occurs, or may even dismiss the scorer is this occurs frequently. This checkset serves to assure that scorers are validly scoring responses according to the criteria of the scoring rubrics and sample student responses on which they were trained and certified. These are quality control steps typically taken during scoring of student written responses.

It was alleged in the anonymous letter that it was possible to quickly enter the two scores on a student's essay for the two dimensions such that the second score given not only was entered into the second field on the computer but also replaced the first score. If true, this would result in the same two scores being applied to each student for which this glitch occurred. And, because the range of possible scores on the second dimension of the Writing assessments was only 0-4, fewer students would receive high scores (5s and 6s) on the first dimension of the Writing assessments.

Steps Taken by CTB and IDOE to Investigate

The same anonymous letter was received by the Indiana Department of Education (IDOE) on November 25, 2015 and forwarded to CTB for response on November 30, 2015. By the time of the newspaper story, the IDOE had already begun with an investigation of the allegation through the ISTEP+ program contractor CTB. The director of the ISTEP+ program had sent CTB a number of questions and requests for data that might be used to investigate this situation.

On behalf of CTB, Ellen Haley responded on December 8, 2015 to the anonymous letter, indicating that it was first apprised of the issue on April 22, 2015, and that the issue was brought to the attention of scoring management the same day. Ms. Haley writes:

"[s]scoring management immediately looked into the keypad issue but had difficulty reproducing it. Only when the scoring director, after many tries, entered two scores in very rapid succession, before the second trait screen had fully loaded, did the issue occur. The director had to enter the two scores nearly simultaneously to cause the override."

The letter gives a detailed chronology of events in the April 22-30, 2015 period. The letter indicates that on May 1, a fix for the scoring issue was released and began to be used on May 2, 2015. Ms. Haley concludes her letter with this statement:

"In sum, my investigation confirmed the scoring director's opinion that the keypad occurrence did not impact the scoring. If an evaluator entered successive scores too quickly, he or she would see the overwrite/scores change and could go back and re-enter both scores. As soon as the issue was reported, evaluators were instructed not to use the keypad and to use the mouse for score entries. In addition, quality metrics – check

sets, read behinds, inter-rater reliabilities for responses read twice, pattern of "split scores" versus same scores by trait – none of these indicated an issue for student scores. The issue was not a common occurrence, was actually difficult to create in live scoring, and was fixed quickly – both in on-the-floor instructions and then technically, in the PEMS software. Based on CTB's quality control tests, there was no need to rescore any tests as a result of the keypad issue."

Additional information about actions taken by IDOE and CTB are shown in the attached Appendix. It shows the chronology of steps taken to investigate the scoring issue in the appendix of this report.

Steps Taken by Expert Panel to Investigate

Upon learning of this issue on December 13, 2015, the experts suggested several ways to investigate the issue. Their questions and requests for data were forwarded to CTB by SBOE staff, and CTB responded promptly with answers to the questions, data sets as requested, and interpretations of the data for expert panel review.

The panel started by assembling a chronology of the scoring Incident, shown in Figure 1.

Figure 1. Timeline of Investigation of the Scoring Incident

Date	**Activity**
April 8	ISTEP scoring begins
April 22	A "Writing" scorer reports keypad issue to their supervisor
	Supervisors were instructed to direct scorers to stop using the numerical keypad
April 30	Regular meeting with all scoring supervisors (ISTEP and other programs). Supervisors were reminded the keypad was not to be used and told that a fix would soon be deployed
May 1	At about 9:30 pm, a software update that eliminated the use of the numerical keypad in CTB scoring system was deployed.
November 25	Anonymous letter received at IDOE
November 30	Anonymous letter sent by IDOE to CYB for response
December 8	CTB responds to IDOE
December 13	Indianapolis Star prints story about ISTEP+ scoring
December 13-21	Expert panel suggests methods to determine whether the glitch occurred systematically and if so, what impacts it had

During the December 13-20, 2015 period, the expert panel was provided with the data sent to the IDOE as well as additional data requested by the SBOE on behalf of the expert panel. Two types of data were especially important for the expert panel to review. These are 1) differences in the proportion of student responses given high scores (5s and 6s) on the extended writing prompt before and after the glitch was fixed, and 2) whether the percent of students given identical first and second scores went down after the glitch was corrected. These two sets of data are important in detecting whether a second score over-riding a first score changed the response that should have been given to students.

