Score Distances of Technology Enhanced Items

Wenhao Wang

Jessica T. Loughran

University of Kansas

Paper presented at the annual meeting of the National Council on Measurement in Education

Chicago, IL. April, 2015
Introduction

With the advent of the Common Core State Standards (CCSS), state assessments now include more technology-enhanced (TE) item types. TE item types include ordering, select text, drag-and-drop and matching. TE items are usually scored with partial credit. The most common scoring rule for these TE items is that every correct answer has the same score value and the item score is the multiplication of the number of correct answers and the single score value. This scoring rule means that students will earn the same score if they only give two correct answers for an item with four correct answers no matter which two correct answers they give.

In this study, the item score distance is defined as the difference between two adjacent scores. Therefore, for TE items with this scoring rule, the observed item score distances are equal across different adjacent score categories and are equal to the single score value. For example, for a TE item that has four correct answers and each correct answer scores 1, the observed item score distance is 1 for this item.

In order to ensure the accuracy of this scoring rule, item writers constructed these items with the goal that the latent item score distances are equal across different adjacent latent score categories. If the latent score distances are also equal, the item will provide exact measurements of the latent proficiency (Wakita, Ueshima, & Noguchi, 2012). However, the latent score distances for some TE items are not equal because there are differences between the expectation of the item writers and the real behavior of test takers. As a result, TE item scoring rules might lead to inaccurate observed score distributions.

The purpose of this study was to examine TE items that have scoring rules that lead to equal observed score distance and identify the TE items without equal latent score distances. We also studied the effect of including these unequal latent distance items on the observed score distribution. If the effect is large, the reasons for unequal latent score distances and the suggestions on changing item scoring rules will be provided.

Latent Score Distance

The psychological distance method proposed by Wakita, Ueshima, and Noguchi (2012) was used to calculate the latent score distance for the TE items. The latent score distance is the scoring system
supported by the data. The psychological distance method is suitable to indicate systematic errors of measurement (Wakita, etc., 2012). In order to calculate the latent score distances, the generalized partial credit model (GPCM, Muraki, 1992) was fitted to TE items and the item parameters were calibrated. The psychological distance method assumes that the latent continuum exists behind the latent score categories and is divided into several intervals for the latent score categories. The intervals for both ends of the score categories are open intervals. A border to divide the continuum is assumed at the midpoint between the adjacent latent score categories. Thus the latent score category has a certain range of length on the latent continuum. The border, the midpoint between the adjacent latent score categories, is defined as the point representing score category parameters. We assume the latent continuum scale values for score categories are normally distributed according to category parameters. Figure 1 illustrates the borders and the scale value distribution on the latent continuum.

Figure 1. Psychological distance method.
The latent continuum scale values for score categories were defined as the expectation of the interval on the latent continuum. For example, the expectation of the interval between \(-\infty\) and the first category parameter \((b_1)\) is defined as the latent continuum scale value of the first category \((SV_1)\). The equation to calculate the latent continuum scale values of score categories is:

\[
SV_p = \int_{b_{p-1}}^{b_p} x \times \frac{f(x)}{\int_{b_{p-1}}^{b_p} f(x) dx} \, dx = \frac{f(b_{p-1}) - f(b_p)}{\int_{b_{p-1}}^{b_p} f(x) dx}
\]  

(1)

In this equation, \(b_p\) represents the category parameters for score category \(p\) of the GPCM. For calculating the latent continuum scale value for the first category, \(b_0\) is \(-\infty\) and \(f(b_0)=0\). For calculating latent continuum scale value for the last, \(m\)th, category, \(b_m\) is \(+\infty\) and \(f(b_m)=0\). Then, the latent score distance is the differences of the latent continuum scale value between two adjacent score categories.

**Methods**

The 54 TE items examined in this study were English language arts (ELA) items that appeared on summative assessments for Grades 3 to high school (i.e., Grades 9-12). About 10,000 students took each item. These items had maximum scores of either 1, 2 or 3. The number of possible scores for one item ranged from 3 to 9. The latent continuum scale values of these items were calculated using the psychological distance method. Then, we compare the observed and the latent score distances by comparing the latent continuum scale values and observed scores on the same scale. If the differences between the observed and latent score distances are larger than one quarter of the observed score distance, the items will be identified as ‘non-equal latent score distance items’.

