Summer 5-21-2014

Linear and Nonlinear Modeling of Item Position Effects

Chansuk Kang
University of Nebraska-Lincoln

LINEAR AND NONLINEAR MODELING OF ITEM POSITION EFFECTS

By

Chansuk Kang

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Arts

Major: Educational Psychology

Under the Supervision of Professor Anthony D. Albano

Lincoln, Nebraska

August 2014

LINEAR AND NONLINEAR MODELING OF ITEM POSITION EFFECTS

Chansuk Kang, M.A.

University of Nebraska, 2014

Adviser: Anthony D. Albano

Item parameter invariance is one of the properties of item response theory (IRT) that enables computerized adaptive testing (CAT) for test administration. The possible influence of item position on test performance is one of the severe threats to the property of item parameter invariance within IRT. This study examines how different representations of item position, i.e., using categorical, linear, and quadratic terms, can impact how the relationship between item position and item difficulty is expressed. An explanatory IRT model is formulated for estimating item position effects. The model is demonstrated using data from the Program for International Student Assessment (PISA) 2009 reading data for the U.S., wherein the same items appeared in four different positions across item clusters. Methods of choosing the best model to detect item position effects are discussed as well as preliminary item analysis for the estimation of item position effects.

Table of Contents

List of Tables

List of Figures

Chapter I: Introduction

Item parameter invariance is one of the properties of item response theory (IRT) that enables computerized adaptive testing (CAT) for test administration (Meyers et al., 2009). The influence of item position on test performance is one of the severe threats to the property of item parameter invariance within item response theory (Leary & Dorans, 1985). In an era of rapid proliferation of CAT, item position effects on item difficulty and test performance is an important issue to examine. Since the position effects may cause unintended impact on test performance, test developers do not want the item position effects to be present in their tests. In that sense, item position effects on test performance could negatively impact the validity of test score interpretations and uses.

Since the early work of Mollenkopf (1950), the unintended effects of item order on test performance have been studied thoroughly. Early studies of item order or context effects were mainly focused on their impact on item difficulty. These studies are roughly summarized as two types of effects: a fatigue effect and a learning effect. Fatigue effects occur when items become more difficult as they are located later in the test. Even though there could be many reasons for the decrease of the test performance, such as lower motivation or simply running out of time, studies tend to refer to decreases in performance as fatigue effects. On the other hand, a learning effect occurs when items become easier as they are located at the end of the test. The explanation of fatigue effects can also apply to the learning effects. However, it must be noted that previous studies were mainly conducted on paper and pencil tests.

The development of IRT models provided more opportunities for researchers to study item position effects with more advanced statistical techniques. IRT models were used to detect item position effects in the CAT format. Recent empirical studies have mainly examined item position effects at the item level. Debeer and Janssen (2013) studied item position effects in the framework of descriptive and explanatory item response models. Albano (2013) utilized a hierarchical generalized linear model (HGLM) framework for modeling item position effects. In that framework, as in explanatory/descriptive IRT, item, position, and person effects are estimated simultaneously. Li et al. (2012) analyzed the impact of item position on item difficulty and item intensity. For the ordering of items, Li et al. used random ordering, which means that all items have equal chance of assignment to a given position.

The purpose of this study is to extend previous work by comparing item position effects estimated at the item level (Albano, 2013) with other related approaches to estimating item position effects. Within multilevel modeling frameworks, item position effects can be estimated in a number of ways, e.g., using categorical, linear, and quadratic model terms. These different terms result in different descriptions of the relationship between item position and test performance at the item level. This study compares item position effects models from four recent studies (i.e., Albano, 2013; Debeer & Janssen, 2013; Li et al., 2012; Meyers et al., 2009) using reading data from the U.S. cohort of the Program for International Student Assessment (PISA) 2009. Debeer and Janssen (2013) analyzed item position effects within an IRT framework using the PISA 2006 data. This study applies explanatory IRT models to the PISA 2009 reading data for the U.S. to compare categorical, linear, and quadratic item position effects. The research questions of

this paper are:

1. How does item position impact test performance?

2. Which expression of item position, categorical, linear, or quadratic, functions best?

In the following section, this paper reviews previous studies on the effects of item position on test performance, including studies utilizing categorical, linear, and quadratic item position terms. These three different approaches to expressing item position are then applied within a series of models of the PISA 2009 reading data for the U.S. The remaining sections of the paper present and discuss the results of item analyses and model comparisons of the PISA 2009 data.

Chapter II: Literature Review

Studying Position Effects

Concerns about the effects of item order and context on test performance are not new to psychometricians (Leary & Dorans, 1985, p. 388). Leary and Dorans (1985) summarized earlier trends of studies on within-test context effects up to the mid-1980s starting with studies conducted by Mollenkopf in 1950. In the 1950s and early 1960s, researchers paid attention to the simple main effect relationship between item position and test performance. Studies during the late 1960s and 1970s mainly focused on interaction effects between item arrangement and test taker characteristics. During the 1980s, several studies published by Plake and her colleagues presented analyses of item

position effects using parameter estimates from IRT models (Plake, 1980; Plake et al., 1981; Plake et al., 1982). Although previous studies found the existence of context effects, Leary and Dorans (1985) summarized that "the evidence does not suggest that the effect of test-item or test-section rearrangement is so detrimental as to invalidate the test theory or practice that is dependent on the assumption of item parameter invariance" (p. 410).

Later studies of position effects tended to utilize an IRT framework. Eignor and Cook (1983) found fatigue effects, where item difficulty increased as items appeared later on the test, presumably because individuals experienced more fatigue and less focus and motivation with time. The reverse case is presented as learning effect, as posited by Davis and Ferdous (2005) given their studies with math and reading tests. Results of these two studies show that when item position changes, the difficulty of the item often changes. Several other studies also found significant effects for item position in terms of equating results (Brennan, 1992; Kolen & Harris, 1990; Zwick, 1991). Studies have also claimed that the impact of item position on test performance could be influenced by other factors such as the number of position changes, the direction of item change, and the ability of the examinees (Way et al., 1992; Wise et al., 1989).

Recent studies have paid attention to methodological issues in the modeling of item position effects. Regarding levels of analysis, the item position effects can be analyzed at both the item level and test level. Traditionally, the detection of the item position effects has been based on a two-step model (Yen, 1980; Whitely & Dawis, 1976). Once item difficulties are estimated for each test form, the item difficulties between the two test forms are examined by position.

IRT-based studies focus on the item level analysis of the item position effects. For example, Debeer and Janssen (2013) paid attention to a one-step model, which is an item-level modeling of item position effects, and noted that the one-step model has several advantages of modeling item position effects compared to the traditional two-step IRT procedures: it is applicable to more complex designs than the two-step approach; it has more explanatory power for the found effects than the two-step modeling; and it is easier to generate new test forms of similar conditions than in the two-step procedure (p. 168-169). The IRT framework is useful for examining item position because position effects can be examined as properties of items (Debeer & Janssen, 2013).

IRT models also can be formulated as multilevel item-response models (Kamata, 2001). Albano (2013) conducted a study to analyze item position effects with hierarchical generalized linear model (HGLM). He adopted the HGLM framework due to its flexibility of handling complex data structures. The HGLM also can use partial information pooling to the estimation of the position effects.

Modeling Position Effects

Recent studies show three different ways of modeling item position effects:

1. If position effects do not have a linear change, *categorical* variables, i.e., indicator variables for position, may be most appropriate (e.g., Alexandrowicz & Matschinger, 2008; Debeer & Janssen, 2013; Pomplun & Ritchie, 2004).

2. If position effects have a linear change, *continuous* variables may best estimate the position effects (Albano, 2013; Davey & Lee, 2011).

