Multilevel Modeling of Item Position Effects

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Anthony D. Albano

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Doctor of Philosophy

Michael C. Rodriguez

May, 2012

© Anthony D. Albano 2012

ALL RIGHTS RESERVED

Acknowledgements

My thanks go, first and foremost, to my advisor, Michael Rodriguez, for his thoughtful mentoring and his commitment to my training and success as a scholar. His dedication to education and educational measurement will always be an inspiration in my own work.

Thanks also go to Mark Davison, David Weiss, and Leah McGuire, members of my dissertation committee, for their valuable feedback throughout the various iterations of my dissertation study.

This study is an extension of a project which I began while working at the American Registry of Radiologic Technologists. My thanks also go to the Registry, including Ben Babcock, Michael Yoes, Dan Anderson, and Lauren Wood, for their encouragement and support in my exploration of this topic.

Finally, I would like to thank Educational Testing Service, including Tim Davey and Yi-Hsuan Lee, for granting me access to data from the GRE.

Dedication

To my family: my parents, examples of persistence and optimism; my fabulous wife, an example of patience and dedication; and my children, examples of happiness.

Abstract

In many testing programs it is assumed that the context or position in which an item is administered does not have a differential effect on examinee responses to the item, or at least that any differential effect is negligible. Violations of this assumption may bias item response theory estimates of item and person parameters. This study examines the potentially biasing effects of item position.

Previous work has approached position effects in testing from a variety of methodological perspectives, resulting in a variety of findings. This study presents a hierarchical generalized linear model, a type of multilevel model, for estimating item position effects. Previous approaches to estimating and modeling position effects are described within a multilevel framework, and an extension of these approaches is demonstrated, one which incorporates item position as a continuous variable. Position effects are estimated as interactions between the position and the item, in other words, as slopes or changes in item difficulty per shift in the position of the item within the test form.

The model is demonstrated using real and simulated data. Real data came from two sources: a K-12 reading achievement test administered to over

90,000 students in which pilot items were included in random positions; and pilot sections of the GRE administered to roughly 1,800 examinees, where the same items appeared in different positions across the form. Data were simulated to have item-position effects similar to those found in the real data studies and in previous research. A base model and two position effect models were then compared in terms of parameter recovery and fit to the simulated data. Practical applications of the model are discussed.

Contents

List of Tables

List of Figures

Chapter I: Introduction

Large-scale testing programs commonly involve multiple versions of a test which are statistically linked to a single measurement scale. Examples are K-12 tests such as the MAP (Northwest Evaluation Association) and STAR (California Department of Education), admission tests such as the GRE, ACT, and SAT, and certification/licensure tests such as the NAPLEX (National Association of Boards of Pharmacy) and the NCLEX (National Council of State Boards of Nursing). The overarching measurement objective for these testing programs is to maintain a common metric, so as to allow for comparisons of performance across different examinees and different test forms over time.

Maintaining a stable measurement scale becomes complicated when alternate test forms include different items and/or different orders of item administration for different groups of examinees. The extent of form differences depends on the scope and structure of the testing program. In small-scale applications, such as classroom assessment, a small number of forms may be used, each with a different arrangement of the same items. Similarly, in larger-scale applications, computerized testing facilitates scrambled item

administration (Leary & Dorans, 1985), where the item set remains unchanged, but the item sequence is randomized for each examinee. These item rearrangement strategies are often used for test security reasons, to limit the effectiveness of memorizing items and/or copying answers from other examinees during administration. Computerized testing also facilitates computerized adaptive testing (CAT), where an algorithm estimates examinee ability and presents precalibrated items of corresponding difficulty in real-time. In CAT, the test form is created at the time of administration, and the item set and sequence may differ significantly across examinees.

In these testing applications it is assumed that the context or position in which an item is administered does not have a differential effect on examinee perception of or response to the item, or at least that any differential effect is negligible; in other words, the interaction between the item and the examinee does not depend on the location of the item in the form. Wainer and Kiely (1987) defined this differential effect as a context effect, "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" (p. 187). Defined in this way, any test form can be affected by the context in which items are administered; however, with a fixed form where position does not vary, each item context effect would be the same for all examinees. When item sequences differ across examinees, the

effects are potentially differential, which may bias estimates of item and person parameters (Leary & Dorans, 1985).

Research from the educational and psychological measurement literatures has addressed the effects of context in a variety of operational settings, both under the classical test theory and the item response theory (IRT) models. Chapter 2 reviews this previous research, focusing on conceptualizations and definitions of context effects, and methods of estimating and modeling context effects. Although conceptualizations and estimation methods differ somewhat from one study to the next, the main conclusion derived from this body of work is that changes in item position and context can affect test performance. As a result, testing programs may not be justified in assuming that item-level form differences do not have an effect on testing results.

This study presents a new approach to conceptualizing and modeling the effect of item position across test forms. A hierarchical generalized linear model (HGLM), also referred to as a multilevel measurement model, is formulated in Chapter 3. The model is demonstrated with real and simulated data, as outlined below, in order to answer the research question, does item position bias the estimation of item parameters? Chapter 4 presents the results of the real and simulated data studies. Chapter 5 concludes with a discussion of the results and implications for educational and psychological measurement.

Chapter II: Literature Review

Introduction

Leary and Dorans (1985) presented a comprehensive review of the literature on context effects, beginning in the 1950s and ending in the early 1980s. Over this thirty year period, substantial changes were made in psychometric theory and applied measurement, especially as educational testing became more prevalent and as computers influenced test design and administration. Leary and Dorans (1985) noted that the context effect research questions addressed across this period corresponded closely to the practical psychometric issues of the time. A similar trend is evident in the research from the 1980s onward. The present review organizes the literature on context effects by the major psychometric issues which drive the research, issues associated with the use of multiple test forms and the creation and maintenance of a measurement scale across these multiple forms.

Research has demonstrated that in some situations item context can negatively impact estimation and score interpretation. However, previous work has approached context effects in testing from a variety of perspectives, with a

variety of different findings. Brennan (1992) discussed the lack of agreement in the literature, noting that, "models need to be developed that explicitly incorporate the possibility that context effects exist" (p. 259). More recently, Davey and Lee (2011) came to a similar conclusion, suggesting that, "one possible direction for future analysis is to employ some IRT model that can incorporate item position as a predictor, as well as unique parameters" (p. 40). The present study demonstrates such a model.

This chapter reviews previous research so as to develop a clear conceptualization of context and position effects, and provide background for a multilevel model that builds upon previous studies by incorporating position effects into the estimation of item and person parameters. The first section below conceptualizes context effects by highlighting the differences between context and position and discussing the item arrangement schemes which are expected to lead to each type of effect. The second section below reviews the research on context effects, including early studies, which are characterized by an emphasis on item arrangement and order of administration, with context effects estimated in terms of change in the proportion of examinees getting an item or set of items correct; and later work, which is characterized by an emphasis on issues related to equating, with context effects estimated mainly in terms of change in the IRT item location or difficulty parameter. Finally, in the

third section below, three studies which examine item position effects within an IRT framework are discussed in detail. In Chapter 3, these and other position effect models are presented as different forms of HGLM.

Conceptualization

Defining Context

Wainer and Kiely (1987) refer to a context effect in a broad sense 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" (p. 187). Defined in this way, the terms "influence" and "interpretation" reference the effect, which is acquired by the item, and the term "relationship" references the cause or source of the effect, its context, or, the other items in the test.

The context, or the other items in the test, may be described in numerous ways. The other items may differ in type and difficulty, and may be designed to measure different content at a different level of cognitive understanding. For example, items may vary in the number of options, response type, task type, and content assessed, such as algebra or physical science. Items may also be more or less difficult than one another, overall. The items may require different amounts of time and effort, and there may simply be a larger or smaller total number of items. Simpler items may require only a few seconds to complete, whereas

complex items may require more reasoning or compound constructed responses. Finally, items may differ in quality and in their ability to discriminate between examinees of different ability levels. As any one of these features, and others not listed here, changes, the context of an item within the test form changes as well.

Context is typically defined in reference to the items preceding a target item, that is, the portion of the test form which examinees have completed and which will likely have the most impact on their response to the target item. In some cases, such as when an examinee is free to skip and then come back to items or go back to review previous responses, the items which follow a target item may also contribute to the definition of context and position. Additionally, with speeded tests, the number of items remaining in the test may impact performance on the target item, depending on how much testing time remains. This discussion focuses primarily on preceding items, under the assumption that the items following a target item will have minimal impact in most practical testing situations, which are typically not speeded, and no impact in computerized adaptive testing when the number of items remaining is unknown to the examinee. A few studies, referenced below, have focused on highly speeded testing conditions, where examinees are not expected to reach the end of the form within the time limit. However, this review and this study are concerned primarily with effects under power or slightly speeded conditions, as

are often found in practice.

As will be discussed below, the majority of research on context effects has focused more specifically on the effects of item position. However, a distinction between the terms context and position has rarely been made, likely because the two are not easily disentangled from one another. In this discussion, context effects are attributed specifically to the kind or quality of items that precede a target item or set of items, whereas position effects are attributed primarily to the quantity of items, i.e., the amount of testing, preceding a target item, though position effects may be attributed to the context created by the preceding items as well, depending on the study design.

The designs employed in the studies reviewed below are most often determined by the conditions available in a particular testing program. Context and position have rarely been manipulated purposively, by design. Focusing specifically on context would require that the target item position be fixed across forms, with only the quality of the adjacent items changing. On the other hand, focusing specifically on position would require that the number of items preceding a target item vary across forms, but not their quality, as discussed above. A subset of studies have attempted to describe one without the other; however, most have been unable to identify context or position alone as the cause of an effect.

Defining Effect

Previous studies have taken a variety of approaches to defining and estimating the effect of context. Most often, an effect has been defined as a statistically significant change in item difficulty (e.g., Brennan, 1992). Across two test forms, the effect may be estimated as a difference in proportion correct or logit score for the item, where context, position, or both have changed across the forms, and where a t-test, for example, is used to determine statistical significance. Across multiple forms, with each form treated as a group within an analysis of variance model, the effect may be estimated as a variance in item difficulties across the forms, where an F-test, for example, is used to determine statistical significance. Context effects have also been defined as change in average performance across sets of items, or at the test level, given change in context or position.

The review in the following section provides an overview of the study designs and approaches to defining and estimating context effects in the literature. As in previous research, the present study defines a context effect as a change in item difficulty. This will be referred to as bias in item difficulty or location (Kingston & Dorans, 1984). Other considerations, such as item discrimination, are also important. However, the focus on item difficulty in the

literature and in this study is justified by the fact that no other statistical property of an item has a greater impact on the estimation of person ability or person location on the measurement scale.

