Determining item position effects in a computer-based test

Gary Skaggs
Virginia Tech,
145 Dixon Drive,
Hardy VA 24101, USA
Email: gskaggs@vt.edu

Abstract: If items become more or less difficult or discriminating depending on their position within a test form, and if different examinees respond to the same items in different positions, the fairness of test scoring for examinees is undermined. Such context effects have been investigated in the past with mixed results. This study investigates whether item position effects are present in a computer-administered certification test in which items are presented in random order. The results showed few consistent significant position effects for either item difficulty or discrimination for the entire test. However, there was a tendency for items to be slightly more difficult when they were administered in the last five positions at the end of each test form. For individual items, there was no obvious position effect for either item difficulty or discrimination. For practitioners working in computer-based testing programs, the findings support the use of random item ordering, but it is still recommended that each testing program conduct an item position investigation for items.

Keywords: item position effects; random item ordering; computer-based testing.


Biographical notes: Gary Skaggs is an Associate Professor in the Educational Research and Evaluation Program at Virginia Tech. He teaches courses in introductory and intermediate quantitative research methods, measurement theory, and item response theory. His recent research focuses on test equating, standard setting and cognitive diagnostic models.

1 Introduction

Large-scale testing has depended to a large extent on the assumption that the statistical properties of test items are stable quantities. That is, the difficulty and discrimination of items are presumed to be uninfluenced by the context in which items appear on a test, particularly the order of their administration to examinees. This assertion is made regardless of whether the scaling and scoring are based on classical observed score methods or on item-centred latent trait models. As a result, research into determining whether contextual effects exist due to item position has been an ongoing area of study for several decades. In particular, there have been two major concerns:
1. Whether items appearing near the end of a test become more or less difficult due to the presence of fatigue or practice effects, respectively.

2. Whether items appearing near the beginning of a test are more difficult than when they appear later in the test due to a respondent’s ‘warming up’ effect.

Concerns about item position effects have centred around their potential impact on test scaling and scoring in several contexts. First, items that are field tested may appear in a different position and/or a different order on the operational test form. Determining the psychometric properties of a test from field test data could therefore be compromised by item position effects. Second, items measuring specific content that are used to equate test forms may appear in markedly different positions in their respective forms. Or, if the common item design is used, common items may appear in different positions on the two forms to be equated. Third, item properties could vary depending on which items precede it and which items follow it. Most latent trait models assume local item independence, but this assumption could be challenged in tests in which items are associated with passages. Finally, educational achievement test batteries are often vertically scaled in order to provide a developmental score scale. In that situation, the common items between levels may consist of some of the most difficult items at one test level and may be among the easiest items at the next level. As such, item positions may vary systematically, for example, with common items at the lower level appearing near the end of the test form and those same items appearing near the beginning of the higher level. In all of the above situations, the presence of item position effects constitutes a threat to the validity and fairness of test scores.

A great deal of research has dealt with determining item position effects on intact test forms. Position effects for computer-based tests (CBTs) have only recently received attention. CBT administration offers several potential advantages over paper-and-pencil administration, including increased test security and immediate test scoring. CBTs can be administered adaptively, using latent trait methodology, or as an intact test form. When an intact test form is administered via computer, a popular test administration practice in recent years is to administer the items in random order, which is the focus of this study. In testing centres, random ordering of items is intended to reduce cheating. However, it is also assumed that the order of item administration has no effect on the properties of the test items and as a result no impact on the score an examinee receives. That is, if there are differences for some items, it is assumed that there is a compensatory effect, items that become more difficult are balanced by items that become less difficult. If there are any systematic effects, such as items becoming more difficult near the end of the test, these would be expected to be the same for all examinees.