3. If position effects have a curvilinear change, polynomial terms for continuous variables, such as *quadratic* form, can also improve the estimation of position effects (Meyers et al., 2009).

First, in order to estimate non-linear trend, position effects could be modeled using categorical variables. Pomplun and Ritchie (2004) and Alexandrowicz and Matschinger (2008) estimated position effect using categorical variables. In Pomplun and Ritchie (2004), each position effect is applied to a specific item. However, in Alexandrowicz and Matschinger (2008), each position effect is applied to all items. Therefore, the total number of position effects in Pomplun and Ritchie (2004) is much greater than the total number of position effects in Alexandrowicz and Matschinger (2008). Debeer and Janssen (2013) treated item position effects as a person effect using a categorical variable. As a part of the longitudinal model, it focuses on the growth of test performance for each person across time. Since the PISA 2006 data that were analyzed in the Debeer and Janssen (2013) has only four different positions of blocks, the estimation of categorical change in their model was not complex.

Second, in order to estimate linear trend, position effects could be modeled as continuous variable. According to Albano (2013), "if position effects are estimated as a continuous variable, the interaction effects of item-position become a slope and it enables researchers to estimate the average linear change of getting correct answers across all positions" (p. 413). Compared to other models, the linear position effects model is relatively parsimonious and requires fewer parameters. Therefore, the linear model can provide a simple and efficient summary of item position effects on item difficulty. Albano (2013) treated item position effect as item effect. However, he also cautioned

about the potential of missing or misrepresenting important trends in the data with the application of the linear/continuous model.

Third, in order to estimate curvilinear trends, position effects could be modeled using quadratic terms. For example, in cases where the item difficulty increases in the beginning, remains stable in the middle then decreases at the end, a linear effect would misestimate the effect. In such cases, the polynomial model can provide better estimation than the linear model for the impact of item position on item difficulty. Meyers et al. (2009) identified a quadratic relationship between item position change and item difficulty (note that this was a two-step analysis). The item position effects on difficulty were estimated using a quadratic model to capture the curvilinear relationship between the two. In their study, difficult items were located in the middle of the test, and easy items were located in the beginning and end of the test. After controlling for other components affecting item difficulty, the study assumed that the relationship between item position change and item difficulty change followed the quadratic regression model**.**

In contrast to the majority of previous research, Li et al. (2012) found that item position effects were negligible. Three different types of IRT models and three different types of response time models were used. Position effects were treated as random item effects rather than as fixed effects, as in Albano (2013).

Previous research shows that in the prediction of certain trends, some models are better than others; each approach has its advantages and disadvantages. Selection of an approach should be based on data visualization and model fit to the data. The following section presents three different models for estimating item position. The data and analyses used to compare these models will then be explained.

Chapter III: Methods

Model Specification

The model specification presented below is closely related to research questions of this study. These questions focus on how item-level performance varies due to ability, item difficulty, and item position. All three of these factors may impact item responses. Person effects and item effects were considered as random effects because this study is not interested in specific fixed estimates of each. Since there are multiple random effects in this study (person and item), multilevel modeling is more appropriate for the estimation of item position effects. Since the logic behind the multilevel model is relatively similar to that of the IRT model, this study refers to model specification examples used in Debeer and Janssen (2013) whose study analyzed item position effects using the IRT framework.

In logit form, the base model is expressed as:

$$logit\,[Y_{pik} = 1] = \theta_p - \beta_{ik}.$$

The base model (M0) is a Rasch model. In this base model, the effect of item difficulty (β_{ik}) and ability of persons (θ_p) are estimated at the same time. This model represents the probability of a correct answer for person p (p =1, 2 …, P) to item i (i = 1, 2, …, I) in position k (k= 1, 2, …, K) as a function of item difficulty (β_{ik}) and ability of persons (θ_p) for item i at position k. There are no position effects in this base model.

The next model includes an additional main effect of position. It decomposes β_{ik} from M0 into two components:

$$logit\,[Y_{pik} = 1] = \theta_p - (\beta_i + \delta_k^{\beta}).$$

In the categorical main (fixed) position effects model (M1A), β_i represents the difficulty of item *i* in the reference position and δ_k^{β} stands for the position parameter (1, 2, …, *N*). Note that there will be a position effect δ_k^{β} for each position *k*. In this model, main position effects were applied to all items because the categorical position parameter is only position dependent.

The main position effects model can also be assumed as a linear position effect on difficulty:

$$logit\,[Y_{pik} = 1] = \theta_p - [\beta_i + \gamma_1(k - 1)].$$

In the linear main position effects model (M1B), γ_1 represents the linear weight of the position and β_i represents the difficulty of item *i* in position 1. In this model, main position effects were applied to all items because the linear position parameter ($\gamma_1(k - 1)$) is only position dependent. The main position effects stand for the overall change of the mean in performance across all items. Therefore, there is only a single average slope for all items.

The next model includes additional interaction effects of item-position:

$$logit\,[Y_{pik} = 1] = \theta_p - (\beta_i + \delta_k^{\beta} + \delta_{ik}^{\beta}).$$

In the categorical interaction (random) position effects model (M2A) above, δ_{ik}^{β} represents a position parameter of interaction effect and it is item dependent. It models the difference of item difficulty of position *k* from reference position. The δ_k^{β} represents a position parameter of main effect.

The interaction position effects model can also be assumed as a linear position effect on difficulty:

$$logit\,[Y_{pik} = 1] = \theta_p - [\beta_i + \gamma_1(k-1) + \gamma_{1i}(k-1)].$$

In the linear interaction position effects model (M2B) above, γ_{1i} represents the linear weight of the position parameter that is item dependent (interaction effect). Therefore, in this model each item has its own slope. This model also includes the linear position parameter of the main effect ($\gamma_1(k - 1)$).

If the linear main position effects model, M1B, does not accurately represent nonlinear relationship between item difficulty and position, application of the quadratic main position effects model, M3, may be necessary:

$$logit\,[Y_{pik} = 1] = \theta_p - [\beta_i + \gamma_1(k-1) + \gamma_2\,(k-1)^2].$$

In this model, γ_2 represents the quadratic weight of the main position effects and it is applied to all items. Similar to M1B, the quadratic main position effects, γ_2, stands for the overall change of mean in performance across all items. Therefore, there is only a single curve-shape slope for all items. The model, M3, is like an extension of the previous model of M1B. This model also includes the linear position parameter of the main effect ($\gamma_1(k - 1)$).

For the estimation of quadratic relationships between position effects and item difficulty, the next model includes additional interaction effects of item-quadratic position:

$$logit\,[Y_{pik} = 1] = \theta_p - [\beta_i + \gamma_1(k-1) + \gamma_{1i}(k-1) + \gamma_2(k-1)^2 + \gamma_{2i}\,(k-1)^2].$$

In the quadratic interaction position effects model (M4), γ_{2i} represents the quadratic weight of the position parameter that is item dependent (interaction effect of quadratic

position). Therefore, in the interaction position effects model, each item has its own curve-shape slope. This model also includes (1) the linear position parameter of the main effect ($\gamma_1(k$ - 1)); (2) the linear position parameter of the interaction effect ($\gamma_{1i}(k-1)$); (3) the quadratic position parameter of the main effect ($\gamma_2(k-1)^2$).

Model Fit and Comparison

Comparison of the seven models is not simple. Some models are nested between each other and some are not. Comparison of the seven models was done using of three sets of nested model comparisons and one set of non-nested model comparison. For the comparisons of the nested models, Chi-square likelihood ratio tests were conducted. The AIC and BIC were also used to determine model fit. The order of the nested model comparisons was:

(a) M0 vs. M1A; M1A vs. M2A.

(b) M0 vs. M1B; M1B vs. M3; M3 vs. M4.

(c) M0 vs. M1B; M1B vs. M2B; M2B vs. M4.