Context Effects Research

Item Arrangement

Early studies of context effects involved relatively simple designs where the same items were used across test forms and only their arrangement in the forms differed. Since forms contained the same items, these designs did not involve equating. Instead, in equating terms, the forms consisted entirely of common or anchor items. Examinee groups taking each form were assumed to be comparable in ability, typically through random assignment. As a result, any differences across the score distributions were attributed to the rearrangement of items from one form to the next.

Item Arrangement Research

Monk and Stallings (1970) compared performance across multiple pairs of classroom assessment forms, where the items within each of the two forms were the same but were ordered randomly. Total scores were examined across eleven pairs of forms, with statistically significant total score mean differences found for two of the form pairs: a 7-point score difference for one pair ($t_{254} = 5.98$,

$p < .001$), and a 5-point score difference for the other ($t_{176} = 2.28$, $p < .01$). In this way, context was conceptualized at the test level, rather than at the item level. These findings provided some evidence of what is referred to here as a *test position effect*.

Mollenkopf (1950) compared performance based on a controlled arrangement, rather than a random assignment, of items. Each form in the study consisted of the same three item sets, where the first and last sets were swapped across two forms; thus, the items at the beginning of one form appeared at the end of the other, and vice versa. The two swapped item sets were designed, using pilot data, to be comparable in content and to cover a range of item difficulties. The middle portion, which remained unchanged across forms, contained items of average difficulty. This approach made it possible to break down the overall test position effect examined by Monk and Stallings (1970) into slightly more specific effects based on location or position. As a result, context was conceptualized at the position level. Mollenkopf (1950) found that under speeded conditions, verbal item difficulty was lower when items appeared at the beginning of the form. On the other hand, under unspeeded, i.e., power conditions, verbal item difficulty was higher in the beginning section of the test than in the end; item-level proportion correct decreased on average by about .05, with values for some items decreasing by as much as .10. These

findings provided evidence of what is referred to here as an *item-position effect*.

Other early studies compared mean performance for forms constructed via specific item arrangement schemes, including arrangements increasing in difficulty, decreasing in difficulty, and with a balancing of item difficulty across the form (e.g., Brenner, 1964; Flaugher, Melton, & Myers, 1968; Sax & Cromack, 1966). The general findings from these studies were that easy-to-hard sequencing resulted in higher average performance when the test was speeded, with hard-to-easy sequencing resulting in lower average performance, whereas performance differences were generally not found for unspeeded tests.

Additional studies during the 1960s and 1970s investigated more complex relationships between item arrangement and other variables. The majority of these studies examined the interaction between position and examinee anxiety (e.g., Berger, Munz, Smouse, & Angelino, 1969; Smouse & Munz, 1968, 1969; Towle & Merrill, 1975). As a result, context was conceptualized as both a position effect and a position interaction effect. Overall, and especially in unspeeded conditions, significant interactions between position and anxiety were not found. Thus, the results did not provide evidence for *position interaction effects*.

In contrast to these position effect studies, Sax and Carr (1962) and Huck and Bowers (1972) compared average performance for items which remained in

the same position, but which were preceded by items of differing content and difficulty. The items were again the same across forms, with only their arrangement being modified. Context was conceptualized specifically by the immediate context of an item, rather than by its position, since, for the items in question, position was fixed. As a result, changes in performance on an item could be attributed to the changing context in which the item appeared. The findings of these studies were mixed, with statistically significant effects found in one case (Sax & Carr, 1962) but not the other (Huck & Bowers, 1972). This type of effect is referred to here as a *context effect*, because of the emphasis on the characteristics of the items preceding the target item, or the item that is common across forms.

Equating Designs

Early studies of context effects focused on the rearrangement of items across test forms, which did not necessitate special procedures for equating to a common measurement scale. As testing programs have grown, test forms have begun to differ in more than just the arrangement of items. A testing program's equating design delineates the structure of forms, including differences in item context across forms. Research on context effects from the last 30 years has dealt primarily with the common-item nonequivalent groups equating design,

referred to here as the common-item design.

Rather than involving the same items across all test forms, a common-item design involves forms with some items unique and some common. Examinees taking one test form are expected to differ in ability from examinees taking another form, and the common items embedded in each form are used to statistically control for this ability difference. When embedding common items, it is recommended that their position and context among the unique items be the same across forms (Cook & Paterson, 1987). For example, if a common item appears in position five in one form, it will also appear in position five in the other, and the items preceding it will be as similar as possible, across forms, in content and difficulty; that is, the fourth item in one form will be similar to the fourth item in the other, etc. This matching of common items across forms is expected to reduce the effects of position and context.

Due to practical constraints, the organization of the common item set may differ substantially across forms. Differences are perhaps most dramatic under a precalibration or pre-equating design (Bejar & Wingersky, 1982). In precalibration, the common items in a new test form are selected from a pool of items previously administered and calibrated via one or more other forms. Because these items have already been calibrated, they can be used to estimate person parameters as soon as the test is completed. The items unique to a new

form are referred to as pilot items, as they are typically not used for scoring, and they may serve as common items in later forms. As a result of this structuring of items, context and position for a given item can change substantially from one form to the next.

A precalibration design is often employed in computerized adaptive testing, where an algorithm estimates examinee ability and presents precalibrated items of corresponding difficulty in real-time. The test form is created at the time of administration, and the items and their sequence are typically unique to the examinee. Another common application of precalibration is scrambled item administration (Leary & Dorans, 1985). With item scrambling, the items are the same within a given form but their sequence is randomized for each examinee. Precalibrated common items are often embedded in scrambled forms as well, so as to provide the examinee with immediate results and calibrate the remaining items to the measurement scale. As with CAT, the scrambled test form is created at the time of administration.

Both CAT and scrambled forms may consist of items common to one or more previous test forms (items previously calibrated, or pre-equated in an earlier form) and pilot items which will be utilized in a later test form (items currently calibrated, or pre-equated in the current form). In each case, parameters for the set of common items are estimated at one administration and

then utilized for anchoring and scoring in another. If the position of the item in the precalibration administration biases the item parameter estimate, say, to make it appear more difficult than it actually is, parameter estimates for other items may be biased to appear more difficult and person parameter estimates may be biased to indicate higher ability. For example, parameter estimates for a pilot item appearing at the beginning of a form may differ meaningfully from estimates that would be obtained had the item appeared at the end of the form. For scrambled or computer adaptive tests, the position of this pilot item is free to vary when used operationally as a common item. If the item were later implemented at the end of an operational form, biased item parameter estimates would be used to calibrate new items and locate people on the measurement scale.

In some CATs, such as the GMAT and previous versions of the GRE, precalibrated operational items used to estimate ability are presented separately from new, uncalibrated test items. Aside from the test section or sections that count toward their score, students take an entire, separate section of pilot items and their responses are used to calibrate the items for future use. This is referred to as section pre-equating (Holland & Thayer, 1985). In section pre-equating, the pilot section is designed to cover the same content as one of the operational sections, in part so that examinees cannot distinguish between

the two and adjust their performance accordingly. Because the same material is tested twice, once for a score and once to collect pilot data, performance on whichever of the sections comes first may serve as a practice test for the section administered second. Thus, parameter estimates for the second section may be biased by position effects, most likely as practice effects (Brennan, 1992).

Equating and IRT-Based Research

The common-item design, and, in particular, the precalibration and section pre-equating designs, have been utilized in the majority of studies from the 1980s onward. Although studies seem to agree that context effects can lead to biased estimates, there is no consensus regarding appropriate methods of detecting and estimating the effects of context. This may in part be due to a lack of agreement across conceptualizations of context. Brennan (1992) notes that "because context effects are seldom defined, the literature does not usually provide an unambiguous basis for judging whether or not a context effect exists" (p. 236).

Some studies have examined changes in item statistics and examinee performance resulting from a general context change across forms. These studies did not attempt to disentangle context from position; effects are attributed to a variety of contextual factors. For example, Whitely and Dawis (1976) and Yen

(1980) found statistically significant variability in item difficulty when position was somewhat stable but item context, including the type and quality of adjacent items, was altered. Harris (1991) examined the effects of item order on equating with the ACT, in terms of equated scale scores. Examinee scaled scores were compared based on equating conversions derived from different arrangements of a base item set. Item arrangement was not controlled. Harris found that up to half of the examinees would have received a different score if the equating were based on the base form, as opposed to the examinee's own rearranged form. Studies such as these suggest that changes to an item's context, overall, can impact results, both at the item and person levels.

Zwick (1991) summarized the conclusions regarding what has been referred to as the 1986 NAEP reading anomaly. The anomaly was found in a comparison of national reading proficiency scores for 17-year-olds and 9-year-olds from 1984 to 1986. For both age groups, average scores decreased significantly in the span of two years, whereas in previous years average scores had remained stable. A thorough investigation, including a smaller-scale replication of the original 1984 and 1986 administrations using item subsets, determined that the score decreases were due to the placement of the items common across the two reading forms. Both the position of the items in the form and the adjacent items and the content covered by adjacent items had changed across administrations.

Studies have also examined position effects by taking into account item position change, whether earlier or later, from one form to the next. Results are sometimes referred to as location effects or sequence effects. Eignor and Cook (1983) found statistically significant differences in item difficulty for items shifting from the beginning to the end of a test section; difficulty decreased, indicating a potential negative practice effect, or fatigue effect. Wightman (1981), Swinton, Wild, and Wallmark (1983), and Wightman and Leary (1981) all found evidence of position effects for certain item types, including fatigue effects for some and practice effects for others. Results were reported as test position effects (e.g., Monk & Stallings, 1970), that is, changes in performance across all items of a given type given change in test section location.

Kingston and Dorans (1984) examined the effects of shifting entire item sections of the GRE. Location effects were reported in the logit metric as differences across section location in terms of standardized mean item difficulty. Section was defined as being earlier, as an operational section, or last in the test, as a non-operational section. As a result, position change for an individual item was always at least the total number of items in a section, ranging from a minimum of 30 to a maximum of 80. Four subscales containing a total of 10 item types were examined: verbal (analogies, antonyms, sentence completion, reading comprehension); quantitative (quantitative comparisons, data

interpretations, regular mathematics); analytical 1 (analysis of explanations); and analytical 2 (logical diagrams, analytical reasoning). Change in standardized mean item difficulty was obtained for each of these item types. Kingston and Dorans (1984) found that reading comprehension items decreased in difficulty on average, with an increase as large as 0.14 logits when items went from an early section to the end of the test. Analytical items, requiring examinees to make an analysis of explanations, increased in difficulty on average, with a logit decrease as large as 0.30.

Davis and Ferdous (2005) examined the relationship between change in item position and change in estimates of IRT item difficulty across two state test forms in reading and two in math, administered to third and fifth graders. Correlations between change in item position and change in IRT item difficulty from one form to the other were $-.10$ and $-.32$ for grade three reading and math, and $-.65$ and $-.45$ for grade five reading and math (the last three of these values were statistically significant at $\alpha = .05$). These correlations indicate an overall practice effect, a decrease in item difficulty, or increase in performance on an item, as its position in the form increased.