2. Background

Leary and Dorans (1985) summarised research carried out before the mid-1980s. These authors examined a number of studies in terms of type of context effect, test content, examinee characteristics, and power versus speeded testing. Much of this research was experimental in nature, that is, item orderings were manipulated in specific ways within an experimental design. These studies often produced conflicting findings, but Leary and Dorans were able to draw several conclusions. First, under speeded conditions, arranging
items from difficult to easy produced lower test scores than from arranging them from easy to difficult. Second, under power conditions, random ordering of items produced no consistent effect at the test score or item levels. Third, aptitude tests were more likely to be subject to position effects than achievement tests, particularly for items based on reading passages. Overall, Leary and Dorans noted the presence of item position effects in some studies, but concluded that these effects were not large enough to invalidate either classical or IRT scaling and scoring.

More recently, several investigators have examined item position effects for specific testing programs, including NAEP (Zwick, 1991), PISA (Adams and Wu, 2000), ACT (Brennan, 1992; Harris, 1991; Kolen and Harris, 1990; Pommerich and Harris, 2003), SAT (Eignor and Stocking, 1986), TOEFL (Way et al., 1992) and state assessment programs (Meyers et al., 2009). Much of this research explored differences in item difficulty and discrimination for items appearing in different positions on alternate forms or on pretests versus operational tests. This research has generally found a greater incidence of position effects than suggested by the studies reviewed by Leary and Dorans (1985). These tests were all essentially power or near-power tests, and, in many cases, items administered late in an exam appeared to be more difficult than when administered earlier in the test. These observed effects indicated the need for any large-scale testing program to investigate whether position effects exist for its items.

Some recent research has attempted to estimate systematic effects from item position. Meyers et al. (2009) regressed the change in Rasch item difficulty on item position change, test content, and time between item administrations for reading and math tests from a state assessment program. They found that, when items were arranged from easy to difficult based on their field test item difficulties, difficult items became more difficult when placed near the end of a test, and easy items became easier when placed near the beginning of a test. As a result, high ability examinees were advantaged and low ability examinees were disadvantaged when operational forms were equated.

Alternatively, several researchers have explored modelling item position effects through Fisher’s (1973) linear logistic test model (LLTM). An advantage of this approach is that specific kinds of position effects can be modelled, such as fatigue, content learning, speededness, and ‘warming-up’ effects, and there is a straightforward way to test the significance of these effects (Kubinger, 2008). Practical applications of the LLTM have shown mixed results for the presence of significant position effects. A small but significant fatigue effect was found for a fourth grade mathematics assessment for items presented in random order within print-based test booklets (Hohensinn et al., 2009). However, position effects were not found for a test of logical reasoning ability (Hahne, 2008). Furthermore, for this methodology, issues of sample size, type I error rate, and statistical power remain to be investigated.

Some recent research has begun to look at item position effects for computer-administered tests. Bowles et al. (2008) investigated these effects for a computer adaptive test (CAT) format of a licensure examination. Their study used the multifaceted Rasch model to create an item position parameter and then estimated this parameter with item response data. They observed two consistent position effects. First, the initial few items administered contained a significant position effect, making these items more difficult than had they appeared later in the test. Second, there was a small effect of increasing difficulty the later items appeared in the exam, but this effect was too small to have any practical impact on ability estimates.
The purpose of this study is to determine if item position effects exist for a particular method of test administration in a CBT, namely, in which items from an intact test form are administered in random order, a practice offered to testing programs by organisations that offer computer-based testing. To the best of the author’s knowledge, a recent study by Li et al. (2012) is the only published study to date that has investigated this form of test administration. Their item response data came from a medical licensure exam, and the authors used several Rasch models to estimate position effects within intact item blocks of 50 items. Unlike, the Bowles et al. (2008) study, no significant position effects were found with this method of test administration.

The current study is similar to the Li et al. (2012) study in that it focused on items administered in random order, for a high-stakes test, with sufficient testing time to be considered a power test. In that sense, this study is a replication of the Li et al. study. However, it differs in two important respects. First, the test studied here is much longer and therefore could provide a sterner test of position effects. Second, this test was not developed according to a latent trait model, and therefore a different method, using classical item indices, was used to detect position effects. Finally, this study examined position effects for item discrimination in addition to item difficulty. Since it is not an item parameter in Rasch models, discrimination has not been typically studied in research using a Rasch model to investigate item position effects.

