Spring 2019

Are all cognitive items equally prone to position effects? Exploring the relationships among item features and position effects

Thai Quang Ong
James Madison University

Are All Cognitive Items Equally Prone to Position Effects?

Exploring the Relationships Among Item Features and Position Effects

Thai Quang Ong

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

For the degree of

Doctor of Philosophy

Department of Graduate Psychology

May 2019

FACULTY COMMITTEE:

Committee Chair: Dena A. Pastor, Ph.D.

Committee Members/Readers:

Deborah L. Bandalos, Ph.D.

Christine E. DeMars, Ph.D.

Monica K. Erbacher, Ph.D.

Acknowledgements

First, I would like to thank my wonderful mentor, Dr. Dena Pastor. Dena, I don't think you know this but you are the number one reason why I stayed in the Ph.D. program. You allowed me to grow and develop without pressuring me to be the "best" student. You allowed me to complain and vent to you when times were tough. You allowed me to put my mental health first in place of a few extra lines on my curriculum vitae. I am so grateful to have had the opportunity to work with you, learn from you, and be your advisee. Thank you for everything you've done for me both personally and professionally. I did it! We did it!

Second, I would like to thank my dissertation committee members: Dr. Deborah Bandalos, Dr. Chistine DeMars, and Dr. Monica Erbacher. Your feedback on this dissertation was invaluable and each of you contributed uniquely to this dissertation. Thank you for all of the time you spent thinking about position effects! I am honored to have had the opportunity to work and learn from each of you in the Ph.D. program.

Third, I would like to thank some of the amazing friends I have made in the Ph.D. program: Kristen Smith, Scott Strickman, Courtney Sanders, Madison Holzman, and Aaron Myers. You all truly made my graduate career delightful. Thank you for letting me "do me" and "be me". We are all pink starbursts and don't you ever forget that!

Finally, I would like to thank my family. Mom and dad, thank you for supporting me emotionally and financially throughout the last five years. I could not have done this without you both. We now officially have a doctor in the family. I hope you are proud! Vy and Rashid, thank you for always telling me I can do it. Dinner is on me for the next 5 (ish) years!

Table of Contents

List of Tables

List of Figures

Abstract

One type of context effect is a position effect, which implies parameters of an item are influenced by the position of the item on the test. Researchers often discuss two types of position effects: negative position effects and positive position effects (e.g., Albano, 2013; Debeer & Janssen, 2013). Items exhibiting negative position effects become harder when placed later on the test, whereas items exhibiting positive position effects become easier when placed later on the test. Researchers have primarily examined the underlying causes of position effects through an item or person perspective (e.g., Bulut, 2015; Kingston & Dorans, 1984; Qian, 2014). Researchers who adopted an examinee perspective on position effects exclusively studied the relationships among person variables and position effects. Researchers who adopted an item perspective on position effects exclusively studied the relationships among item variables and position effects. These two perspectives are limiting because they do not encourage researchers to consider the potential interactions among person variables, item variables, and position effects.

In this dissertation, I examined the underlying causes of position effects through an integrated perspective, where I studied the relationships among person variables, item variables, and position effects simultaneously. I conducted a true experiment in which I administered items from two low-stakes assessments in different order to two groups of examinees, examined the presence of position effects, and evaluated the degree to which position effects were moderated by different item (item length, number of response options, mental taxation, and graphic) and person variables (effort, change in effort, and gender). I modeled position effects and their relationships with item and person variables under the generalized linear mixed modeling (GLMM) framework.

On both assessments, I found items exhibited significant negative linear position effects on both assessments, with the magnitude of the position effects varying from item to item. Items became harder when placed later on the assessments but the extent to which they became harder differed slightly across items. Additionally, I found the position effects to be moderated by item difficulty and item length but not number of response options, mental taxation, or graphic. Easier and longer items were more prone to position effects than harder and shorter items; however, items varying in mental taxation, items containing a graphic, and items varying in response options were similarly prone to position effects. More so, I found examinee effort levels, change in effort patterns, and genders did not moderate the relationships among position effects and item features. Based on these findings, testing practitioners should be cautious about administering long or easy items in different order across forms and/or administrations.

CHAPTER 1

INTRODUCTION

Context Effects

The context in which an item is administered is defined in relation to the characteristics of the set of items preceding the item. As the characteristics of the set of items preceding the item change (e.g., difficulty, format, discrimination), the context of the item also changes (Albano, 2013). In both small- and large-scale testing programs, there are many scenarios in which items are administered in different contexts across test forms and examinees. When test security is of concern, different test forms, with items scattered across them, may be administered to examinees. When adaptive testing is used, the same item may be administered to two examinees but in a different order. When pretesting items, the same set of pretest items may be administered on different test forms but in different general locations. Across these scenarios, the contexts of the items are different because the characteristics of the set of items preceding each item changes across test forms. It is assumed under the scenarios above that contexts in which the items are administered have no direct influence on the item parameters; that is, when the same item appears in a different context across two test forms, it is assumed the parameters of that item are stable across the two test forms. This assumption, however, may not always be true due to the presence of context effects, which are defined as "any influence or interpretation that an item may acquire purely as a result of its relationship to the other items making up a specific test" (Wainer & Kiely, 1987, p.187). Based on the definition above, context effects comprise a group of effects that influence item responses on a test.

Position Effects

One type of context effect is a position effect, which implies parameters of an item are influenced by the position of the item on the test. Researchers often cite and discuss two types of position effects: negative position effects and positive position effects (e.g., Albano, 2013; Debeer & Janssen, 2013). Items exhibiting negative position effects become harder when placed later on the test, whereas items exhibiting positive position effects become easier when placed later on the test. Although a position effect is categorized as one type of context effect, Albano (2013) noted the importance of clearly distinguishing between position and context for research purposes. He defined context in relation to the characteristics of the set of items preceding the item (e.g., item type) and position in relation to the quantity of the set of items preceding the item. Due to the interdependent nature of context and position, the distinction between context and position is rarely made explicit in the literature, which has led researchers to use the two terms interchangeably.

Consequences of Position Effects

There are a number of reasons why psychometricians working in operational testing organizations should be concerned with position effects. First, it is common for test items to be administered in different order across different subsets of examinees. For example, in computerized testing, examinees may be administered the same set of items but the items may be administered in a randomized order (scrambled) for security purposes. Additionally, in computerized adaptive testing (CAT), examinees receive different sets of items (depending on their ability) and the different sets of items are administered in a different orders. In both scenarios, the presence of item position effects can heavily impact the scores and outcomes of examinees. For example, let us assume Examinee A and Examinee B are of equal ability and Item X has a negative position effect. In a computerized testing scenario where items are administered in a random order, Examinee A may be administered Item X at the beginning of the test, whereas Examinee B may be administered Item X at the end of the test. Because Examinee A is administered Item X at the beginning of the test, Examinee A would have a higher probability of getting Item X correct than Examinee B, despite Examinee A and Examinee B being equal in true ability. The reverse would be true if we assume Item X has a positive position effect instead of a negative position effect. Although I only highlighted the impact of a single item's position effect on examinee performance in the example above, it is possible for a set of items with varying position effects to exist within a single test. Thus, the degree to which examinees' scores and outcomes are impacted by position effects depends on the number of items with position effects, the magnitude of the position effects, and the direction of the position effects.

Second, it is common for testing organizations to field test newly developed items to obtain information about item performance. The new items (i.e., pre-test items) are typically embedded in an operational test; however, the way in which the items are embedded may vary across testing organizations and field testing approaches. For example, psychometricians may administer all pre-test items on a single test form or split them across multiple test forms and place them either at the beginning of the test (before the operational items), end of the test (after the operational items), or scattered throughout the test (mixed with the operational items). Regardless, once the operational test with the embedded pre-test items is administered, psychometricians can evaluate the item statistics of the pre-test items and use the information to determine which pre-test items should be included in future test forms. If, however, the new item statistics are influenced by the position of the pre-test items, then consequently, the decisions to select or use certain pre-test items would also be influenced by the position of the pre-test items. Thus, psychometricians may make different decisions about the pre-test items, depending on where the pre-test items are placed on the operational test. Additionally, because the position of the pre-test items may change when moving from a field test to an operational test, the item statistics of the pre-test items may differ substantially on the operational test compared to the field test.

Third, given most operational testing organizations rely on item response theory (IRT) for scoring and equating, psychometricians often calibrate pre-test items to place them on the same scale as the operational items shortly after field testing – a process known as precalibration. One of the advantages of precalibration is psychometricians can immediately place the pre-test items in the item pool and use them as anchor items in future non-equivalent anchor test (NEAT) equating designs or as operational items in future pre-equating designs (Kolen & Brennan, 2004). The position of the pre-test items on the field test may, however, bias the initial IRT parameter estimates. For example, a pre-test item may appear to be easier when administered at the beginning of the field test than at the end of the field test. Because the pre-calibrated pre-test items may be given in different positions or scrambled across examinees in future test forms, psychometricians may potentially be using biased item parameter estimates to calibrate the new items, which would result in biased new item and person parameter estimates.

Finally, in NEAT equating designs, psychometricians use the performance on anchor items (or common items across the two test forms) to adjust for minor differences in ability across different groups of examinees (Kolen & Brennan, 2004). Thus, the anchor items used should be similar in characteristics across forms – similar not only in parameter estimates but also in context and position (Cook & Paterson, 1987; Kolen & Brennan, 2004). For example, if an anchor item is placed in the 10th position on Form X, then the same anchor item should be placed in the 10th position on Form Y (assuming all other aspects of item context are equal). Due to practical constraints, the latter may not always be satisfied in practice. If the anchor items differ in their positions across Form X and Form Y, the parameter estimates of the anchor items may not be invariant across forms, resulting in biased equating relationships.

In summary, the presence of position effects is threatening to almost all aspects of the testing process and should be empirically investigated. If left uninvestigated, the presence of position effects can influence the estimation of item parameters, test scoring and equating, examinee ability estimates, and decisions about examinees. This may, in turn, have major consequences for both low-stakes and high-stakes testing programs. For low-stakes testing programs (e.g., higher education assessments), where test scores are often used for accountability purposes and program improvements, inaccurate decisions about examinees may lead to inaccurate conclusions about a status of a university or academic program. For high-stakes testing programs (e.g., medical licensures, medical certifications), where test scores are often used to ensure one has the prerequisite knowledge to practice a particular profession, inaccurate decisions about examinees may lead to the endangerment of the general public and/or to unfair decisions made about testtakers. Thus, ignoring position effects may have high-stakes consequences for both low-stakes and high-stakes testing programs.

Person and Item Perspectives on Position Effects

The potential negative consequences of position effects in testing have led researchers to explore why position effects occur. Some researchers viewed position effects from the examinee perspective, where they assumed position effects are a function of the examinees. These researchers focused on exploring how different person variables were related to position effects (e.g., Bulut, 2015; Hambleton & Traub, 1974; Klosner & Gellman, 1973; Munz & Smouse, 1968; Qian, 2014). In contrast, other researchers viewed position effects from the item perspective, where they assumed position effects are a function of the items. These researchers focused on exploring how different item variables were related to position effects (e.g., Kingston & Dorans, 1984; Le, 2007). By adopting either an item or person perspective on position effects, researchers were able to focus their research on a particular set of person or item variables. These two perspectives, however, have limited the research on position effects in several ways. First, researchers who adopted an examinee perspective on position effects exclusively studied person variables, whereas researchers who adopted an item perspective on position effects exclusively only studied item variables. Thus, these researchers failed to consider how position effects may be related to both person and item variables. Second, related to the latter point, researchers adopting either one of the perspectives also failed to consider the potential interactions among person variables, item variables, and position effects.

An Integrated Perspective on Position Effects

Given the limitations above, I argue researchers should instead adopt an integrated perspective on position effects, where position effects are viewed as a function of both the examinees and items. There are several advantages to adopting an integrated perspective on position effects. First, researchers are encouraged to examine both item and person variables and their relationships to position effects within a single framework under this perspective, which may help further the research on position effects. Second, researchers are encouraged to examine potential interactions among position effects, item variables, and person variables under this perspective, which may help uncover the complexity of position effects. For example, a researcher adopting a person perspective or item perspective on position effects may find position effects to be related to fatigue or item type, respectively; however, a researcher adopting an integrated perspective on position effects may find position effects to be related to both fatigue and item type, with fatigue moderating the relationship between item type and position effects. To date, however, no researchers have examined position effects through an integrated perspective.

Purpose of Study

The general purpose of my dissertation is to investigate why position effects occur through an integrated perspective, where position effects are considered a function of both the examinees and items. First, I evaluated item position effects on two cognitive assessments administered in low-stakes testing conditions. I manipulated the order of the items on the two assessments and administered the assessments to two samples of undergraduate students at a mid-sized university. Second, I evaluated how different item and person variables were related to the position effects. Because I administered my assessments in low-stakes testing conditions, I chose to evaluate item and person variables most relevant to position effects in the low-stakes testing context. Third, I evaluated the consistency of the results across the two assessments, which differed in content.

CHAPTER 2

REVIEW OF LITERATURE

In my dissertation, I simultaneously investigated the relationships among position effects, item variables, and person variables in two cognitive tests administered in low-stakes testing conditions. To help the reader understand why there is a need to examine the relationships among position effects, item variables, and person variables, I discuss the following points. First, I review previous research on position effects. Second, I discuss previous research examining the relationships among position effects and person/item variables. Third, I argue for the need to examine the relationships among position effects, item variables, and person variables simultaneously, particularly in low-stakes testing contexts. Finally, I further elaborate on the purpose of my study and present my primary research questions.

Previous Research on Position Effects

The research on position effects has not been "clean" given that context and position effects are hard to disentangle from one another. In fact, it is rather difficult for researchers to manipulate only one of these factors from a research design (and logistical) perspective. Researchers who want to explore the sole effect of context would need to fix the position of an item and manipulate the formats of prior items across forms, whereas researchers who want to explore the sole effect of position would need to manipulate the position of an item and fix the formats of prior items across forms (Albano, 2013). Although the research designs I summarized above seem relatively straightforward, they can rarely be employed in practice due to logistical reasons[1]. Instead, researchers have engaged in alternative research designs when studying position effects, such as classifying items into blocks and reordering item blocks across different test forms. These research designs, however, are limited in that they do not isolate position effects because the contexts in which the items are administered are not kept constant. Thus, in the following sections where I review previous and current research on position effects, the reader should acknowledge the findings I summarize may be a function of context effects, position effects, or both (see Table 1 for summary of studies discussed below).

Impact of Position Effects on Test Performance. Early research on position effects focused primarily on examining the impact of position effects on test performance[2] (Leary & Dorans, 1985). These researchers manipulated the position of items across test forms and compared examinee performance at the test level. In these studies, researchers manipulated the position of items by either randomizing the order of the individual items (Monk & Stallings, 1970), categorizing the items into separate item blocks and randomizing the item blocks (Klein, 1981; Mollenkopf, 1950), or ordering the items based on their item difficulties (Brenner, 1964; Hambleton & Traub, 1974; Lane, Bull, Kundert, & Newman, 1987; MacNicol, 1956; Sax & Cromack, 1966). In general, researchers found mixed results on the impact of different item order arrangements on test performance. Some researchers found different item order arrangements had a

[1] In order to fully isolate position from context for an item, the researcher must administer the same number of items prior to the item of interest but allow the item of interest to be in a different position across forms. This is, however, not possible because the number of items prior to the item of interest must vary so the position of item of interest can vary across forms.

[2] In their 1985 article, Leary and Dorans provided a comprehensive overview of the research on context effects, including position effects, from the 1950s until 1980s. For more information on the studies I summarized in this section, please refer to their article.

significant effect on test performance (Hambleton & Traub, 1974; MacNicol, 1956, Sax & Cromack, 1966), with test performance most negatively impacted when items on a partly-speeded test were arranged from hardest to easiest (Leary & Dorans, 1985). In contrast, other researchers found different item order arrangements had no significant effect on test performance (Brenner, 1964; Klimko, 1984; Lane et al., 1987; Marso, 1970; Monk & Stallings, 1970). Thus, the impact of different item order arrangements on test performance was not consistent, but rather dependent on the item arrangement (e.g., random item scrambling, random section scrambling, and item order by difficulty) and the conditions under which the test was administered (e.g., speeded versus power).

Impact of Position Effects on Item Difficulty and Equating. With the growth in testing organizations adopting IRT, researchers have shifted from studying the impact of position effects on test-level performance to studying the impact of position effects on item difficulty. Unlike previous researchers, these researchers examined the change in item difficulty when items were placed in different positions across two or more test forms (e.g., field test versus operational test). In general, researchers found changes in item positioning across two or more test forms sometimes led to changes in item difficulty. When items were placed later on one test form compared to earlier on another test form, some researchers found negative differences in *P* indices (proportion correct) or positive differences in *b*-parameters (ability level at which the examinee has a 50% probability of obtaining correct response) across the two test forms, suggesting items were harder for examinees when placed later on the test than earlier on the test (negative position effects; Davis & Ferdous, 2005; Eignor & Cook, 1983, Kingston & Dorans, 1984; Meyers, Miller, & Way, 2009). In contrast, some researchers found positive differences in *P* indices or negative differences in *b*-parameters across the two test forms, suggesting items were easier for examinees when placed later on the test than earlier on the test (positive position effects; Kingston & Dorans, 1984). In other instances, some researchers found no differences in *P* indices or *b*-parameters across the two test forms (Huck & Bowers, 1972). In the studies above, researchers not only found item position changes led to differences in item difficulty but also led to differences in equating results. Although these researchers studied different equating methods, they all generally found different item order arrangements had a significant impact on equating results (Harris, 1991; Kingston & Dorans 1984; Kolen & Harris, 1990; Yen, 1980; Zwick, 1991).

Modeling Position Effects. In recent years, researchers have applied generalized linear mixed models (GLMM), IRT models, and structural equation models to study position effects (Albano, 2013; Bulut, Quo, & Gierl, 2017; Debeer & Janssen, 2013; Weirich, Hecht, & Böhme, 2014). Unlike previous researchers, these researchers used complex statistical models to empirically investigate the relationship between item position and item performance. Researchers differed in how they parameterized position effects across the different models (e.g., position effects varied across items versus position effects did not vary across items); however, they typically specified the relationship between item position and item performance as linear, which allowed them to interpret the position effect as the change in log-odds of getting an item correct for every one position or block increase on the test. In general, researchers found item performance was correlated with item position. In some studies, researchers found item performance and item position were negatively correlated, suggesting item performance decreases as item position increases (negative position effects; Albano, 2013; Bulut et al., 2017; Davey & Lee, 2011; Debeer, Buchholz, Hartig, & Janssen, 2014; Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Le, 2007, Weirich, Hecht, Penk, Roppelt, & Böhme, 2017). In other studies, researchers found item performance and item position were positively correlated, suggesting item performance increases as item position increases (positive position effects; Kingston & Dorans, 1984). Contrary to the studies above, other researchers found item performance was not correlated with item position (Hahne, 2008; Hohensinn, Kubinger, Reif, Schleiber, & Khorramdel, 2011). Thus, the relationship between item performance and item position may be moderated by factors related to the set of examinees, items, or both.

In summary, researchers have studied position effects in a variety of contexts. Across these studies, researchers found position effects may impact examinees' responses, which can lead to changes in test performance, item difficulty, and equating results. Although these findings are informative as they show how position effects can impact various testing outcomes, they fail to address the more important questions at hand: why do position effects occur and what variables relate to position effects? To that end, researchers have explored why position effects occur by studying different variables related to position effects. I discuss and elaborate on this in the following section, starting with person variables and ending with item variables.

Variables Related to Position Effects

There are two types of variables we can examine when studying why position effects occur: person and item. Person variables are those related to characteristics of examinees, such as fatigue, whereas item variables are those related to the characteristics of items, such as item type. Although position effects may be related to both person and item variables, the majority of the research on position effects has focused primarily on person variables. Thus, not surprisingly, it is common for researchers to attribute position effects to characteristics of the examinees rather than characteristics of the items.