Meyers, Miller, and Way (2008), estimated changes in item difficulty across two test forms and modeled this change as a function of the position change for the item. The approach was similar to that of Davis and Ferdous

(2005). Although only two forms were used, considering the results across all items gives a rough indication of how item difficulty can change when an item can appear anywhere in the form. Change in item difficulty y was best fit by a cubic function of item position p. The resulting regression model for a math test was

$$y = 0.00329(p) + 0.00002173(p^2) + 0.00000677(p^3), \quad (1)$$

and for a reading test the model was

$$y = 0.00845(p) - 0.00008343(p^2) + 0.00001135(p^3). \quad (2)$$

The three coefficients in each model are expressed as change in the logit metric. Thus, for 40 items on the math test, the logit change would be estimated to range from 0, for a position change of 0, to

$$0.00329(40) + 0.00002173(40^2) + 0.00000677(40^3) = 0.5996,$$

for a position change of 40. For reading, the range would extend from 0 to 0.9309 logits.

Summary

The results from Meyers et al. (2008) indicate that item difficulty can change as much as a full logit when estimated for an item at the beginning versus the end of a 40-item form. Interpreted in this way, ignoring the effect of

item context could have a significant impact on the estimation of item difficulty. However, this conclusion is based on two assumptions. The first is that position effects appear at the item level; person effects, including person-item and person-position interactions, are ignored. When modeling change in item difficulty across position, the observations at each position for a given item come from different people. To interpret the coefficients in Equations 1 and 2 across multiple positions, predicting change in difficulty, say, from the beginning to the end of the test, we essentially treat each item as having repeated measures across position. In longitudinal modeling terms, position would be the covariate for time, and the person(s) at a given position would be a repeated measure for the item. As with a longitudinal model, the repeated measures, i.e., persons, are assumed to be equivalent.

The second assumption is that the effect of position is the same across items. In this way, Equations 1 and 2 involve what is referred to as *complete pooling* (Gelman & Hill, 2007) across items, where group indicators for items are not included in the model. In Meyers et al. (2008), complete pooling may have been necessary because each item appeared in only two different positions, across two forms. Thus, the regression models in 1 and 2 aggregated changes in difficulty and position across items. As a result, the effect of position on a given item could not be determined.

As in Meyers et al. (2008), other previous studies are limited in that they have typically involved a small set of test forms where item position varies according to a specific scheme. To model position effects, information is completely pooled in some way, whether across items, people, or both. Including grouping variables for both items and people, as is typically done in IRT, along with position covariates, would require a more complex study design and data set than those utilized in the studies reviewed above. Situations where an item can take numerous or any position within a form for an individual examinee have not been thoroughly addressed in the literature. As a result, the extent of bias in the estimation of difficulty for a specific item, or the extent to which a specific item may lead to violations of the IRT assumption of local independence, is not well understood.

Modeling Position Effects

Local Independence in IRT

Controlling the organization of items within test forms can help ensure that they function similarly across groups (Kolen & Brennan, 2004). This is especially important in IRT applications because of the reliance on item level data and because of the IRT assumption of local independence, which requires that the interaction between an item and person be fully characterized by a set

of parameters for that item and person. In the unidimensional Rasch (1960) model, the probability of correct response to item i for person j is modeled as:

$$P(y_{ij} = 1|b_i, \theta_j) = \frac{1}{1 + e^{-(\theta_j - b_i)}}. \quad (3)$$

Here, it is assumed that no other characteristics of item i, beside its difficulty b_i, and no other trait or ability for person j, beside θ_j, are needed to describe the relationship between the two in terms of the probability that $y_{ij} = 1$. If a variable were missing from the linear component of the model, $\theta_j - b_i$, such as an additional effect for the format or position of the item, and such that $P(y_{ij} = 1)$ were somehow dependent on $P(y_{i\prime j} = 1)$, the probability of correct response on another item $i\prime$, this assumption would be violated. Local independence requires that dependencies across items and people are accounted for by the parameters in the model. Context and position effects become problematic when the locations of the items within the form introduce unexplained dependencies in the item responses. The following section reviews three models related to Equation 3 that have been used to examine position effects.

Logistic Regression Models

As recommended by Kingston and Dorans (1984), "the development of more general models that incorporate parameters such as item familiarity or item position should proceed" (p. 154). This section reviews three such models.

All are based on logistic regression. The first, from Pomplun and Ritchie (2004), combined the estimation of position effects with the estimation of item difficulties. The second, from Davey and Lee (2011), involved both logistic regression and IRT modeling. The third, from Alexandrowicz and Matschinger (2008), examined the estimation of position effects within an IRT framework, where item difficulty, person ability, and position effects were estimated concurrently.

Pomplun and Ritchie (2004) investigated position effects for items with position randomized within testlets. The testlets ranged from 2 to 5 items in length and were nested within full tests ranging from 35 to 41 items in length, depending on grade level. Separate logistic regression models were fit to dichotomous item responses for each item across persons. Thus, there were as many regression models as there were items, and each addressed as many positions as there were items within the corresponding testlet. Position effects were found for about ten percent of the items, with the majority of these items getting more difficult as position increased, indicating possible fatigue effects.

The logistic regression model can be described as a reformulation of the Rasch model where the dependent variable is the log-odds, rather than the

probability, of correct response:

$$\log \frac{P}{(1-P)} = \eta_{ij} = \theta_j - b_i. \tag{4}$$

In hierarchical linear modeling notation (Kamata, 2001) the item parameter is expressed as γ_i, and the person parameter as u_j:

$$\eta_{ij} = \gamma_i + u_j. \tag{5}$$

Here, the log-odds η_{ij} are modeled as a summation of the item and person parameters, rather than as a difference. As a result, γ_i represents the easiness of the item, rather than its difficulty as in Equations 3 and 4. Thus, larger values for γ_i are associated with easier items and higher log-odds of correct response.

When dichotomous responses are modeled for a single item, as in Pomplun and Ritchie (2004), ability u_j cannot be estimated and must be supplied as a covariate, in this case, the total score W. Within a testlet of 3 items, the model could be written as:

$$\eta = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \delta W. \tag{6}$$

Position effects are included here as main effects, using categorical indicator variables where $X_1 = 1$ and $X_2 = 0$ when the item is in position 1, and vice-versa when it is in position 2. Thus, the intercept γ_0 is the average easiness of the item at position 3, adjusted for δ, the estimated impact of ability when

$X_1 = X_2 = 0$. The terms γ_1 and γ_2 are expected changes in the item easiness based on its location in the testlet, controlling for ability. This resembles Equation 5, where the item effect γ_i is broken into effects based on position, and ability is controlled through the inclusion of the total score W in place of u.

Note that the model used by Pomplun and Ritchie (2004) was not displayed in their article; however, based on their descriptions, position effects appear to have been modeled as shown above. Equation 6 corresponds to a logistic regression model of differential item functioning (DIF; Swaminathan & Rogers, 1990), where group membership is defined based on the position in which the item appeared. The terms γ_1 and γ_2 test for uniform DIF, and additional terms for the interactions X_1W and X_2W would test for non-uniform DIF.

Davey and Lee (2011) manipulated non-operational quantitative and verbal pilot sections of the GRE to study change in item difficulty across different item orderings. Quantitative items were arranged according to 13 different ordering schemes, where each of the 28 items appeared with roughly equal frequency in each of several locations across the form. Thirty verbal items were arranged similarly into 7 different forms. The forms were administered to over $10,000$ examinees (over $5,000$ for quantitative and $5,000$ for verbal).

Davey and Lee (2011) also examined position effects using logistic regression, among other analyses. The model was similar to that of Pomplun

and Ritchie (2004):

$$\eta = \gamma_0 + \gamma_1 p. \tag{7}$$

This model appears to have been fit to each item, with γ_0 estimating the average easiness of the item at position $p = 0$, and γ_1 estimating the linear change in item easiness for a 1-unit change in position. In contrast to Pomplun and Ritchie (2004), position is included as a continuous variable in Equation 7.

Of the 28 quantitative items, 8 were estimated to have statistically significant position effect slopes; only 4 of the 30 verbal item position slopes were statistically significant (effects were tested using likelihood ratios, comparing fit of the model with and without the additional position slope term). The largest slopes were $-.032$, for quantitative items, and $-.025$, for verbal items. These indicate expected decreases in item easiness, or increases in difficulty, of .032 and .025 logits for a change in item position of 1, on the quantitative and verbal forms, respectively. The majority of slope estimates were negative.

Alexandrowicz and Matschinger (2008) also applied logistic regression to estimate position effects. However, in this case, a mixed logistic regression model, a type of linear logistic test model (Kubinger, 2008), was fit to the entire item response vector, that is, all N items and all J persons. As described by Kamata (2001) in HGLM terms, the combination of items and persons in a single response vector requires additional categorical variables to indicate the

specific item which corresponds to a given item response:

$$\eta_{ij} = \gamma_0 + \sum_{q=1}^{N-1} \gamma_q X_{qij} + u_j. \tag{8}$$

In this model, $X_{qij} = 1$ when $q = i$ and $X_{qij} = 0$ otherwise (the index q is added to accommodate missing data). Position effects have not yet been added. Thus, this is an illustration of the log-odds Rasch model in Equation 5, where the intercept γ_0 represents the item location for the reference item, here, the last item in the form. For item $q = i$, the model reduces to

$$\eta_{ij} = \gamma_0 + \gamma_q + u_j, \tag{9}$$

and for the reference item, the model reduces to

$$\eta_{ij} = \gamma_0 + u_j. \tag{10}$$

Alexandrowicz and Matschinger (2008) examined a conditional version of this model, with additional main effects for position. Again, in log-odds form:

$$\eta_{ij} = \gamma_0 + \sum_{q=1}^{N-1} \gamma_q X_{qij} + \sum_{r=1}^{N-1} \gamma_{r+N} X_{(r+N)ij} + u_j. \tag{11}$$

Here, the position effects are specified as in Pomplun and Ritchie (2004), where $X_{(r+N)ij} = 1$ when item i appears in position r, and $X_{(r+N)ij} = 0$ otherwise. However, the model differs in that ability u_j is estimated, rather than supplied as a covariate, and position effects across all items are estimated simultaneously.

Position $r = N$ is omitted as a reference position. The parameter γ_r is interpreted as the additive change in η_{ij} associated with all items appearing in position r. Alexandrowicz and Matschinger (2008) demonstrated a similar version of this model using both simulated and real data examples.

A hypothetical data matrix for fitting a model such as in Equation 11 is included in Table 1. The data are in long format, with one row per item response and thus $N \times J$ total rows, assuming complete data. Columns 1 through 4 contain the item responses and the indices for items, persons, and position. The indicator variables for the item effects (minus the reference item) are displayed in the next three columns. These would correspond to X_{qij} in Equation 11. The position indicator variables, as conceived by Pomplun and Ritchie (2004) and Alexandrowicz and Matschinger (2008), are displayed in the remaining columns.