3 Method

3.1 Instrument and data

The data for this study came from one of four separately scored tests of a certification examination program for knowledge of a specific type of accounting. There were two forms for each exam, both of which were used here. Each form consisted of 80 multiple choice items with an internal anchor set of 29 items that the two forms had in common. The test forms were administered via computer at commercial testing centres located in major cities in the USA. For each form, items were administered in random order, but the order of the answer choices was not altered. A time limit was enforced, but the allotted time was considered to be sufficient for most examinees to complete the entire exam.

Cases for this study were limited to examinees who were tested for the first time and who completed enough items to receive a reportable test score. First-time examinees were randomly assigned to one of the two forms at the testing centres. The data used in this study were collected during the sponsoring organisation’s 2008 testing cycle. For the first form, which I call form A, there were 841 examinees, and there were 928 examinees who took form B. These sample sizes are typical of the number of examinees taking the exams each year.

Descriptive statistics for both forms are shown in Table 1. The raw score means and standard deviations of the anchor items from each form were very similar, indicating that the samples taking each form were likewise very close in their accounting knowledge. However, the overall level of difficulty of the two forms differed by about a fourth of a standard deviation, with form A being the more difficult form.
This exam was developed and scored using classical test theory methods. As a result, the passing score is a specific number correct score on form A and its equivalent observed score on form B using linear equating. Much previous research has used a latent trait model, usually the Rasch model, to estimate position effects. In this study, the item response data did not fit any of the most popular latent trait models particularly well. For example, this exam measured mastery of several content areas, and as a result, the data were clearly multidimensional. Finally, since the passing score was in terms of number of items answered correctly, there were very few missing item responses as examinees had an incentive to guess when in doubt of the correct answer.

### 3.2 Item position effects

The research previously described typically defined item position effects in terms of item difficulty. Earlier research viewed such position effects as a symptom of speededness, that is, as an increase in item difficulty due to lack of time or fatigue. More recent research using the Rasch model has viewed the effect as a measurement facet that can increase or decrease the difficulty of an item. Li et al. (2012) modelled two kinds of position effects. One was a fixed effect due to the position of the item that affects all items in the same way. Another was a random, or interaction, effect in which individual items could behave differently to a specific position. In this study, I used classical item indices to estimate both of these types of effects.

First, fixed item position effects were calculated for each possible item position based on both item difficulty and discrimination. For item difficulty, mean residual p-values (proportion correct) for each item position \(i\) were calculated across all items of the test as follows:

\[
RESID_i = \overline{p}_i - \overline{p}
\]

where \(\overline{p}_i\) is the mean p-value across all items for item position \(i\), and \(\overline{p}\) is the overall mean p-value for the entire test. Positive residual values indicated item positions for which all items are relatively easier than they appear overall. Negative values indicated item positions in which items were more difficult. A 95% confidence interval was formed around a zero residual value using the following formula:

\[
CI = \pm t_{0.025, df=n-1} * \sqrt{ \left( \frac{\overline{p}_i(1-\overline{p}_i)}{n} \right) }
\]

If item difficulty effects are minimal, approximately 95% of the residuals would be expected to occur within this confidence interval. Additionally, item standard deviations were calculated for each item position. Greater than average standard deviations would
Determining item position effects in a computer-based test

indicate item positions that produce a wider range of item difficulties, while the reverse would be true for smaller standard deviations. Finally, item biserial correlations between each item position and total raw score (corrected for spuriousness) were calculated. Larger biserials would indicate item positions in which items are more highly related to total test scores. The reverse would be true for smaller biserials. If there were no discrimination position effects, these biserials would roughly be the same across item positions.