Person Variables. The two most common person variables researchers attribute position effects to are fatigue and practice (e.g., Hohensinn et al., 2011; Kingston & Dorans, 1984). Fatigue is often associated with negative position effects (items placed later on the test are harder because examinees become fatigued) whereas practice is often associated with positive position effects (items placed later on the test are easier because examinees become familiarized with the test items). Both fatigue and practice are plausible explanations for position effects; however, there is surprisingly limited empirical research on the relationships among practice/fatigue and position effects. Thus, researchers who often cite fatigue and practice as explanations for position effects do so without much empirical support for their claims. In fact, Debeer and Jannsen (2013) noted simply attributing position effects to either fatigue or practice can be considered as "tautological as it is a relabeling of the phenomenon rather than giving a true cause" (p. 169).

It is also plausible for different subgroups of examinees to be more susceptible to position effects than other subgroups of examinees. For example, some examinees may be more affected by position effects while other examinees may be less affected by position effects due to individual differences in fatigue, practice, and/or other person variables. Although no researchers have specifically examined fatigue or practice, researchers have examined other person variables as potential moderators of the relationship between item position and performance at both the test and item level. These researchers studied the following person variables: test anxiety, gender, ability, and motivation.

Test Anxiety. Munz and Smouse (1968) and Smouse and Munz (1968, 1969) were the first researchers to study the impact of item position, test anxiety, and their interaction on test performance in three different studies. Across these studies, they explored whether the impact of three different item order arrangements (easy-to-hard, hard-to-easy, and random order) on test performance differed across different testing conditions (anxiety-provoking and normal; Munz & Smouse, 1968) and achievement anxiety types[3] (Facilitators and Deliberators; Smouse & Munz, 1968, 1969). They found a significant interaction between item order arrangement and test anxiety, with different types of achievement anxiety groups performing differentially across different item order arrangements, in two of their three studies. Other researchers, however, who either attempted to replicate their findings or studied the impact of item order arrangements and test anxiety on test performance, failed to obtain similar results (Berger, Munz, Smouse, & Angelino, 1968; Hambleton & Traub, 1974; Marso, 1970; Plake, Ansorge, Parker, & Lowry, 1982; Plake, Thompson, & Lowry, 1981; Towle & Merrill, 1975). The latter scenario may be explained by differences among the studies conducted by these other researchers and Smouse and Munz. For example, Smouse and Munz (1968, 1969) studied an achievement test (high-stakes), whereas Berger et al. (1968) studied an aptitude test

[3] Smouse and Munz (1968, 1969) measured achievement anxiety types using the Achievement Anxiety Test (AAT; Alpekt & Habek, 1960). The AAT is a 19-item scale comprised of two subscales: Facilitating and Deliberating. Examinees who score high on the Facilitating subscale are thought to perform better in anxiety-provoking situations. Examinees who score high on the Deliberating subscale are thought to perform worse in anxiety-provoking situations. Smouse and Munz (1968, 1969) used the AAT scores to categorize examinees as Facilitators, Non-Affecters, Deliberators, or High-Affecters.

(low-stakes). Smouse and Munz (1968, 1969) studied a psychology test, whereas Towle and Merrill (1975) studied a mathematics test (high-stakes). Thus, the degree to which test anxiety moderates the relationship between item position and test performance may be dependent on other factors related to the testing condition, such as testing stake.

Gender. Researchers have also studied the impact of item position, gender, and their interaction on test performance. Plake et al. (1982) found a significant interaction between gender and item position (even after controlling for test anxiety and knowledge of item order arrangement), with male examinees outperforming female examinees in two of three item order arrangements (random and easy-to-hard) on a mathematics test. They found the gender difference in test performance was most substantial when items were arranged from easiest to hardest – male examinees scored 13 points higher (on a 48-item test) than female examinees in the easy-to-hard item arrangement but scored less than 5 points higher in the other two item arrangements (random and spiral[4]). Thus, they found female examinees performed similarly and male examinees performed differentially across the three item order arrangements, suggesting male examinees were more susceptible to position effects then female examinees. Similar to Plake et al. (1982), Hambleton and Traub (1974), Plake and Ansorge (1984), and Plake, Patience, and Whitney (1988) also examined gender differences in test performance across different item order arrangements. Hambleton and Traub (1974) studied gender differences across two item order arrangements (easy-to-hard and hard-to-easy) on a mathematics test, Plake and Ansorge (1984) studied gender differences across three item order arrangements (easy-to-hard, random, and spiral) on an educational psychology test, and Plake et al.

[4] Items were grouped and arranged easy-to-hard in different blocks. Item blocks are then ordered such that each subsequent item block increased in overall difficulty.

(1988) studied gender differences across three item order arrangements (easy-to-hard, easy-to-hard within content, and spiral) on a General Education Development (GED) mathematics test. In contrast to Plake et al.'s (1982) findings, they all found male and female examinees performed similarly across different item order arrangements in their studies.

Researchers were also interested in the relationships among item position, gender, and item performance. Plake et al. (1988) compared the *b*-parameters of GED mathematics test items across male and female examinees in three item order arrangements (spiral, easy-to-hard, and easy-to-hard within content). They found only one and two of the *b*-parameters (out of twenty) significantly differed across male and female examinees in the spiral and easy-to-hard within content conditions, with these items having higher *b*-parameters for male examinees than female examinees. Unlike Plake et al. (1988), other researchers statistically modeled the relationships among item position, gender, and item performance using complex statistical models (e.g., GLMM). Qian (2014) explored the moderating effect of gender on the relationship between item position and item performance in two 2007 NAEP writing assessments (Grade 8 and Grade 12). Across both assessments, he found the essays administered exhibited negative position effects across male and female examinees, with scores on an essay being lower when the essay was administered later in the test period. He found, however, the negative position effects were stronger for male examinees than female examinees, suggesting male examinees were more susceptible to position effects than female examinees.

Bulut (2015) and Ryan and Chiu (2001) explored the moderating effect of item position on the relationship between gender and item performance. Ryan and Chiu found different item order arrangements (random and easy-to-hard within content) had little impact on gender DIF for items on the Midwestern Mathematics Placement Exam (MMPE). Bulut studied the impact of different test booklets of a verbal reasoning test (where the same set of item blocks was used but administered in different order) on gender DIF. His study has two major findings. First, he found the number, magnitude, and direction of DIF due to gender varied across different test booklets. For example, he found certain items favored male examinees in some test booklets but not in others. Second, he found the number, magnitude, and direction of DIF due to test booklet varied across males and females. For example, he found different items to be easier/harder in some test booklets than in others within each gender group but the flagged items were not the same across gender groups. Thus, the degree to which gender moderates the relationship between item position and test performance may be dependent on other factors related to the testing condition.

Ability. In addition to test anxiety and gender, researchers have also explored the impact of item position, ability, and their interaction on performance at the test and item level. At the test level, Klosner and Gellman (1973) found high- and low-ability examinees performed similarly on an achievement test across different item order arrangements. At the item level, several researchers have used complex statistical models to study the relationships among ability, item position, and item performance. Though not the purpose of their study, Weirich et al. (2017) found high-ability examinees were more susceptible to position effects than low-ability examinees. Debeer and Jannsen (2013) conducted two separate applied studies on position effects. In their first study, they examined position effects of items on a listening comprehension test. In their second study, they examined position effects of items on the 2006 PISA (math, reading, and science). Across both studies, they found the items, on average, exhibited negative position effects but not all examinees were equally susceptible to the position effects. Contrary to Weirich et al. (2017), they found high-ability examinees were less susceptible to the position effects than low-ability examinees. The results above provide some support for the moderating effect of ability on the relationship between item position and item performance. The direction of this relationship, however, remains unclear.

Hartig and Buchholz (2012) also examined position effects of items from the 2006 PISA (science). They conducted separate analyses for PISA science data obtained from 10 different countries and compared their findings across the different countries. In general, they found items exhibited negative position effects and examinees varied in their susceptibility to the position effects in all countries. They did not find a consistent relationship between ability and position effects, but rather found the relationship between ability and position effects differed across high- and low-performing countries. In low-performing countries (those with lower national PISA science average scores), they found high-ability examinees were actually more susceptible to position effects (more negative) than low-ability examinees. In high-performing countries (those with higher national PISA science average scores), they found ability was not correlated with position effects. Debeer et al. (2014) conducted a similar study where they examined position effects of items from the 2009 PISA (reading) across 65 countries. In all countries, they found negative position effects with examinees varying in their susceptibility to the position effects. Similar to Hartig and Buchholz (2012), they also found the relationship between ability and position effects differed across high- and low-performing countries. In high-performing countries, they found ability was positively related to position effects, whereas in low-performance countries, they found ability was negatively related to position effects. Thus, similar to test anxiety and gender, the influence of ability on position effects may also be dependent on other factors related to the testing condition.

Effort. Recent researchers studying position effects have started to examine person variables related to position effects in low-stakes testing conditions, such as test-taking motivation. The underlying idea behind this is simple: if position effects are truly due to an increase in fatigue (which is most plausible in low-stakes testing conditions), then we would expect variables related to fatigue, such as test-taking effort, to moderate the relationship between item position and item performance (i.e., position effects). Weirich et al. (2017) tested the latter hypotheses by statistically modeling the interactions among self-reported initial effort/change in effort, item position, and item performance on a low-stakes assessment. Although they did not find initial effort to moderate the relationship between item position and item performance, they did find change in effort moderated the relationship between item position and item performance. Examinees who exhibited a greater decrease in effort were more susceptible to position effects than examinees who exhibited a lesser decrease in effort. They found, however, change in effort did not fully explain the differences in position effects across examinees in their study. Thus, there may be other moderating person or item variables that were omitted from their study.

Qian (2014) also studied the moderating effect of effort on item position and item performance in a low-stakes assessment, where he used examinees' self-reported importance ratings (the importance of doing well on the test) as a proxy for effort. He found examinees who reported doing well on the test was very important to them were less susceptible to position effects than examinees who reported doing well on the test was not very important to them. Although Qian examined item performance, his results provided some support for Hambleton and Traub's (1974) original hypothesis about the relationships among item position, test importance, and test performance, in which they hypothesized that "the effect of item order on test performance is directly related to the importance a student attaches to the test" (pg. 40). Thus, at least in low-stakes testing conditions, researchers have found examinee effort (and proxies of examinee effort) moderated the relationship between item position and item performance.

In summary, researchers studying person variables and position effects have found inconsistent results. Although it is plausible for male examinees, examinees high in test anxiety, examinees low in ability, and examinees low in effort to be most susceptible to position effects, the inconsistencies of the results make such claims questionable. Additionally, because these researchers studied test items of varying content and type, it is plausible for certain item variables to also moderate the relationship between position effects and test/item performance, in addition to person variables. The researchers above did not examine item variables in their studies, nor did they consider the potential interactions among person variables, item variables, and position effects.

Item Variables. Because most researchers focused on person variables, there is limited research on the relationships between item variables and position effects. Of the studies conducted, researchers focused on studying the relationships among item type, item content, and position effects. Kingston and Dorans (1984) compared the change in item difficulty when items were placed later versus earlier on the test across different item types on the Graduate Record Examination (GRE) and found verbal and analytical items exhibited greater changes in difficulty compared to quantitative items. Le (2007) conducted a similar study on the 2006 PISA science items and found open-response items and "knowledge about science" items[5] exhibited greater changes in item difficulty compared to other item types (multiple-choice, complex multiple choice, closed response) and item content ("knowledge of science"). Davis and Ferdous (2005) also found only the reading items but not the math items on a standardized state achievement test exhibited significant changes in item difficulty. Based on these findings, item content and item type are potential moderators of the relationship between position effects and test performance.

Besides the studies I summarized above, the research on item variables and position effects has not been well established. These initial studies provide support for the need to further examine item type/content and position effects; however, they do not provide us with a comprehensive understanding of the underlying mechanisms of these relationships. For example, why are items of certain type and content more prone to position effects compared to other items? Are items of a certain type and content more cognitively demanding, making them more susceptible to position effects? Researchers may need to examine additional item variables in order to fully answer the latter two questions. Similar to those who studied person variables, the researchers above did not

[5] Knowledge about science items include those related to physical systems, living systems, and earth and space systems, whereas knowledge of science items include those related to scientific enquiry, scientific explanations, and science and technology.

examine person variables in their studies, nor did they consider the potential interactions among person variables, item variables, and position effects.

Position Effects and Low-Stakes Testing

Previous research indicates certain person and item variables may be related to position effects; however, depending on the stakes of the test, certain person and item variables may be more or less relevant. For example, if a test is administered in a high-stakes testing condition, any position effects found on the test are unlikely due to low effort, as examinees taking a high-stakes test are likely to put forth considerable effort. In contrast, if a test is administered in a low-stakes condition, any position effects found on the test are unlikely due to high test anxiety, as examinees taking a low-stakes test are likely to have low test anxiety. Thus, depending on the testing stakes, certain person and item variables may serve as more or less plausible explanations for the position effects. Previous researchers have not considered this distinction when studying the relationships among item/person variables and position effects, which may be one reason for the inconsistent results found across the many studies.

In the context of low-stakes testing, researchers have found effort to decline for at least some examinees as the test progresses (e.g., Bovaird, 2002; Pastor, Ong, & Strickman, in press), with change in effort being related to position effects (Weirich et al., 2017). If change in effort is truly related to position effects in low-stakes testing, then we would expect certain person and item variables that are related to effort in low-stakes testing to also be related to position effects. To that end, several researchers have identified different item variables related to effort in low-stakes testing, which may help shed light on why some items might be more susceptible to position effects than others.

These researchers focused on examining the relationships among various item features (e.g., item length) and the amount of effort examinees put forth on items. Across these studies, researchers used solution behavior (SB) indices as a measure for effort. SB indices are created by dichotomizing the item response time distributions into either rapid-guess responses, which are responses so fast that the item could not be fully read or considered, or solution behavior responses, which characterizes all other responses.

Bovaird (2002) used different item features to predict the proportion of rapid-guessing responses on each item on an Abstract Reasoning Test. In all conditions (power and speeded), he found item position was a consistent significant predictor of rapid-guessing responses on each item, and in some conditions (speeded), he found item difficulty and item working memory load (i.e., a measure of the number of rules required of the examinee to answer the item correctly) were significant predictors of rapid-guessing responses on each item. Thus, examinees were likely to rapidly guess on items placed later on the test, difficult items, and mentally taxing items. Wise (2006) used the average SB index for each item (or the average proportion of examinees engaging in SB for each item; response time fidelity [RTF]) as the dependent variable in his analyses, with various item features serving as independent variables. He found item position and item length significantly predicted RTF of items. Setzer, Wise, van den Heuvel, and Ling (2013) conducted a similar study and found not only item position and item length but also ancillary reading material (e.g., the presence of diagrams or charts) significantly predicted RTF of items. Across both studies, examinees were likely to put forth more effort on short items, items placed earlier on the test, and items containing no ancillary reading materials.

Unlike Wise (2006) and Setzer et al. (2009), Wise, Pastor, and Kong (2009) used the SB index for each item as the dependent variable in a GLMM, with various person and item variables as the independent variables. They found item length, item position, item graphic, and number of response options significantly predicted effort, with examinees putting forth more effort on items at the beginning of the test, easy items, short items, items containing a graphic, and items with small number of response options. Interestingly, they also found examinees put forth more effort on items at the end of the test if they contained graphics, which suggested examinees were likely to put similar amount of effort on items with graphics regardless of item position. Although the researchers above focused on examining the relationships among item features and effort on an item, they provide some insight into why some items might be more or less prone to position effects, particularly in low-stakes testing. For example, the magnitude and direction of position effects found in low-stakes testing may depend on both the motivation levels of the examinees and the item features associated with the set of items.

Several researchers have also identified different person variables related to effort in low-stakes testing. The most prevalent person variable that has been heavily studied is gender. DeMars, Bashkov, and Socha (2013) did a systematic review of studies examining gender and effort and found female examinees were more likely to put forth more effort than male examinees in low-stakes testing conditions. Thus, the magnitude and direction of position effects in low-stakes testing may not only depend on the motivation levels of the examinees and the item features associated with the set of items but also depend on the gender of the examinees. To more directly investigate the latter hypothesis, researchers should consider the interplay among item features, position effects, and examinee characteristics, including their genders and levels of motivation.

Need for Study

Despite the abundance of research on position effects, there are still several gaps in the literature. First, because of the lack of focus on item variables, additional research on how other item variables (such as item features) relate to position effects is warranted. Certain items may be more prone to position effects than others because they share similar item features. Thus, it is plausible for certain item features to be related to position effects, particularly in the context of low-stakes testing. Second, researchers had only exclusively examined person or item variables in previous studies but never both within a single comprehensive study. It is plausible the relationships among person variables and position effects are moderated by certain item variables, and vice versa. For example, we may find a significant relationship between position effects and item length, such that longer items are more susceptible to position effects, but find the latter relationship differs across different subgroups of examinees, such as those with different levels of motivation. Thus, in order to study these complex relationships, researchers may need to examine both person and item variables simultaneously when studying the causes of position effects, particularly in the low-stakes testing context.

In an attempt to fill the current gaps in the literature, in my dissertation I investigated the relationships among item variables, person variables, and position effects simultaneously on item responses obtained from two low-stakes assessments. I addressed the first gap in the literature by empirically investigating the relationships between item features and position effects. I addressed the second gap in the literature by empirically investigating whether certain person variables moderate the relationships between item features and position effects. Specifically, I had three primary research questions. I discuss each of them below.

Research Question One: How do certain item features relate to position effects? Although researchers found certain item types were more prone to position effects than other item types (e.g., reading and verbal items), no researchers have actually explored the underlying reasons as to why this might be. One possible reason is certain item types may share similar item features related to position effects. For example, Kingston and Dorans (1984) and Davis and Ferdous (2005) both found reading items, which are generally lengthier items compared to other items, most susceptible to position effects. Thus, item length, along with other item features, may be related to position effects, particularly in low-stakes testing. To evaluate the latter hypotheses, I manipulated the order of items on two assessments, administered the two assessments to undergraduate students in low-stakes testing conditions, and evaluated the relationships among four different item features and position effects using a GLMM. The four item features I focused on in my dissertation were a) the total word count in the item stem and options (Item Length), b) the number of response options (Number of Options), c) the perceived amount of mental taxation required (Mental Taxation) to complete the item, and d) the presence of graphics (e.g., graphs or figures; Item Graphic). I chose to study these specific item features because they are universal item features that are applicable to all items, regardless of the content or type of the specific items studied.

If I find certain item features are related to position effects, then my findings will have several implications for psychometricians. First, psychometricians could fix the position of these items across forms when developing new test forms. Second, psychometricians could fix the position of these items across field and operational forms. Third, psychometricians could limit the use of these items in CAT or CBT, where they are likely to be administered in different order across examinees. Finally, when equating or pre-equating is necessary, psychometricians could limit the use of these items as anchor items.

Research Question Two: How do person variables moderate the relationships among item features and position effects? Given previous research on item features, item position, examinee effort, and gender in low-stakes testing conditions, it is plausible for these variables to have moderating effects on one another, similar to the moderating effect reported in Wise et al. (2009)'s study. For example, we may find the effects of position and combined effects of position and item features to differ across different subgroups of examinees. Certain subgroups of examinees, particularly those differing in gender or their levels of motivation, may be more or less susceptible to position effects, and even more or less susceptible to the impact of item features on position effects, than other subgroups of examinees. Thus, in addition to exploring the relationships among certain item features and position effects, I also explored whether these relationships were moderated by gender and effort in the same GLMM.