Comparing Models

Although the models in Equations 6 and 11 are both forms of logistic regression, and both utilize categorical variables for position, the position effect estimates are not the same. In Equation 6 (Pomplun & Ritchie, 2004), the position effect is estimated for a specific position and item. When using this approach there may be as many position effects as there are unique item-position combinations. For example, if $N = 20$ items were distributed into

2 testlets, each containing 10 of the items, a model containing terms for 9 position effects would be fit once per item, resulting in 180 total position effects (excluding intercepts). Without the nesting within testlets, the total number of position effects in this approach would be the number of items changing position times the number of available positions, at most totaling to $N \times N$ effects.

On the other hand, in Equation 11 (Alexandrowicz & Matschinger, 2008), each position effect is estimated across all items. When using this approach the total possible number of position effects will always be the number of items in the test, minus the reference position, and each position effect will apply to all items appearing in the corresponding position. Thus, the number of position effects decreases from Equation 6 to Equation 11; however, an effect is no longer specific to the item.

Equation 7 estimates a single item-specific position effect per item. This provides an efficient summary of the impact of position on item difficulty. The cost comes in assuming that the change is linear. Any nonlinear change will not be captured by γ_1 in Equation 7. Nonlinear change may result, for example, in decreasing difficulty at the beginning of the test, a leveling-off of difficulty at the center, and an increase toward the end. The linear component of such a change could be estimated to be zero, misrepresenting the actual change due to position. In this case, the position effect may more appropriately be modeled

using polynomial terms, as in Meyers et al. (2008), or using position indicators, as in Pomplun and Ritchie (2004) and Alexandrowicz and Matschinger (2008). Thus, the estimation of linear change as in Equation 7 is less complex, requiring fewer parameters than the nonlinear methods; however, important trends in the data may be missed or misrepresented. Visual inspection of the observed data, including plots of item proportion correct by position, should inform the choice of estimation method, whether position effects are linear, nonlinear, or categorical.

Other limitations of Equation 7 are that it does not control for person parameters, as do Equations 6 and 11, and it does not benefit from the information sharing that is provided by a multilevel modeling framework, as does Equation 11. The approaches taken by Pomplun and Ritchie (2004) and Davey and Lee (2011) represent what is referred to as *no pooling* (Gelman & Hill, 2007), the opposite of complete pooling as described above for Meyers et al. (2008). Complete pooling involves the exclusion of grouping variables (e.g., items) from a model. Instead, no pooling involves an individual model for each level of the grouping variable (e.g., a model for each item). Information sharing via *partial pooling*, as utilized in multilevel modeling, can be considered a compromise between the two. Information is pooled to the extent that data is lacking for a given level of a grouping variable. For further details on each type

of pooling and the benefits of partial pooling, see Gelman and Hill (2007).

Table 1: Example of Indicator Coding for Item and Position Effects

				Items			Positions		
y_{ij}	i	j	p_{ij}	X_{1ij}	X_{2ij}	X_{3ij}	X_{4ij}	X_{5ij}	X_{6ij}
1	1	1	1	1	0	0	1	0	0
0	2	1	2	0	1	0	0	1	0
1	3	1	3	0	0	1	0	0	1
1	4	1	4	0	0	0	0	0	0
0	1	2	4	1	0	0	0	0	0
1	2	2	3	0	1	0	0	0	1
0	3	2	2	0	0	1	0	1	0
0	4	2	1	0	0	0	1	0	0

Note: This data set includes information for 2 people taking 4 items, where person $j = 2$ sees the items in reverse order. X_{1ij}, X_{2ij}, and X_{3ij} are indicator variables for item index i, used to estimate item effects for items 1, 2, and 3 (item 4 is omitted as the reference item). X_{4ij}, X_{5ij}, and X_{6ij} are indicator variables for the position index p_{ij}, used to estimate position effects (position 4 is omitted as the reference position).

Summary

Research over the past 50 years has demonstrated that in certain situations item context and position can have a negative, potentially biasing impact on the estimation of item and person parameters. Early research described this impact in terms of change in average performance, at both the item and the item-set or test levels. Later research focused on the implications of item context and position from an IRT perspective, at the item-set or test level and, in a few cases, at the person level. Missing from the literature are

demonstrations of models for estimating item-level position effects. As a result, little is known about the extent to which item-level bias resulting from position effects may impact the estimation of IRT item and person parameters.

The necessity of incorporating the estimation of context and position effects into the IRT model has been highlighted in the literature. Brennan (1992) stated that "models need to be developed that explicitly incorporate the possibility that context effects exist" (p. 259). Kingston and Dorans (1984) recommended that "the development of more general models that incorporate parameters such as item familiarity or item position should proceed" (p. 154). And Davey and Lee (2011) suggested that "one possible direction for future analysis is to employ some IRT model that can incorporate item position as a predictor, as well as unique parameters" (p. 40). A handful of studies have moved in this direction (e.g., Alexandrowicz & Matschinger, 2008; Pomplun & Ritchie, 2004).

The present study demonstrates a new approach to modeling context and position effects, one which builds upon previous work. The first objective in this approach is to estimate the item, position, and person effects simultaneously. A hierarchical generalized linear modeling framework, similar to the one used by Alexandrowicz and Matschinger (2008), is adopted because of the flexibility it affords in the estimation of these effects and others across complex data

structures.

The second objective in this approach is to provide estimates of item-specific position effects, referred to here as item-position effects. Using only categorical indicator variables in the HGLM framework, this would require an interaction between the $N-1$ item indicator variables and the $N-1$ position indicator variables, resulting in an unreasonable number of parameters for anything but very short tests. Instead, position is included as a continuous variable, as in Meyers et al. (2008) and Davey and Lee (2011), potentially reducing the number of item-position interactions to $N-1$. With position as a continuous covariate, the item-position interaction effects become slopes, estimates of the average linear change in log-odds of correct response across all the positions in which an item appears. Rather than describe the overall effect of each position on any item presented in that position, these effects index change in the difficulty of each item per shift in the position of the item within the test form.

Similar to Meyers et al. (2008) and Davey and Lee (2011), with position as a continuous variable the HGLM resembles a longitudinal model for items, where the item-position slope estimates change in item difficulty across positions and people. However, unlike Meyers et al. (2008) and Davey and Lee (2011), the model accounts for differences in person ability at each position by also

including person effects. Position effects are then conceptualized as item characteristics or item parameters, and the assumption of equivalence in person ability across position is not made.

As with any of the various position effect conceptualizations discussed above, significant values for one or more item-position effect within a test form would suggest bias in the estimation of the remaining model parameters. When position effects are conceptualized and estimated as item parameters within the HGLM, they can be used to examine item-level sources of bias and potential item-level violations of the assumption of local independence. Thus, this approach can be used to determine the appropriateness of a given item for future implementations in a CAT or scrambled test form.

The following chapter outlines the method employed in this study to examine position effects, including the research questions addressed and the formulation of the item-position effect HGLM. Real and simulated data were used to demonstrate the model.

Chapter III: Method

Research Questions

The purpose of this study was to demonstrate a hierarchical generalized linear model for estimating item-position effects. The HGLM was first applied to data from two operational testing programs, as described in the next section. These real data analyses served as an initial trial of the HGLM, providing an indication of the extent to which item-position effects are present and detectable with actual test scores.

Results of the real data analysis and findings from previous research informed the design of a simulation study involving item-position effects in varying degrees. Simulated item responses were used to assess the HGLM recovery of the simulated data generating parameters. Three main research questions were addressed:

1. How is recovery of item, item-position, and person parameters influenced by the number of examinees taking the test?

2. How is recovery of item, item-position, and person parameters influenced by the number of items in the test?

3. How is recovery of item, item-position, and person parameters influenced by the distribution of position effects?

Although all of the generating parameters were considered, the examination of parameter recovery focused on bias introduced by item position and the extent to which item difficulty estimates from three different HGLM were impacted by the simulated position bias. The next section presents the formulation of these three and other related HGLM. The remainder of the chapter describes the real and simulated data, the simulation design, estimation methods, and approach to analyzing the results.

Model Formulations

Base Model: M0

In HLM terminology (Kamata, 2001; Raudenbush & Bryk, 2002), Equations 8 and 11 would be considered combined or mixed models, with item responses nested within persons, and item parameters fixed across persons. The

unconditional model in Equation 8 is written in hierarchical form as:

$$\begin{aligned}
\eta_{ij} &= \beta_{0j} + \sum_{q=1}^{N-1} \beta_{qj} X_{qij} \\
\beta_{0j} &= \gamma_{00} + u_{0j} \\
\beta_{1j} &= \gamma_{10} \\
&\vdots \\
\beta_{(N-1)j} &= \gamma_{(N-1)0},
\end{aligned} \tag{12}$$

where the combined form is obtained through substitution. This HGLM, or multilevel logistic regression, is equivalent to the Rasch model in Equation 3. As in other IRT models, effects for items and persons are estimated simultaneously. The model is flexible in that it can also accommodate repeated measures (e.g., Pastor & Beretvas, 2006) and additional variables at both the item and person levels (e.g., Doran, Bates, Bliese, & Dowling, 2007; Kamata, 2001).

In Equation 12, the random term u_{0j} is the person parameter, considered the ability or trait level of person j. The interpretation of γ_{00} and γ_{q0} depends on the values in the item indicator variables X_{qij}. Using dummy coding, as in Table 1, the intercept γ_{00} is the parameter for the selected reference item $qref$ and γ_{q0} is the parameter for item q expressed as a difference from the reference. Alternative coding schemes exist and may be preferable. For example, with what is referred to as effect coding, $X_{qij} = 1$ when $i = q$, as with dummy coding;

however, $X_{qij} = 0$ only when $i \neq qref$. For $i = qref$, $X_{qij} = -1$. As a result, γ_{00} becomes the mean item parameter and the remaining parameters are expressed as differences from the mean.

In practice, the Rasch logit scale is often centered on the mean item effect as opposed to the effect of a particular item. Centering on the mean item effect aids in the interpretation of other effects, such as effects for item position, as described below. The following models are presented under the assumption that effect coding is used in X_{qij}.