Next, the effect of the ‘non-equal latent score distance items’ was studied by comparing the observed score distributions. The two observed score distributions being compared are: 1) the original observed score distribution which includes the ‘non-equal latent score distance items,’ and 2) the new observed score distribution which includes the same items with non-equal observed score distances reflecting the non-equal latent score distances. If the difference between the two observed score distributions results in students being classified into different performance categories, the items and all their correct answers will be examined carefully by content experts.
The content experts for this study were individuals with extensive experience related to ELA assessments (e.g., ELA item writing, ELA item review, experience teaching ELA courses). We presented the reasons for the unequal latent score distances provided by the experts and the suggestions on changing the scoring rules, the correct answers, or the item stems from the experts.

**Results**

Among 42 items, 21 items were identified as the ‘non-equal latent score distance items’. The following four figures shows the differences between the latent and observed score distances of three ‘non-equal latent score distance items’ and one item with equal latent and observed score distances.

*Figure 2. Grade 3 TE item with unequal latent score distance*
Figure 3. Grade 5 TE item with unequal latent score distance

Figure 4. High school TE item with unequal latent score distance
Figure 5. Grade 6 TE item with equal latent score distance

The three ‘non-equal latent score distance items’ all present the differences between the latent and observed score differences. On the Grade 3 item, there are eight score categories, and some latent score categories are very close. However, for the Grade 6 item with equal latent and observed score differences, we can see that the latent and observed score categories are very similar.

Grade 5 and high school forms were chosen to compare the original and new observed score distributions. The Grade 5 test form had three TE items, two of which were ‘non-equal latent score distance items’. For these two items, we changed the scoring from 0 for 0 correct answers, 1 for 1 correct answer, and 2 for 2 correct answers to 0 for 0 correct answers, 0.5 for 1 correct answer, and 2 for 2 correct answers. Figure 6 displays the new and original observed score distribution.
Grade 5 Results

For Grade 5, the differences between the original and new observed score distribution is small since only two items’ scores were changed. But the differences appear at the observed score range with highest probability. The cut score of Grade 5 for meeting the standard is 41.5. Suppose there are 10,000 students who took this test. Based on the original observed distribution, 6063 students met the cut score. Based on the new observed distribution, 5984 students met the cut score. The new score reflects the latent score distances; with the new score, fewer students met the cut score.

Grade 5 item review. Content experts reviewed the two Grade 5 TE items that displayed different original and new score distances. The first Grade 5 TE item included four sentences. Two of the sentences in the item had one word omitted; students had to choose the correct spelling of the words from drop-down boxes within the sentences (an example of this TE item type is included in Figure 7). On this item, the content experts concluded that students would have much more difficulty identifying the correct
spelling of one omitted word (the word *louder*) in the item compared to the second omitted word (*males*) in the item. Item response data supports this conclusion. Students were nearly four times as likely to identify the correct spelling *males* as the correct spelling of *louder*.

Jane needs help correcting her report about the planet Jupiter. Read the sentences from her report and correct the **two** spelling errors in the report.

Jupiter is the largest planet in our Solar System. Jupiter has a huge red spot which is **actually** a giant storm. This storm is so large that three Earths could fit inside of it. Jupiter is visible the night sky because it reflects light from the sun.

*Figure 7. Sample Grade 5 technology-enhanced ELA item. Students choose the correct spelling of each word from the drop-down menus in the text.*

**High School Results**

The high school form included five TE items; three of which were ‘non-equal latent score distance items’. Table 1 presents the comparison of the original scoring rules and new scoring rules.