AIC and BIC were used for the comparisons of the non-nested models. The order of the non-nested model comparisons was:

(d) M1A vs. M1B; M2A vs. M2B.

Figure 1 shows a visual representation of four sets of model comparisons.

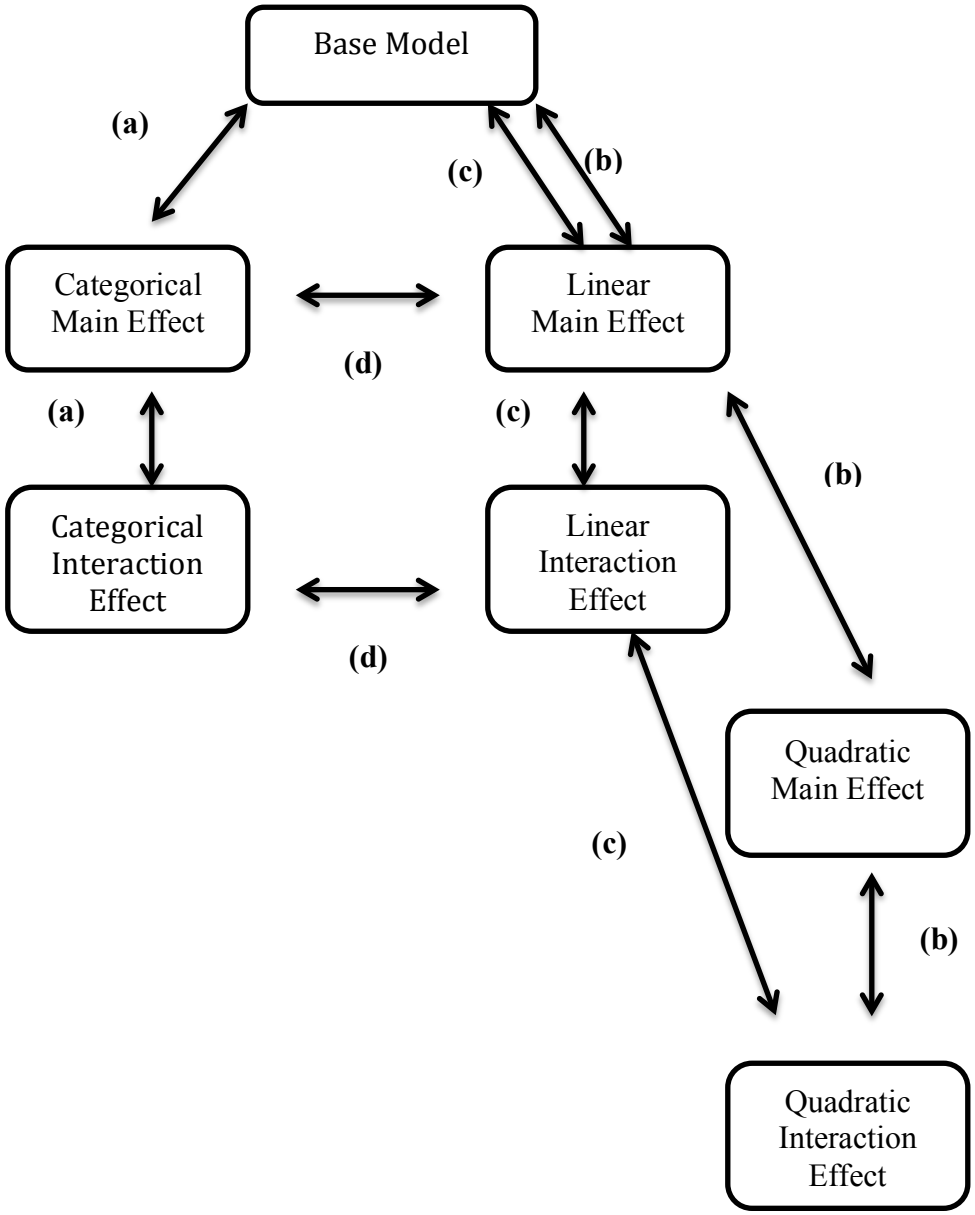

Figure 1. A visual representation of four sets of model comparisons. Vertical comparisons, such as (a), (b), and (c), are comparisons between nested models. Horizontal comparisons, such as (d), are comparisons between non-nested models.

Data

The models presented above were compared using reading data from the US cohort of PISA 2009. PISA is an international test used to measure reading, math, and science literacy of 15-year-olds globally (OECD, 2009). Students from seventy-five

countries, 34 OECD member countries, and 41 partner countries, participated in the 2009 test. The PISA 2009 test items were spread across thirteen clusters: seven reading clusters (R1-R7), three science clusters (S1-S3), and three mathematics clusters (M1-M3). Each cluster was comprised of twelve to sixteen test items and administered in thirty minutes of test time. For the test administration, a balanced incomplete blocking design was used. Each item cluster was located in each of the four possible positions within a booklet once and students were randomly assigned to one of the thirteen test booklets. Since item order within each item cluster was fixed, there were only differences in the position of the item clusters (from position 1 to position 4). Since there are only four positions in the PISA 2009, modeling position effect is not relatively complex compared to other cases with many position effects (e.g., Albano, 2013). This is the reason why the application of categorical variable for the item position effect is possible in this study. Other than categorical change, this study also applies continuous/linear change and continuous/quadratic change. More information about the test design of the PISA 2009 can be found in the PISA 2009 Technical Report (OECD, 2012).

The cluster rotation design of the PISA 2009 data is vulnerable to context effects because each student randomly assigned to a booklet does not take the same item clusters as other students. For example, it is possible for student A to take a booklet of four reading clusters and student B to take a booklet of reading, math, and science clusters. If this is the case, it could be possible for the two students to take completely different test contents. This context effect can also have an impact on test scores above and beyond any item position effects. Because of their complexity, context effects were ignored in this study. Clusters of mathematics and science were treated as missing and were removed

from the data set and were not modeled. Table 1 shows how the reading clusters were distributed across booklets, with empty cells where the math and science clusters were. All seven clusters of reading appeared once in each position (from position 1 to position 4).

Table 1
Revised Cluster Rotation Design Used to Form Test Booklets

Booklet ID	Position 1	Position 2	Position 3	Position 4
B1		R1	R3A	
B2	R1		R4A	R7
B3		R3A		
B4	R3A	R4A		R2
B5	R4A		R5	
B6	R5	R6	R7	R3A
B7	R6			R4A
B8	R2			R6
B9			R6	R1
B10		R5		
B11		R7	R2	
B12	R7			
B13		R2	R1	R5

Preliminary Analysis

Descriptive statistics. In this study, the total 101 reading items in the PISA 2009 data for USA were grouped in seven clusters. A total sample of 5231 students in the U.S. was randomly assigned to take the reading test and completed it. For the handling of omitted items and not-reached items, this study adopted the PISA scoring, which treats these two types of responses as missing. Table 2 shows descriptive statistics of total scores for each reading cluster. Among seven item clusters, cluster 7 (R7) has the highest mean. Cluster 3 (R3A) has the highest standard deviation of the seven clusters; therefore,

items in R3A have greater variability in difficulty than items in other reading item clusters.

Table 2
Descriptive Statistics of Total Scores for Each Item Cluster

Cluster	Mean	Median	Min.	Max.	S.D.	N of item
1 (R1)	7.71	8.00	0	12.00	2.84	12
2 (R2)	8.34	9.00	0	14.00	3.73	14
3 (R3A)	8.10	8.00	0	15.00	3.84	15
4 (R4A)	9.21	9.00	0	16.00	3.82	16
5 (R5)	7.78	8.00	0	15.00	3.76	15
6 (R6)	8.97	9.00	0	15.00	3.43	15
7 (R7)	9.57	10.00	0	14.00	2.96	14

Item analysis. In general, item analysis is used for the improvement of test items and for the detection of bias or unfair items. Albano (2013) recommended examining proportion correct by item position as a preliminary analysis, prior to modeling position effects. As a part of the preliminary analysis, this study also conducted item analysis for each cluster. In the item analysis for each cluster, position effect of each item and reliability of each cluster were analyzed.