Equation 12 serves as a base model (M0) from which many of the context and position effect models discussed in Chapter 2 can be developed. For example, Monk and Stallings (1970) examined the effect of item position for the same items ordered differently across two test forms. Such a test position effect may be obtained from M0 by including an indicator variable for group membership, W_{1j}, in the level 2 intercept model:

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \sum_{q=1}^{N-1} \beta_{qj} X_{qij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} W_{1j} + u_{0j} \\ \beta_{1j} &= \gamma_{10} \\ &\vdots \\ \beta_{(N-1)j} &= \gamma_{(N-1)0}. \end{aligned} \tag{13}$$

Here, $W_{1j} = 1$ if examinee j belongs to the target group and $W_{1j} = 0$ otherwise, and γ_{01} estimates the test position effect, the average overall change in log-odds for the target group. The same grouping variable could also be included in the remaining level 2 models to estimate item-position effects for the two test forms, as in Mollenkopf (1950):

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \sum_{q=1}^{N-1} \beta_{qj} X_{qij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} W_{1j} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} W_{1j} \\ &\vdots \\ \beta_{(N-1)j} &= \gamma_{(N-1)0} + \gamma_{(N-1)1} W_{1j}. \end{aligned} \tag{14}$$

Here, the terms $\gamma_{01}, \gamma_{11}, \ldots, \gamma_{(N-1)1}$ index the effect of the item rearrangement from one form to the other on each item. For example, if item 1 were estimated to be easier for the target group than for the reference group, γ_{11} would be a positive value. Additional person-level variables could also be added, such as a measure of examinee anxiety level, to estimate position interaction effects, as in Smouse and Munz (1968).

When item position can take on a wide range of values, for example, any position in the test form, item-position interaction effects can be estimated using continuous position covariates in each level 2 item-effect model, in place of W_{1j}

in Equation 14. However, in each level 2 model, the position covariate would change based on the specific item to which it pertains. The nature of the item position variable, which varies by both items and persons, reveals that it could equivalently enter the model at level 1. This equivalence across levels suggests that typical item-response data, with each item taken by more than one person, are more appropriately considered to have a crossed or partially crossed data structure (Doran et al., 2007), rather than a structure where items are nested within persons (Kamata, 2001; Raudenbush & Bryk, 2002). This distinction is discussed further below.

Position Effect Model: M1

Building upon M0, the following model (M1) contains a main effect for position (β_{Nj}) at level 1. Here, p_{ij} is the position $(1, 2, \ldots, N)$ of item $q = i$ for

examinee j:

$$\begin{aligned}
\eta_{ij} &= \beta_{0j} + \sum_{q=1}^{N-1} \beta_{qj} X_{qij} + \beta_{Nj} p_{ij} \\
\beta_{0j} &= \gamma_{00} + u_{0j} \\
\beta_{1j} &= \gamma_{10} \\
&\vdots \\
\beta_{(N-1)j} &= \gamma_{(N-1)0} \\
\beta_{Nj} &= \gamma_{N0}.
\end{aligned} \tag{15}$$

The fixed effect for position γ_{N0} is the mean change in log-odds across all positions, after controlling for the individual item difficulties.

Item-Position Effect Model: M2

The following model (M2) contains a main effect for position at level 1 and $N-1$ additional item-position interaction terms (β_{N+q}) at level 1 for the

item-position effects:

$$\begin{aligned}
\eta_{ij} &= \beta_{0j} + \sum_{q=1}^{N-1} \beta_{qj} X_{qij} + \beta_{Nj} p_{ij} + \sum_{q=1}^{N-1} \beta_{(N+q)j} X_{qij} p_{ij} \\
\beta_{0j} &= \gamma_{00} + u_{0j} \\
\beta_{1j} &= \gamma_{10} \\
&\vdots \\
\beta_{(N-1)j} &= \gamma_{(N-1)0} \\
\beta_{Nj} &= \gamma_{N0} \\
\beta_{(N+1)j} &= \gamma_{(N+1)0} \\
&\vdots \\
\beta_{(2N-1)j} &= \gamma_{(2N-1)0}.
\end{aligned} \tag{16}$$

Model M2 results in the intercept again varying across persons at level 2, with the position main effect, $N-1$ item effects, and $N-1$ interaction effects fixed at level 2.

In M2R, a reformulation of M2, the item by position interaction effects

are parameterized equivalently as level 2 covariates:

$$\begin{aligned}
\eta_{ij} &= \beta_{0j} + \sum_{q=1}^{N-1} \beta_{qj} X_{qij} + \beta_{Nj} p_{ij} \\
\beta_{0j} &= \gamma_{00} + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11} p_{1j} \\
&\vdots \\
\beta_{(N-1)j} &= \gamma_{(N-1)0} + \gamma_{(N-1)1} p_{(N-1)j} \\
\beta_{Nj} &= \gamma_{N0}.
\end{aligned} \tag{17}$$

Here, each of the position covariates enters only at the appropriate level-2 item effect model, making the item subscript on p unnecessary. In the combined form of M2R, with the position terms moved to the end of the equation,

$$\begin{aligned}
\eta_{ij} = \gamma_{00} + u_{0j} + \gamma_{10} X_{1ij} + \cdots + \gamma_{(N-1)0} X_{(N-1)ij} \\
+ \gamma_{11} X_{1ij} p_{1j} + \cdots + \gamma_{(N-1)1} X_{(N-1)ij} p_{(N-1)j} + \gamma_{N0} p_{ij},
\end{aligned} \tag{18}$$

which differs from the formulation in M2 only by the subscripts on p and γ for the item-position interactions. In the combined form of M2,

$$\begin{aligned}
\eta_{ij} = \gamma_{00} + u_{0j} + \gamma_{10} X_{1ij} + \cdots + \gamma_{(N-1)0} X_{(N-1)ij} \\
+ \gamma_{(N+1)0} X_{(N+1)ij} p_{ij} + \cdots + \gamma_{(2N-1)0} X_{(2N-1)ij} p_{ij} + \gamma_{N0} p_{ij}.
\end{aligned} \tag{19}$$

Reduced to a single item q, M2R becomes

$$\eta_{ij} = \gamma_{00} + \gamma_{q0} + u_{0j} + \gamma_{N0} p_{ij} + \gamma_{q1} p_{qj}, \tag{20}$$

and M2 becomes

$$\eta_{ij} = \gamma_{00} + \gamma_{q0} + u_{0j} + \gamma_{N0} p_{ij} + \gamma_{(N+q)0} p_{ij}. \tag{21}$$

In both Equations 20 and 21 the log-odds of correct response for item q are modeled as a linear combination of the mean item effect, the additional effect for item q, the random effect for the examinee, a main effect for position, and the item-position interaction effect for item q.

Models M2 and M2R are algebraically equivalent. When fit as nested models, however, results may differ for incomplete data sets, depending on the multilevel modeling software and specifications used. With missingness on any covariate at level two, a person may be removed, by default, from all level 2 modeling. On the other hand, missingness at level 1 results in exclusion of only the data for person j on item i. For this reason, the formulation in M2 may be preferred.

Other Position Effect Models

Model M2 is most appropriate when position varies randomly across items and persons, such as in the scrambled testing administrations discussed above. It may also be used when items appear in roughly each position across a large number of forms. In these contexts, M2 resembles a longitudinal model, with items as a type of repeated measure, where the intercepts in each level 2

model control for differences in item difficulty and the position slopes control for linear changes in the item difficulty across the test form, in other words, linear growth. This perspective suggests that position effects may also be conceptualized as person parameters:

$$\begin{aligned}
\eta_{ij} &= \beta_{0j} + \sum_{q=1}^{N-1} \beta_{qj} X_{qij} + \beta_{Nj} p_{ij} \\
\beta_{0j} &= \gamma_{00} + u_{0j} \\
\beta_{1j} &= \gamma_{10} \\
&\vdots \\
\beta_{(N-1)j} &= \gamma_{(N-1)0} \\
\beta_{Nj} &= \gamma_{N0} + u_{Nj}.
\end{aligned} \tag{22}$$

This position effect model (M3) is an extension of M1, with a main effect for position, β_{Nj}, which varies across persons at level 2. In combined form and reduced to item q, M3 becomes

$$\eta_{ij} = \gamma_{00} + \gamma_{q0} + u_{0j} + \gamma_{N0} p_{ij} + u_{Nj} p_{ij}. \tag{23}$$

The term u_{Nj} represents the change in log-odds for examinee j, across the test form, associated with the sequence in which the items were presented, and γ_{N0} is the average linear change across all examinees, as in M1. The remaining components are also interpreted as in M1 and M2.

Item-position effects may also be conceptualized as slopes which vary by items. In this case, both items and persons are considered to be random effects, resulting in a cross-classified item-position effect model (M4), shown here in combined form:

$$\eta_{ij} = \gamma_0 + t_{0i} + u_{0j} + \gamma_1 p_{ij} + t_{1i} p_{ij}. \tag{24}$$

In M4, random effects for items are denoted with t. The log-odds are modeled as a function of an intercept γ_0, which varies by items (with random effect t_{0i}) and by persons (with random effect u_{0j}), and a position effect γ_1 which varies by items (with random effect t_{1i}). The inclusion of random intercepts for items results in a change of scale, in comparison to the models above where only person effects are random. In M1, M2, and M3, the intercept γ_{00} is the mean item effect, conditional on position, that is, at $p = 0$. In M4, without fixed effects for items, γ_0 is simply the mean log-odds, conditional on position.

Unlike fixed effects, individual random effects such as t_{0i} are typically not tested for statistical significance. Instead, the variability of t_{0i} would be examined as an overall index of position effect bias. In this way, modeling with M4 is similar to the approach of Yen (1980), who examined variability in item difficulty due to change in item context.

Finally, model M4 can be reduced further for certain types of item response data. In some testing programs, a limited number of items, usually

pilots, vary randomly in position across the form. The remaining items may appear in fixed positions, or may be presented adaptively, as in a CAT. When the number of pilot items is sufficient to estimate person parameters, any of models M1, M2, M3 and M4 may be fit to the pilot item responses, either with the remaining item parameters fixed or the remaining items removed from the data set. In the first case, pilot item parameters will be anchored to the scale defined by the remaining items. In the second, the pilot items will define the scale. With few pilot items, where accurate estimation of u_{0j} is not possible, the data set can be reduced to one observation per examinee. This would require removal of the random person effects from the model:

$$\begin{aligned} \eta_{hi} &= \beta_{0i} + \beta_{1i} p_{hi} \\ \beta_{0i} &= \gamma_{00} + t_{0i} \\ \beta_{1i} &= \gamma_{10} + t_{1i}. \end{aligned} \tag{25}$$

In this model (M5), η_{hi} is the log-odds of correct response for observation h on item i. The subscript h indicates a nesting of responses or observations within items. The subscript h could equivalently indicate a nesting of item positions within items. As in M4, random effects for items are denoted by t. Rather than controlling for person ability u_{0j}, as in previous models, in M5 it is assumed that the different groups responding to each item in each position are equivalent in terms of ability. In other words, β_{0i} is assumed to be fixed across samples of

people.

Summary

Table 2 displays the components of models M0, M1, M2, M3, M4, and M5 in combined reduced form. Model M0 is the base model, equivalent to the Rasch model. Model M1 is considered a position effect model, as it contains only a main effect for position, in addition to item and person effects. Models M2, M3, M4, and M5 are considered item-position effect models, as they contain additional item-specific position effects. In the framework presented by De Boeck and Wilson (2004), M0 would be considered a doubly descriptive, i.e., non-explanatory IRT model; M1, M2, M4 and M5 would be different types of item explanatory models; and M3 would be a person explanatory model.