Three time periods were examined in order to determine the impact of the glitch on scoring. These are:

- Period 1 (Beginning of scoring to April 22) – Scoring that took place before the glitch was discovered
- Period 2 (April 23 to May 1) – Scoring that occurred during the time when scorers were told to not use their number keypads use for data entry but the glitch had not been fixed
- Period 3 (May 2 to end of scoring) – Scoring that took place after the glitch was corrected by CTB.

The expert panel looked at both the mathematics and ELA data supplied to it, but felt that given the

score scales used to score the Writing prompts, this content area was most likely to show differences if the glitch occurred on a wide-scale basis during scoring. Scores for writing quality are scored on a scale of 0 to 6 point scale, and these would have been reduced by the second score, which was scored on a 0 to 4 point scale.

If the glitch resulted in the first score (which could range up to 6) being replaced by the second score (that could only go as high as 4), then there should be fewer 5s and 6s in the Period 1 and perhaps in Period 2 as well (especially if all scorers did not switch to scoring using pull-down menus as alleged by the anonymous letter writer), when compared to the number of 5s and 6s in Period 3 (when the number keypads were de-activated).

Table 1 shows the percent of high scores data for all of the ISTEP+ Writing prompts.

Table 1. Scoring Trends Before, During and After Discovery of Error – All Writing Prompts

Grade	RIB #	Rubric Score	Percent of Students Receiving Score			% Period 3– % Period 1
			Period 1	Period 2	Period 3	
3	1	5	4.96	5.12	4.64	-0.32
3	1	6	1.08	1.06	0.70	-0.38
4	1	5	4.05	4.16	3.38	-0.67
4	1	6	0.91	0.87	0.30	-0.61
6	1	5	7.28	7.91	7.93	0.65
6	1	6	0.62	0.47	0.39	-0.23
7	1	5	5.96	8.38	3.29	-2.67
7	1	6	0.71	0.79	0.33	-0.38
7	2	5	5.42	4.88	4.08	-1.34
7	2	6	1.21	0.36	0.20	-1.01

Table 1 shows that with one exception (grade 6), a slightly smaller proportion of 5s and 6s were received in Period 3 than in Period 1. There appears to be little or no evidence that the glitch, if it did occur, caused scores to be lowered in Period 1.

The expert panel also used a second type of data to investigate whether the glitch impacted students' scores. Had the glitch been occurring on a wide-scale basis during Period 1, not been fully corrected in Period 2 (because some scorers were continuing to use the number keypads in spite of CTB directions not to do so, as alleged by the anonymous letter writer), and then eliminated in Period 3, then the data should also show significantly large percentages of exact agreement in Period 1 versus Period 3 for the two scores given to each Writing response.

Table 2 shows the number of exact scores given to student responses on the two dimensions used to score Writing responses during the three time periods (before and after the glitch was discovered and corrected). The comparisons shown are for assessment items in which significant numbers of student responses were scored during each of the three scoring periods.

Table 2. Percent of Students Receiving Duplicate Scores Before, During and After Discovery of Error

	Percent Exact Agreement First/Second Scores			Period 3 –
Writing Task	Time Period 1	Time Period 2	Time Period 3	Period 1
3 Writing RIB 1	47.35	43.29	48.72	1.37
4 Writing RIB 2	42.38	27.91	46.27	3.89
6 Writing RIB 2	38.94	50.00	50.94	12.00
7 Writing RIB 1	36.34	48.56	33.00	-3.34
7 Writing RIB 2	35.86	36.52	35.49	-0.37

As can be seen, there is virtually no pattern of higher exact agreement on the first and second dimension scores in Period 1 versus Period 3 across the five prompts for which such comparisons are possible. Three prompts showed more exact agreement between the first and second dimension scores in Period 3 than Period 1, and two prompts showed less exact agreement in Period 3 versus Period 1. However, none of these are large differences and could be accounted for by differences in the students scored in each time period. Since a clear pattern is not evident, evidence of a glitch that changed students' scores is not evident here, either.

Conclusions

The IDOE, CTB, SBOE, and the expert panel took steps to investigate this situation. Unfortunately, since scoring of the ISTEP+ tests concluded over six months earlier, some of the steps that might have been possible to examine as the glitch was supposedly occurring aren't possible so many months after the conclusion of scoring. In addition, looking at large data files does not mean that each student was scored accurately, only whether large effects of the glitch can be discerned. Finally, although the expert panel investigated data from three periods (before the glitch was discovered, when scorers were not to use the number keypads, and after the number keypads were disabled), the lack of comparability between the students scored during each time period hampers the cross-period analyses.