If I find gender and effort moderate the relationships among certain item features and position effects, then my findings will have a number of implications for researchers. First, although researchers have broadly studied the variables above, they had not examined these variables in single comprehensive modeling framework. Thus, the findings will provide researchers with the first empirical examination of these variables in a single comprehensive modeling framework. Second, the findings will help further the research on examinee effort in low-stakes testing, potentially uncovering another negative consequence of low effort in low-stakes testing. Thus, interventions used to increase effort may also be potentially used to mitigate position effects. Finally, the findings will help further the research of position effects, particularly in understanding the cause or multiple causes of position effects.

Research Question Three: How do these relationships differ across two tests of varying content? Across the studies I summarized in this chapter, researchers found inconsistent results on the impact of position effects on various outcomes, such as test performance and item difficulty. One possible reason for this inconsistency may be due to differences in the tests (and testing conditions) they studied. For example, researchers studied the impact of position effects in tests of various stakes (e.g., low-stakes versus high-stakes), conditions (e.g., power versus speeded), and content (e.g., mathematics versus verbal), which could all theoretically have contributed to their inconsistent results. To evaluate the latter hypothesis, I evaluated whether the relationships above varied across the two assessments (which differ in content) in my dissertation.

Unlike previous studies, where test content, stakes, and conditions were often confounded, I administered the two assessments under the same conditions (power) and stakes (low-stakes) in my study. Thus, if I find differences between the two sets of results, I could attribute those differences to other differences between the tests, such as test content. If I find the two sets of results differ from one another, then my findings will have two implications for researchers. First, the findings will further uncover the complexity of position effects. Positions effects may be test-specific and researchers may need to consider other test characteristics in addition to item features, effort, and gender when studying position effects. Second, the findings will potentially provide an empirical explanation for the inconsistent results found in previous studies.

CHAPTER 3

METHOD

I collected and analyzed data from undergraduate students at a mid-sized public university to answer my three primary research questions. In the following sections, I first describe my data collection procedure. Second, I describe the set of cognitive and non-cognitive measures I administered. Third, I describe the design of my study. Fourth, I describe the sample of the undergraduate students in my study. Finally, I describe in detail my data analytic plan.

Data Collection

I collected data from undergraduate students during an institution-wide, mandatory testing session known as Assessment Day. Students at the university are required to participate in Assessment Day twice during their academic career: once as incoming freshmen and again once they have completed 45 to 70 credit hours. On each Assessment Day, students are exempt from classes and randomly assigned to testing sessions based on the last four digits of their student identification number. Within each testing session, students are asked to complete a battery of cognitive and non-cognitive assessments over the course of two hours. Trained proctors are present in every testing session to provide standardized instructions to the students and to ensure the testing condition is as consistent as possible across testing sessions (i.e., students are quiet and putting forth effort). The scores obtained from examinees on Assessment Day are considered low-stakes because they are only used for accountability purposes and do not have any impact on the students' academic transcripts or graduation requirements.

For my study, I administered two different cognitive tests followed by a set of non-cognitive measures to undergraduate students during the Fall 2018 Assessment Day, which consisted mainly of incoming freshmen students. I administered the set of cognitive and non-cognitive measures at the start of each testing session. Thus, students had not completed any cognitive or non-cognitive measures prior to completing my set of cognitive and non-cognitive measures. I describe the cognitive and non-cognitive measures I administered below.

Measures

At the start of each testing session, students completed either the American Experience test (AMEX) or the Environmental Stewardship Reasoning and Knowledge Assessment (ESRKA) followed by a set of non-cognitive items used to measure their effort (5-item subscale from the Student Opinion Scale; Sundre & Moore, 2002) and change in effort (single item). I administered the set of cognitive and non-cognitive measures in paper-and-pencil format. Thus, students saw multiple items on each page and recorded their responses on a separate scantron provided to them by the trained proctors.

AMEX. The AMEX (version 4) is a 40-item, multiple-choice test used to assess students' knowledge of American history, politics, and society. Faculty at the university wrote the items to align with the General Education Cluster Four (social and cultural processes) learning objectives. During the Fall 2018 Assessment Day, students were given 50 minutes to complete the test. All students completed the AMEX within 41 through 47 minutes of the allotted time. The reliability estimate (Cronbach's alpha) of the AMEX for this sample was .81.

ESRKA. The ESRKA (version 3) is a 45-item, multiple-choice test used to assess students' environmental stewardship reasoning and knowledge abilities. Faculty at the university created the test to support the university's strategic emphases on environmental stewardship. They wrote the items to align with General Education Cluster Three (natural world) learning objectives and the Office of Environmental Stewardship learning objectives. During Fall 2018 Assessment Day, students were given 60 minutes to complete the test. All students completed the ESRKA within 50 through 60 minutes of the allotted time. The reliability estimate (Cronbach's alpha) of the ESRKA for this sample was .78.

Student Opinion Scale (SOS). The Student Opinion Scale (SOS; Sundre & Moore, 2002) is a 10-item measure comprised of two subscales: Effort and Importance. Because I was interested in measuring examinee effort, I only administered the 5-item Effort subscale to the undergraduate students. Students were asked to respond to items that assess the amount of effort they gave on the test (e.g., "I gave my best effort on this test") using a scale of 1 (*Strongly Disagree*) to 5 (*Strongly Agree*), with high scores indicating high levels of effort. The reliabilities (Cronbach's alpha) of the effort scores were .79 and .76 for the AMEX and ESRKA, respectively.

Change in Effort. I used a single item to measure the extent to which examinee effort changes across the testing session. Students were asked to choose from three response options that best describe their level of effort during the test. The three response options available to students were a) my effort level did not change during the test, b) I put forth less effort as the test progressed, and c) I put forth more effort as the test progressed.

Gender. I used the university records of students to identify students' self-reported gender. Students were only able to choose between two gender groups (male and female). Thus, I was unable to include other gender groups.

Research Design

Item Order, Test Forms, and Form Administration. There are two common methods researchers have used to manipulate the order of items when studying position effects. The first method is to create different test forms with items randomly scrambled across test forms. The second method is to first group the items into blocks and then create different test forms with item blocks in different orders across test forms. For the purpose of my dissertation, I chose to adopt the second method. My decision was based on three primary reasons. First, due to logistical restrictions, I could only administer the two tests in my study in paper-and-pencil formats. Thus, the first method was neither feasible nor possible. Second, by grouping items into blocks, it allowed me to keep the (local) context of some items the same across test forms and examinees. Thus, this reduced the contamination of context effects on position effects. Third, by administering item blocks in different order across test forms, I could administer the items in all block positions with a relatively small number of test forms.

I categorized the AMEX and ESRKA items into four different blocks and manipulated the order of the item blocks to create different test forms. There were 24 possible combinations of item blocks across test forms, resulting in 24 possible test forms for each test[6]. To ensure adequate sample sizes, I chose to only create four different test

[6] I calculated this number by taking the factorial of the number of item blocks, which represented the number of different ways the blocks can be arranged without repeating any block combination. I had four item blocks so the factorial of four was 24.

forms based on 4 out of the 24 possible block order combinations (see Table 2). I chose these four specific block order combinations because it allowed all item blocks to appear in every possible block position (first, second, third, and fourth portions of the test) at least once across test forms. This ensured individual items were administered in different portions of the test at least once across students during each test administration.

AMEX. I categorized the 40 AMEX items into four different item blocks, with each item block comprised of 10 AMEX items. I categorized items 1 through 10 into an item block (A), items 11 through 20 into an item block (B), items 21 through 30 into an item block (C), and items 31 through 40 into an item block (D). During the Fall 2018 Assessment Day, the trained proctors administered the four AMEX test forms in a spiral order to students to ensure the student-form ratios were approximately equal (e.g., first student was given AMEX Form A, second student was given AMEX Form B, etc.).

ESRKA. I categorized the 45 ESRKA items into four different item blocks, with each item block comprised of 11 ESRKA items. Because the ESRKA consisted of 45 items, I had to keep the position of the first item constant (first position) across all test forms to ensure an equal number of items within each item block. I then categorized items 2 through 12 into an item block (A), items 13 through 23 into an item block (B), items 24 through 34 into an item block (C), and items 35 through 45 into an item block (D). During the Fall 2018 Assessment Day, the trained proctors administered the four ESRKA test forms in a spiral order to students to ensure the student-form ratios were approximately equal (e.g., first student was given ESRKA Form A, second student was given ESRKA Form B, etc.).

Item Features. Recall I was interested in examining four different item features: item length, number of options, presence of graphics, and mental taxation. For item length, number of options, and presence of graphics, I individually inspected and recorded the specific features above for each item. For mental taxation, I and four other raters inspected each item and rated how much mental effort we perceived is required to answer the item correctly using the method proposed by Wolf, Smith, and Birnbaum (1995).

Item Length. I defined item length as the number of words in the item stem and response options. For the AMEX items, the minimum and maximum item lengths were 26 and 230, respectively. For the ESRKA items, the minimum and maximum item lengths were 17 and 278, respectively.

Number of Options. For the AMEX, the items had between 4 and 5 response options. For the ESRKA, all items had 4 response options. Thus, I could not examine the relationship between number of options and position effects for the ESRKA.

Presence of Graphics. I defined presence of graphics as a dichotomous variable indicating whether an item was presented with a graphic (e.g., tables, charts, graphs). For the AMEX items, five items contained some sort of graphic. For ESRKA items, three items contained some sort of graphic.

Mental Taxation. I defined mental taxation using the definition proposed by Wolf et al. (1995). They defined mental taxation as the amount of mental effort an examinee must put forth to achieve the correct answer on an item. Although mental taxation is correlated with item difficulty, they argued the two concepts are theoretically independent. A multiple-choice item may be considered low in difficulty and low in mental taxation; however, the same multiple-choice item with an added graph may still be considered low in difficulty but high in mental taxation. That is, the two multiple-choice items may be testing the same concepts, which may be easy, but the second multiple-choice item requires the examinee to examine a graph, which may be more mentally taxing.

To determine the mental taxation of AMEX and ESRKA items, I and four other raters adopted the approach proposed by Wolf et al. (1995). We independently inspected and rated each item using the following criteria: "Rate each question based on the mental energy required to solve it. Use a 10-point scale ranging from low (1) to high (10). Consider how much mental energy a student would have to expend to come to a correct answer." The raw mental taxation ratings for all items are presented in the Appendix. The mean correlations of ratings across all pairs of raters were .654 and .731 for the AMEX and ESRKA, respectively. I computed the mean mental taxation rating for each item and used them in the primary analyses.

Participants

A total of 1,028 undergraduate students completed the AMEX and 1,092 undergraduate students completed the ESRKA during the Fall 2018 Assessment Day (see Table 3 for final samples size by form). Of those who completed the AMEX, 58% self-identified as female and 87% self-identified as White/Caucasian. Of those who completed the ESRKA, 60% self-identified as female and 85% self-identified as White/Caucasian. The demographics of the two samples were representative of students at the university.

Data Analytic Plan

Generalized Linear Mixed Models. Position effects are commonly examined within IRT models, specifically the one parameter logistic (1PL) model, parameterized in a generalized linear mixed modeling (GLMM) framework. By specifying the 1PL model within a GLMM framework, researchers can enter in predictors (e.g., person and item characteristics) of item responses as either fixed or random effects. There are two common GLMM parameterizations of the 1PL model often seen in position effect research. I first describe the two common GLMM parameterizations of the 1PL. Then, I discuss the specific GLMM parameterization of the 1PL I used to address my three primary research questions.

Persons Random and Items Fixed. The first GLMM parameterization of the 1PL treats item responses as nested within persons. In the most simplistic model, only random effects for persons and fixed effects for items are included[7]. This GLMM parameterization is specified as (with no position effect included in the model):

$$\eta_{ij} = \ln\left[\frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)}\right] = \theta_i + \beta_j \tag{1}$$

[7] To clarify what is meant by “random effects for persons” and “fixed effects for items” it is helpful to think about “persons” and “items” as categorical predictors. Because these predictors are categorical, they can be represented in the model as a series of dummy coded variables: a dummy coded variable for each person and a dummy coded variable for each item. In the model in Equation 1 the “person” predictor is considered a random predictor, thus the effects associated with the “person” predictor are random effects (and to simplify notation, only the random effect for person *i*, not the predictor itself, is shown). In contrast, the “item” predictor is considered a fixed predictor, thus the effects associated with the “item” predictor are fixed effects. In Equation 1, both the item dummy-coded variables and item fixed effects are shown.

$$\theta_i \sim N(0,\sigma_\theta^2)$$
$$\beta_j = \sum_{q=1}^{Q} \beta_q X_{jq},$$

where η_{ij} is the log odds of obtaining a correct response to item j by person i, Y_{ij} is the response of person i to item j, θ_i is the random effect for person i, β_j is fixed effect for item j, σ_θ^2 is the variance of the random effects for persons, Q is the total number of dummy codes (with total number of dummy codes = total number of items), q is the specific dummy code associated with item j, β_q is effect associated with X_{jq}, and X_{jq} is a dummy-coded variable indicating the item response associated with item j where $X_{jq} = 1$ when $q = j$ and $X_{jq} = 0$ otherwise. A hypothetical data matrix for fitting the GLMM parameterization above is included in Table 4 (adapted from Albano, 2013).

This GLMM parameterization produces parameter estimates that align with the 1PL model. The random effects for persons are analogous to the theta estimates obtained from the 1PL model, and the fixed effects for items are analogous to the negative of the item difficulty estimates obtained from the 1PL model. Equation 1 can be extended to include a position parameter to examine the influence of position on item responses:

$$\eta_{ijk} = \ln\left[\frac{P(Y_{ijk}=1)}{1-P(Y_{ijk}=1)}\right] = \theta_i + \beta_j + \delta$$

$$\theta_i \sim N(0,\sigma_\theta^2) \tag{2}$$
$$\beta_j = \sum_{q=1}^{Q} \beta_q X_{jq},$$

where η_{ijk} is now log odds of obtaining a correct response to item *j* at position k (with positions ranging from k =1 to *K*) by person i and δ is the position parameter, where position can be entered into the model either as a categorical or continuous variable.

When item position is entered into the model as a categorical variable, the δ parameter is specified as:

$$\delta = \sum_{r=2}^{K} \gamma_r P_r, \tag{3}$$

where P_2 through P_K are the position dummy-coded variables with the first position serving as the reference position, β_j is now the item easiness parameter for item *j* in the reference position, and γ_r is the difference in the log odds of obtaining a correct response at position *r* relative to the reference position. For example, if there are four block positions, three dummy-coded position variables (P_2, P_3, P_4) are included in the model and three position effects are estimated, with each effect representing the change or difference in the log odds of obtaining a correct response in the respective block position relative to the reference block position (see Table 4 for example of hypothetical data matrix).

When item position is entered into the model as a continuous variable, the δ parameter is specified as:

$$\delta = \gamma P, \tag{4}$$

where γ is the linear effect of position, P is a single variable with values equal to position k - 1, and β_j is now the item easiness parameter for item *j* in the first position. For example, if there are 4 block positions, one continuous position variable (P) is included in the model and only one position effect is estimated, which represents the change in the log odds of obtaining a correct response for every one unit increase in block position (see Table 4 for example of hypothetical data matrix). The position effect can be specified as non-linear by including polynomials, which would suggest a non-linear relationship between item responses and item position.

In Equations 3 and 4 there are three main effects in the model: a main effect for persons (represented by the random effects for persons), a main effect of items (represented by the fixed effects for items), and a main effect for position (represent by one or more fixed effects for position). Because no interactions among these predictors are included, the position effect in Equations 3 and 4 is considered to be the same for all persons and items. To allow the position effect to vary across persons, the interaction between position and persons would need to be included in the model. For example, *P* in Equation 4 would be multiplied by the dummy-coded variables for persons, which are not explicitly shown in Equation 4, to create the position by person interaction terms. The coefficients associated with the interaction would be random effects and can be described as “person by position random effects”. The addition of these random effects would allow the linear or non-linear position effect to vary across persons. If position effects are specified to vary across persons, researchers can ascertain the correlation between ability and examinee-specific position effects (i.e., the correlation between person random effects and the person by position random effects).

Researchers are also able to allow the position effect to vary across items by including additional interaction terms into the model, making the position effect item dependent. For example, Equation 3 can be extended to allow item-specific position effects by including the interaction terms among each item-specific dummy code (e.g.,

X_{j1} - X_{j4} in Table 4) and position-specific dummy code (e.g., $P_2 – P_4$, in Table 4) in the model. Equation 4 can be extended to allow item-specific position effects by including the interaction terms among each item-specific dummy code and the position variable (e.g., P in Table 4) in the model.

Because the GLMM parameterizations represented by Equations 2 and 3 and Equations 2 and 4 include item-specific dummy codes in the model, researchers are unable to include other item characteristics as predictors[8] (Meulders & Xie, 2004). To explore other item characteristics in addition to item position requires a GLMM parameterization where the effect of both persons and items are specified as random effects. I discuss this GLMM parameterization below.

Persons Random and Items Random. The second GLMM parameterization models both persons and items as random effects[9]. This GLMM parameterization is specified as (with no position effect included in the model):

[8] To understand why this is true, we can think about a simpler example, where one is predicting student math scores using a regression model and the school a student attends is included as a predictor in the regression model (using a series of school-specific dummy codes). Because school is represented by the dummy-codes, all information about schools is captured by the set of school dummy codes. Thus, no other school-level predictors can be included in the regression model. In the GLMM example, all information about items is captured by the set of item dummy codes. Thus, no other item-level predictors can be included in the GLMM.

[9] To clarify what is meant by "random effects for persons" and "random effects for items" it is helpful to think about "persons" and "items" as categorical predictors. Because these predictors are categorical, they can be represented in the model as a series of dummy coded variables: a dummy coded variable for each person and a dummy-coded variable for each item. In the model in Equation 5 the "person" predictor is considered a random predictor, thus the effects associated with the "person" predictor are random effects. The "item" predictor is also considered a random predictor, thus the effects associated with the "item" predictor are random effects. To simplify notation, neither the dummy-coded variables for persons or items are included in Equation 5.

$$\eta_{ij} = \ln\left[\frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)}\right] = \theta_i + \beta_j$$

$$\begin{aligned} &\theta_i \sim N(0,\sigma_\theta^2) \\ &\beta_j \sim N(0,\sigma_\beta^2), \end{aligned} \tag{5}$$

where η_{ij} is the log odds of obtaining a correct response for item j by person i, θ_i is the random effect for person i, β_j is the random effect for item j, σ_θ^2 is the variance of the person random effects (i.e., variance of thetas), and σ_β^2 is the variance of the item random effects (i.e., variance of items' easiness values). Equations 1 and 5 are similar, with the exception of how items are specified. In Equation 1, items are specified as fixed effects, whereas in Equation 5, items are specified as random effects. Equation 5 can also be extended to include a position effect parameter to examine the influence of position on item responses:

$$\eta_{ijk} = \ln\left[\frac{P(Y_{ijk}=1)}{1-P(Y_{ijk}=1)}\right] = \theta_i + \beta_j + \delta$$

$$\begin{aligned} &\theta_i \sim N(0,\sigma_\theta^2) \\ &\beta_j \sim N(0,\sigma_\beta^2), \end{aligned} \tag{6}$$

where σ_β^2 is variance of item easiness at the initial or reference position (depending on how position is entered into the model below). The parameter δ is modeled as a function of item position and differs in interpretation depending on whether item position is included as categorical or continuous variable (same as the first GLMM parameterization).