Table 2: Reduced Position Effect Model Formulations

Model	Reference	Item	Person	Position	Item-Position
M0	γ_{00}	γ_{q0}	u_{0j}		
M1	γ_{00}	γ_{q0}	u_{0j}	$\gamma_{N0}p_{ij}$	
M2	γ_{00}	γ_{q0}	u_{0j}	$\gamma_{N0}p_{ij}$	$\gamma_{(q+N)0}p_{ij}$
M2R	γ_{00}	γ_{q0}	u_{0j}	$\gamma_{N0}p_{ij}$	$\gamma_{q1}p_{qj}$
M3	γ_{00}	γ_{q0}	u_{0j}	$\gamma_{N0}p_{ij}$	$u_{Nj}p_{ij}$
M4	γ_{0}	t_{0i}	u_{0j}	$\gamma_{1}p_{ij}$	$t_{1i}p_{ij}$
M5	γ_{00}	t_{0i}		$\gamma_{10}p_{hi}$	$t_{1i}p_{hi}$

Note: Fixed effects are denoted by greek letters, i.e., γ, and random effects are denoted by arabic letters t for items and u for people. The item-position effect for M3 is actually a person-position effect. The item position covariate is denoted by p.

When using effect coding in M2, the $qref$ item effect γ_{qref0} and item-position effect $\gamma_{(N+qref)0}$ are not estimated directly by the model. Because γ_{q0} and $\gamma_{(N+q)0}$ are deviations from the mean effects γ_{00} and γ_{N0}, the item and item-position effects for item $qref$ can be obtained indirectly as

$$\gamma_{qref0} = -\sum_{q=1}^{N-1} \gamma_{q0}, \tag{26}$$

and

$$\gamma_{(N+qref)0} = -\sum_{q=1}^{N-1} \gamma_{(N+q)0}. \tag{27}$$

Here, the negative sum of the $N-1$ estimated parameters is the unestimated reference effect.

Models M2 and M4 differ only in their treatment of items as fixed versus random. In general, the literature recommends that the choice between fixed versus random effects be based on whether or not the grouping factor of interest represents a sample from a larger population of groups (Raudenbush & Bryk, 2002; Gelman & Hill, 2007). In the models above, the grouping factor is items, and the items that constitute a test form could be thought of as a sample from a larger bank of items. This would suggest that items should be treated as random effects. However, in addition to the sample-population relationship, the purpose of the model must also be considered. The reason for estimating item effects and item-position interaction effects with a subset of items is not to generalize

inferences to a broader item population. Instead, the specific estimates are of interest, as they can inform decisions regarding future administrations of the same items. Thus, for the purposes of this study, model M2 seemed more appropriate than M4.

Since people are considered a sample from a broader population, they too can be modeled as random effects, and in this case the decision seems justified. Although the specific ability estimates obtained at the individual level, that is, u_{0j}, are typically of interest, the particular levels of the persons grouping factor represented in a given sample do not have meaning beyond the sample. That is, any random sample from the examinee population would suffice for the purposes of the item-position effects models described above. Again, the purpose of the model is to examine position effects as item parameters which may impact the usefulness of an item in future test administrations, that is, with future samples of examinees.

Real Data Study

Existing item response data sets were obtained from two large-scale, national testing programs. Each program followed recommended item writing and test development guidelines and reported reliability and validity information for the tests discussed below.

Reading Test Data Set

The first data set, referred to as RT (reading test), came from a reading CAT administered to first graders. Items were multiple choice with four response options. A precalibration design was used, wherein a subset of non-adaptive, non-operational pilot items was interspersed among operational adaptive items for each examinee. Position of the pilot items was randomized at the time of administration.

Each examinee only responded to a small subset of the total number of pilot items. Thus the data set was reduced to a single item response per person; for examinees seeing multiple items, one response was randomly selected and the others discarded. The final data set included 93,238 examinees, each responding to one of 50 pilot items in a randomized position.

GRE Data Sets

The second data set came from the GRE, as described in Davey and Lee (2011). These data were collected prior to the release of the current revised version of the GRE. At the time of data collection the GRE included three separate, timed, test sections, two of which consisted of adaptive multiple-choice items. The adaptive sections tested quantitative and verbal reasoning. The third section was non-operational and non-adaptive, and contained pilot items from

either the quantitative or verbal item banks. This design, described in Chapter 2, is referred to as section pre-equating (Holland & Thayer, 1985).

Davey and Lee (2011) manipulated the items and item orderings for 6 non-operational pilot sections of the GRE: 3 quantitative and 3 verbal. Items were selected to be representative of the GRE item bank. Data from the 3 quantitative item sets were utilized in this study. These are referred to as GRE1, GRE2, and GRE3. Each contained 28 items rearranged into 13 fixed forms with different item orders, where each item appeared with roughly equal frequency in each of several general locations across the form. Table 3 displays the different item orderings for the 13 quantitative forms, with one column per form. These same item positions were used for each of the three GRE item sets. In each row are the positions of an item across the 13 forms. Table 4 displays the number of examinees taking each form and item set.

Real Data Analysis

Data sets RT and GRE1, GRE2, and GRE3 each contained unique item and person identifiers for differentiating among responses within the data sets, dichotomous scored responses for each item-person interaction, and the position in which an item appeared for each item-person interaction. The data were all reformatted to contain one item response per row prior to modeling. A series of

Table 3: GRE Item Positions Across 13 Forms

1	2	3	4	5	6	7	8	9	10	11	12	13
1	27	25	23	21	19	16	14	12	10	7	5	4
2	26	22	18	14	9	4	28	24	20	15	11	7
3	1	27	25	23	21	18	16	14	12	9	8	6
4	28	24	20	16	12	6	2	26	22	17	13	9
5	3	1	27	25	23	20	18	16	14	12	10	8
6	2	26	22	18	14	9	4	28	24	19	15	11
7	5	3	1	27	25	22	20	18	17	14	12	10
8	4	28	24	20	16	11	7	2	26	21	17	13
9	7	5	3	1	27	24	22	21	19	16	14	12
10	6	2	26	22	18	13	9	5	28	23	19	15
11	9	7	5	3	1	26	25	23	21	18	16	14
12	8	4	28	24	20	15	11	7	3	25	21	17
13	11	9	7	5	3	1	27	25	23	20	18	16
14	10	6	2	26	22	17	13	9	5	28	23	19
15	13	11	9	7	24	3	1	27	25	22	20	18
16	12	8	4	28	6	19	15	11	7	2	26	21
17	15	13	11	10	8	5	3	1	27	24	22	20
18	14	10	6	2	26	21	17	13	9	4	28	24
19	17	15	14	12	10	7	5	3	1	26	24	22
20	18	16	15	13	11	8	6	4	2	27	25	23
21	16	12	8	4	28	23	19	15	11	6	2	26
22	20	19	17	15	13	10	8	6	4	1	27	25
23	19	14	10	6	2	25	21	17	13	8	4	28
24	23	21	19	17	15	12	10	8	6	3	1	27
25	21	17	12	8	4	27	23	19	15	10	6	2
26	22	18	13	9	5	28	24	20	16	11	7	3
27	25	23	21	19	17	14	12	10	8	5	3	1
28	24	20	16	11	7	2	26	22	18	13	9	5

Note: The same ordering scheme was used with GRE1, GRE2, and GRE3. Each column includes the position of items 1 through 28 for the corresponding form. Form 1, in column 1, can be considered a base item ordering. Each row contains the positions of a given item across forms. For example, in the first row, item 1 appeared in positions 1, 27, 25, etc., for forms 1, 2, 3, etc.

Table 4: GRE Sample Sizes Across 13 Forms

Form	GRE1	GRE2	GRE3
1	146	163	143
2	140	128	166
3	126	145	138
4	134	141	154
5	142	137	131
6	122	152	145
7	138	140	136
8	129	122	137
9	150	146	143
10	133	161	139
11	153	159	155
12	139	129	119
13	145	129	137
Total	1797	1852	1843

nested models were then fit to each data set.

With one item response per examinee, model M5 seemed most appropriate for the reading test. This model involved complete pooling across persons, where person effects were excluded from the model. The base model for the reading test was a random intercepts model, with item responses nested within items. In the second reading test model, a main effect for position was added at level 1. In the final model, M5, position effect slopes varied by item. A chi-square likelihood ratio test, defined below, was used to compare deviance statistics for M5 and the less complex position effect model. A statistically significant value for this test would suggest the presence of non-negligible

variability in position slopes across items. If nonsignificant, a z-test would follow to determine the necessity of the main effect for position. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were also considered in the model comparisons.

Three models were also fit to GRE1, GRE2, and GRE3: the base model, M0; the position main effect model, M1; and the full item-position effect model, M2. Model M0 contained only item effects and random intercepts for persons. A fixed effect for position was added in model M1. Item by position interaction effects were added in model M2. A chi-square likelihood ratio test was used to compare deviance statistics for M2 and M1. This served as an omnibus test for the item-position effect interaction terms. If statistically significant, inspection of the individual interaction terms would be justified. Otherwise, M1 would be investigated, based on the statistical significance of the main effect for position. The AIC and BIC were also considered as indices of model fit. These steps were repeated for each of the three quantitative GRE item sets.

Item, item-position, and person effects were also examined individually for each model. Visual inspections accompanied each analysis. The results of these analyses, reported in Chapter 4, were used to inform the design of a simulation study.

Simulation Study

Simulation Factors

Three factors were varied in the simulation of item responses: sample size (S), test length (T), and the distribution of item-position effects (P). Sample size and test length were chosen as simulation factors because of their expected impact on sampling error and accuracy of estimation of item, position, and person parameters. Item-position effect distribution was chosen as it was the main focus of the study. Each of these factors contained multiple levels, as described below.

Sample size consisted of two levels: S1, with 500 examinees; and S2, with 1000 examinees. These levels corresponded to small and medium samples, and were chosen to represent two common sizes that may be obtained in practice. At each replication of the simulation a new group of respondents was sampled from a normally distributed population with mean 0 and standard deviation 1 (as discussed in Chapter 4, ability was normally distributed for each of the GRE samples). Thus, S1 and S2 came from the same population ability distribution and only the size of the sample changed across conditions. With a new group sampled at each simulation replication, persons were treated as random effects, corresponding to the formulation in models M0, M1, and M2.

Item parameters were chosen based on two tests lengths: T1, a short test, with 20 items; and T2, a medium-length test, with 40 items. T1 corresponded to an assessment that might be used in low-stakes decision making, such as a classroom assessment or progress monitoring measure. The reading assessment would be considered such a test. Brief tests such as these are often used with younger student populations, for example, students in kindergarten or first grade, where time for testing is limited. T2 corresponded to a longer progress monitoring measure or high-stakes test such as the GRE or GMAT, which contain between 30 and 40 items per section.