Cluster 1 (R1). R1 is comprised of twelve items. Table 3 shows results of reliability analysis of all twelve items in R1. *p*-value stands for the item difficulty, i.e., the proportion correct for each item. Higher *p*-values indicate easier items and lower *p*-values indicate more difficult items. The *p*-values by position in Table 3 show changes in item difficulty by position. In Table 3, the mean *p*-value of each position decreases from position 1 to position 3; however, it increases in position 4 (.69, .67, .60, .62). This means

that, in R1, items may be getting harder as positions move from 1 to 3; however, they are getting a little bit easier as positions move from 3 to 4. Item discrimination, in the form of corrected point-biserial correlations (CPB), indexes the relationship between individual items and the total test scores. The mean CPB of R1 is .43. Since the correlation is above the .30, overall CPB in R1 appears good. The alpha-if-item-deleted (AID) stands for change in coefficient alpha (internal consistency of all items) if an item is deleted. If AID of an item is higher than alpha, the item normally needs to be removed from the test. In R1, AID of all items are lower than alpha, therefore, it was not necessary to remove any of the items in R1. Alpha of .79 in R1 is close to the ideal minimum level of internal consistency of .80. Therefore, items in R1 have good internal consistency.

Cluster 2 (R2). R2 is comprised of fourteen items. Table 4 shows results of reliability analysis of all fourteen items in R2. In Table 4, as position moves from 1 to 4, the mean *p*-value of each position decreases except at position 3 (.64, .60, .62, .53). This means that, in R2, items are getting harder as position move from 1 to 3, except at the position 3. When position moves from 2 to 3, items are getting easier, however, when position moves from 3 to 4, items are getting harder again. The mean CPB of R2 is .48. Since it is above the .30, overall CPB in R2 are good. In R2, AID of most items, except two items (r104q02, r227q01), are lower than alpha. The AID of two items (r104q02, r227q01) are same as alpha. Alpha is .84 in R2, indicating good internal consistency.

Cluster 3 (R3A). R3A is comprised of fifteen items. Table 5 shows results of reliability analysis of all fifteen items in R3A. In Table 5, as position moves from 1 to 4, the mean *p*-value of each position continuously decreases (.59, .54, .53, .50). This means that, in R3A, items are getting harder as position moves from 1 to 4. The mean CPB of

R3A is .44. Since it is above the .30, overall CPB in R3A are in good shape. However, one item (r447q05) has a CPB of .25. In R3A, AID of four items (r414q11, r447q05, r452q03, r458q04) are same as alpha. AID of other eleven items are lower than alpha. Alpha is .82 in R3A, indicating good internal consistency.

Cluster 4 (R4A). R4A is comprised of sixteen items. Table 6 shows results of reliability analysis of all sixteen items in R4A. In Table 6, mean *p*-value of position 1 and 2 are same, however, when position moves from 2 to 4, the average *p*-value decreases (.62, .62, .55, .52). This means that, after position 2, items in R4A are getting harder as items move to later positions. The mean CPB of R4A is .44. Since it is above the .30, overall CPB in R4A are in good shape. No items in R4A have a CPB lower than .30. In R4A, AID of two items (r083q02, r101q01) are same as alpha. Alpha is .83 in R4A, indicating good internal consistency.

Cluster 5 (R5). R5 is comprised of fifteen items. Table 7 shows results of reliability analysis of all fifteen items in R5. In Table 7, as position moves from 1 to 4, mean *p*-value decreases progressively (.58, .54, .53, .42). This means that items in R5 are getting harder as items move to later positions. The mean CPB of R5 is .44. Since it is above the .30, all CPB in R5 are in good shape. In R5, AID for two items (r424q02t, r424q03) are same as alpha. Alpha is .82 in R5, indicating good internal consistency.

Cluster 6 (R6). R6 is comprised of fifteen items. Table 8 shows results of reliability analysis of all fifteen items in R6. In Table 8, as position moves from 1 to 4, mean *p*-value progressively decreases (.64, .62, .61, .53). This means that items in R6 are getting harder as items move to later positions. The mean CPB of R6 is .41. Since it is above the .30, most CPB in R6 are in good shape. However, CPB of one item (r412q06t)

is much lower than the threshold of .30. Alpha is .80 in R6, indicating good internal consistency.

Cluster 7 (R7). R7 is comprised of fourteen items. Table 9 shows results of reliability analysis of all fourteen items in R7. In Table 9, as position moves from 1 to 4, mean *p*-value continuously decreases (.73, .72, .68, .62). This means that items in R7 are getting harder as items move to later positions. The mean CPB of R7 is .45. Since it is above the .30, most CPB in R7 are in good shape. However, CPB of one item (r466q03t) is much lower than the threshold of .30. In R7, AID of one item (r466q03t) is higher than alpha. Alpha is .82 in R7, indicating good internal consistency.

Figure 2 shows proportion correct by position of all 101 items of the PISA 2009 reading test for the U.S. In the figure items are getting harder as items move to later positions. Overall, *p*-values in position 1 are the highest and *p*-values in position 4 are the lowest.

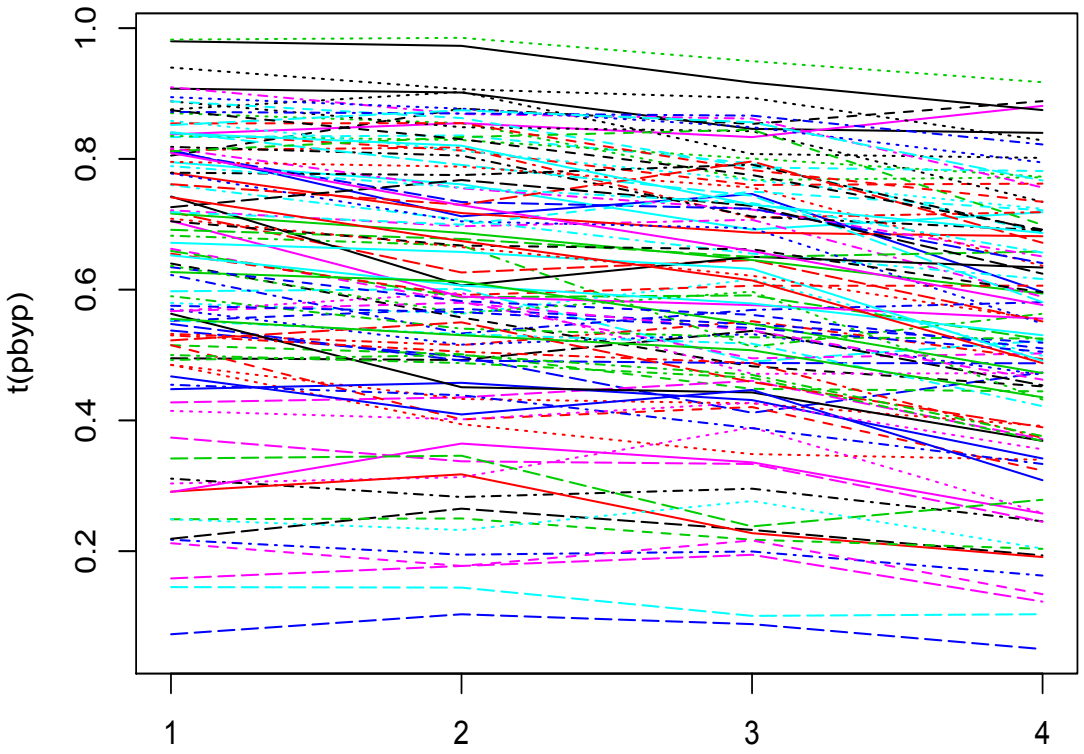

Figure 2. *p*-value change by position of all 101 items in the PISA 2009 reading test for the U.S. X-axis stands for item position and y-axis stands for proportion correct (*p*-value).