When item position is entered into the model as a categorical variable, the δ parameter is specified as:

$$\delta = \gamma_0 + \sum_{r=2}^{K} \gamma_r P_r, \qquad (7)$$

where γ_0 is the log odds of obtaining a correct response when an item is in the reference position, and γ_r and P_2 through P_K are the same as in Equation 3. When item position is entered into the model as a continuous variable, the δ parameter is specified as:

$$\delta = \gamma_0 + \gamma_1 P, \qquad (8)$$

where γ_0 is the log odds of obtaining a correct response when the item is in first position, $\gamma_1$is the linear effect of position, and P is a single variable with values equal to position $k - 1$. An important difference between Equations 2 and 6, where item position is included in the models, is the specification of the item effects. In Equation 2, item effects are specified as fixed effects. In Equation 6, item effects are specified as random effects.

Similar to the previous parameterizations, where persons are random and items are fixed, there are three main effects included in Equations 7 and Equation 8: a main effect for persons (represented by the random effects for persons), a main effect of items (represented by the random effects for items), and a main effect for position (represent by one or more fixed effects for position). Researchers are also able to allow the position effect to vary across items under this GLMM parameterization by including the item by position interaction effects. For example, Equations 7 and 8 can be extended to allow for item-specific position effects by adding an item by position random effect term into the equations:

$$\begin{gathered} \delta = \gamma_0 + \sum_{r=2}^{K} \gamma_r P_r + \varepsilon_{jr} P_r \\ \varepsilon_{jr} \sim N(0, \sigma^2_{\varepsilon_r}) \end{gathered} \qquad (9)$$

$$\delta = \gamma_0 + \gamma_1 P + \varepsilon_j P$$
$$\varepsilon_j \sim N(0, \sigma_\varepsilon^2) \qquad (10)$$

In Equation 9, γ_r is now the average difference in the log odds of obtaining a correct response at position *r* relative to the reference position, ε_{jr} is the deviation of item *j* from the position *r* specific position effect, $\sigma_{\varepsilon_r}^2$ is the variance of the position *r* specific position effect across items, and γ_0 and P_2 through P_K are the same as Equation 7. In Equation 10, $\gamma_1$is now the average linear effect of position, ε_j is the deviation of item *j* from the average position effect, σ_ε^2 is the variance of the position effect across items, and *P* is the same as in Equation 8.

The item by position random effect in Equations 9 and 10 can be further specified to correlate with the item random effect (i.e., the random effect of position for each item can be correlated with the easiness for each item). For example, if the item by position random effect (ε_j) is specified to be correlated with the item random effect (β_j), the item by position and item random effects are assumed to follow a multivariate normal distribution:

$$\delta = \gamma_0 + \gamma_1 P + \varepsilon_j P$$
$$\begin{pmatrix} \beta_j \\ \varepsilon_j \end{pmatrix} \sim MVN(0, \textstyle\sum_\beta)$$
$$\textstyle\sum_\beta = \begin{pmatrix} \sigma_\beta^2 & \\ \sigma_{\beta\varepsilon}^2 & \sigma_\varepsilon^2 \end{pmatrix}, \qquad (11)$$

where $\sigma_{\beta\varepsilon}^2$ is the covariance between the item by position and item random effects. When no additional predictors are included in the model, the $\sigma_{\beta\varepsilon}^2$ term has a meaningful interpretation and represents the relationship between item easiness at the initial position and the item-specific position effects. When additional predictors are included in the model, the $\sigma^2_{\beta\varepsilon}$ term has a less meaningful interpretation and represents the relationship between item easiness at the initial position and item-specific position effects once controlling for the predictors in the model.

Preliminary Data Analysis. For all GLMMs below, I included random effects for persons, items, and items by position and allowed the covariance(s)[10] between the item random effects and item by position random effects to be freely estimated. Prior to estimating primary GLMMs of interest below, I evaluated the nature and significance of the position effect of each test by estimating two GLMMs with position as the only predictor in the models. I treated position as both a continuous and categorical predictor and compared the two model results to determine how position should be specified in subsequent analyses.

If the GLMM with categorical position variable is preferred, it would imply the form of the position effect is non-linear, suggesting the change in the log odds of obtaining a correct response is not constant across positions. Of particular interest in this model is significance of the K -1 main effects of the position dummy codes. A significant negative position effect would indicate the log odds of obtaining a correct response decreases from the respective position relative to the reference position. A significant positive position effect would indicate the log odds of obtaining a correct response increases from the respective positive relative to the reference position.

[10] When position is treated as linear, only one covariance is estimated. When position is treated as categorical, more than one covariance is estimated.

If the GLMM with continuous position predictor is preferred, it would imply the form of the position effect is linear, suggesting the change in the log odds of obtaining a correct response is constant across positions. Of particular interest in this model is the significance of the main effect of position. A significant negative linear position effect would indicate the log odds of obtaining a correct response decreases by a constant amount for every one unit increase in position. A significant positive linear position effect would indicate the log odds of obtaining a correct response increases by a constant amount for every one unit increase in position. I refer to the GLMM with position as a continuous predictor as M1 and the GLMM with position as categorical predictor as M2 in the results.

Research Question One: **How do certain item features relate to position effects?** To answer my first research question, I estimated a GLMM with position, all item features (item length, mental taxation, number of response options, and graphic), and all two-way interactions between position and item features as predictors of AMEX and ESRKA item responses. For example, a GLMM with just linear position, item length, and their interaction as predictors is specified as:

$$\eta_{ijk} = \ln\left(\frac{P(Y_{ijk}=1)}{1-P(Y_{ijk}=1)}\right) = \theta_i + \beta_j + \varepsilon_j P + \gamma_0 + \gamma_1 P + \gamma_2 ItemLength + \gamma_3 P * ItemLength$$

$$\theta_i \sim N(0,\sigma_\theta^2) \quad \begin{pmatrix} \beta_j \\ \varepsilon_j \end{pmatrix} \sim MVN(0, \textstyle\sum_\beta) \tag{12}$$

$$\textstyle\sum_\beta = \begin{pmatrix} \sigma_\beta^2 & \\ \sigma_{\beta\varepsilon}^2 & \sigma_\varepsilon^2 \end{pmatrix}.$$

Of particular interest in this model is the significance of the two-way interaction effect between position and item length. A significant two-way interaction effect between position and item length would indicate the magnitude of the position effect differs across items of different lengths.

In the full GLMM, person ability, item easiness, and item by position were specified as random effects, and position, item length, mental taxation, number of response options, graphic, and all possible interactions between item features and position were specified as fixed effects. I refer to the GLMM with position, all item features, and all possible two-way interactions of interest as M3 in the results.

Research Question Two: **How do person variables moderate the relationships among item features and position effects?** To answer my second research question, I added each person characteristic (examinee effort, change in effort, and gender) separately as a predictor in the model and estimated three different GLMMs[11]. For example, a GLMM with just linear position, item length, effort, and their interactions as predictors is specified as:

$$\eta_{ijk} = \ln\left(\frac{P(Y_{ijk}=1)}{1-P(Y_{ijk}=1)}\right) = \theta_i + \beta_j + \varepsilon_j P + \gamma_0 + \gamma_1 P + \gamma_2 ItemLength + \gamma_3 Effort + \gamma_4 P * ItemLength + \gamma_5 ItemLength * Effort + \gamma_6 P * Effort + \gamma_7 P * ItemLength * Effort \qquad (13)$$

[11] I initially estimated a GLMM that included position, all item features, all person characteristics, all two-way interactions of interest, and all three-way interactions of interest as predictors of item responses. Not surprisingly, given the number of estimated effects, I ran into scaling issues, even after standardizing all continuous predictors, which ultimately resulted in convergence issues. To address this issue, I estimated three separate GLMM for each assessment, where each person characteristic was entered individually in three separate models.

$$\theta_i \sim N(0,\sigma^2_\theta) \quad \begin{pmatrix} \beta_j \\ \varepsilon_j \end{pmatrix} \sim MVN(0,\textstyle\sum_\beta)$$

$$\textstyle\sum_\beta = \begin{pmatrix} \sigma^2_\beta & \\ \sigma^2_{\beta\varepsilon} & \sigma^2_\varepsilon \end{pmatrix}.$$

Of particular interest is the significance of the three-way interaction effect between position, effort, and item length. A significant three-way interaction effect would indicate the two-way interaction effect between position and item length differs across high-effort versus low-effort examinees.

In the full GLMMs, person ability, item easiness, and item by position were specified as random effects, and position, item length, mental taxation, number of response options, graphic, effort, change in effort, gender, all possible two-way interactions between position, item features, and person characteristics, and all possible three-way interactions between position, item features, and person characteristics were specified as fixed effects. I refer to the GLMM with position, all item features, effort, all two-way interactions of interest, and all three-way interactions of interest as M4 in the results. I refer to the GLMM with position, all item features, change in effort, all two-way interactions of interest, and all three-way interactions of interest as M5 in the results. I refer to the GLMM with position, all item features, gender, all two-way interactions of interest, and all three-way interactions of interest as M6 in the results.

Research Question Three: **How do these relationships differ across two assessments of varying content?** To answer my third research question, I compared the AMEX GLMM model results to the ESRKA GLMM model results.

Estimation. I estimated all GLMMs using maximum likelihood (ML) estimation based on the Laplace approximation in R Version 3.4.2 (R Core Team, 2017) via the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014). I chose to use the ML estimator with the Laplace approximation because it is the only estimation method that can be used to estimate a GLMM where items and persons are both specified as random effects. Under this estimation method, the Laplace approximation is used to approximate the likelihoods of the fixed and random effects in the model, which are then optimized to obtain the approximate values of the maximum likelihood estimates for the fixed and random parameters (Doran, Bates, Bliese, & Dowling, 2007). The likelihood ratio test can be used to compare nested GLMMs that differ in fixed effects; however, the likelihood ratio test cannot be used to test the significance of the variance components when using ML with the Laplace approximation (De Boeck et al., 2011).

In summary, I estimated six separate GLMMs (M1 – M6) for each assessment to answer my primary research questions (see Table 5 for summary of GLMMs). I interpreted the models in the following order. First, I interpreted the main effects from M1 and M2 results to evaluate significance of the position effect and compared M1 and M2 via a likelihood ratio test (LRT) to determine how the position effect should be specified in subsequent analyses (linear or non-linear) on the AMEX and ESRKA (preliminary). Second, I interpreted the two-way interactions among item features and position from M3 to evaluate the combined effects of item features and position on item responses on the AMEX and ESRKA (first research question). Finally, I interpreted the three-way interactions among position, item features, and each person characteristic from

M4, M5, and M6 to evaluate the combined effects of position, item features, and person characteristics on item responses on the AMEX and ESRKA (second research question).

CHAPTER 4

RESULTS

For each assessment, I estimated a total of six GLMMs to address my three primary research questions (M1-M6; see Table 5). I present the results in the following order. First, I present the descriptive statistics and preliminary model results, where I examined the nature and significance of the position effects on AMEX and ESRKA performance (M1 and M2). Second, I present the model results associated with my first research question, where I examined whether position effects on the AMEX and ESRKA were related to item characteristics (M3). Third, I present the model results associated with my second research question, where I examined whether the relationships among position effect and item characteristics were moderated by person characteristics (M4, M5, and M6). Finally, I compare and contrast the model results across the AMEX and ESRKA to address my third research question.

Preliminary Analyses

Descriptive Statistics. With respect to the item variables, there were ~5 examinees who had missing data on at least one item for the AMEX (~ .01%) and ESRKA (~ .01%). During the item scoring procedure, missing responses were scored as incorrect, which aligns with the standard scoring procedure for the AMEX and ESRKA. With respect to the person variables, there were 47 (5%) and 16 (2%) examinees with missing data on the change in effort and effort variables, respectively, for the AMEX and 25 (2%), 1 (<.01%), and 13 (1%) examinees with missing data on the change in effort, gender, and effort variables, respectively, on the ESRKA. Missing data on the person variables were left unaltered because the estimation method used assumes the missing data mechanism is missing at random (De Boeck et al., 2011). Thus, parameter estimates were based on only the non-missing data provided by each examinee.

Potential issues with multicollinearity among item variables were considered by inspecting the correlations among the item variables. The correlation values for the item variables for the AMEX and ESRKA are presented in Table 6 and Table 7, respectively. The patterns of correlations among the item variables were similar across the two assessments and therefore will be discussed in tandem below. With the exception of mental taxation, all item variables were minimally correlated with one another (*r*'s < |.10|). Mental taxation was moderately positively correlated with item length (AMEX = .603; ESRKA = .703) and graph (AMEX = .373; ESRKA = .512). Raters assigned higher mental taxation ratings to longer items and items with a graph than shorter items and items without a graph. The patterns of correlations indicated no potential issues with multicollinearity for the item variables. Thus, I retained all item variables and used them as predictors in subsequent analyses.

To evaluate the relationships among all person variables, I conducted a series of regression and chi-square analyses. The relationships among all person predictors were similar across both assessments and therefore will be discussed in tandem below. Gender explained less than < .001% of variance in effort scores on both tests, whereas change in effort explained 17% and 15% of variance in effort scores on the AMEX and ESRKA, respectively. Gender was also not related to change in effort for the AMEX [$\chi^2(2)$ = 1.583, p = .453] and ESRKA [$\chi^2(2)$ = 0.828, p = .661]. Because I entered each person variable individually (rather than all at once) into M4, M5, and M6, multicollinearity was not an issue; however, these results indicated person variables were not highly related to one another.

Out of the 1,028 examinees who completed the AMEX, 669 (65%) examinees reported their effort level remained constant, 220 (21%) examinees reported their effort level decreased, and 92 examinees (9%) reported their effort level increased during the testing period. Out of the 1,092 examinees who completed the ESRKA, 713 (65%) examinees reported their effort level remained constant, 261 (24%) examinees reported their effort level decreased, and 93 (9%) examinees reported their effort level increased during the testing period. The mean effort score on the AMEX and ESRKA was 20.364 (*SD* = 3.411) and 19.759 (*SD* = 3.271), which indicated examinees for the most part put forth high effort but varied moderately in their effort on both tests. Taken together, the majority of examinees reported putting forth a consistently high amount of effort during the testing period, with about ~20% of examinees putting forth less effort and ~10% of examinees putting forth more effort as the tests progressed across both tests.

Position Effect. I evaluated the nature and significance of the position effect on each assessment by estimating two GLMMs, where I only entered in position as a predictor of item responses (M1 and M2). I treated position as a continuous variable (uncentered with values equal to 0, 1, 2, and 3) in M1 and as a categorical variable via three dummy coded variables (with the reference position being block one) in M2. I compared the two sets of results to evaluate the nature (functional form) of the position effect on each assessment using LRTs. The M1 and M2 parameter estimates for the AMEX and ESRKA are presented in Table 8.

AMEX. In M1, the linear position effect was negative and statistically significant (-0.065), which implied the log odds of obtaining a correct response decreased by -0.065 for every one unit increase in block position on the AMEX. The linear position effect varied from item to item, with 95% of items in the population having position effects ranging from -0.190 to 0.06 (see Figure 1). The correlation between item random effects and item by position random effects was –0.360, which suggested the position effect was pronounced (more negative) for easier than harder items (at first position) on the AMEX. In M2, the change in the log odds of obtaining a correct response from block one to block two was negative but non-statistically significant (Position2 = -0.041, 95% CI [-0.180, 0.098]); however, the changes in the log odds of obtaining a correct response from block one to block three and block one to block four were both negative and statistically significant (Position3 = -0.120, 95% [-0.321 – 0.091]; Position4 = -0.190, 95% [-0.577 – 0.197]). Thus, the log odds of obtaining a correct response was lower when the same items were placed in the third or fourth block relative to the first block on the AMEX[12]. The changes in the log odds of obtaining a correct response on adjacent blocks were similar in magnitude and in the same direction, which suggested visually the position effect is linear in nature (see Figure 2). The LRT results indicated M2 did not fit the data statistically significantly better than M1, which suggested statistically the position effect is linear in nature [$\chi^2(9)$ = 3.473, p = .943]. Thus, in all subsequent analyses for the

[12] To test the overall significance of the categorical position effect in M2 for AMEX and ESRKA, I compared M2 to a null model (with no predictors) via a LRT. For both tests, M2 fit significantly better than the null model, which indicated the overall categorical position effect was statistically significant [AMEX = $\chi^2(12)$ = 55.253, p = < .001; ESRKA == $\chi^2(12)$ = 88.722, p = < .001].

AMEX (M3-M6), I treated position as a continuous variable and modeled the position effect as linear.

ESRKA. In M1, the linear position effect was negative and statistically significant (-0.059), which implied the log odds of obtaining a correct response decreased by 0.059 for every one unit increase in position on the ESRKA. The linear position effect varied from item to item, with 95% of items in the population having position effects ranging from -0.214 to 0.096 (see Figure 3). The correlation between item random effects and item by position random effects was –0.140, which suggested the position effect was pronounced (more negative) for easier than harder items (at first position) on the ESRKA. In M2, the change in the log odds of obtaining a correct response from block one to block two was negative but non-statistically significant (Position2 = -0.063, 95% [-0.424 – 0.298]); however, the changes in the log odds of obtaining a correct response from block one to block three and block one to block four were both negative and statistically significant (Position3 = -0.099, 95% [-0.466 – 0.268]; Position4 = -0.184, 95% [-0.631 – 0.263]). Thus, the log odds of obtaining a correct response was lower when the same items were placed in later blocks than the initial block on the ESRKA. Similar to the AMEX, the changes in the log odds of obtaining a correct response on adjacent blocks were similar in magnitude and in the same direction, which suggested visually the position effect was linear in nature (see Figure 2). The LRT results indicated M2 did not fit the data statistically significantly better than M1, which suggested statistically the position effect is linear in nature [$\chi^2(9)$ = 13.270, p = .151]. Thus, in all subsequent analyses for the ESRKA (M3-M6), I treated block position as a continuous variable and modeled the position effect as linear.

Primary Analyses

Research Question One: How do certain item features relate to position effects? For each assessment, I estimated a GLMM that included position, all item features, and all possible two-way interactions among position and item features as predictors of item responses (M3). I treated position, item length, and mental taxation as continuous predictors and graph (0 = *does not contain graph*; 1 = *contains graph*) and option[13] (0 = *4 response options*; 1 = *5 response options*) as categorical predictors in the model. With the exception of position, I standardized all continuous predictors prior to entering them into the model to help with model convergence[14]. The M3 parameter estimates for the AMEX and ESRKA are presented in Table 9. I examined the significance of the two-way interactions among position and item characteristics for each test to address my first research question.

AMEX and ESRKA. Across both tests, the position by item length interaction effect was negative and statistically significant. The position by item length interaction

[13] For M3 through M6, I only included option as a predictor for only the AMEX because all ESRKA items had the same number of response options.

[14] The *lme4* package (Bates et al., 2014) uses nonlinear optimizers during estimation to obtain the variance-covariance matrices of the random effects. Given the complexity of such algorithms, the authors of the package noted it is difficult to evaluate their convergence and therefore possible for the user to obtain a false positive convergence warning. To check whether a convergence warning is a false positive, they recommend estimating the model using different nonlinear optimizers and comparing the model results. If the model results are consistent across the different nonlinear optimizers (e.g., similar parameter estimates), the convergence warning should be considered a false positive. I obtained a convergence warning for the following models: M3 – M6 for AMEX and M4 – M6 for ESRKA. I adopted the recommendation above and estimated the seven models above using five different nonlinear optimizers. The model results were nearly identical across the five different nonlinear optimizers for all models. Thus, I considered the convergence warnings to be false positives and proceeded forward with interpreting the parameter estimates.

effect was estimated to be -0.037 and -0.055 for the AMEX and ESRKA, respectively. To better interpret these two interaction effects, I plotted the probability of obtaining a correct response at each position across three different length values on the AMEX (see Figure 4) and ESRKA (see Figure 5). The three item length values represent a short-, medium-, and long-length item on each test, which I determined after visual inspection of the AMEX and ESRKA item length distributions. As depicted in Figure 4 and Figure 5, the position effect appeared to be more pronounced (more negative) for longer items compared to shorter items on both tests.