With these disclaimers aside, however, the expert panel did not see evidence, in the form of either reduced scores or higher exact agreements among scores for the same responses, that supports the allegations of the anonymous letter writer. There does not appear to be a discernable impact of this scoring glitch on overall scores given to students' responses.

APPENDIX

Chronology of Steps Taken to Investigation of the ISTEP+ Scoring Issue

As summarized in our report, there were a number of steps taken by IDOE, CTB, the SBOE, and the expert panel to investigate the allegations contained in the anonymous letter sent to both the Indianapolis Star newspaper and the Indiana Department of Education (IDOE). This chronology of the steps taken is provided to assure educators and the public that the allegations were carefully and thoroughly investigated, using all available data. Several pieces of information are cited here:

- Anonymous letter sent to the IDOE and the IDOE (attached)
- December 8 letter from CTB, which was a response to the anonymous letter sent to it on November 30 by IDOE (excerpted in the expert panel report and attached)
- December 15 CTB response to the second communication from IDOE dated Dec 10
- December 17 CTB response to the expert panel questions sent on December 13
- December 18 CTB response transmitted with additional data provided to the expert panel
- CTB responses attached to the December 18 transmittal containing CTB interpretations of the additional data provided to the expert panel

Anonymous Letter
This letter is attached.

December 8 Letter from CTB to the IDOE
This letter is attached.

December 15 CTB Response to December 10 E-Mail from IDOE
Based on the December 8 letter from CTB, IDOE responded on December 10, 2015 via e-mail with a series of additional questions for CTB. CTB responded to these questions (in italics) on December 15, 2015:

- "The number of test items that were scored between April 13 and April 22 (all dates inclusive). *Answer: 72 items had at least one student response scored. See "ISTEP Part 1" attachment.*
- The number of students whose tests were scored between April 13 and April 22 (all dates inclusive). *Answer: 227,373 students had at least one item scored in this time frame. See "ISTEP Part 1" attachment.*
- The schools and school corporations that had their tests scored between April 13 and April 22 (all dates inclusive) *Answer: 1,503 schools had at least one item scored in this time frame. See "ISTEP Part 1" attachment.*
- The number of test items that were scored between April 23 and April 30 (all dates inclusive). *Answer: 63 items had at least one student response scored. See "ISTEP Part 1" attachment.*
- The number of students whose tests were scored between April 23 and April 30 (all dates inclusive). *Answer: 354,059 students had at least one item scored in this time frame. See "ISTEP Part 1" attachment.*
- The schools and school corporations that had their tests scored between April 23 and April 30 (all dates inclusive). *Answer: 1,822 schools had at least one item scored in this time frame. See "ISTEP Part 1" attachment.*
- Within the date ranges in question (April 13-22 and April 23-30), were the test items being scored narrowly focused to one item type (i.e., writing prompt)? Please identify the item type(s) being scored at that time. *Answer: All types of constructed response items were scored during this time period (Math, ELA CR, ELA ER and Writing). See "ISTEP Part 1" attachment.*
- What specific assurances can CTB provide that the tests scored between April 13 and April 30 were scored accurately? Are these assurances based on statistical analysis after the fact, real-time data generated during testing, current review of the actual tests themselves, or some other manner? *Answer:*

See six quality reports attached. Based on these reports, generated daily during scoring, with three days highlighted here for comparison purposes, and a careful review of all quality data, we see no evidence of any impact from the keypad occurrence and can assure Indiana that student scores are correct.