<table>
<thead>
<tr>
<th>Table 1</th>
<th><em>High school original and new scoring rule comparison</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td></td>
</tr>
<tr>
<td>Scoring Rule</td>
<td>Score 1</td>
</tr>
<tr>
<td>Original</td>
<td>0</td>
</tr>
<tr>
<td>New</td>
<td>0</td>
</tr>
<tr>
<td>Item 2</td>
<td></td>
</tr>
<tr>
<td>Scoring Rule</td>
<td>Score 1</td>
</tr>
<tr>
<td>Original</td>
<td>0</td>
</tr>
<tr>
<td>New</td>
<td>0</td>
</tr>
<tr>
<td>Item 3</td>
<td></td>
</tr>
<tr>
<td>Scoring Rule</td>
<td>Score 1</td>
</tr>
<tr>
<td>Original</td>
<td>0</td>
</tr>
<tr>
<td>New</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 8 displays the new and original observed score cumulative distribution. The reason for plotting the cumulative distribution instead of density distribution is that there is a large number of possible observed scores and some observed scores have very low probability.

![Graph showing cumulative distribution](image)

*Figure 8. High school observed score cumulative distribution.*

For high school, the differences between the original and new observed score distribution is small since only three items’ scores were changed. The high school cut score for meeting the standard is 45.5.

Suppose 10,000 students took this test. Based on the original observed distribution, 4,340 students would meet the standard. Based on the new observed distribution, 4,246 students would meet the standard. The new score reflects the latent score distances; with the new score, fewer students met the cut score.

**High school item review.** Item 2 included a paragraph of three sentences. Each of the sentences had one word omitted; students had to choose the word/phrase that best fit the purpose and audience of the paragraph. This TE item was similar in structure to the example in Figure 7; the item contained three drop-down menus in which students chose the correct word. Content experts noted that the second drop-down menu in the item contained word choices that were considerably more difficult than the first and third drop-down menu choices. Indeed, of all the possible answer option patterns, a majority of the
students answered the first and third drop-down menu options correctly, while answering the second drop-down menu item incorrectly. This may explain why, according to the “new” scoring rule in Table 1, students only received 1.33 on the item (rather than 2) for answering two of the three vocabulary question of the TE item correctly.

Item 3 was similar to the TE item described above. This item contained four sentences, three of which had drop-down boxes in which students would choose the best word/phrase. Content experts noted that, on the third drop-down menu, students would be able to quickly rule out three of the five distracters, thus increasing their odds of answering the item correctly. The distracters on the third drop-down menu were not as plausible as distracters on the first and third drop-down menu. Indeed, a majority of students answered the third drop-down menu item correctly while incorrectly answering either the first or second drop-down item. This may help explain why, according to the “new” scoring rule in Table 1, students only received 0.667 credit for answering one of the vocabulary questions correctly on this TE test item.

**Discussion**

This study revealed that, when the scoring rule ensuring the equal observed score distance was used, half of the TE items had some differences between the observed and latent score categories. Items with different observed and latent score categories have an effect on observed score distribution and student classification.

As evidenced by content experts’ review of Grade 5 and high school items, item content and item answer choices influenced the scoring rules. In this study, the original and new observed score distributions were compared for Grade 5 and high school items. For Grade 5, the new score rules indicated that students would only receive 0.5 for 1 correct answer, rather than 1 for 1 correct answer. Content experts’ review of a Grade 5 TE item revealed that one component of the item was much easier to answer than the second component of the item. This helps explain why scoring for answering only one component of the TE item was reduced under the new scoring rules.

At the high school level, content analyses of two of the TE items with different ‘old’ and ‘new’ scoring rules revealed similar results. In these two high school items, the components of the high school
TE items were not equal in terms of difficulty. In Item 2, one component was much more difficult for students to answer correctly. In Item 3, one component was much easier than the other two components because students could quickly rule out two of the five distracters.

These results are helpful for understanding TE items and their scoring rules. From these results, we suggest that tests can use multiple scoring rules for the TE items; the selection of the scoring rules depends on the content and correct answers of the items. Future research should further investigate TE item scoring rules and the effects that these rules have on student outcomes. Specifically, future research should continue to identify non-equal latent score distance items, qualitatively examine the content of those items, and use this information to make more informed decisions about the appropriateness of TE item scoring rules.

References