For the AMEX, the position by mental taxation, position by option, and position by graph interaction effects were not statistically significant, which indicated the position effect did not differ across items of varying mental taxation, items with a graph, and items with varying number of response options. For the ESRKA, the position by mental taxation and position by graph interaction effects were statistically non-significant, which indicated the position effect did not differ across items of varying metal taxation and items with a graph.

Research Question Two: How do person variables moderate the relationships among item features and position effects? For each assessment, I estimated three separate GLMMs, where each person characteristic was entered individually in three separate models (M4-M6). I included position, all item features, a person characteristic, all possible two-way interactions among position, item features, and person characteristics, and all possible three-way interactions among position, item features, and person characteristic as predictors of item responses in each of the three GLMM. Again, with the exception of position, I standardized all continuous predictors prior to entering them into the models.

In M4, I entered in effort and treated effort as a continuous predictor in the model. In M5, I entered in change in effort and treated change in effort as a categorical predictor (0 = *did not change*; 1 = *decrease in effort*; 2 = *increase in effort*) in the model, where chEffort1 equaled 1 when change in effort equaled 1 (*decrease in effort*), 0 otherwise, and chEffort2 equaled 1 when change in effort equaled 2 (*increase in effort*), 0 otherwise. Thus, I used the *did not change* category as the reference category. In M5, I entered in gender and treated gender as a categorical predictor (0 = *female*; 1 = *male*) in the model. The M4, M5, and M6 parameter estimates for the AMEX and ESRKA are presented in Table 10 and Table 11, respectively. I examined the significance of the three-way interactions among position, item characteristics, and person characteristics on each assessment to address my second research question.

AMEX and ESRKA. Out of all three-way interactions estimated across M4 through M6 for the AMEX and ESRKA, I found none to be statistically significant, which implied effort, change in effort, and gender did not moderate the two-way interaction effects among item characteristics and position on either test. In other words, the moderating effects of item characteristics on the relationship among position and item responses did not vary across examinees of different effort levels, patterns of change in effort, and gender on both tests.

Although unrelated to the primary research question, a few two-way interactions among position and person characteristics were statistically significant on the ESRKA, but not the AMEX. Specifically, the position by effort interaction effect was statistically significant and positive (0.035). To better interpret this interaction effect, I plotted the probability of obtaining a correct response at each position across five different levels of effort (see Figure 6). As depicted in Figure 6, the position effect was more pronounced (more negative) for low-effort examinees than high-effort examinees on the ESRKA. The position by chEffort1 interaction effect was also statistically significant and negative (-0.064). To better interpret this interaction effect, I plotted the probability of obtaining a correct response at each position for examinees reporting no change in effort and examinees reporting a decrease in effort (see Figure 7). As depicted in Figure 7, the position effect was more pronounced (more negative) for examinees reporting a decrease in effort than examinees reporting no change in effort.

Research Question Three: How do these relationships differ across two tests of varying content? The AMEX and ESRKA results were consistent with one another. Across both tests, there was a significant negative position effect and easier items tended to be more prone to the position effect than harder items. Additionally, item length was the only significant moderator of the position effect, with the direction of the moderating effect being similar in nature. Moreover, effort, change in effort, and gender did not moderate the two-way interactions among item characteristics and position effects.

CHAPTER 5

Discussion

Previous researchers have primarily examined the underlying causes of position effects through an item or person perspective (e.g., Bulut, 2015; Kingston & Dorans, 1984; Qian, 2014). These researchers tended to focus on exploring the relationships between position effects and item or person variables but none have examined the relationships among position effects, item variables, and person variables simultaneously. In this dissertation, I evaluated the underlying causes of position effect in a low-stakes testing context through an integrated perspective, where I viewed position effect through both the item and person perspectives. I administered items from two assessments in different orders to two groups of examinees, examined the presence of position effects, and evaluated the degree to which position effects were moderated by different item and person variables.

In the following sections, I first discuss the general findings of each research question. Then, I discuss the implications of my study from a measurement perspective. That is, how does my study contribute to position effect research? Then, I discuss the implications of my study from a practitioner perspective. That is, how does my study inform the practice of current psychometricians in industry? Finally, I discuss the limitations of my study and provide a few recommendations for future directions in this area of research.

General Findings by Research Question

Preliminary. In this study, I evaluated the significance and form of position effects on items from the AMEX and ESRKA in a low-stakes testing context. Similar to other researchers who studied position effects in low-stakes testing (e.g., Weirich et al., 2017), I found items exhibited significant negative linear position effects on both assessments, with the magnitude of the position effects varying from item to item. The practical significance of the variability in item difficulty at first position and position effects across items can be seen in Figures 1 and 3 for the AMEX and ESRKA, respectively. There were large differences in item difficulty at the first position for both assessments (i.e., intercepts varied considerably across items); however, there were small differences in position effects across items (i.e., slopes did not vary much across items) for both assessments. These results indicated items varied substantially in difficulty at first position and certain items were more prone to position effects than other items but only to a small degree on both assessments.

Interestingly, I found item difficulty at first position to be negatively correlated with position effects on both assessments. Easier items were more prone to position effects than harder items. There are three plausible explanations for this. The first explanation is related to examinee guessing behavior and only possible if we assume guessing behavior increases as the test progresses. If easy and hard items are placed at the beginning of the test and we assume the majority of examinees are not guessing, the item difficulty will not be influenced by guessing– easy items will appear easy and hard items will appear hard. In contrast, if easy and hard items are placed at the end of the test and we assume examinees guess more on items at the end of the test, the item difficulty is influenced by guessing – hard items will still appear hard but easy items will now appear harder. In fact, when random responders (i.e., examinees who randomly respond to items) were present, Mislevy and Verhlest (1990) found randomly responding had small impact on the Rasch difficulties of hard items but large impact on the Rasch difficulties of easy items. The second explanation is related to the impact of correct guessing on item difficulty estimates. For example, if correct guessing is present (regardless of position), the change in the log odds of correct response for easy versus hard items differs across positions, with larger differences in the log odds for easy items and smaller differences in the log odds for hard items. The third explanation is related to item difficulty and features. It is plausible easier items have particular features in common. To explore this, I examined the relationships among item difficulty and the four item features to evaluate whether easier items tended to share certain features. Item difficulty was correlated with the four items features in expected ways (i.e., easier items tended to less lengthy and mentally taxing); however, all correlations were small in nature (correlations ranged from |.5| through |.3|). Thus, I did not find item difficulty to be strongly related to the four item features across the two assessments.

To better understand why items varied in position effects, I examined the features of three items with the most negative position effects on the AMEX and ESRKA. For the AMEX, the item with the largest negative position effect contains no graphic and is shorter in length (72 words), harder (*P* = .51), and less mentally taxing (average rating = 3.4), whereas the items with the second and third largest negative position effect contain no graphics but are longer in length (> 150 words), easier (*P* = ~.70), and less mentally taxing (average rating ~ 4) compared to other items. For the ESRKA, the item with the largest and third largest negative position effect have similar features, such that they both contain no graphic and are longer in length (> 200words) and more mentally taxing (average rating = 7.4) compared to other items. They differ in difficulty though - the item with the largest negative position effect (*P* = .37) is harder compared to other items, whereas the item with the third largest negative position effect (*P* = .61) is easier compared to other items. The item with the second largest negative position effect contains no graphic and is shorter in length (63 words), easier (*P* = .63), and more mentally taxing (average rating = 6.4) compared to other items.

Based on these observations, it appears four out of six items with the most negative position effects on the two assessments are long, easy, and contain no graphic. These observations align with the correlation findings (as discussed above) and partially align with the findings in research question one (as discussed in the next section). Interestingly, the other two items have features that do not align with the trend above. For example, these items tend to be of shorter length compared to the other four items. Yet, these items still have either the largest or second largest negative position effect on the two assessments. Thus, it is plausible for these items to contain features not examined in this study that may make them more prone to position effects than other items.

Research Question One: How do certain item features relate to position effects? In this study, I examined whether four different item features (item length, number of response options, mental taxation, and graphic) moderated the degree to which AMEX and ESRKA items were impacted by position effects in the low-stakes testing context. Recall I chose to study these specific item features for two reasons. First, these item features are universal to all items – they are not content specific. Second, researchers have found these item features to be related to effort in low-stakes testing contexts. Thus, given effort has been found to be related to position effects, I expected item features related to effort (such as those above) to also be related to position effects. Based on the relationships between effort and item features (as previously reviewed), I expected items of longer length, items with more response options, items requiring more mental taxation, and items containing no graphic to be more susceptible to (negative) position effects than items of shorter length, items with fewer response options, items requiring less mental taxation, and items containing a graphic due to low examinee effort or decrease in examinee effort across the testing period.

Contrary to my expectations, I found item length to be the only significant moderator of position effects on the two assessments, with longer items being more prone to the position effects than shorter items. The practical significance of the moderating effect of item length on position effects can be seen in Figures 4 and 5 for the AMEX and ESRKA, respectively. For a typical short item, the position effect is only slightly negative for the AMEX and is almost null for the ESRKA; however, for a typical long item, the position effect is much more negative for both the AMEX and ESRKA. Thus, at least with this sample and testing context, lengthier items are more prone to position effects than other items on the two assessments. This finding closely aligns with the findings from Kingston and Dorans (1984) and Davis and Ferdous (2005), where they found reading items to be most prone to position effects. It is possible for the reading items in their studies to be lengthier than other items that were studied. Thus, the reason they found reading items to be more prone to position effects may be because they shared a common item feature: they are all long items.

It was surprising number of response options, mental taxation, and graphic were not significant moderators of position effects on the two assessments. With respect to number of response options and graphic, the non-significant effects may be due to the small number of items with those features on the two assessments. There were only five and three items on the AMEX and ESRKA, respectively, with some sort of graphic and there were only two items on the AMEX with five response options (as opposed to four response options). The small variability of items with these features may have limited the statistical power to detect any potential moderating effects. With respect to mental taxation, the non-significant effect may be due to the validity of the mental taxation ratings. The average correlations across all raters were moderately high; however, there were considerable inconsistencies in ratings between pairs of raters. For example, the lowest correlations between any two raters were .401 and .519 for the AMEX and ESRKA, respectively. Thus, these pairs of raters might have based their ratings on different criteria, which would be problematic from a validity perspective.

Research Question Two: How do person variables moderate the relationships among item features and position effects? In this study, I examined whether three person variables (change in effort, effort, and gender) moderated the degree to which item features were related to position effects in the low-stakes testing context. I chose to study effort and change in effort because both variables have been found to be related to both position effects and item features in low-stakes testing contexts (e.g.,Weirich et al., 2017). Thus, I expected effort to moderate the relationships among item features and position effects. I chose to study gender because gender has been found to be related to effort, which in turn, has been found to be related to position effects in low-stakes testing contexts. Thus, I expected gender to also moderate the relationships among item features and position effects.

Contrary to my expectations, change in effort, effort, and gender were non-significant moderators of item features and position effects. The relationships of item features and position effects did not differ across examinees with varying effort levels, change in effort patterns, and genders. There are a few explanations for the non-significant results. With respect to effort and gender, examinees had median effort scores of 21 and 22 for the AMEX and ESRKA, with only minor gender differences in effort scores across the two tests. The latter is interesting because previous researchers have found significant gender differences in effort in low-stakes testing (DeMars et al., 2013); however, this was not true in my study. The small to moderate variability in effort scores and gender differences in effort scores may have limited the statistical power to detect any effect regarding effort, change in effort, and item features. With respect to change in effort, examinees were only asked about their change in effort using a single item at the end of the test. It is questionable whether examinees were able to accurately recall or honestly convey their effort pattern.

Research Question Three: How do these relationships differ across two tests of varying content? In this study, I compared the results from research questions one and two across the two tests to evaluate the stability of the relationships explored. Contrary to previous researchers who had reported mixed findings on position effects, I found the relationships explored in my study to be stable across the two tests with respect to statistical significance. For example, the relationships among item features and position effects varied in their magnitude, which is to be expected, but were similar in their statistical significance on both tests. I found items exhibited significant linear negative position effects, with position effects being stronger for easier and longer items than harder and shorter items. These results provided some support for the generalizability of the findings across tests of varying content. It should be noted the administration of the tests was similar but not identical. Thus, the stability of the results across the tests may be due to other common testing factors, such as both tests being administered in the low-stakes context and/or both tests being administered to undergraduate students.

Implications

The results from my study should not be viewed as limited but rather informative from both a measurement and practitioner perspective. From a measurement viewpoint, my study uniquely contributes to position effects research. From a practitioner viewpoint, my findings may be used to inform future practices in testing. I further discuss these implications below.

Measurement Viewpoint. My study contributes to position effects research in three primary ways. First, I explored the underlying causes of position effects through an integrated perspective, which no researchers have previously done. Unlike the item or person perspective, the integrated perspective combines the latter perspectives and allows researchers to consider both item and person variables as potential underlying causes of position effects. Second, I demonstrated the utility of a GLMM parameterization that aligns with the integrated perspective, where both person and item variables can be included in the model. This allows researchers to explore the relationships among person variables, item variables, and position effects simultaneously within a single modeling framework. The GLMM parameterization I used can be extended to allow position effects to vary across examinees rather than items, which would enable researchers to examine variability in position effects across examinees and potential variables that may explain that variability. Third, I examined the moderating effects of different item and person variables on position effects in a low-stakes testing context. Previous researchers have focused primarily on how person variables, such as test anxiety, were related to position effects, whereas in my study, I focused primarily on how item variables, such as item length, were related to position effects and whether these relationships were moderated by three person variables important in a low-stakes testing context. No other researchers have explored these relationships before.

Practitioner Viewpoint. My study informs the practice of testing practitioners in two primary ways. First, although I studied position effects in low-stakes testing context, the item features I studied were not specific to low-stakes testing contexts but universal to all items. I found easier and longer items to be more prone to position effects than harder and shorter items; therefore, testing practitioners should be cautious about administering these particular items in different positions across different forms or administrations. In contrast, I found items differing in number of response options, amount of mental taxation, and the presence of a graphic to be similarly prone to position effects;; therefore, testing practitioners should feel more comfortable administering these particular types of items in different positions across different forms or administrations. If item scrambling is required for test security purposes, testing practitioners may consider fixing the positions of long and easy items and scrambling the positions of the remaining items across different forms or administrations[15]. Following the latter procedure would likely help mitigate the adverse impact of position effects as those items are likely to be most prone to position effects than other items. I should note the significant effects (position main and position by item length interaction effects) across both tests were small in magnitude. Thus, it is questionable whether these effects would actual make a practical difference in the estimation of person ability estimates. Researchers should consider the magnitude of the significant effects when adopting the recommendation above.

Second, testing practitioners may not only use the different GLMMs described in my study to detect position effects and their relationships to different item and person variables, but they can also use them to obtain more accurate person ability estimates. For example, if a practitioner finds a significant position effect, they can explicitly include the position effect in the GLMM to statistically control for that effect. This would allow testing practitioners to obtain person ability estimates adjusted for position effect, which would be more trustworthy. Additionally, if a practitioner finds the significant effect to be moderated by a person or item variables, they can include those additional person or item variables into the model to statistically control for those effects. This would allow testing practitioners to person ability estimates adjusted for the combined effects of position, item variables, and person variables, which, again, should be more

[15] In this scenario, there could still be other context effects that may influence the item statistics of long or easy items; however, by fixing the positions of long or easy items across forms or administrations, it would least reduce the position effects detected found in this study.

trustworthy[16]. Note the latter would result in person ability estimates that are adjusted based on certain examinee characteristics, which may not be defensible in practice. Thus, I recommend the former approach (including position effect in the model) for potential operational purposes and the latter approach (including position effect, item variables, and person variables in the model) for potential research purposes. For research purposes, testing practitioners may want to compare statistically-adjusted person ability estimates to those obtained from a traditional Rasch model to evaluate the impact of position effects (and interactions of position effects and other variables) on estimation of person ability estimates and classifications of examinees. Doing the latter would allow testing practitioners to examine the impact of position effects on examinee outcomes.

Limitations and Future Directions

Limitations. My findings should be interpreted with some reservations. First, because I only studied the relationships among item features, person characteristics, position effects under one testing context and sample, the generalizability of my results should be questioned. It is plausible for the relationships among the item features and position effects studied in my study to differ across high- and low-stakes testing contexts. The person characteristics I studied were also chosen specifically based on previous research in low-stakes testing, which may not be applicable in the high-stakes testing

[16] To evaluate the impact of position effects on ability estimates, I compared ability estimates obtained from the GLMM with only item and person as predictors (traditional Rasch model) and GLMM with items, persons, position (linear), item length, and position by item length interaction as predictors. For both the AMEX and ESRKA, I found the rank-order of examinees based on ability estimates was nearly identical across the two GLMMs (*r*'s > .99), with average differences in ability estimates being < .001 and none larger than .03. Thus, at least for the two assessments studied in my dissertation, the position effects had essentially no impact on ability estimates.

contexts. Thus, my general results and recommendations for testing practitioners above should be taken with some caution.

Second, I only examined how each item feature was related to position effects individually. I did not consider the potential combined effects of different item features on position effects. It is plausible for items with different combinations of features to be more susceptible to position effects than others (e.g., a long item with a graphic may be more prone to position effect than a long item without a graphic). The latter relationships can be explored by including the interactions between two or more item features and position effects in the model as predictors of item responses (i.e., M3). Although it is possible to examine these relationships, they require modeling complex interactions, which may be difficult to interpret in a meaningful way. Moreover, if there are only a small number of items with certain combinations of features, researchers should be concern about statistical power (i.e., large standard errors for those interaction effects).

Third, I certainly did not study an exhaustive list of item features and person variables. I only examined four item features; but as discussed above, it is plausible for other item features, such as the linguistic features of items, to be related to position effects. I only examined three person characteristics; but as discussed before, it is plausible for other person characteristics, such as fatigue and interest, to be related to position effects. Given the latter possibilities, the GLMMs estimated in the study may be potentially misspecified because I may have failed to include all relevant variables related to position effects.

Fourth, similar to other previous studies, I did not completely isolate position effects from context effects with my research design. I attempted to control for local context effects by keeping the positions of items within each block constant; however, this did not completely control for context effects overall. For example, when item blocks were administered in different order, both the item characteristics (context) and quantity (position) of the set of items preceding each item block were different across forms. Thus, I did not fully isolate position effects because the contexts of the items were not kept constant across forms. It is possible for the results in my study to be attributable to context effects, position effects, or both.

Finally, I did obtain convergence warnings for seven out of the 12 GLMMs estimated. Although I concluded these convergence warnings were most likely false positives, it is best practice to evaluate this further. One possible approach would be to estimate the same models in another statistical program (e.g., Stata) and compare the results across the different statistical programs. If the results are comparable across different statistical programs, this will serve as additional evidence for the convergence warning to be false positives.

Future Directions. Based on the limitations above, I encourage future researchers studying position effects to continue to adopt an integrated perspective as demonstrated in this study. Under this perspective, future researchers should aim to replicate my findings and evaluate how other item and person variables are related to position effects under different tests, contexts, and samples. If future researchers find the relationships among item variables, person variables, and position effects to vary across different tests, contexts, and samples, it would imply the relationships between position effects and external variables are test specific, context specific, and/or sample specific.

This information can then be used to inform the specific item and person variables that should be examined in future studies.

Conclusion

Previous researchers have only either adopted an item or person perspective to position effects, where they focused on exploring the relationships among position effects and item or person variables separately. Unlike previous researchers, I adopted an integrated perspective to position effect, where I focused on exploring the relationships among position effects, item variables, and person variables simultaneously. It is through this perspective that I discovered easy and long items were most prone to position effects in the low-stakes testing context regardless of examinee gender and effort.