- *Each day, scoring quality was monitored to ensure valid and reliable scoring was taking place. As part of the quality monitoring process, the following statistics were captured for each item every day that item was being scored*
    - *Validity: Pre-scored responses sent at random to readers to test adherence to scoring rubrics and guidelines*
    - *Inter-Rater Reliability: 5 percent of all documents were scored twice, allowing tracking of reader agreement rates*
    - *Score point distribution: Percent of responses scored at each score point*
- *As part of the validity process, readers were given immediate feedback if they miss scored a response. This included showing the reader the scores they provided as well as the correct scores. Had there been an issue with the first trait score not being captured correctly, this would have been noted immediately by the reader. With hundreds of readers scoring, thousands of validity responses were being scored each day.*
- *Validity statistics for items that were being scored prior to the issue being resolved were in expected range, and were comparable to the validity statistics for items scored after the issue was corrected.*
- *Inter-rater reliability statistics for the first trait of items do not indicate an issue with reader agreement, which we would see if first trait scores were being overwritten. IRR stats for items scored prior to the issue were comparable to similar items that were scored after the issue was corrected.*
- *Score point distribution for multi-trait items do not indicate issues with the first trait being overwritten by the 2nd. While split scores are less common in the writing items, and thus the SPD of the 2 traits align (this is the case both for items scored before the issue was corrected as well as after), this is expected. For math, however, SPD of the 2 traits are relatively independent, and this is reflected in both the items that were scored prior to the issue being corrected as well as the items scored after.*
- *Also as a note, when the keys were hit in such a way to make the defect occur, the score changes visible on screen. No reader noted this occurring prior to 4/22 despite hundreds of thousands of item reads that were completed to that point, indicating this was not a common occurrence."*

<u>December 17 CTB Response to Expert Panel Questions from December 13</u>

Several questions were posed by the expert panel by December 13, 2015. Both the expert panel questions and responses from CTB (which were received on December 17, 2015) are shown below:

Question: On the QA charts, which items are the items in question.

Answer: ISTEP items with 2 dimensions:

a. All Math Items (score ranges of 0-2/0-2 or 0-3/0-3)
b. All Writing items (score range of 1-6/1-4)
c. All ELA ER - RIB 4 and RIB 8 for each grade (score range of 1-4/1-4)

Question: content area the rater who reported was scoring.
Answer: Supervisor that reported the issue was overseeing scoring of Writing.

Question: Definitions of terms that are not defined such as "red flags" or "yellow flags."

Answer: Red flags and yellow flags – a reader will, on average, take between 10 and 15 checksets per day. Each day, the reader is expected to maintain a certain level of exact agreement against the key scores for the checkset responses that they score. A reader falling below this standard received a red flag, which results in corrective action being taken up to and including removal from scoring for that item and resetting of responses. A yellow flag is given if the reader is above the required quality standard, but below the standard required for qualification on the item.

Question: Number of dimensions scored on each of the open-end items

Answer: All items on ISTEP are either single dimension or 2 dimension items. The 2-dimension items are listed in #1 above.

Questions with Answers to come later today, or Friday morning:

- The pre-disabling and post-disabling data on levels of exact agreement among the various dimensions on items where there are two or more dimensions for all of the items.

- As I understand the issue, if the second score over-rides the first and shows up as both the first and the second score, there should be a higher level of exact agreement for the scores students received who were scored prior to disabling the key pads versus those scored after the key pads were disabled. This could be a cumulative score report from scoring prior to disabling the key pads (what the official date of this May 3 or May 4?) and a cumulative one for scoring done *after* the key pads were disabled (not cumulative through all scoring). I did not see this information, but perhaps I am not reading the reports correctly.

- The RIB reports seem to show exact agreement between scorers - either of the check sets or the read-behinds. This is different from the information that I requested and is only tangentially related to intra-student score agreement that I am interested in.

- The two windows provided are from the time the issue was identified and before and the time the issue was identified until the fix was put into place. I/we need to see how this compares with AFTER the fix was in place. Same numbers provided, but for AFTER the issue was corrected.

- The comparison (number and percentage) of 5s and 6s awarded in these windows and for all three windows.

- Scores (or score distribution) up to keypad being disabled ("was released on May 1, 2015 at 9:30 p.m") and after -- at a minimum from beginning of scoring through last shift on May 1 (Friday) and from May 4 to the end of scoring. Need it only for live items, even week by week.

- Read behind data on items with identical scores (4/4, 3/3, 2/2, 1/1)."

<u>December 18 CTB Response with Additional Data Provided to the Expert Panel</u>

The following response from CTB was forwarded to the ISBE staff and the experts on December 18, 2015:

"Please find attached the data to answer your remaining questions, noted below. You requested GA summary data, but the data belongs to that customer, and they did not give me consent to share it with you. I believe the data here and in my previous two emails should answer your questions for Indiana.