Second, item effects for individual items were investigated in a different fashion. Every item appeared in every possible item position, from first to eightieth, but not in great enough frequency to provide a stable item analysis for each item in each position. As a result, two types of analyses were carried out. One, items were grouped into several levels based on item difficulty or item discrimination, and the above analyses were repeated for these distinct groups of items. Two, item positions were organised into eight intervals of ten positions, first to tenth, 11th to 20th, and so forth. Within these position intervals, each item appeared in each interval approximately a hundred times. Mean item p-values and point biserial correlations were calculated separately within each item position interval. To determine statistical significance for item difficulty, a 95% confidence interval was formed around each item’s overall p-value or point biserial, and then the p-values or point biserials within each position interval were examined to see if they fell into the item’s confidence interval. For point biserials, the confidence interval was created using Fisher’s r-to-z transformation.

4 Results

4.1 Effects for item positions

Figure 1 shows the mean residual p-values for each item position for both forms. The 95% confidence interval for each mean residual is superimposed on the graph. In each form, there was no obvious visual pattern of residuals across item positions. In each form, three item positions out of 80 had residuals outside the confidence interval. The fact that the percentage of significant residuals was within the nominal 5% type I error expectation and that there was no apparent pattern of significant residuals suggests that these significant residuals are indistinguishable from type I errors.

To further investigate any trends by item position, moving averages of the residuals were calculated from the data from Figure 1 and plotted. Each plotted data point is the average of the mean residual of each position and the mean residuals for the two adjacent item positions. These are shown in Figure 2 for each form. There was still a random-appearing pattern of residual averages across the zero line. However, in both forms, the trend line declined for the last five or six item positions, indicating an overall increasing difficulty of the items when they appeared at the end of the test. On the other hand, this trend was still within the confidence interval, suggesting that if an end-of-test position effect existed, it was very small in magnitude.

The standard deviations of item p-values across item positions are plotted in Figure 3. For both forms, the standard deviations are essentially the same regardless of item position. Finally, in Figure 4, biserial correlations are shown by item position. As is well known, item discrimination indices tend to be less stable than item difficulty indices, and
this instability can be seen in the jaggedness of the plot lines. However, as before, there is no obvious systematic position effect due to item discrimination.

**Figure 1** Residual mean p-values of item positions for (a) form A and (b) form B

![Figure 1](image1.png)

(a)

![Figure 1](image2.png)

(b)

**Figure 2** Residual moving mean p-values of item positions for (a) form A and (b) form B

![Figure 2](image3.png)

(a)
Determining item position effects in a computer-based test

Figure 2  Residual moving mean p-values of item positions for (a) form A and (b) form B

Figure 3  Standard deviations of item p-values by item position for (a) form A and (b) form B
4.2 Item position effects for individual items

Figure 5 shows mean item p-values across item position intervals for subgroups of items within specific difficulty intervals, based on p-values. For both forms, there was no apparent interaction between item difficulty and position interval. That is, for the items grouped by difficulty levels, there was no consistent position effect across forms for any group of items. For form A, there was a small trend of moderately difficult items becoming more difficult and difficult and easy items becoming less difficult near the end of the exam. However, this trend was not replicated in form B. In fact, nearly the opposite was true. Easy items became more difficult, and difficult items became less difficult in the last two or three item positions. These observed trends are therefore likely to be chance results.

Similarly, Figure 6 shows the results for item biserials for three subgroups of items based on their biserial values. As was the case in Figure 4 for all items combined, there is some instability across item positions, more so for the items with lower biserials, but there is no apparent systematic position effect on item discrimination.
Figure 5  Mean p-values of items by item difficulty for (a) form A and (b) form B

(a)

(b)

Figure 6  Biserials of item position levels for (a) form A and (b) form B

(a)