Each test was constructed to cover the same section of logit difficulty, ranging from -1 to 2 (as discussed in Chapter 4, the GRE items spanned a similar range). Items were equally spaced across this range, with T1 consisting of the sequence $(-1.00, -0.84, -0.68, \ldots, 1.68, 1.84, 2.00)$, and T2 the sequence $(-1.00, -0.92, -0.85, \ldots, 1.85, 1.92, 2.00)$. As a result, the means of T1 and T2 were the same, 0.5, and the standard deviations similar, 0.93 and 0.90. Mean item difficulty, or in this case easiness, was slightly higher on the logit scale than the mean ability 0. A mean item logit of 0.5 converts to a mean item proportion correct of .62 for $u = 0$, indicating that a person of average ability 0 is predicted on average to have a 62% chance of correct response.

Unlike person ability, the item parameters in T1 and T2 were fixed across

other conditions for a particular level of test length. This corresponded to the fixed item effect formulation in models M0, M1, and M2, and it allowed for an examination of parameter recovery, as described below. The same was true of the item-position generating parameters.

Item-position generating parameters were chosen based on results from the real data analysis and findings of previous research. P1 corresponded to small or negligible position effects, P2 to practice effects, and P3 to fatigue effects. For each level, parameters were sampled from a normal distribution; however, P1, P2, and P3 differed in terms of mean, standard deviation, and the constraints imposed on the minimum and maximum values and on the correlation with item parameters in the corresponding level of T, r_{TP}.

In T1-P1, a set of 20 parameters was sampled from a normal distribution with mean 0 and standard deviation 0.01. The parameters were constrained to fall within the range $[-0.02, 0.02]$, and to have an absolute correlation with the corresponding item difficulties less than .01. In T2-P1, these same values were duplicated to obtain a set of 40 parameters. Thus, each parameter in T1-P1 was simply included twice in T2-P1.

A similar procedure was used to obtain parameters for P2 and P3. In T1-P2, 20 parameters were sampled from a normal distribution with mean 0.01 and standard deviation 0.01. The parameters were constrained to fall within the

range $[-0.01, 0.03]$, and to correlate with the T1 item difficulties within the range $[-.62, -.58]$. For T2-P2, these values were duplicated, as in T2-P1. In P3 the same procedures were used as in P2, but the population mean was -0.01 and values were constrained to fall within the range $[-0.03, 0.01]$. The correlation r_{T1P3} was again constrained to fall within the range $[-.62, -.58]$.

The positive slope values in P2 corresponded to practice effects, where item logit tends to increase and items tend to become easier with position increases. The negative slope values in P3 corresponded to fatigue effects, where item logit decreases and items tend to become harder with position increases. In each case, the correlation between item and item-position effects of roughly $-.6$ indicated that easier items, with higher logits, tended to have lower slopes, and harder items, with lower logits, tended to have higher slopes. These values and magnitudes were similar to those found in the literature and in the real data results presented below.

Table 7 summarizes the generating parameters for T and P. Because they were fixed sequences, generating parameters in T1 and T2 were essentially the same in mean, standard deviation, skewness, kurtosis, minimum, and maximum (compare rows 1 and 5 in Table 7). Generating parameters for P were designed to differ in mean, minimum, maximum, and for P1 versus P2/P3, correlation with T. These objectives appear to have been met (compare rows 3, 4, and 5).

Finally, because parameters in P were reused as test length increased, P1, P2, and P3 were very similar for a given level of P across T (compare rows 2 and 6, 3 and 7, 4 and 8).

Tables 5 and 6 contain all of the fixed generating parameters T1 and T2, and the corresponding generating parameters for P1, P2, and P3 for each test length. The generating parameters are represented visually in Figures 1 and 2.

Table 5: T1 Item and Item-Position Generating Parameters

Item	T	P1	P2	P3
1	−1.000	−0.005	0.022	0.007
2	−0.842	0.016	0.006	−0.006
3	−0.684	−0.011	0.018	−0.002
4	−0.526	0.001	0.021	0.006
5	−0.368	0.005	0.014	−0.013
6	−0.211	−0.017	0.027	−0.020
7	−0.053	0.011	0.017	−0.002
8	0.105	−0.001	0.012	−0.003
9	0.263	−0.002	0.007	−0.012
10	0.421	−0.006	−0.009	−0.017
11	0.579	0.007	0.013	−0.021
12	0.737	−0.001	0.015	−0.005
13	0.895	−0.016	0.006	−0.012
14	1.053	0.005	0.010	−0.023
15	1.211	0.007	−0.003	−0.024
16	1.368	0.013	−0.001	−0.018
17	1.526	−0.003	0.006	−0.015
18	1.684	0.003	−0.001	−0.007
19	1.842	−0.005	0.012	−0.018
20	2.000	−0.006	−0.002	−0.013

The largest absolute value for an item-position effect was 0.027, in T1-P2

Table 6: T2 Item and Item-Position Generating Parameters

Item	T	P1	P2	P3	Item	T	P1	P2	P3
1	−1.000	−0.005	0.022	0.007	21	0.538	0.007	0.013	−0.021
2	−0.923	−0.005	0.022	0.007	22	0.615	0.007	0.013	−0.021
3	−0.846	0.016	0.006	−0.006	23	0.692	−0.001	0.015	−0.005
4	−0.769	0.016	0.006	−0.006	24	0.769	−0.001	0.015	−0.005
5	−0.692	−0.011	0.018	−0.002	25	0.846	−0.016	0.006	−0.012
6	−0.615	−0.011	0.018	−0.002	26	0.923	−0.016	0.006	−0.012
7	−0.538	0.001	0.021	0.006	27	1.000	0.005	0.010	−0.023
8	−0.462	0.001	0.021	0.006	28	1.077	0.005	0.010	−0.023
9	−0.385	0.005	0.014	−0.013	29	1.154	0.007	−0.003	−0.024
10	−0.308	0.005	0.014	−0.013	30	1.231	0.007	−0.003	−0.024
11	−0.231	−0.017	0.027	−0.020	31	1.308	0.013	−0.001	−0.018
12	−0.154	−0.017	0.027	−0.020	32	1.385	0.013	−0.001	−0.018
13	−0.077	0.011	0.017	−0.002	33	1.462	−0.003	0.006	−0.015
14	0.000	0.011	0.017	−0.002	34	1.538	−0.003	0.006	−0.015
15	0.077	−0.001	0.012	−0.003	35	1.615	0.003	−0.001	−0.007
16	0.154	−0.001	0.012	−0.003	36	1.692	0.003	−0.001	−0.007
17	0.231	−0.002	0.007	−0.012	37	1.769	−0.005	0.012	−0.018
18	0.308	−0.002	0.007	−0.012	38	1.846	−0.005	0.012	−0.018
19	0.385	−0.006	−0.009	−0.017	39	1.923	−0.006	−0.002	−0.013
20	0.462	−0.006	−0.009	−0.017	40	2.000	−0.006	−0.002	−0.013

Table 7: Descriptive Statistics for Generating Parameters

Condition	$Mean$	SD	$Skew$	$Kurt$	Min	Max	N	r_{TP}
T1	0.500	0.934	0.000	−1.381	-1.000	2.000	20	
T1-P1	0.000	0.009	−0.113	−0.800	-0.017	0.016	20	−0.003
T1-P2	0.010	0.009	−0.115	−0.881	-0.009	0.027	20	−0.615
T1-P3	−0.011	0.009	0.397	−0.971	-0.024	0.007	20	−0.592
T2	0.500	0.899	0.000	−1.290	-1.000	2.000	40	
T2-P1	0.000	0.009	−0.118	−0.682	-0.017	0.016	40	−0.003
T2-P2	0.010	0.009	−0.119	−0.768	-0.009	0.027	40	−0.614
T2-P3	−0.011	0.009	0.413	−0.863	-0.024	0.007	40	−0.591

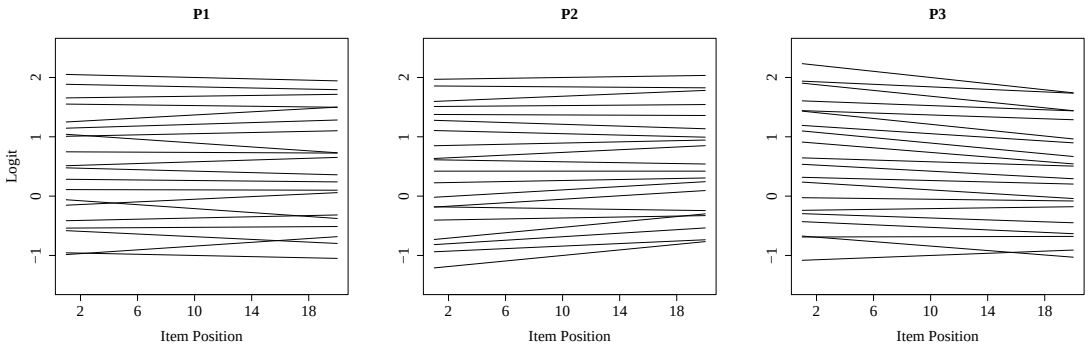

Figure 1: T1 item and item-position generating parameters for P1, P2, and P3, expressed as logits across position. Each line represents an item.

and T2-P2. In M2 this value is a linear slope, representing a change of 0.027 in the logit metric for a change in position of 1. Thus, for the small test T1, the maximum value under P3 would result in a predicted difference of $0.027 \times 20 = 0.54$ logits from the beginning to the end of the test. For T2, this difference would be $0.027 \times 40 = 1.08$ logits. Smaller values, such as the largest absolute value for P1, 0.017, result in a smaller relative impact, $0.017 \times 20 = 0.34$ and $0.017 \times 40 = .68$.