Table 1
Summary of Sample of Previous Research on Position Effects

Study	General Purpose	Sample	Content	Condition	Stake	Item Type	General Results
Impact of Position Effects on Test Performance							
Brenner (1964)	Examined the effects of five item arrangements (easy-to-hard, hard-to-easy, 10 easiest items placed in increasing difficulty order and the 30 items placed in random order, 10 hardest items placed in increasing difficulty order and the 30 items placed in random order, and in random order) on test performance.	Undergraduate college students	Psychology	Power	High	MCQ	Item arrangements had no significant impact on mean test performance.
Hambleton & Traub (1974)	Examined the effects of two item arrangements (hard-to-easy, easy-to-hard), test anxiety, and their interactions on mean test performance	11th grade students	Mathematics	Power	High	MCQ	Mean test performance on the easy-to-hard form was significantly higher than mean test performance on the hard-to-easy form. The interaction between item order and test anxiety had no significant impact on mean test performance.
Klimko (1984)	Examined the effects of three item arrangements (hard-to-easy, easy-to-hard, random), gender, test anxiety, cognitive characteristics, and their	Undergraduate college students	Psychology	Power	High	MCQ	Item arrangements (main effect and all interaction effects) had no significant impact on

Lane, Bull, Kundert, & Newman (1987)	**Study 1:** Examined the effects of five item arrangements, gender, and their interaction on mean test performance. Items were grouped based on their cognitive difficulty (application, comprehension, and knowledge items) and statistical difficulty (*P* indices). The five item arrangements were cognitive increasing/statistical increasing, cognitive decreasing/statistical decreasing, cognitive decreasing/statistical increasing, cognitive increasing/statistical decreasing, random order (statistical difficulty levels were ordered within the cognitive difficulty levels). **Study 2:** Examined the effects of six item arrangements, knowledge of item type (application, comprehension, and knowledge items), gender, and their interactions on mean test performance. The six item arrangements were cognitive increasing/statistical increasing (with and without item labels), cognitive decreasing/statistical decreasing (with and without item labels), and random order (with and without item labels).	Undergraduate college students	Education	Power	High	MCQ	**Study 1:** Item arrangements and their interaction with gender had no significant impact on mean test performance. **Study 2:** Item arrangements and their interactions with gender and label had no significant impact on mean test performance.

MacNicol (1956)	Examined the effects of three item arrangements (easy-to-hard, hard-to-easy, random) on mean test performance.	High school students	Verbal	Power	Unknown	MCQ	Mean test performance on the easy-to-hard form was significantly higher than mean test performance on the hard-to-easy form. Mean test performance on the random form was not significantly different than mean test performance on the easy-to-hard arrangement.
Mollenkopf (1950)	Examined the effect of item block arrangement on item statistics (*P* indices and *r* indices) under power and speeded conditions on two tests. Items were grouped into three item blocks and rearranged across two forms, such that the first item block was administered first in one form and last in another form, and vice versa, for each test.	11th and 12th grade students	Verbal and mathematics	Power and speeded	Unknown	MCQ	Under power conditions, only verbal items were found to be less difficult when placed earlier rather than later on the test (no position effect for mathematic items). Under speeded conditions, verbal and math items were found to be more difficult and discriminating when placed later rather than earlier on the test.
Monk & Stallings (1970)	Examined the effect of randomizing items on test performance by comparing 11 different pairs of test forms - each pair of test form had the same test items but the test items were ordered randomly across	Undergraduate college students	Geography	Power	High	MCQ	Randomization of items across test forms had no significant impact on mean test performance.

	the two forms.						
Sax & Cromack (1966)	Examined the effects of four item arrangements (easy-to-difficult, difficult-to-easy, easy items interspersed throughout, and random) on mean test performance under power and speeded conditions.	Undergraduate college students	Henmom-Nelson (mental ability)	Power and speeded	Low	MCQ	Under power conditions, no differences were found among mean test performance across all four forms. Under speeded condition, mean test performance on the easy-to-difficulty form was significantly higher than mean test performances on the other three forms.
Impact of Position Effects on Item Difficulty and Equating							
Davis & Ferdous (2005)	Examined the effect of item order on item difficulty (*P* indices and *b* parameters) during field and live testing, where the same items were first grouped in two equal-item blocks and administered in different positions in the two administrations.	3rd and 5th grade students	Reading and mathematics	Power	Unknown	MCQ	For Grade 3 math items, Grade 5 math items, and Grade 3 reading items, no significant differences in *P* indices and *b* parameters were found when items were administered in different positions. For Grade 5 reading items, a significant difference in the *P* indices and *b* parameters was found, with items becoming

							more difficult when moving from higher positions on the field test to lower positions on the live test.
Harris (1991)	Examined whether the equating relationships (equipercentile and IRT) of a new form back to an anchor form would depend on which version of the new form was used, with one version having the items in a set order and the other two versions scrambling the items in a random order for security purposes.	High school students	ACT (general)	Power	High	MCQ	Equating relationships depended on the new form version used in the equating, which indicated the item invariance assumption had been violated because of the different positioning of items across forms.
Huck & Bowers (1972)	Examined the effect of item order on item difficulty (*P* indices), where a balanced Latin design was used to rearrange the items, in two separate studies.	Undergraduate college students	Psychology	Power	High	MCQ	In both studies, item difficulty (*P* indices) did not significantly differ by item order.

Kingston & Dorans (1984)	Examined the effect of item order on item difficulty (*b* parameters) by item type (verbal, quantitative, and analytical) and equating across two test forms (A and B). Form A and Form B both contained four operational sections and one nonoperational section (pretest). Form A was administered to examinees with each receiving one of six different versions of the nonoperational section (which contained ~ one-half of items from the operation sections in Form B). Form B was administered to examinees with each receiving one of six different versions of the nonoperational section (which contained ~ one-half of items from the operation sections in Form A).	Undergraduate college students + others	GRE (verbal, quantitative, and analytical)	Power	High	MCQ	Verbal, quantitative, and analytical items all exhibited some degree of position effect when they were placed later rather than earlier on the test. The direction and magnitude of the position effects varied by item types, with some being more and some being less difficult when placed later on the test. Analytical items and verbal items exhibited the greatest amount of position effect than quantitative items. Form B was equated to Form A twice: once using Form B parameters obtained when items appeared in their operational location and once using Form B parameters obtained when items appeared in the nonoperational location (last section). The equating results (IRT true score) were most different between the two equating approaches for the analytical section,

							which indicated the impact of position effect on equating.
Kolen & Harris (1990)	Examined and compared preequating and postequating results on the ACT mathematic test.	High school students + others	Mathematics	Power	High	MCQ	Preequating methods resulted in more error than postequating methods. Item statistics on the pretest forms were different from the operational forms due to context/position effects, which explained the equating errors.
Meyers, Miller, & Way (2008)	Examined the effect of item position change on item difficulty (*b* parameters) during field and live testing, where items were administered in different positions in the two administrations.	3rd – 8th grade students	Reading and mathematics	Power	Low	MCQ	Change in item position was a significant predictor of change in item difficulty and explained > 50% of the variance in change in item difficulty. As change in item position increases, change in item difficulty also increases.

Yen (1980)	Examined the effect of context on item parameters and equating. The same sets of items were administered in similar and different contexts across multiple forms with a set of common items. The forms were equated back to a reference form and ICCs and TCCs of the manipulated items were compared.	4th and 6th grade students	Reading and mathematics	Power	Unknown	MCQ	ICCs and TCCs of manipulated items were more different when they were administered in different contexts than similar contexts.
Modeling Position Effects							
Albano (2013)	Examined position effects of quantitative and verbal items using GLMMs, where position effects were specified as linear and to vary across items.	Undergraduate college students + others	GRE (quantitative and verbal items)	Power	High	MCQ	Quantitative and verbal items exhibited negative position effects but varied in their magnitude across items.
Bulut, Quo, & Gierl (2017)	Examined position effects of reading passages and individual items via SEM, where position effect was specified as linear and to vary across items.	3rd grade students	Reading	Power	Unknown	MCQ	Reading passages and individual items exhibited negative position effects but varied in their magnitude across items.
Davey & Lee (2011)	Examined position effects of quantitative and verbal items. Item *P* indices were computed when items were placed in different positions and compared.		GRE (quantitative and verbal items)	Power	High	MCQ	Quantitative and verbal items exhibited negative position effects.

Debeer, Buchholz, Hartig, & Janssen (2014)	Examined position effect of reading items across 65 different countries using GLMMs, where position effects were specified as linear and to vary across examinees and schools.	High school students	Reading	Power	Low	MCQ	Reading items exhibited a negative position effect across all countries but varied in their magnitude across examinees and schools. At the examinee level, position effects were only slightly negatively correlated with ability, whereas at the school level, position effects were positively correlated with ability. Schools with lower average reading ability were more prone to the position effects than schools with higher average reading ability.
Debeer & Janssen (2013)	**Study 1:**Examined position effect of listening comprehension items via GLMMs, where position effect was specified as linear, quadratic, and cubic and to vary across students. **Study 2:** Examined position effect of PISA items via GLMMs, where position effect was specified as linear and to vary across students.	Study 1: 8th grade students Study 2: 8th grade students	Step 1: French listening comprehension Study 2: Mathematics, reading, and science	Power	Low	MCQ	Across both studies, the items exhibited negative position effects but not all examinees were equally susceptible to the position effects.

Weirich, Hecht, Penk, Roppelt, & Böhme (2017)	Examined whether initial and change in effort moderated position effect using GLMMs, where position effect was specified as linear and to vary across students/classrooms. Items were grouped into 31 item blocks and used to construct 31 test booklets, with each containing six blocks. Effort was measured twice during the testing period.	9th grade students	Scientific literacy	Power	Low	MCQ and open response	Scientific literacy items exhibited negative position effect, with the position effect moderated by change in effort but not initial effort. The position effect was less pronounced for those who decreased less in effort.
Hahne (2008)	Examined position effect of logic reasoning items using a LLTM, where position effect was specified as linear and same across items and examinees.	High school students	Logic reasoning	Power	Low	MCQ	No position effects were found.
Hohensinn, Kubinger, Reif, Schleiber, & Khorramdel (2011)	Examined position effect of mathematic items using a LLTM, where position effect was specified as linear and same across items and examinees. LLTM without the linear position effect and LLTM with the linear position effect were estimated and compared.	4th grade students	Mathematics	Power	Low	MCQ	The LLTM without the linear position effect fit the data as well as he LLTM with the linear position effect, which suggested no evidence for position effects.

Table 2

AMEX and ESRKA Block Positions Across Test Forms

		Block Position			
	Form	1	2	3	4
AMEX (40 items)	Form 1	Block A	Block B	Block C	Block D
	Form 2	Block D	Block A	Block B	Block C
	Form 3	Block C	Block D	Block A	Block B
	Form 4	Block B	Block C	Block D	Block A
	Form	1	2	3	4
ESRKA[a] (45 items)	Form 1	Block A	Block B	Block C	Block D
	Form 2	Block D	Block A	Block B	Block C
	Form 3	Block C	Block D	Block A	Block B
	Form 4	Block B	Block C	Block D	Block A

[a] I kept the position of item 1 (first position) on the ESRKA constant across all forms.

Table 3

Sample Sizes Across AMEX and ESRKA Forms

Test	Form 1	Form 2	Form 3	Form 4
AMEX	262 (25%)	257 (25%)	255 (25%)	254 (25%)
ESRKA	279 (25%)	273 (25%)	270 (25%)	270 (25%)

Table 4

Hypothetical data matrix for two examinees taking four items

					Item Dummy Codes				Position Dummy Codes		
Y_{ij}	i	j	k	P	X_{j1}	X_{j2}	X_{j3}	X_{j4}	P_2	P_3	P_4
1	1	1	1	0	1	0	0	0	0	0	0
0	1	2	2	1	0	1	0	0	1	0	0
1	1	3	3	2	0	0	1	0	0	1	0
1	1	4	4	3	0	0	0	1	0	0	1
1	2	1	4	3	1	0	0	0	0	0	1
1	2	2	3	2	0	1	0	0	0	1	0
0	2	3	2	1	0	0	1	0	1	0	0
1	2	4	1	0	0	0	0	1	0	0	0

Note. This data matrix aligns with a scenario where two examinees completed four items with the second examinee being administered the items in reverse order. Y_{ij} is the response of person i to item j. X_{j1}, X_{j2}, X_{j3}, and Xj_4 are the dummy-coded item variables indicating the item response associated with item j used to estimate the item effects for items one through four. $P_{2,}$ $P_{3,}$ and P_4 are the dummy-coded position variables indicating the administration of items in positions 2, 3, and 4 (i.e., k = 2, 3, 4) respectively and used to estimate the position effects for block position two, three, and four, with block position one as the reference position.

Table 5

Description of GLMMs Estimated for AMEX and ESRKA

Name	Predictor	Associated Research Question
M1	Linear Position	Preliminary
M2	Categorical Position	Preliminary
M3	Linear Position + Item Characteristics (including all interactions of interest)	1
M4	Linear Position + Item Characteristics + Effort (including all interactions of interest)	2
M5	Linear Position + Item Characteristics + Change in Effort (including all interactions of interest)	2
M6	Linear Position + Item Characteristics + Gender (including all interactions of interest)	2

Table 6

Descriptive and Correlation Statistics for AMEX Item Variables

	1	2	3	4
1. Length	1			
2. Mental	0.603	1		
3. Graph	-0.029	0.3733	1	
4. Option	0.085	0.0605	0.087	1
M	83.950	3.275	0.125	0.950
SD	44.294	1.422	0.331	0.218

Note. All descriptive and correlation statistics were computed based on the unstandardized variables.

Table 7

Descriptive and Correlation Statistics for ESRKA Item Variables

	1	2	3
1. Length	1		
2. Mental	0.703	1	
3. Graph	-0.053	0.5116	1
M	67.756	3.058	0.067
SD	45.792	1.537	0.249

Note. All descriptive and correlation statistics were computed based on the unstandardized variables.

Table 8

Fixed and Random Effects for the AMEX and ESRKA Position Effect GLMMs

	AMEX-M1 (*N* = 1,028)				ESRKA-M1 (*N* = 1,092)			
Parameter	Estimate	*SE*	*p-value*		Estimate	*SE*	*p-value*	
Fixed Effects								
Intercept	**0.414**	**0.139**	**.003**		**0.553**	**0.129**	**< .001**	
Position	**-0.065**	**0.014**	**< .001**		**-0.059**	**0.015**	**< .001**	
Random Effects		Corr.				Corr.		
Person Variance	0.541				0.406			
Item Variance	0.741				0.723			
Item by Position Variance	0.004	-0.360			0.006	-0.140		
	AMEX-M2 (*N* = 1,028)				ESRKA-M2 (*N* = 1,092)			
Fixed Effects								
Intercept	**0.403**	**0.139**	**.004**		**0.551**	**0.126**	**< .001**	
Position2	-0.041	0.034	.218		-0.063	0.041	.122	
Position3	**-0.115**	**0.036**	**.001**		**-0.099**	**0.041**	**.015**	
Position4	**-0.190**	**0.045**	**< .001**		**-0.184**	**0.045**	**< .001**	
Random Effects		Corr.				Corr.		
Person Variance	0.541				0.407			
Item Variance	0.741				0.722			
Item by Postion2 Variance	0.005	-0.320			0.034	-0.360		
Item by Position3 Variance	0.011	-0.040	0.830		0.035	-0.180	0.620	
Item by Position4 Variance	0.039	-0.400	0.960	0.830	0.052	-0.010	0.480	0.700

Table 9

Fixed and Random Effects for the AMEX and ESRKA Position Effect + Item Characteristics GLMMs

	AMEX-M3 (*N* = 1,028)			ESRKA-M3 (*N* = 1,092)		
Parameter	Estimate	*SE*	*p-value*	Estimate	*SE*	*p-value*
Fixed Effects						
Intercept	**-1.042**	**0.525**	**.047**	**0.491**	**0.138**	**< .001**
Position	**-0.103**	**0.063**	**.101**	**-0.056**	**0.013**	**< .001**
Length	0.107	0.157	.496	0.342	0.244	.162
Mental	-0.316	0.168	.061	-0.429	0.286	.134
Graph	**1.037**	**0.408**	**.011**	0.927	0.832	.265
Option	**1.397**	**0.541**	**.010**			
Position*Length	**-0.037**	**0.019**	**.049**	**-0.055**	**0.023**	**.016**
Position*Mental	0.018	0.020	.353	< .001	0.026	.987
Position*Graph	-0.049	0.048	.304	-0.038	0.077	.627
Position*Option	0.046	0.064	.478			
Random Effects		Correlation			Correlation	
Person Variance	0.542			0.406		
Item Variance	0.533			0.688		
Item by Position Variance	0.003	-0.482		0.002	-0.245	

Note. With the exception of position, all predictors were standardized prior to being entered into the model.

Table 10

Fixed and Random Effects for the AMEX Position Effect + Item Characteristics + Person Characteristics GLMMs (N = 1,028)

AMEX-M4				AMEX-M5				AMEX-M6			
Parameter	Estimate	*SE*	*p-value*	Parameter	Estimate	*SE*	*p-value*	Parameter	Estimate	*SE*	*p-value*
Fixed Effects				Fixed Effects				Fixed Effects			
Intercept	-1.028	0.531	.053	Intercept	-0.925	0.539	.087	Intercept	**-1.250**	**0.527**	**.018**
Position	-0.096	0.063	.126	Position	**-0.150**	**0.071**	**.036**	Position	-0.055	0.076	.473
Length	0.112	0.156	.473	Length	0.120	0.159	.449	Length	0.162	0.159	.309
Mental	-0.318	0.168	.058	Mental	-0.321	0.170	.059	Mental	**-0.365**	**0.171**	**.032**
Graph	**1.010**	**0.408**	**.013**	Graph	**0.996**	**0.414**	**.016**	Graph	**1.100**	**0.413**	**.008**
Option	**1.385**	**0.548**	**.011**	Option	**1.366**	**0.555**	**.014**	Option	**1.404**	**0.543**	**.010**
Effort	0.104	0.093	.262	ChEffort1	-0.459	0.237	.053	Gender	**0.492**	**0.183**	**.007**
Position*Length	**-0.039**	**0.019**	**.039**	ChEffort2	-0.345	0.333	.300	Position*Length	-0.043	0.022	.053
Position*Mental	0.017	0.020	.377	Position*Length	-0.031	0.022	.146	Position*Mental	0.031	0.023	.187
Position*Graph	-0.038	0.048	.433	Position*Mental	0.011	0.022	.621	Position*Graph	-0.061	0.057	.279
Position*Option	0.042	0.065	.517	Position*Graph	-0.028	0.055	.614	Position*Option	0.030	0.078	.698
Position*Effort	0.027	0.048	.577	Position*Option	0.111	0.074	.130	Position*Gender	-0.108	0.097	.263
Length*Effort	0.030	0.027	.276	Position*ChEffort1	0.164	0.122	.177	Length*Gender	**-0.140**	**0.055**	**.011**
Mental*Effort	-0.006	0.028	.842	Position*CHEffort2	0.246	0.174	.157	Mental*Gender	**0.119**	**0.056**	**.034**
Graph*Effort	-0.064	0.068	.346	Length*ChEffort1	-0.036	0.066	.587	Graph*Gender	-0.144	0.137	.294
Option*Effort	0.157	0.093	.091	Mental*ChEffort1	0.006	0.068	.935	Option*Gender	-0.009	0.183	.961
Position*Length*Effort	< .001	0.015	.995	Graph*ChEffort1	0.211	0.166	.204	Position*Length*Gender	0.017	0.029	.570
Position*Mental*Effort	-0.007	0.015	.643	Option*ChEffort1	0.116	0.236	.624	Position*Mental*Gender	-0.029	0.030	.328
Position*Graph*Effort	0.024	0.037	.519	Length*ChEffort2	0.246	0.174	.157	Position*Graph*Gender	0.025	0.073	.737
Position*Option*Effort	-0.005	0.050	.925	Mental*ChEffort2	0.049	0.093	.601	Position*Optionl*Gender	0.028	0.099	.777
				Graph*ChEffort2	-0.198	0.228	.385				
				Option*ChEffort2	0.443	0.333	.183				
				Position*Length*ChEffort1	0.006	0.035	.873				
				Position*Mental*ChEffort1	0.008	0.036	.822				
				Position*Graph*ChEffort1	-0.049	0.089	.585				
				Position*Option*ChEffort1	-0.224	0.125	.073				
				Position*Length*ChEffort2	-0.069	0.053	.194				
				Position*Mental*ChEffort2	0.021	0.053	.686				
				Position*Graph*ChEffort2	-0.004	0.132	.979				
				Position*Option*ChEffort2	-0.295	0.179	.099				
Random Effects		Corr.		Random Effects		Corr.		Random Effects		Corr.	
Person Variance	0.459			Person Variance	0.515			Person Variance	0.514		
Item Variance	0.526			Item Variance	0.533			Item Variance	0.536		
Item by Position Variance	0.003	-0.440		Item by Position Variance	0.003	-0.450		Item by Position Variance	0.003	-0.500	

Note. With the exception of position, all continuous predictors were standardized prior to being entered into the model.