Our findings on the attached set of data are noted in 1 and 2 at the bottom of this email. We do not see any evidence of the keypad overwrites, and we see no impact on student scores in this or any of the data.

Scores (or score distribution) up to keypad being disabled ("was released on May 1, 2015 at 9:30 p.m.") and after -- at a minimum from beginning of scoring through last shift on May 1 (Friday) and from May 4 to the end of scoring. Need it only for live items, even week by week.

Read behind data on items with identical scores (4/4, 3/3, 2/2, 1/1)

The pre-disabling and post-disabling data on levels of exact agreement among the various dimensions on items where there are two or more dimensions for all of the items. And As I understand the issue, if the second score over-rides the first and shows up as both the first and the second score, there should be a higher level of exact agreement for the scores students received who were scored prior to disabling the key pads versus those scored after the key pads were disabled. This could be a cumulative score report from scoring prior to disabling the key pads (what the official date of this May 3 or May 4?) and a cumulative one for scoring done after the key pads were disabled (not cumulative through all scoring). I did not see this information, but perhaps I am not reading the reports correctly.

The comparison (number and percentage) of 5s and 6s awarded in these windows and for all three windows.

There are 2 major indicators in this particular set of data:

1. Looking at the ISTEP Writing, a similar percentage of responses were given 5's and 6's for trait A in all 3

time periods for which we gathered data (the time before the defect was discovered, the time between discovery and the fix, and the time after the fix). Since the 2nd trait has a max score of 4, we would see a lower percentage of 5's and 6's for trait 1 had the first trait been overwritten by the trait 2 scores.

2. Looking across Writing, ELA ER and Math, the percentage of responses receiving the same score for trait A and trait B likewise was comparable when measured across the three time periods. Since the effect of the defect would be to artificially increase the number of responses receiving the same score for both traits, we would see a larger percentage of these in the earlier time period had the defect impacted scoring, and we do not see this.

We do not see any evidence of the overwriting occurring. The data is very clean. We do not see any impact on student scores. "

December 18 CTB Responses Attached to the Transmittal E-Mail Containing Additional Data Provided to the Expert Panel

Three additional responses to the questions posed listed above were attached to the December 18, 2015 CTB letter. These responses serve as explanations of the accompanying data files sent to the expert panel, as well as CTB's interpretation of what each data file shows. These explanations are:

"ISTEP_13_5s_6s_Comparison

There are 3 tabs: "BOS to EOD 4-22" represents data from the start of scoring until the end of day on 4-22. $-22 was when the issue was discovered and readers were told to score using the mouse instead of the keyboard. "BOD 4-23 to EOD 5-1" represents data from scoring on the beginning of the day on 4-23 to the end of the day on 5-1. The fix was put in place after scoring ended on 5-1. "BOD 5-2 to EOS" represents data from the beginning of the day on 5-2 to the end of scoring. This is scoring that occurred after the issue had been fixed.

RIBNAME represents the name of the item (all items on this report are Writing items

ITEM represents the item number and data point (all data points on this report are for trait A (the first of the two traits)

ITEMGUID represents the internal item number

ITEMTRSITGUID represents the internal trait number

"Score" represents the score given to the trait. This report shows the number of responses given either a 5 or a 6.

Count is the number of responses given the listed score point

Total count is the number of responses scored in the given time frame

Percent is the percent of responses scored that were given the score during the given time frame.

Interpreting this report: This report is intended to show how often students on the extended writing essay received scores of 5 or 6 in the three time periods in question. The reason why this is important is that the second trait for the extended writing has a maximum score of 4. so, if the first trait scores were being overwritten due to the defect, we would likely see a lower number of scores of 5 or 6 on the first trait, as a scorer that intended to score a 5-4, for example, would have instead had the score recorded as a 4-4. This would show in the statistics as fewer students receiving scores of 5 and 6 in the time period that the defect was present vs. the number of 5's and 6's given in the period when the defect had been corrected.

Observations on the data in the report: Looking at items which had significant amounts of scoring in more than one time period, we do not see any patterns which indicate fewer 5's and 6's were being given during the period when the defect was present. For example, 4 Writing RIB 2 had 4.05 percent of responses (out of 55597) receive a score of 5 and 0.91 percent of responses scored as a 6 during the first time period. During the 2nd time period, this was 4.16 percent 5's (out of 20240) and 0.87 percent 6's. During the third time period, we see 3.38 percent 5's and 0.3 percent 6's (out of 1330). The first time period is when we would expect fewer5's and 6's had the defect been impacting the score data, but we do not see this. The same hold for the other items which have significant amounts of scoring taking place in multiple time periods. There is no indication that fewer 5's and 6's were being given when the defect was present."