Chapter IV: Results

The research models in this study were examined within a generalized linear mixed model (GLMM) fit by maximum likelihood ('glmerMod') with a binomial or logit link. The first part of the results section is focused on the estimation of the position effects on item difficulty. The second part is focused on the model fit comparison. For the

comparisons of the model fit, this study conducted four sets of model comparisons among the seven models that were specified in the methods section.

Estimation of the Position Effects on Item Difficulty

Based on the model specifications of this study, a higher item effect indicates a more difficult item and a lower effect indicates an easier item. Table 10 shows position effects for the base model (M0), two categorical models (M1A, M2A), two linear models (M1B, M2B), and two quadratic models (M3, M4).

First, Table 10 shows the results of position effects of the two categorical models (M1A, M2A). Dummy coding was used to address the categorical position effects. In M1A and M2A, the intercept was coded as position 1. Other position effects (*position2, position3,* and *position4*) were coded as binary variables. The intercept in M0, which has no predictor variables, is the estimated mean performance across items and people. In M1A and M2A, the intercept is the estimated mean performance in the reference position (position 1), and scores in other predictors (*position2, position3,* and *position4*) are differences in average performance compared to the reference position. In the categorical fixed position effects model (M1A), the average difficulty of items in position 1 (the reference position) was .8349 logits. The change of position from 1 to 2 was associated with a .0894 unit increase from the reference position in the expected log odds of average item difficulty. The change of position from 2 to 3 and 3 to 4 was associated with a .3288 unit increase of log odds of average item difficulty from the reference position and a .6602 unit increase of log odds of average item difficulty from the reference position, respectively. Therefore, in the categorical fixed position effects model (M1A), average

item difficulties continued to increase when items moved to the later position. In other words, items became difficult when items were located in the later position.

In the categorical random position effects model (M2A), item-position interaction effects represented item-specific difficulties. The item-specific difficulties increased as items were placed in the later position. The average item specific difficulty in position 1(reference position) was .8667 logits. The item position change from 1 to 2 was associated with a .1272 unit increase of log odds of average item difficulty, compared to the reference position (position 1). The position changes from 2 to 3 and 3 to 4 were also associated with the increase of log odds of average item difficulty. In other words, items became more difficult when located in later positions. The increase of item difficulties was greater in the interaction position effects model (M2A) than the main position effects model (M1A).

Second, Table 10 also shows the results of position effects of the two linear models (M1B, M2B). In modeling linear position effects, the intercept estimated the mean performance at position 1. A predictor, *position*, indicated the average change of the estimated log odds of mean performance with one unit increase in the linear position effects. In the linear main position effects model (M1B), the intercept was .9051 logits. In M1B, a one unit increase in linear position was associated with a .2227 unit increase in the expected log odds of mean item difficulty. This estimate indicated an overall increase of difficulty from position 1 to 4 in the linear main position effects model (M1B). In other words, overall, items were getting difficult when they were located in later positions. In the linear interaction position effects model (M2B), the average item difficulty in position 1 (the intercept) was .9206 logits. In M2B, a one unit increase in

linear position was associated with a .2292 unit increase in the expected log odds of mean item difficulty. This estimate indicated an overall increase in item difficulty from position 1 to 4 in the linear interaction position effects model (M2B). In other words, items became difficult when located in later positions. The increase of item difficulties was greater in the interaction position effects model (M2B) than the main position effects model (M1B).

Third, results of position effects of the two quadratic models (M3, M4) can also be seen in Table 10. In modeling quadratic position effects, the intercept was coded as the mean difficulty at position 1. Another predictor, *position square,* indicated the average change of the estimated log odds of average item difficulty for a one unit increase in the quadratic position effects. In the quadratic main position effects model (M3), item difficulty was continuously increased when items were placed in the later position. In M3, the average difficulty of items in the position 1 (the intercept) was .8393 logits. In the model (M3), a one unit increase in the linear position was associated with a .0448 unit increase in the expected log odds of mean item difficulty. M3 also had predictors for the quadratic position effects. A one unit increase in the quadratic position was associated with a .0590 unit increase in the expected log odds of mean item difficulty. In the quadratic interaction position effects model (M4), the mean difficulty of items in position 1 (the intercept) was .8659 logits. A one unit increase in the linear position was associated with a .0738 unit increase of the expected log odds of mean item difficulty. M4 specialized in the modeling of item specific quadratic position effects. A one unit increase in the quadratic item specific position was associated with a .0524 unit increase of the expected log odds of mean item difficulty. Therefore, in M4, item specific

difficulty was continuously increased when items were placed in the later position. In other words, overall, items became difficult when located in later positions. The increase of item difficulties was greater in the interaction position effects model (M4) than the main position effects model (M3).

In each of the three types of the position effects models (categorical, linear, and quadratic), items became more difficult when located in later positions. Conversely, the interaction position effects modes (M2A, M2B, and M4) estimated greater increases in item difficulty than the main position effects models (M1A, M1B, and M3).

Table 10
Estimation of the Position Effects on Item Difficulty in the PISA 2009 Data for the USA

		Categorical			Linear	Quadratic
Model	Intercept	Position2	Position3	Position4	Position	Position Square
M0	.5853					
M1A	.8349	.0894	.3288	.6602		
M2A	.8667	.1272	.3562	.6935		
M1B	.9051				.2227	
M2B	.9206				.2292	
M3	.8393				.0448	.0590
M4	.8659				.0738	.0524

Model Fit and Comparison

The first set of model fit comparisons was focused on comparisons of a base model (M0) and two categorical models (M1A, M2A). Table 11 shows the results of model fit comparisons among the three models. For the decision of model fit in multilevel modeling, χ^2, AIC, and BIC could be used (McCoach & Black, 2008). The categorical main position effects model (M1A) had smaller AIC and BIC than the base model (M0). What is more, χ^2 between the two models was statistically significant at α = .001, supporting the inclusion of the M1A categorical fixed position effects in each

dataset. Therefore, M1A had a better model fit than M0. The categorical interaction position effects model (M2A) had smaller AIC and BIC than the categorical main position effects model (M1A). The χ^2 between the two models was also statistically significant at α = .001, supporting the inclusion of the M2A interaction terms in each dataset. Therefore, M2A had a better model fit than M1A. Among the three models in the first set of model fit comparison (M0, M1A, and M2A), M2A had the best model fit for the detection of position effects in the categorical setting.

Table 11
Model Fit Comparing M0 with M1A and M1A with M2A

Model	*df*	AIC	BIC	logLik	deviance	χ^2	$\chi^2 df$	*p*
M0	3	164013	164043	-82004	164007			
M1A	6	162706	162766	-81347	162694	1313.8	3	<.001
M2A	15	162523	162673	-81247	162493	200.36	9	<.001

The second set of model fit comparisons included the comparisons of a base model (M0), a linear main position effects model (M1B), a quadratic main position effects model (M3), and a quadratic interaction position effects model (M4). Table 12 shows the results of model fit comparison among the four models. The linear main position effects model (M1B) had smaller AIC and BIC than the base model (M0). The χ^2, between the two models, was statistically significant at α = .001. Therefore, M1B had a better model fit than M0. The quadratic main position effects model (M3) had smaller AIC and BIC than the linear main position effects model (M1B). The χ^2 between the two models was statistically significant at α = .001. Therefore, M3 had a better model fit than M1B. The quadratic interaction position effects model (M4) had smaller AIC and BIC than the quadratic main position effects model (M3). The χ^2 between the two models was

statistically significant at α = .001. Among four models of the second set of model fit comparison (M0, M1B, M3 and M4), the quadratic interaction position effects model (M4) had the best model fit for the detection of position effects in the fixed linear or quadratic setting.