(b)
Position effects for item difficulty and discrimination were also examined for individual items. As described above, to achieve a stable difficulty estimate, mean p-values were calculated for each item within eight position intervals (1–10, 11–20, etc.). A 95% confidence interval was then constructed around each item’s overall p-value, and mean position interval p-values were compared to this interval. Eighty items, each with eight position intervals, meant that 640 significance tests were performed on each form. At a 0.05 level of significance, we would expect about 32 significant results by chance. Form A had 16 significant results while form B had 20 significant results, much less than the nominal type I error rate. Table 2 summarises the significant results for each form. It shows the number of items with significant results for each position interval. The plus sign column indicates the number of items in which the interval p-value was greater than the overall p-value, in other words, items that were less difficult in that interval. Likewise, the minus column indicates the number of items with interval p-values lower than the overall p-value, that is, were more difficult in that interval. There is no obvious pattern for the significant item difficulty results across position intervals. Moreover, an examination of these items according to overall difficulty, content area, and reading load required did not reveal any obvious reason why item position would affect difficulty in particular intervals.

Table 2  Number of Items with p-values outside 95% confidence interval

<table>
<thead>
<tr>
<th>Position interval</th>
<th>Form A</th>
<th>Form B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>1–10</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>11–20</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>21–30</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>31–40</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>41–50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>51–60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>61–70</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>71–80</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Four of the items with significant results in form A and eight of the items in form B were anchor items that also appeared in the other form. With two exceptions, these items appeared in approximately the same position on the other form. For no items were significant results in one form replicated in the other form. All of this evidence suggests that the significant findings for individual items are consistent with chance results.

Likewise, point biserial correlations for each item were calculated when they appeared in the eight position intervals. The point biserials in each interval were compared to the item’s overall point biserial by constructing a 95% confidence interval around the overall point biserial using Fisher’s r-to-Z transformation. Table 3, constructed in a manner similar to Table 2, shows the results for item discrimination interval effects. Nineteen position effects were significant on form A, while 12 were significant on form B. As it was for item difficulty, the proportion of significant point biserial intervals was considerably less than 0.05.
Determining item position effects in a computer-based test

Table 3  Number of items with point biserials outside 95% Fisher r-to-z confidence interval

<table>
<thead>
<tr>
<th>Position interval</th>
<th>Form A</th>
<th></th>
<th>Form B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>1–10</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11–20</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>21–30</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>31–40</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>41–50</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>51–60</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>61–70</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>71–80</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Interestingly, there were approximately twice as many significant positive point biserial intervals as negative point biserial intervals, suggesting that the Fisher transformation for some reason did not entirely eliminate the skewness in the sampling distribution of the correlations. In further examining the significant point biserial intervals, as with item difficulty, there was no apparent pattern of point biserial effects according to item content, difficulty level, or reading load. Six of the significant point biserial intervals on form A were from common items, but only one of these results were replicated on form B. Likewise, four of the significant point biserial intervals on form B came from common items, but only the one result was replicated. Again, these findings suggest that the pattern of significant results could easily be due to chance.

5 Discussion

Context effects due to the order of administration of items have been a concern for test developers for many years. Prior research on these effects has shown mixed results for print-based test forms with a fixed length and fixed item order, but little research has been conducted to date on fixed-length tests administered by computer. When item position has been experimentally altered in power tests in technical subjects such as mathematics, few consistent position effects have been demonstrated. Effects have sometimes been found in speeded and/or language-intensive tests. On the other hand, position effects have sometimes been found, usually at the beginning or end of tests. However, there has been no consistent practice or fatigue effect observed across testing programs, and so there are no specific guidelines for item ordering. For the test developer, these findings have generally meant that context effects must be investigated by each testing program.

For CBTs, administering items in random order could tend to mitigate position effects. If there are any such effects, they would be expected to be equal for all examinees. For example, if the last several items (no matter which items they are) tend to be more difficult, this difficulty effect would be expected to be the same for all examinees. As a result, the practice of random item ordering has become quite popular in recent years, especially for certification/licensure exams. On the other hand, systematic position effects would be a challenge to fair test scoring if certain types of items based on
format, content, or difficulty level exhibited either practice or fatigue effects, and they would impact individual examinees’ test scores and subsequently pass/fail decisions if those items happened to be placed near the beginning or end of the test, respectively. Such a situation would undermine the fairness of random item ordering.