"ISTEP_09_Exact_Agreement

There are 3 tabs: "BOS to EOD 4-22" represents data from the start of scoring until the end of day on 4-22. $-22 was when the issue was discovered and readers were told to score using the mouse instead of the keyboard. "BOD 4-23 to EOD 5-1" represents data from scoring on the beginning of the day on 4-23 to the end of the day on 5-1. The fix was put in place after scoring ended on 5-1. "BOD 5-2 to EOS" represents data from the beginning of the day on 5-2 to the end of scoring (there is a typo here, spreadsheet says 5-22 instead of 5-2). This is scoring that occurred after the issue had been fixed.

RIBNAME is the name of the item

TotalResponseCount is the number of responses scored during the given time period

ExactResponseCount is the number of responses scored during the given time period where the score for the first trait and the score for the second trait were the same numerical value (0-0, 1-1, 2-2, 3-3 or 4-4).

ExactResponse percent is the percentage of responses scored during the given time period where the score for the first trait and the score for the second trait were the same numerical value.

Interpreting this report: This report shows how often the score for the 2 traits/dimensions for an item matched. For example, on a Math CR item, the score range for trait A is 0-2and the score range for trait B is 0-2, so possible numerical score combinations are 0-0, 0-1, 0-2, 1-0, 1-1, 1-2, 2-0, 2-1 and 2-2. If the defect were impacting scores, we would tend to see more scores where the score for trait A and the score for trait B matched, so this percentage would be higher during the earlier time period when the defect was present.

Observations on the data: We do not see any trends where an item is showing a higher percentage of matching A/B scores in the earlier time periods, thus showing that the defect was not having an impact on the students' scores. To see this, we would look at items which had significant amounts of scoring in multiple time periods, and compare the percent of A/B matching scores. For example, 4 math RIB 2 had 26380 responses scored in the first time period, 33779 in the second time period, and 14116 in the third time period. The percent of matching A/B scores was 38.44 percent in the first time period, 40.90 in the second time period, and 38.29 percent in the third time period. Had the defect been impacting scores, we would see a higher percent of matching scores in the first time period of scoring that took place before the defect was noted. This holds true as you look at all of the items scored in the earlier time period. We do not see any trend of higher percentage of matching A/B scores for scores applied before the defect was notice or before it was corrected."

"ISTEP_02_ScoreFreq

There are 3 tabs: "BOS to EOD 4-22" represents data from the start of scoring until the end of day on 4-22. 4-22 was when the issue was discovered and readers were told to score using the mouse instead of the keyboard. "BOD 4-23 to EOD 5-1" represents data from scoring on the beginning of the day on 4-23 to the end of the day on 5-1. The fix was put in place after scoring ended on 5-1. "BOD 5-2 to EOS" represents data from the beginning of the day on 5-2 to the end of scoring. This is scoring that occurred after the issue had been fixed.

RIBNAME is the name of the item

Item is the item number and trait

ITEMGUID is the internal item number

ITEMTRAIGUID is the internal trait number

Zero-Six and A-E - This shows the number of responses scored at each numerical score point and each condition code during the given time frame.

Interpreting this report: It is more difficult to use this report to make a claim about the impact of the defect, but the defect, if it were impacting students' scores, would show an impact in the score distributions for trait A (score distributions would be different during the earlier time period when the defect was present).

Observations about the data: There is no indication of different score distributions for trait A in the earlier time period vs. the later time periods.

Anonymous Letter Sent to IDOE and the Indianapolis Star

To Whom It May Concern:

A new system (PEMS) was utilized this spring for scoring the 2015 spring ISTEP test. A major flaw in the system was brought to management's attention after 8 days of scoring. The flaw was that the system changed students' scores when utilizing the numerical keypad.

Specifically, on those items in which 2 scores were assigned, the second score entered on the numerical key pad would override the first score. For example, if a student was to receive a score of (4, 2) and a score of (4, 2) was entered too quickly when utilizing the numerical keypad, the system would assign a score of 2 to both parts since 2 was the last number entered. Thus the student would receive a score of (2, 2) rather than (4, 2). This problem was common knowledge among evaluators, team leaders and supervisors, all of whom were very concerned.