Table 12
Model Fit Comparing M0 with M1B, M1B with M3, and M3 with M4

Model	*df*	AIC	BIC	logLik	deviance	χ^2	$\chi^2 df$	*p*
M0	3	164013	164043	-82004	164007			
M1B	4	162746	162786	-81369	162738	1269.2	1	<.001
M3	5	162704	162754	-81347	162694	43.968	1	<.001
M4	10	162519	162619	-81250	162499	194.9	5	<.001

The third set of model fit comparisons included comparisons of a base model (M0), a linear main position effects model (M1B), a linear interaction position effects model (M2B), and a quadratic interaction position effects model (M4). Table 13 shows the results of model fit comparisons among the four models. As mentioned previously, the linear main position effects model (M1B), had a better model fit than the base model (M0). The linear interaction position effects model (M2B) had smaller AIC and BIC than the linear main position effects model (M1B). The χ^2 between the two models was statistically significant at α = .001. Therefore, M2B had a better model fit than M1B. The AIC and BIC for the model fit comparison between the quadratic interaction position effects model (M4) and the linear interaction position effects model (M2B) conflicted: AIC favored M4 but BIC favored M2B. However, the χ^2 between the two models was statistically significant at α = .001. Therefore, it supported model M4. Among the four models in the third set of model fit comparisons (M0, M1B, M2B, and M4), the quadratic

interaction position effects model (M4) had the best model fit for the detection of position effects in the linear or random quadratic setting.

Table 13
Model Fit Comparing M0 with M1B, M1B with M2B, and M2B with M4

Model	*df*	AIC	BIC	logLik	deviance	χ^2	$\chi^2 df$	*p*
M0	3	164013	164043	-82004	164007			
M1B	4	162746	162786	-81369	162738	1269.2	1	<.001
M2B	6	162556	162616	-81272	162544	194.44	2	<.001
M4	10	162519	162619	-81250	162499	44.428	4	<.001

Model fit comparison between non-nested models (M1A with M1B, M2A with M2B) could also show useful information on the determination of the best model fit among the seven models. AIC and BIC could be used for the determination of this. In the main position effects model (categorical or linear), a main effect for position was applied to all items. Therefore, there was only a single slope for all items. Table 14 shows the results of a model fit comparison between the categorical main position effects model (M1A) and the linear main position effects model (M1B). In the table, M1A had smaller AIC and BIC than M1B. This indicated that M1A had a better model fit than M1B. Since there were only four positions in the dataset, M1B, which expressed position as a slope, might not have a significant advantage over the categorical fixed position effects model. On the other hand, in the interaction position effects model (categorical or linear), due to the inclusion of item-position interaction terms, each item has its own slope. Table 14 also shows the results of a model fit comparison between the categorical interaction position effects model (M2A) and the linear interaction position effects model (M2B). The table shows conflicted results: AIC favored M2A but BIC favored M2B. From the perspective of parsimony, a model with less number of parameters could be better than a

model with more number of parameters. In that sense M2B could have a better model fit than M2A.

Results of the model fit comparisons between non-nested models (M1A with M1B, M2A with M2B) showed that the categorical main position effects model (M1A) had a better model fit than the linear main position effects model (M1B). For the comparison between M2A and M2B, M2B could have a better model fit than M2A because M2B had fewer parameters and smaller BIC than M2A.

Table 14
Goodness-of-Fit Statistics for the Seven Estimated Models

Model	*df*	AIC	BIC	logLik
M0	3	164013	164043	-82004
M1A	6	162706	162766	-81347
M1B	4	162746	162786	-81369
M2A	15	162523	162673	-81247
M2B	6	162556	162616	-81272
M3	5	162704	162754	-81347
M4	10	162519	162619	-81250

Results of the four sets of model comparisons do not provide sufficient information about the decision of the best model among the seven models of predicting the relationship between item position and test performance in the PISA 2009 reading data for the U.S. In order to find the best model, this study used AIC and BIC results again. Table 14 also shows AIC and BIC results of the all seven models. The AIC and BIC conflicted again: AIC favored the quadratic interaction position effects model (M4) but BIC favored the linear interaction position effects model (M2A). In general, the BIC penalizes the number of parameters more strongly than does the AIC to prevent the possibility of increasing the likelihood by adding more parameters. Since M4 has more parameters than M2B, the BIC penalized M4 more than M2B. From the perspective of

parsimony, a model with fewer parameters could be better than a model with more parameters. Given that logic, M2B could be the best model among the seven models of predicting the relationship between item position and test performance.

Chapter V: Discussion

The purpose of this study was to extend previous research, specifically Albano (2013) and Debeer and Janssen (2013), in the examination of multilevel models of item position effects. In order to achieve the purpose, this study analyzed position effects using seven different models. This study demonstrated that different relationships between item position and item difficulty were dependent on how the position effects were coded (categorical, linear, and quadratic). Results achieved in this study also provided indications as to which model among the seven position effects models had the best model fit to describe the relationship between item position and item difficulty.

This study made contributions to the literature in four main ways. First, this study extends the original study of Albano (2013). After some revision, this study applied Albano's item-specific position effects model to a different dataset, the PISA 2009 Data. As in Albano (2013), the multilevel models used here also detected item-position interaction effects in this different dataset. Therefore, as shown in previous studies, this study also statistically supported the finding that ignoring item position effects could lead to biased item parameter estimates.

Second, this study used three different types of indicators for position simultaneously: categorical indicator for position (M1A, M2A), linear indicator for

position (M1B, M2B), and quadratic indicator for position (M3, M4). Previous studies did not include the three different types of indicators for position in one paper at the same time (Albano, 2013; Debeer & Janssen, 2013; Kingston & Dorans, 1984). This study also analyzed both main effects (M1A, M1B, and M3) and interaction effects (M2A, M2B, and M4) of each types of position indicator in one paper. In this way, this study analyzed variability of the position effects.

Third, all models but the base model (M0) had similar patterns of performance change: test performance continuously decreased as position increased (highest at position 1 and lowest at position 4). These results are similar to previous studies, such as Albano (2013), Davey and Lee (2011), and Kingston and Dorans (1984). The increase of item difficulties and decrease of performance when items were located in later positions could be related to fatigue effect, which assumes decreased motivation, concentration, and/or energy levels (Davis & Ferdous, 2005).

Fourth, although this study did not conduct follow-up studies of detecting specific items with problematic position effects, results from item analysis of all 101 items in this study could also give clues for the detection of specific items with the problematic position effects. In general, item analysis is used for the improvement of test items and for the detection of biased or unfair items. In that sense, the item analysis part of this study could also provide useful information about the item-specific position effects to the test administrators. It can be useful as a screening tool for test development.

The decision process of how the seven models fit the data might be controversial. The three sets of the nested model comparisons, which used χ^2 tests, could tell the fit of models in an absolute sense because it involved statistical significance tests. However,

they also had limited power for the model fit comparison because they did not compare all the seven models of estimating item position effects at the same time. In that sense, χ^2 test could not be a universal factor for the decision of the best model among the seven models of the study. For the model fit comparison of all the seven models, AIC and BIC were used. However, AIC and BIC do not have statistical significance tests. Choosing a model is always a relative decision between conventional or innovative models. There is no absolute protocol to firmly guide selection of the best model among the seven models as each has unique advantages and limitations. From the perspective of parsimony and AIC and BIC, this study decided the linear interaction position effects model (M2B) was the best model in a relative sense. This finding also supports results of a previous study (Albano, 2013).