Fortunately, for the practice of random item ordering for CBTs, little evidence was found in this study for item position effects, and so this method of administration would not be expected to undermine certification decisions. The findings of this study are mostly consistent with those from the recent Li et al. (2012) study. Items in the last few positions appeared to be slightly more difficult. Otherwise, no pattern of consistent item position effects was found. There was no overall trend in items on either form becoming easier or more difficult or more or less discriminating depending on where they were positioned in the test. One initial concern that motivated this study was the potential effect item position could have on equating, due to the fact that some anchor items appear in different positions. For all anchor items, there was no consistent position effect due to item difficulty. For only one of the 29 anchor items was there an item discrimination position effect that was the same on both forms. These findings imply that item positioning would likely have no impact on the equating function between the two forms (assuming that the data to carry out the equating comes from the random-order test administration).

Likewise, the percentage of items with difficulty and discrimination differences outside the 95% confidence interval for no difference according to their position was less than the nominal 5% rate. When the significant items were examined closely, there was no clear pattern in terms of reading load, difficulty, and/or content. When the significant findings occurred for an anchor item, the significance was replicated in the other form only once, for item discrimination. It is therefore likely that these findings for most if not all of the flagged items can be attributed to random error. However, there may have been subtle context effects for some items. When items are administered near the end of a CBT, most of the items have been exposed, and so examinees will have the benefit of familiarity with most of the test, a kind of practice effect. On the other hand, examinees will have seen relatively little of the entire test when answering the first items. This degree of familiarity with the rest of the items or lack thereof could have an impact on item characteristics. Such effects could be difficult to detect statistically. Content experts could examine a suspect item for any enhancing or confusing aspect from previously administered items. Alternatively, think aloud protocols could be obtained from examinees in a situation in which item order was manipulated. These protocols might reveal complex reasoning strategies that could affect item characteristics at the beginning as opposed to the end of tests.

Beyond validating the practice of administering items in random order, this administration method provides an opportunity to study position effects independently of specific items. That is, it permits the investigation of item position regardless of the particular items appearing in fixed positions on an intact form. One contrast between the findings of this study and the Li et al. (2012) study is the slightly increasing difficulty of items at the end of both test forms in this study. That effect was not found by Li et al. One difference is that the test forms in this were much longer, suggesting that test length is an important variable in predicting position effects and should be a consideration in future research.

For practitioners, the results of this study likely generalise to other computer-based testing programs of a similar nature that administer items in random order. It should be
Determining item position effects in a computer-based test

pointed out that this testing program was intended to be a power test with reasonable time limits for examinees. Although relatively lengthy for a single testing session in educational achievement settings, this test length, at 80 multiple choice items, is typical of many certification testing programs. The outcome was high stakes for examinees (achieving or not achieving certification), and as a result there were almost no omitted items. Additionally, the content of these items was technical in nature with a very modest reading load. The findings in this study should provide some degree of reassurance for similar testing programs. On the other hand, the best advice would be for each testing program to conduct its own study of item position effects.

Future research should focus on investigating item position effects for different types of CBTs: tests of different content (particularly in more language-intensive areas), different test lengths, and different degrees of speededness. If a fixed-order print-based version of the test exists, item characteristics from the two administration modes could be compared. In addition, item position effects should be studied for population subgroups, including low versus high ability examinees. One limitation of the current study is that there were insufficient sample sizes to investigate all possible item orderings. It is possible that larger sample sizes and/or some experimental manipulation of item order would be more sensitive to small position effects. A latent trait approach, such as the LLTM, might prove useful in that any consistent position effects could be incorporated into the scoring for individuals. Finally, more qualitative methods, such as think aloud protocols, might reveal complex differences in question-responding strategies. In summary, for CBTs, this study supports earlier research on the overall lack of context effects due to item ordering for power and technical subject areas for print-based tests.

References