A meeting was held a week later on April 30th to discuss this issue. Mike Conarroe chaired the meeting. Mr. Conarroe made two decisions regarding what was happening.

The first decision was that supervisors were to instruct their groups to no longer use the numerical keypad and instead use the mouse. Unfortunately, some evaluators continued to use the keypad anyway out of habit since that is how they were initially taught and had been scoring in this fashion for years. This in fact was always previously encouraged because the method was faster than using the mouse. This benefitted the evaluators because having high production numbers helped to ensure that they would be called in to work on other projects.

The second decision was that all the students who had potentially been assigned incorrect scores would not be re-scored, because it would put the project behind.

My only concern is that students be assigned correct scores. Because of this flaw in the system students scores were obviously compromised. Because the majority of evaluators have always used the keypad, it is safe to say this has had a major impact on the integrity of the results. I certainly hope that you can convince CTB to rescore these student tests since the stakes on this test are so high. However, because of the decisions that were made, and because it is unlikely that you were ever informed about what occurred, I fear that Mr. Conarroe and CTB Management will not be forthcoming about this matter. If so, I hope that others present in the April 30th meeting will be.

Unfortunately, I must communicate this information to you anonymously. Even though I am no longer employed by CTB, I did sign a confidentiality agreement when I was hired.

Sincerely,

2015 CTB Employee

Ellen Haley
Executive Vice President
McGraw-Hill Education
ellen.haley@mheducation.com

December 8, 2015

Michele Walker, Ed.D.
Director of Student Assessment
Indiana Department of Education
South Tower, Suite 600
115 W. Washington Street
Indianapolis, IN 46204

Dear Dr. Walker,

I am writing to confirm our discussion from last week about the issues raised in the anonymous letter sent by someone identifying himself or herself as a former CTB employee, received by your office on November 25, 2015 and forwarded to CTB by e-mail on November 30, 2015 concerning CTB's Performance Evaluation Monitoring System (PEMS.) At the IDOE's request, we conducted a thorough review of the allegations raised in the anonymous letter. This letter summarizes our investigation.

During the investigation, I personally interviewed the employees who were involved in CTB's spring scoring processes and supporting scoring software. These included: Mike Conarroe, Director of Hand Scoring; Derek Adams, Director of PEMS Software Development, Christy Huggins, Director of Software QA, and Brenda Williams, Sr. Director of Operations. I reviewed relevant emails and other documents from the April 22, 2015 time period. In sum, as a result of the investigation, I found that there was a very rare, anomalous, temporary keypad issue, which was resolved immediately in the scoring process and fixed quickly in the software. The issue did not affect student scores, as evidenced by our many quality metrics -- check sets, read behinds, inter-rater reliabilities for responses read twice, pattern of "split scores" versus same scores by trait – all of which are monitored throughout each day of scoring. While we do not know the identity of the author of the anonymous letter, we suspect that it could have been a temporary scoring supervisor, who was terminated recently. If the author was in fact a temporary scorer or scoring supervisor, he or she would not have had access to the above documents or information.

Below I will summarize the sequence of events and details related to the keypad occurrence:

On April 22, 2015, an evaluator notified his scoring supervisor that when using the numeric keypad to enter trait scores for a student's response, in certain instances, the assigned scores in PEMS were being changed after he entered them.

On the same day, the supervisor escalated the issue to scoring management and suggested restricting keypad use until the issue was investigated.

Scoring management immediately looked into the keypad issue but had difficulty reproducing it. Only when the scoring director, after many tries, entered two scores in very rapid succession, before the second trait screen had fully loaded, did the issue occur. The director had to enter the two scores nearly simultaneously to cause the override.

The director found that when the User Interface was transitioning between the traits, if an evaluator were to enter the second trait score while the first trait was still on the screen, the system would enter the second score

20 Ryan Ranch Road | Monterey, CA 93940 | phone: (831) 393.7757

for both traits. But if the evaluator waited for the second trait to finish loading, and the first trait was completely collapsed in the accordion, then the second score would enter properly and did not change the first score. If the scores were entered with a pause, rather than nearly simultaneously, the scores got entered as expected.