Position effects matter in the PISA 2009 reading test for the U.S. When items were located in later positions, item difficulties increased. What is more, when position effects were different across all items, difficulties of items were greater than when position effects were the same across all items. For the prediction of the relationship between position effects and item difficulties, the linear interaction position effects model (M2B) was the best model. However, this was only in a relative sense. For the prevention of the position effects, conducting a screening test with simple item analysis, such as examining proportion correct by position, can be useful.

Limitations and Future Directions

This study also has several issues, which could be improved in future research. First, there are only four positions in the dataset for the estimation of linear/quadratic main position effects or interaction position effects. This might be a limited number of

positions for the generalization of the position effects. If the seven position effects models examined here were applied to different datasets involving more item positions, the impact of item position on the test performance could show different patterns.

Second, this study was conducted by examining item-position effects in a large-scale, low to medium-stakes test (the PISA). For future study, the application of this revised model to high-stakes testing programs could also be informative.

Third, due to the handling of clusters in the PISA 2009 dataset, this study actually conducted cluster position effects analysis, instead of item position effects analysis. It was like a quasi-item position effects analysis. Therefore, inclusion of more positions in the dataset might be required for the analysis of genuine item position effects.

Fourth, according to Albano (2013), position effects are possibly associated with other various factors, such as gender, DIF, test length, and etc. However, this study did not provide meaningful results for the generalization of item position effects to the other areas. This study only focused on the possible bias in item difficulty due to the differences in item position.

Fifth, this study did not seriously conduct comprehensive and direct analysis on the potential causes for item position effects, such as fatigue effect or timing. According to Davis and Ferdous (2005), the decrease of test performance during the test is also related to other factors, such as concentration, motivation, and energy level of test takers.

Sixth, this study decided to delete clusters of science and math and focused on reading tests that had the most clusters among the three subjects in the PISA 2009 data. In this way, this study intentionally ignored context effects. This can be a major

limitation of the study. However, it is a reasonable one because completely avoiding the context effects is not feasible.

References

Albano, A.D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement*, 50(4), 408-426.

Alexandrowicz, R., & Matschinger, H. (2008). Estimation of item location effects by means of the generalized logistic regression model: A simulation study and an application. *Psychology Science Quarterly*, 50, 64–74.

Brennan, R. (1992). The context of context effects. *Applied Measurement in Education*, 5, 225–264.

Davey, T., & Lee, Y. H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised general test* (Research Report 11–26). Princeton, NJ: Educational Testing Service.

Davis, J., & Ferdous, A. (2005). *Using item difficulty and item position to measure test fatigue*. Washington, DC: American Institutes for Research.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, 50(2), 164-185.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

Eignor, D. R., & Cook, L. L. (1983, April). *An investigation of the feasibility of using item response theory in the preequating of aptitude tests*. Montreal, Canada: Paper presented at the meeting of the American Educational Research Association.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 38, 79–93.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for

IRT equating and adaptive testing. *Applied Psychological Measurement*, *8*, 147–154.

Kolen, M., & Harris, D. (1990). Comparison of item pre-equating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement,* 27(1), 27–29.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, 55, 387–413.

Li, F., Cohen, A., & Shen, L. (2012). Investigating the effect of item position in computer-based tests. Journal of Educational Measurement, 49(4), 362-379.

McCoach, D. B., & Black, A. C. (2008). Evaluation of model fit and adequacy. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data (pp. 245– 272)*. Charlotte, NC: Information Age Publishing, Inc.

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education*, 22, 38–60.

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika*, 15, 291–315.

Plake, B. S. (1980). Item arrangement and knowledge of arrangement on test scores. *Journal of Experimental Education*, 49, 56-58.

Plake, B. S., Ansorge, C. J., Parker, C. S., & Lowry, S. R. (1982). Effects of item arrangement, knowledge of arrangement, test anxiety, and sex on test performance. *Journal of Educational Measurement*, 19, 49-58.

Plake, B. S., Thompson, P. A., & Lowry, S. (1981). Effect of item arrangement, knowledge of arrangement, and test anxiety on two scoring methods. *Journal of Experimental Education*, 41, 214-219.

Pomplun, M., & Ritchie, T. (2004). An investigation of context effects for item randomization within testlets. *Journal of Educational Computing Research*, 30, 243–254.

Way, W. D., Carey, P., & Golub-Smith, M. (1992). *An exploratory study of characteristics related to IRT item parameter invariance with the Test of English as a Foreign Language.* (TOEFL Tech. Rep. No. 6.) Princeton, NJ: Educational Testing Service.

Whitely, S. E., & Dawis, R. V. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement*, 36, 329–337.

Wise, L., Chia, W., & Park, R. (1989). *Item position effects for test of work knowledge and arithmetic reasoning*. Paper presentation at the annual meeting of the AERA, San Francisco.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, 17, 297–311.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, 10, 10–16.

APPENDIX A: Tables of Item Analysis of Each Cluster

Table 3
Item Analysis of Cluster 1 (R1)

Item ID	*p*-value					cpb	aid
	Position 1	Position 2	Position 3	Position 4	Mean *p*-value		
r067q01	0.91	0.9	0.85	0.84	0.87	0.42	0.77
r067q04	0.84	0.81	0.76	0.76	0.79	0.47	0.77
r067q05	0.87	0.85	0.77	0.77	0.82	0.53	0.76
r102q04a	0.22	0.19	0.2	0.16	0.19	0.36	0.78
r102q05	0.6	0.6	0.49	0.52	0.55	0.41	0.77
r102q07	0.84	0.85	0.83	0.88	0.85	0.39	0.78
r219q02	0.8	0.88	0.85	0.89	0.86	0.35	0.78
r220q01	0.48	0.39	0.35	0.34	0.39	0.44	0.77
r220q02b	0.68	0.67	0.51	0.56	0.61	0.47	0.77
r220q04	0.54	0.5	0.41	0.47	0.48	0.37	0.78
r220q05	0.8	0.76	0.69	0.72	0.74	0.53	0.76
r220q06	0.66	0.58	0.49	0.5	0.56	0.44	0.77
Mean	0.69	0.67	0.60	0.62	0.64	0.43	0.77

Note: alpha = .79, internal consistency reliability

Table 4
Item Analysis of Cluster 2 (R2)

Item ID	*p*-value					cpb	aid
	Position 1	Position 2	Position 3	Position 4	Mean *p*-value		
r055q01	0.89	0.85	0.84	0.73	0.83	0.47	0.83
r055q02	0.53	0.52	0.55	0.49	0.52	0.51	0.83
r055q03	0.66	0.59	0.59	0.52	0.59	0.56	0.82
r055q05	0.81	0.71	0.75	0.6	0.72	0.6	0.82
r104q01	0.83	0.84	0.79	0.69	0.79	0.44	0.83
r104q02	0.3	0.31	0.39	0.26	0.32	0.25	0.84
r104q05	0.31	0.28	0.3	0.25	0.28	0.45	0.83
r111q01	0.72	0.63	0.65	0.55	0.64	0.52	0.83
r111q02b	0.72	0.69	0.65	0.59	0.66	0.5	0.83
r111q06b	0.55	0.57	0.55	0.52	0.55	0.51	0.83
r227q01	0.58	0.56	0.62	0.5	0.56	0.36	0.84
r227q02t	0.72	0.7	0.71	0.59	0.68	0.47	0.83
r227q03	0.5	0.49	0.54	0.45	0.49	0.53	0.83
r227q06	0.78	0.72	0.69	0.68	0.72	0.52	0.83
Mean	0.64	0.60	0.62	0.53	0.60	0.48	0.83