Table 11

Fixed and Random Effects for the ESRKA Position Effect + Item Characteristics + Person Characteristics GLMMs (N = 1,092)

ESRKA-M4				ESRKA-M5				ESRKA-M6			
Parameter	Estimate	*SE*	*p-value*	Parameter	Estimate	*SE*	*p-value*	Parameter	Estimate	*SE*	*p-value*
Fixed Effects				Fixed Effects				Fixed Effects			
Intercept	**0.489**	**0.136**	**< .001**	Intercept	**0.567**	**0.137**	**< .001**	Intercept	**0.404**	**0.139**	**.004**
Position	**-0.052**	**0.012**	**< .001**	Position	**-0.040**	**0.014**	**.006**	Position	**-0.054**	**0.015**	**< .001**
Length	0.334	0.240	.164	Length	0.388	0.237	.102	Length	0.293	0.239	.221
Mental	-0.421	0.280	.132	Mental	-0.467	0.275	.090	Mental	-0.353	0.277	.204
Graph	0.916	0.807	.256	Graph	0.862	0.785	.273	Graph	0.727	0.791	.358
Effort	**0.206**	**0.025**	**< .001**	ChEffort1	**-0.229**	**0.062**	**< .001**	Gender	**0.221**	**0.054**	**< .001**
Position*Length	**-0.049**	**0.022**	**.026**	ChEffort2	**-0.256**	**0.096**	**.007**	Position*Length	-0.046	0.027	.085
Position*Mental	-0.008	0.026	.769	Position*Length	**-0.068**	**0.026**	**.008**	Position*Mental	-0.006	0.031	.843
Position*Graph	-0.025	0.076	.740	Position*Mental	0.020	0.030	.506	Position*Graph	0.008	0.091	.927
Position*Effort	**0.035**	**0.010**	**< .001**	Position*Graph	-0.018	0.088	.835	Position*Gender	-0.005	0.02	.794
Length*Effort	0.050	0.034	.141	Position*ChEffort1	**-0.064**	**0.023**	**.006**	Length*Gender	0.126	0.068	.062
Mental*Effort	-0.019	0.040	.623	Position*CHEffort2	-0.003	0.036	.923	Mental*Gender	**-0.197**	**0.079**	**.013**
Graph*Effort	-0.156	0.119	.191	Length*ChEffort1	**-0.157**	**0.079**	**.046**	Graph*Gender	**0.527**	**0.241**	**.029**
Position*Length*Effort	0.008	0.018	.666	Mental*ChEffort1	0.101	0.091	.271	Position*Length*Gender	-0.020	0.036	.578
Position*Mental*Effort	-0.003	0.021	.897	Graph*ChEffort1	0.277	0.278	.319	Position*Mental*Gender	0.015	0.042	.722
Position*Graph*Effort	0.119	0.063	.057	Length*ChEffort2	-0.117	0.120	.331	Position*Graph*Gender	-0.120	0.127	.346
				Mental*ChEffort2	0.175	0.141	.217				
				Graph*ChEffort2	-0.106	0.425	.804				
				Position*Length*ChEffort1	0.042	0.043	.328				
				Position*Mental*ChEffort1	-0.060	0.049	.221				
				Position*Graph*ChEffort1	-0.105	0.146	.475				
				Position*Length*ChEffort2	0.078	0.065	.234				
				Position*Mental*ChEffort2	-0.118	0.076	.119				
				Position*Graph*ChEffort2	0.223	0.226	.325				
Random Effects		Corr.		Random Effects		Corr.		Random Effects		Corr.	
Person Variance	0.331			Person Variance	0.387			Person Variance	0.394		
Item Variance	0.678			Item Variance	0.681			Item Variance	0.689		
Item by Position Variance	0.002	-0.110		Item by Position Variance	0.002	-0.150		Item by Position Variance	0.002	-0.250	

Note. With the exception of position, all continuous predictors were standardized prior to being entered into the model.

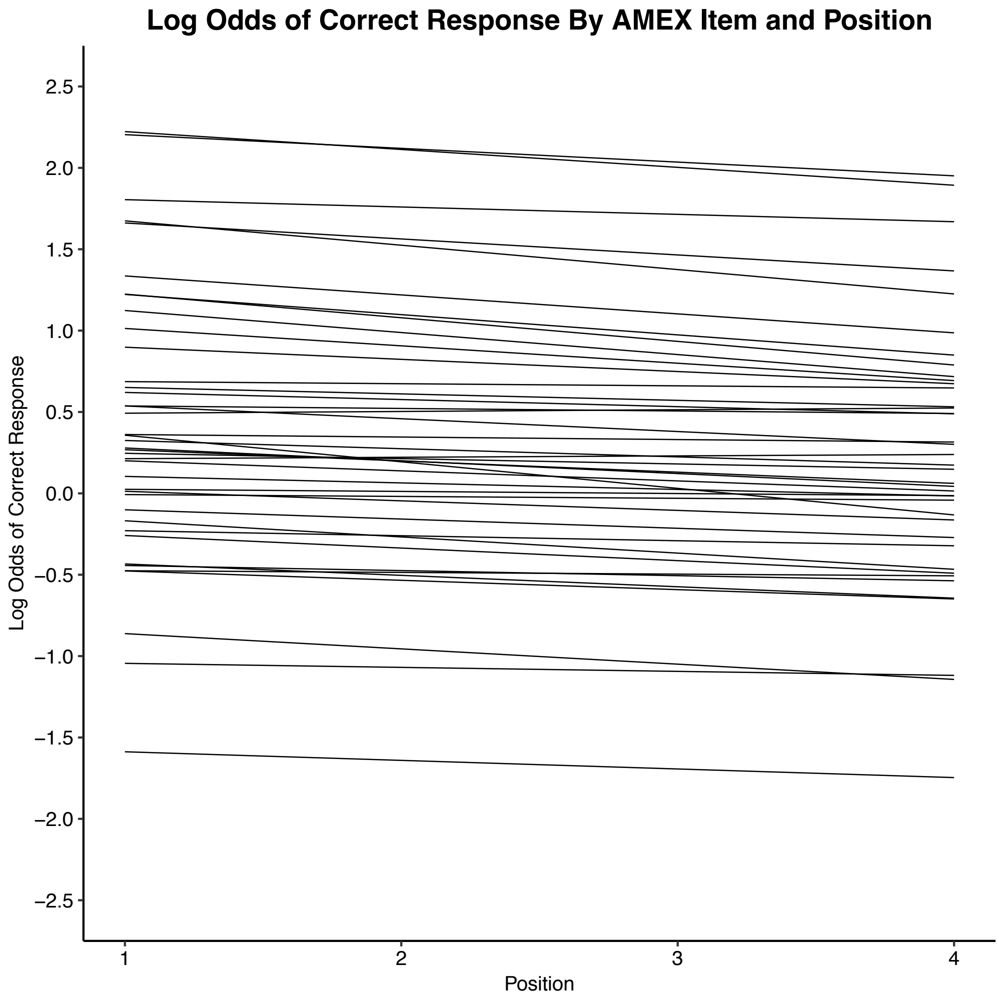

Figure 1. Item-specific position effects of AMEX items

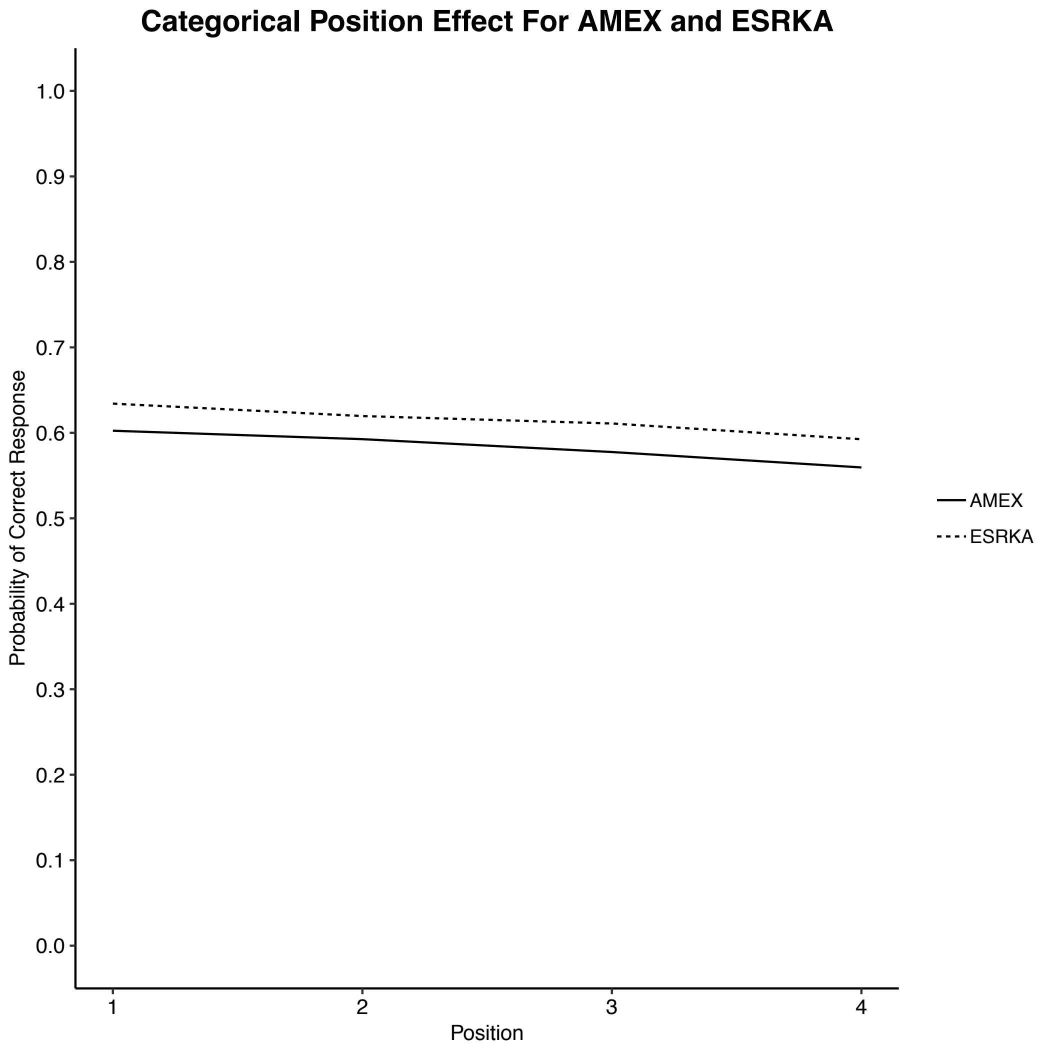

Figure 2. Predicted probability of obtaining a correct response at each position for AMEX and ESRKA.

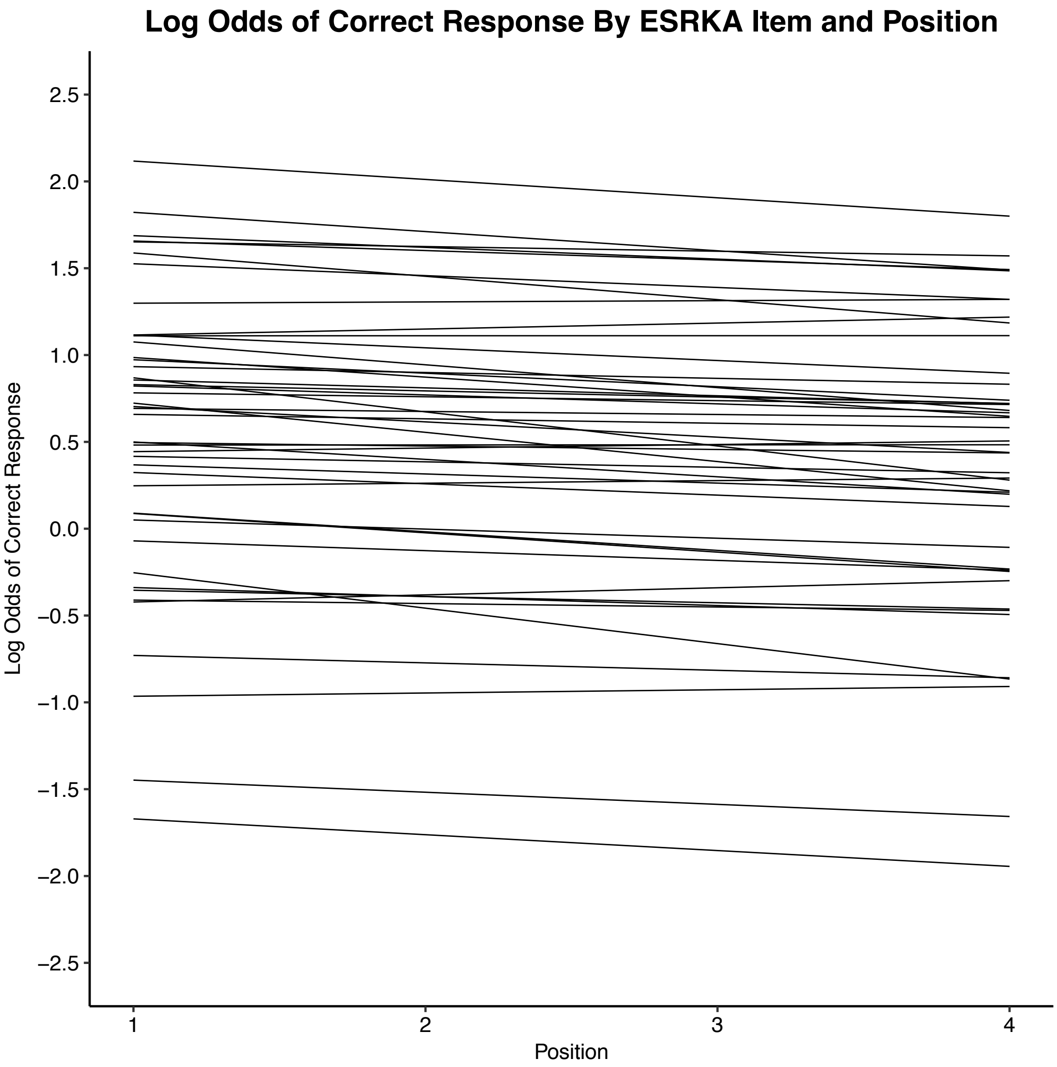

Figure 3. Item-specific position effects of ESRKA items

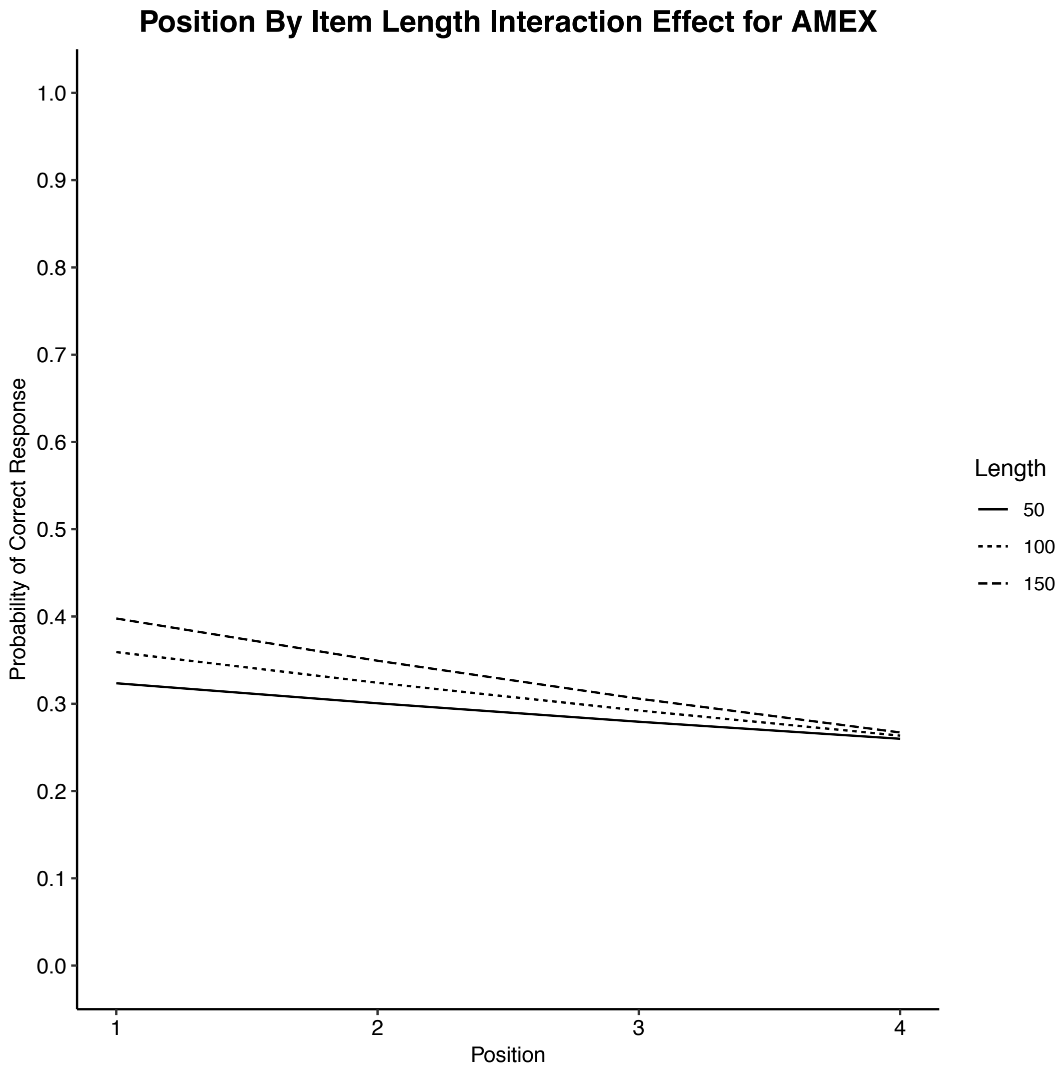

Figure 4. Predicted probability of obtaining a correct response at each position across three different length values on an average mentally taxing item with no graph and four response options on the AMEX.