Management concluded that the override occurrence was very rare. Readers were trained to read the student response for the first trait and provide a score. Then read the same response for the second trait, and enter that score, Even if a reader assigned two traits after one reading, normal keyboard techniques and careful score entry by our trained readers would indicate that the very rapid entry was highly unusual.

Despite the infrequency of the override, on April 22, 2015, scoring supervisors and evaluators were immediately notified of the issue and were directed to not use the keypad but to instead use the mouse to enter the trait scores while CTB investigated the cause of the problem and developed a fix. Using the mouse was a straightforward approach – most scorers use the mouse in any case.

Scoring management submitted the issue as a defect to our Tier 3 technology team, and it was entered as a defect in the tracking system by Tier 3 on April 24, 2015.

On April 28, 2015, the Tier 3 team developed a fix, which disabled the keypad completely, so that the mouse was -- and still is -- the only way to enter a score. The fix for the keypad issue was tested on April 30, 2015 and was released on May 1, 2015 at 9:30 p.m. Scoring had the new release for use when scoring resumed the next day.

I also wanted to specifically address your three questions:

1) Was there a change in the scoring system/tool evaluators used to score the Spring 2015 ISTEP+ assessment? If yes, please provide specific details regarding this change (when was it implemented, what were the *major* differences between the new system and the prior system, etc.).

   In Spring 2015, CTB used our PEMS (Performance Evaluation Monitoring System) hand scoring software. PEMS was developed over the last few years to replace our EHS (Electronic Handscoring System) and to support the complex content introduced by the new state and common core standards of recent years.

   PEMS is a significant upgrade to CTB's hand scoring capabilities and was first used in by CTB in January 2014. EHS was a desktop, client-server system. PEMS is a web-based, distributed scoring system. EHS was able to best handle simple open ended items and prompts. PEMS was designed to score technology-enhanced items, images, complex content, and to handle the volume of this content for consortia states as well as states like Indiana that adopted more rigorous standards and introduced tests with more involved content.

2) What (if any) issues or problems arose with the scoring of student responses using the scoring system (if the system used for Spring 2015 ISTEP+ scoring was not new, please respond to this question based on the existing system as it relates to the allegations in the complaint letter)?

   There were scoring interface and backend issues related to the new system, as is to be expected with any fairly new system. Daily meetings between Scoring and Technology senior staff were held, as they are every spring, to proactively monitor system usage and to discuss and resolve possible technical

topics. We had a "war room" to monitor and check for issues and to resolve any issues detected. This is the same process used every scoring season for all systems.

3) Was there a meeting that took place on April 30, 2015, (or any other date) to discuss issues with the scoring system used for the Spring 2015 ISTEP+ assessment? If so, please provide a copy of the agenda and specific details regarding what was discussed/shared during this meeting as it relates to the allegations in the complaint letter.

   The author of the anonymous letter referenced a meeting on April 30, 2015. I confirmed that there was a site meeting with the scoring supervisors on that date. The meeting was an informational, non-project specific update for our supervisors, working on all projects, not just PEMS projects, at the site. The meeting included approximately 20-22 temporary supervisors and 1-3 CTB regular employees. The April 30, 2015 meeting was not specifically called to discuss the keypad issue, but an update on the keypad issue was one of several topics covered. Specifically, supervisors were informed that the fix was in progress and that evaluators were to continue to use the mouse. There was no written agenda prepared for the meeting.

   In addition to the meeting on April 30, 2015, there were daily quality meetings with all the supervisors – both the day and the evening supervisors. The day supervisors met in the morning and the evening supervisors met in the late afternoon.

In sum, my investigation confirmed the scoring director's opinion that the keypad occurrence did not impact the scoring. If an evaluator entered successive scores too quickly, he or she would see the overwrite/scores change and could go back and re-enter both scores. As soon as the issue was reported, evaluators were instructed not to use the keypad and to use the mouse for score entries. In addition, quality metrics – check sets, read behinds, inter-rater reliabilities for responses read twice, pattern of "split scores" versus same scores by trait – none of these indicated an issue for student scores. The issue was not a common occurrence, was actually difficult to create in live scoring, and was fixed quickly – both in on-the-floor instructions and then technically, in the PEMS software. Based on CTB's quality control tests, there was no need to rescore any tests as a result of the keypad issue.

Sincerely,

Ellen Haley

20 Ryan Ranch Road | Monterey, CA 93940 | phone: (831) 393.7757