Note: alpha = .84, internal consistency reliability

Table 5
Item Analysis of Cluster 3 (R3A)

Item ID	*p*-value					cpb	aid
	Position 1	Position 2	Position 3	Position 4	Mean *p*-value		
r414q02	0.52	0.5	0.45	0.44	0.48	0.46	0.81
r414q06	0.57	0.52	0.53	0.47	0.52	0.56	0.81
r414q09	0.76	0.71	0.66	0.62	0.68	0.45	0.81
r414q11	0.43	0.44	0.46	0.37	0.42	0.31	0.82
r447q01t	0.74	0.61	0.65	0.63	0.66	0.42	0.81
r447q04	0.66	0.59	0.61	0.61	0.61	0.51	0.81
r447q05	0.84	0.83	0.8	0.77	0.81	0.25	0.82
r447q06	0.62	0.53	0.57	0.58	0.58	0.5	0.81
r452q03	0.14	0.14	0.1	0.1	0.12	0.31	0.82
r452q04	0.7	0.59	0.58	0.56	0.61	0.43	0.81
r452q06	0.64	0.56	0.48	0.45	0.53	0.57	0.8
r452q07	0.48	0.43	0.43	0.39	0.43	0.49	0.81
r458q01	0.59	0.54	0.53	0.53	0.55	0.45	0.81
r458q04	0.55	0.49	0.49	0.49	0.5	0.36	0.82
r458q07	0.65	0.61	0.58	0.53	0.59	0.51	0.81
Mean	0.59	0.54	0.53	0.50	0.54	0.44	0.81

Note: alpha = .82, internal consistency reliability

Table 6
Item Analysis of Cluster 4 (R4A)

Item ID	*p*-value					cpb	aid
	Position 1	Position 2	Position 3	Position 4	Mean *p*-value		
r083q01	0.57	0.58	0.54	0.46	0.54	0.46	0.82
r083q02	0.88	0.9	0.81	0.8	0.85	0.3	0.83
r083q03	0.81	0.82	0.71	0.72	0.76	0.37	0.82
r083q04	0.69	0.68	0.65	0.66	0.67	0.42	0.82
r101q01	0.45	0.46	0.43	0.34	0.42	0.35	0.83
r101q02	0.89	0.86	0.79	0.78	0.83	0.43	0.82
r101q03	0.57	0.6	0.47	0.47	0.53	0.54	0.81
r101q04	0.82	0.8	0.71	0.69	0.76	0.5	0.82
r101q05	0.52	0.55	0.46	0.39	0.48	0.4	0.82
r245q01	0.63	0.61	0.55	0.47	0.56	0.39	0.82
r245q02	0.63	0.58	0.56	0.51	0.57	0.48	0.82
r442q02	0.86	0.81	0.77	0.72	0.79	0.4	0.82
r442q03	0.82	0.76	0.72	0.65	0.74	0.57	0.81
r442q05	0.22	0.26	0.23	0.19	0.23	0.49	0.82
r442q06	0.29	0.32	0.23	0.19	0.26	0.49	0.82
r442q07	0.25	0.25	0.22	0.2	0.23	0.45	0.82
Mean	0.62	0.62	0.55	0.52	0.58	0.44	0.82

Note: alpha = .83, internal consistency reliability

Table 7
Item Analysis of Cluster 5 (R5)

Item ID	*p*-value					cpb	aid
	Position 1	Position 2	Position 3	Position 4	Mean *p*-value		
r404q03	0.78	0.7	0.69	0.55	0.68	0.48	0.81
r404q06	0.55	0.53	0.52	0.42	0.51	0.43	0.81
r404q07t	0.37	0.34	0.33	0.24	0.32	0.44	0.81
r404q10a	0.56	0.45	0.44	0.37	0.46	0.58	0.8
r404q10b	0.52	0.4	0.42	0.32	0.42	0.57	0.8
r406q01	0.64	0.56	0.6	0.43	0.56	0.44	0.81
r406q02	0.45	0.44	0.39	0.33	0.4	0.39	0.81
r406q05	0.84	0.79	0.74	0.58	0.74	0.46	0.81
r424q02t	0.29	0.36	0.34	0.26	0.31	0.25	0.82
r424q03	0.7	0.67	0.66	0.6	0.66	0.33	0.82
r424q07	0.8	0.79	0.76	0.64	0.74	0.44	0.81
r455q02	0.5	0.49	0.47	0.37	0.46	0.38	0.81
r455q03	0.81	0.73	0.72	0.64	0.73	0.45	0.81
r455q04	0.67	0.66	0.63	0.49	0.61	0.49	0.81
r455q05t	0.21	0.18	0.22	0.13	0.19	0.42	0.81
Mean	0.58	0.54	0.53	0.42	0.52	0.44	0.81

Note: alpha = .82, internal consistency reliability

Table 8
Item Analysis of Cluster 6 (R6)

Item ID	*p*-value					cpb	aid
	Position 1	Position 2	Position 3	Position 4	Mean *p*-value		
r412q01	0.94	0.91	0.89	0.83	0.89	0.35	0.79
r412q05	0.53	0.51	0.49	0.37	0.47	0.39	0.78
r412q06t	0.34	0.35	0.24	0.28	0.3	0.18	0.8
r412q08	0.47	0.41	0.45	0.31	0.41	0.5	0.78
r420q02	0.72	0.7	0.75	0.72	0.72	0.39	0.78
r420q06	0.41	0.4	0.43	0.36	0.4	0.3	0.79
r420q09	0.78	0.78	0.79	0.69	0.76	0.36	0.79
r420q10	0.76	0.73	0.8	0.67	0.74	0.55	0.77
r437q01	0.56	0.53	0.51	0.44	0.51	0.41	0.78
r437q06	0.58	0.57	0.54	0.5	0.55	0.42	0.78
r437q07	0.25	0.23	0.28	0.2	0.24	0.34	0.79
r453q01	0.91	0.87	0.86	0.76	0.85	0.5	0.78
r453q04	0.73	0.77	0.73	0.62	0.71	0.44	0.78
r453q05t	0.74	0.67	0.61	0.49	0.63	0.48	0.78
r453q06	0.81	0.83	0.84	0.7	0.8	0.48	0.78
Mean	0.64	0.62	0.61	0.53	0.60	0.41	0.78

Note: alpha = .80, internal consistency reliability

Table 9
Item Analysis of Cluster 7 (R7)

Item ID	*p*-value					cpb	aid
	Position 1	Position 2	Position 3	Position 4	Mean *p*-value		
r432q01	0.89	0.88	0.85	0.79	0.85	0.54	0.8
r432q05	0.79	0.76	0.73	0.66	0.73	0.56	0.8
r432q06t	0.16	0.18	0.19	0.12	0.16	0.3	0.82
r446q03	0.98	0.97	0.92	0.87	0.94	0.48	0.81
r446q06	0.85	0.85	0.78	0.73	0.81	0.45	0.81
r456q01	0.98	0.99	0.95	0.92	0.96	0.41	0.81
r456q02	0.87	0.87	0.87	0.82	0.86	0.43	0.81
r456q06	0.85	0.87	0.86	0.77	0.84	0.52	0.8
r460q01	0.81	0.73	0.66	0.58	0.69	0.51	0.8
r460q05	0.87	0.83	0.78	0.69	0.79	0.54	0.8
r460q06	0.71	0.67	0.62	0.55	0.64	0.43	0.81
r466q02	0.49	0.5	0.47	0.38	0.46	0.46	0.81
r466q03t	0.07	0.1	0.09	0.05	0.08	0.08	0.83
r466q06	0.84	0.82	0.73	0.69	0.77	0.55	0.8
Mean	0.73	0.72	0.68	0.62	0.68	0.45	0.81

Note: alpha = .82, internal consistency reliability