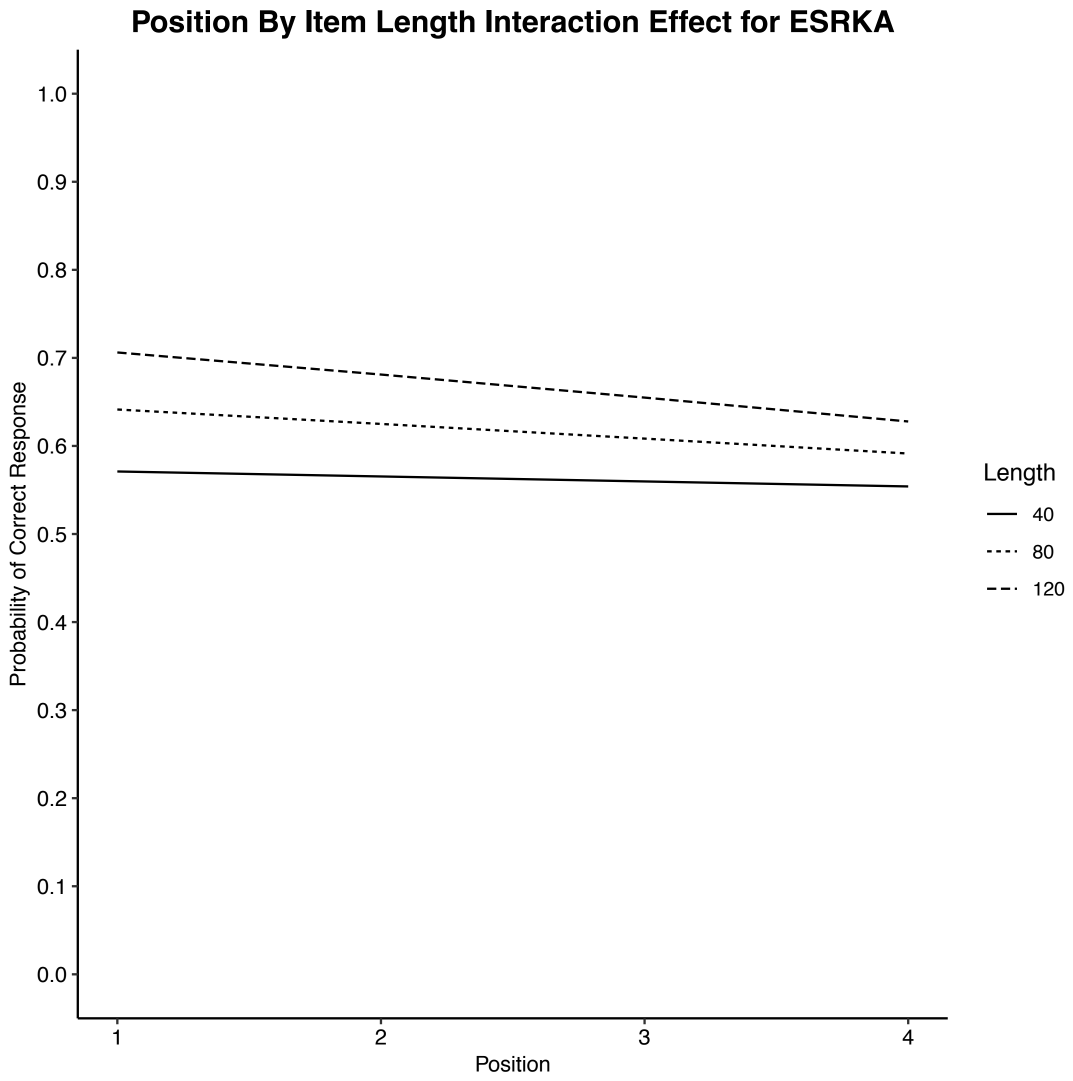

Figure 5. Predicted probability of obtaining a correct response at each position across three different length values on an average mentally taxing item with no graph and five response options on the ESRKA.

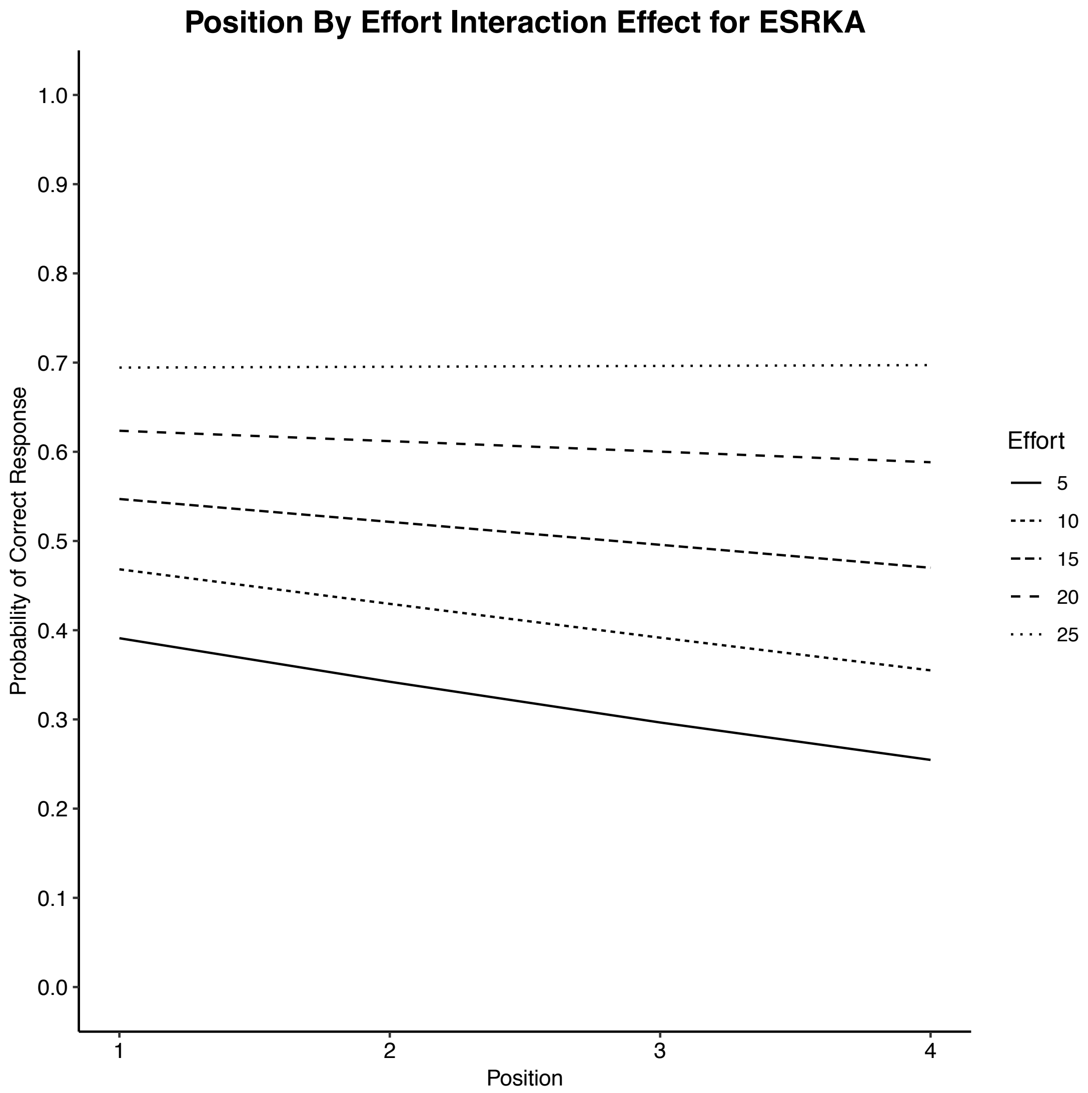

Figure 6. Predicted probability of obtaining a correct response at each position across five different levels of effort on an average mentally taxing item with no graph and five response options on the ESRKA.

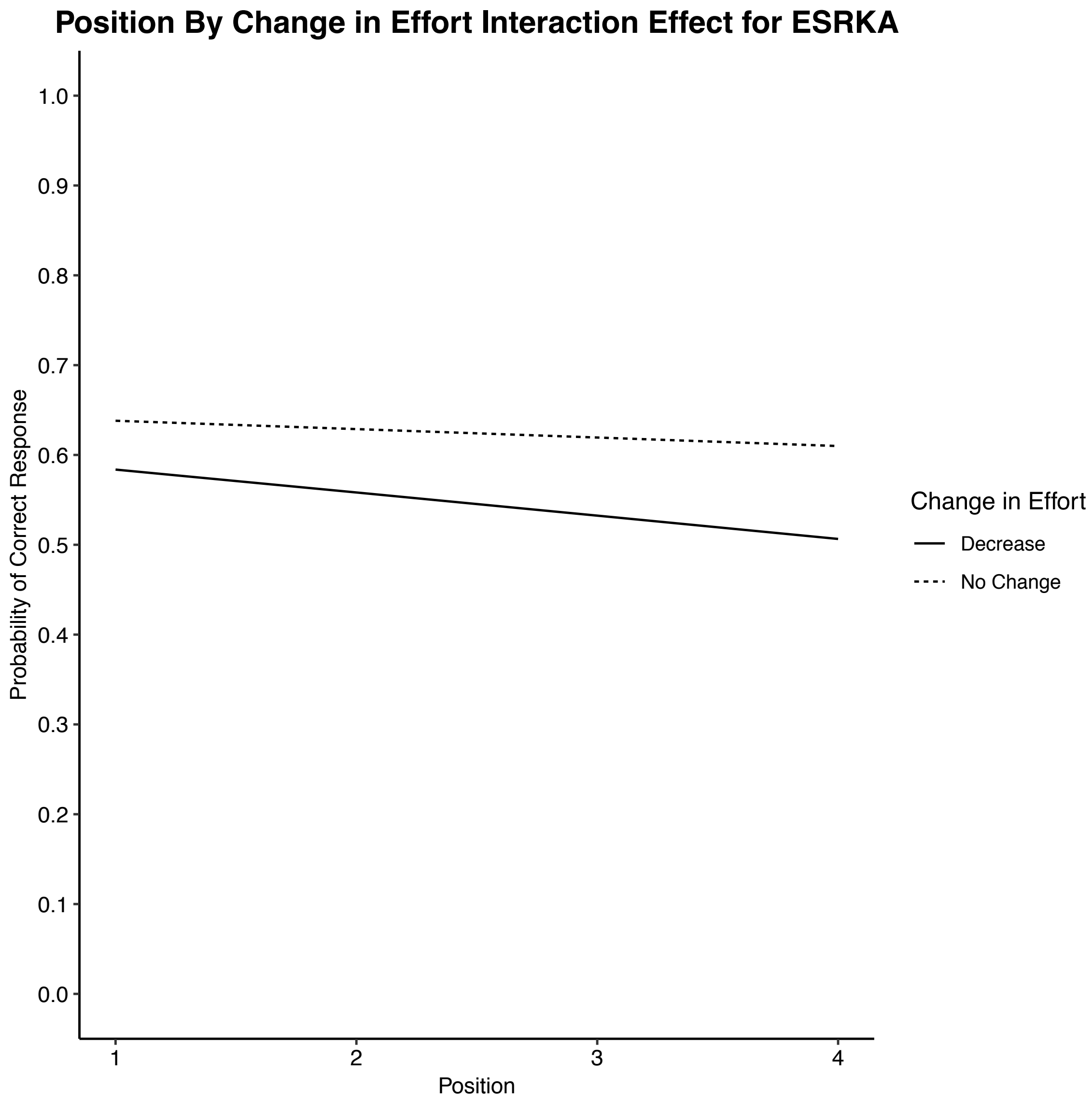

Figure 7. Predicted probability of obtaining a correct response at each position for examinees reporting no change in effort and examinees reporting a decrease in effort on an average mentally taxing item with no graph and five response options on the ESRKA.

Appendix

AMEX Mental Taxation Ratings

Item	Rater A	Rater B	Rater C	Rater D	Rater E
1	5	3	6	6	4
2	4	4	5	5	3
3	7	7	8	7	8
4	6	8	7	5	6
5	5	7	4	7	6
6	3	3	4	3	3
7	4	8	4	2	3
8	5	10	4	2	4
9	4	5	5	4	4
10	2	1	3	2	2
11	3	2	3	3	3
12	8	4	6	6	6
13	2	1	2	1	2
14	2	1	2	1	2
15	4	5	4	4	4
16	3	1	5	1	4
17	3	1	3	3	4
18	4	2	3	3	3
19	4	1	3	4	4
20	3	1	4	3	3
21	2	2	3	4	3
22	4	1	4	4	4
23	2	1	2	2	2
24	3	1	3	1	2
25	4	1	4	5	4
26	1	1	2	3	3
27	3	1	2	2	2
28	1	1	2	1	2
29	3	1	2	3	2
30	4	1	3	4	5
31	3	1	2	2	3
32	4	1	4	4	4
33	4	1	3	5	4
34	4	1	4	3	3
35	6	2	5	4	4
36	1	1	3	2	1
37	7	1	5	4	4
38	4	1	4	1	3
39	1	1	2	2	1
40	3	1	3	2	2

ESRKA Mental Taxation Ratings

Item	Rater A	Rater B	Rater C	Rater D	Rater E
1	8	7	7	8	7
2	2	1	2	4	2
3	1	1	2	2	1
4	4	2	3	3	2
5	2	1	2	1	2
6	3	1	3	1	3
7	2	1	2	1	2
8	5	3	4	5	5
9	3	1	2	1	2
10	1	1	1	2	1
11	4	1	2	1	2
12	3	1	3	3	3
13	7	7	5	6	6
14	5	2	4	2	3
15	6	2	4	4	4
16	5	2	4	4	4
17	4	1	4	3	3
18	1	1	1	1	1
19	1	1	2	1	1
20	2	1	2	2	2
21	4	1	3	2	3
22	2	1	2	3	2
23	5	3	4	5	5
24	1	1	2	2	2
25	5	2	4	2	3
26	3	1	3	1	2
27	3	1	5	3	4
28	9	4	6	6	7
29	3	1	3	2	2
30	2	1	4	1	2
31	3	3	3	2	3
32	4	1	3	3	3
33	7	9	7	6	7
34	6	1	6	3	6
35	5	1	4	2	4
36	4	1	4	2	4
37	5	1	3	3	4
38	3	1	3	2	2

39	6	1	5	4	5
40	2	4	3	2	2
41	3	1	2	2	3
42	5	1	5	7	5
43	4	9	4	4	5
44	5	1	4	3	3
45	3	1	2	1	2

References

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

Alpekt, R., & Habek, R. N. (1960). Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology*, *61*, 207-215.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.0-6). Retrieved from http://CRAN.Rproject.org/package-=lme4.

Berger, V. F., Munz, D. C., Smouse, A. D., & Angelino, H. (1969). The effects of item difficulty sequencing and anxiety reaction type on aptitude test performance. *Journal of Psychology*, *71*, 253-258.

Brenner, M. H. (1964). Test difficulty, reliability, and discrimination as functions of item difficulty order. *Journal of Applied Psychology*, *48*, 98-100.

Bovaird, J. A. (2002). New applications in testing using response time to increase construct validity of a latent trait estimate (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Full Text; ProQuest Dissertations & Theses Global. (305568564).

Bulut, O. (2015). An empirical analysis of gender-based DIF due to test booklet effect. *European Journal of Research on Education, 3*, 7-16.

Bulut, O., Quo, Q., & Gierlm M. J. (2017) A structural equation modeling approach for examining position effects in large-scale assessments. *Large-scale Assessments in Education*, *5*, 1-20.

Cook, L. L., & Paterson, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, *11*, 225-244

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

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

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavior Statistics*, *39*, 502-523.

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

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533-559.

De Boeck , P., Bakker, M., Zwitser, R., Nivard, M., Hofmanm A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, 39, 1-28.

DeMars, C. E., Baskov, B. M., and Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research and Practice in Assessment*, *8*, 69-82.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007) Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, 20, 1-18, Eignor, D. R., & Cook, L. L. (1983). *An investigation of the feasibility of using item response theory in the preequating of aptitude tests.* Paper presented at the meeting of the American Educational Research Association, Montreal, Canada.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374.

Harris, D. J. (1991). Effects of passage and item scrambling on equating relationships. *Applied Psychological Measurement*, *15*, 247-256.

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly*, 50, 379-390.

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. *Journal of Experimental Education*, *43*, 40-46.

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, *54*, 418-431.

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation*, *17*, 497-509.

Huck, S. W., & Bowers, N. D. (1972). Item difficulty level and sequence effects in multiple-choice achievement tests. *Journal of Educational Measurement*, *9*, 105-111.

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.

Klein, S. P. (1981). *The effect of time limits, item sequence and question format on the California bar examination.* A report prepared for the Committee of Bar Examiners of the State of California and the National Conference of Bar Examiners.

Klimko, I. P. (1984). Item arrangement, cognitive entry characteristics, sex, and test anxiety as predictors of achievement examination performance. *Journal of Experimental Education, 52*, 214-219.

Klosner, N. C., & Gellman, E. K. (1973). The effect of item arrangement on classroom test performance: Implications for content validity. *Educational and Psychological Measurement*, *33*, 413-418.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking.* New York, NY: Springer.

Kolen, M. J., & Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement, 27*, 27-39.

Lane, D. S., Bull, K. S., Kundert, D. K., & Newman, D. L. (1987). The effects of knowledge of item arrangement, gender, and statistical and cognitive item difficulty on test performance. *Educational and Psychological Measurement, 47*, 865-879

Le, L. T. (2007). *Effects of item positions on their difficulty and discrimination: A study in PISA Science data across test language and countries.* Paper presented at the 72nd Annual Meeting of the Psychometric Society. Tokyo, Japan.

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.

MacNicol, K. (1956). *Effects of varying order of item difficulty in an unspeeded verbal test.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Marso, R. N. (1970). Test item arrangement, testing time, and performance. *Journal of Educational Measurement*, *7*, 113-118.

Meulders, M., & Xie, Y. (2004). Person-by-items predictors. In P. De Boeck & M. Wilson Centering and Scale Indeterminacy 31 (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 213-240). New York: Springer-Verlag.

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.

Mislevy, R., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*, 195-215.

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.

Monk, J. J., & Stallings, W. M. (1970). Effects of item order on test scores. *The Journal of Educational Research*, 463-465.

Munz, D. C., & Smouse, A. D. (1968). Interaction effects of item-difficulty sequence and achievement-anxiety reaction on academic performance. *Journal of Educational Psychology*, *59*, 370-374.

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (in press). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment.*

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., Patience, W. M., & Whitney D. R. (1988). Differential item performance in mathematics achievement test items: Effect of item arrangement. *Educational and Psychological Measurement*, *48*, 885-894.

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.

Qian, J. (2014). An investigation of position effects in large-scale writing assessments. *Applied Psychological Measurement*, *38*, 518-534.

R Core Team (2017). R: *A language and environment for statistical computing. R Foundation for Statistical Computing,* Vienna, Austria. URL https://www.R-project.org/.

Ryan, E. K., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education*, *14*, 73-90.

Sax, G., & Cromack, T. R. (1966). The effects of various forms of item arrangements on test performance. *Journal of Educational Measurement*, *3*, 309–311.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement of Education*, *26*, 34-39.

Smouse, A. D., & Munz, D. C. (1968). The effects of anxiety and item difficulty sequence on achievement testing scores. *Journal of Psychology*, *68*, 181-184.

Smouse, A. D., & Munz, D. C. (1969). Item difficulty sequencing and response style: A follow-up analysis. *Educational and Psychological Measurement*, *29*, 469-472.

Sundre, D. L. & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update*, 14, 8-9.

Towle, N. J., & Merrill, P. F. (1975). Effects of anxiety type and item-difficulty sequencing on mathematics test performance. *Journal of Educational Measurement*, *12*, 241-249.

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.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185-201.

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement*, *38*, 535-548.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., and Böhme K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, *41*, 115-129.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education*, 19, 95-114.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice.

Applied Measurement in Education, *22*, 185-205.

Wolf, L. F., Smith, J. K., & Birnbaum, M.E. Consequences of performance, test motivation, and mentally taxing items. *Applied Measurement in Education*, 8, 341-351.

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