Simulation Procedures

Item responses were simulated according to the following model:

$$\eta_{ij} = (\gamma_{00} + \gamma_{q0}) + u_{0j} + (\gamma_{N0} + \gamma_{(N+q)0})(p_{ij} - \bar{p}), \tag{28}$$

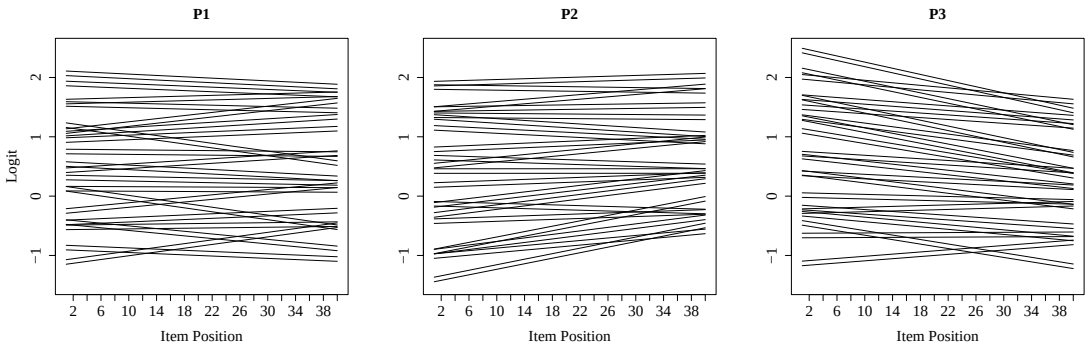

Figure 2: T2 item and item-position generating parameters for P1, P2, and P3, expressed as logits across position. Each line represents an item.

which corresponds to model M2 reduced to item q, as in Equation 21. The combined effect $(\gamma_{00} + \gamma_{q0})$ corresponds to the single generating parameter for item q in T, and the combined effect $(\gamma_{N0} + \gamma_{(N+q)0})$ corresponds to the single generating parameter in P. This generating model relates to the Rasch model formulation as:

$$P(y_{ij}) = \frac{1}{1 + e^{-(\eta_{ij})}}. \tag{29}$$

The vector of positions p_j was randomly sampled from the sequence 1 to N, the number of items in the test, for each examinee. This resulted in simulated scrambled test forms. Position was then mean-centered so that the generating parameters T1 and T2 would reflect true item difficulties at the center of the test, rather than at the beginning. Position $\bar{p}$ was 10.5 for T1 and 20.5 for T2.

The total number of simulation conditions was 2 (sample sizes) × 2 (test

lengths) × 3 (position effect distributions) = 12. Models M0, M1, and M2 were fit under each condition. Thus, the simulation study included three main steps at each replication for a given set of conditions:

1. sample person ability;
2. simulate item responses using ability, item, and item-position generating parameters;
3. fit models M0, M1, and M2 to simulated responses to obtain estimates of item, position, and person effects, and model fit statistics.

Estimation

For the GRE data sets, model M0 was fit in the Rasch modeling software *Winsteps* (version 3.72.2), using joint maximum-likelihood estimation. Models M0, M1, and M2 were then fit in *HLM6* (version 6.02a), using restricted penalized quasi-likelihood (PQL) estimation, and using a Laplace approximation to maximum likelihood (ML). Finally, the models were also fit with the *lme4* package (version 0.999375-41) within the statistical environment *R* (version 2.13.1), using an ML Laplace approximation. The reading test data were modeled only using *lme4*.

Table 8 demonstrates how the GRE1, GRE2, and GRE3 level 1 data file would be constructed in *HLM6*, and, similarly, in *lme4*. Each of these packages

utilizes a generalized linear modeling approach with a binomial or logit link for the dichotomous response vector. Person effects u_{0j} are estimated as varying intercepts which are assumed to be distributed normally with mean 0. Model M5 includes varying intercepts t_{0j} and position slopes t_{1j} for items. These are assumed to be multivariate normally distributed with means of 0.

The results from the three software packages were nearly identical for fitting M0 to the GRE data. Results differed slightly between the *HLM6* PQL and ML estimation methods for M1 and M2; however, the Laplace ML estimates from *HLM6* and *lme4* were essentially the same. The real data results presented in Chapter 4 were obtained with *lme4*.

Data were simulated and analyzed in *R*; however, *HLM6* with Laplace ML was used for model fitting with the simulated data because computation times were significantly less compared to *lme4*.

Parameter Recovery

Parameter recovery was assessed in terms of $Bias$, standard error (SE), and root mean squared error ($RMSE$). In the equations below, true generating parameters are indicated by γ_q for item q, and u_j for person j. To simplify notation, point estimates of these parameters are indicated by $\hat{\gamma}_q$ and $\hat{u}_j$, respectively. Means of these estimates are indicated by $\bar{\hat{\gamma}}_q$, taken for a single

Table 8: Example of Effect Coding for Item and Item-Position Effects

				Items			Item-Positions		
y_{ij}	i	j	p_{ij}	X_{1ij}	X_{2ij}	X_{3ij}	$X_{1ij}p_{ij}$	$X_{2ij}p_{ij}$	$X_{3ij}p_{ij}$
1	1	1	1	1	0	0	1	0	0
0	2	1	2	0	1	0	0	2	0
1	3	1	3	0	0	1	0	0	3
1	4	1	4	-1	-1	-1	-4	-4	-4
0	1	2	4	1	0	0	4	0	0
1	2	2	3	0	1	0	0	3	0
0	3	2	2	0	0	1	0	0	2
0	4	2	1	-1	-1	-1	-1	-1	-1

Note: As in Table 1, this data set includes information for 2 people taking 4 items, where person $j = 2$ sees the items in reverse order. X_{1ij}, X_{2ij}, and X_{3ij} are indicator variables for item index i, used to estimate item effects for items 1, 2, and 3 (the indicator for item 4 is omitted, with responses for item 4 coded as -1 across the remaining items). $X_{1ij}p_{ij}$, $X_{2ij}p_{ij}$, and $X_{3ij}p_{ij}$ are the item indicators multiplied by the position variable p_{ij}, used to estimate the item-position interaction effects.

item across replications, and $\bar{\hat{u}}_r$, taken across all people within replication r. A mean of person parameters, within a replication, was also needed; this mean is indicated by $\bar{u}_r$. Because items were treated as fixed, and persons as random, estimation of parameter recovery differed by parameter type.

Fixed Effects

For the item and item-position parameters, recovery was assessed across replications within a condition for a given item i. *Bias* was estimated as:

$$Bias_i = \bar{\hat{\gamma}}_q - \gamma_q, \tag{30}$$

where γ is a generic term representing one of the true generating parameters γ_{q0} or $\gamma_{(N+q)0}$, $\hat{\gamma}$ represents its estimate $\hat{\gamma}_{q0}$ or $\hat{\gamma}_{(N+q)0}$, and

$$\bar{\hat{\gamma}}_q = \frac{1}{R}\sum_{r=1}^{R}\hat{\gamma}_{rq}, \tag{31}$$

the average of parameter estimates across replications $R = 100$. The average $Bias$ and average absolute $Bias$ ($AbsBias$) were also estimated over all N items in a given condition as $\sum Bias_i/N$ and $\sum |Bias_i|/N$.

SE and $RMSE$ were estimated for item i over replications as:

$$SE_i = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\gamma}_{rq} - \bar{\hat{\gamma}}_q)^2}, \tag{32}$$

and

$$RMSE_i = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\gamma}_{rq} - \gamma_q)^2}. \tag{33}$$

Averages of the SE and $RMSE$ were taken across all items in a given condition as $\sum SE_i/N$ and $\sum RMSE_i/N$.

The equations above are defined in terms of a single generating parameter γ_q for item $q = i$, whether the item difficulty γ_{q0} or the item-position parameter $\gamma_{(N+q)0}$. Recovery of item difficulty was assessed for M0, M1 and M2. Recovery of the item position parameters was assessed for M2.

Parameter recovery was also assessed for each model by considering the true item difficulty across the test form, rather than simply at the center.

Estimation of the centered parameters γ_{q0} in T1 and T2 was expected to be similarly accurate across models. However, recovery toward the beginning and end of the test form was expected to degrade for M0, as the assumption of a singe item difficulty fixed across position would become less reasonable. Model M1, with a single position effect slope applied to all items, was also expected to produce more biased estimates of item difficulty as item position varied from the mean. Variability in M2 estimates was expected to increase, given the relatively larger number of parameters in the model.

The true difficulty of item i at position $p_i = h$ was obtained by combining the item and item-position parameters, $\gamma_{hq0} = \gamma_{q0} + \gamma_{(N+q)0}p_i$, where $\gamma_{(N+q)0}$ represents the entire position effect $\gamma_{N0} + \gamma_{(N+q)0}$, as in Equation 28. Recovery was again assessed in terms of $Bias$, SE and $RMSE$. Thus, Equations 30, 32, and 33 were also defined at the position level. For $Bias$, the mean parameter estimate $\bar{\hat{\gamma}}_q$ was replaced by the mean item difficulty estimate at position h, that is, $\bar{\hat{\gamma}}_{hq0} = \bar{\hat{\gamma}}_{q0} + \bar{\hat{\gamma}}_{(N+q)0}p_i$, as:

$$Bias_{hi} = \bar{\hat{\gamma}}_{hq0} - \gamma_{hq0}. \tag{34}$$

For M0, the item position effect was always 0, as was the mean across replications $\bar{\hat{\gamma}}_{(N+q)0}$. Because the position effect $\gamma_{(N+q)0}$ was not estimated in M0, $Bias_{hi}$ for M0 was always a combination of $Bias_i$ at position $\bar{p}$ and the

entire unestimated component $\gamma_{(N+q)0}p$. For M1, the item position effect was constant across items within a replication. M2 was the least restrictive model, estimating position effects for each item.

SE and $RMSE$ at position $p_i = h$ were estimated as:

$$SE_{hi} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\gamma}_{rhq0} - \bar{\hat{\gamma}}_{hq0})^2}, \tag{35}$$

and

$$RMSE_{hi} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\gamma}_{rhq0} - \gamma_{hq0})^2}. \tag{36}$$

Results are reported for these parameter recovery indices at the beginning and end of the test form, as $Bias\, p_1$, $Bias\, p_N$, $AbsBias\, p_1$, $AbsBias\, p_N$, $SE\, p_1$, $SE\, p_N$, $RMSE\, p_1$, and $RMSE\, p_N$. Recovery was also summarized at the condition level by averaging these values over all items in a given condition.

Random Person Effects

As random effects, person parameters differed for each replication of the simulation. Thus, person parameter recovery was assessed across all J persons in a given replication, rather than across replications for a given parameter. At each replication r, $Bias$ was estimated as:

$$Bias_r = \frac{1}{J}\sum_{j=1}^{J}(\hat{u}_j - u_j). \tag{37}$$

Absolute $Bias$ was estimated as:

$$BiasAbs_r = \frac{1}{J}\sum_{j=1}^{J}|\hat{u}_j - u_j|. \tag{38}$$

And $RMSE$ was estimated as:

$$RMSE_r = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{u}_j - u_j)^2}. \tag{39}$$

Across replications in a condition, $Bias$ was summarized as $\sum Bias_r/R$ and $\sum BiasAbs_r/R$, and $RMSE$ as the average value $\sum RMSE_r/R$. The mean absolute $Bias$ ($AbsBias$) was also summarize as $\sum |Bias_r|/R$.

The difference between $Bias$ estimates for fixed effects, in Equation 30, and random effects, in Equation 37, is subtle. In the first case, multiple estimates are obtained for a single parameter γ_q. In the second case, only one estimate is available for each parameter, since the person ability generating parameters are resampled at each replication.

For both the fixed and random effects, absolute values were used to summarize $Bias$ whenever $Bias$ was combined across estimates of differing generating parameters. For fixed effects, this occurred at the condition level, across items; within an item, across replications, the generating parameter was fixed and the absolute value not taken. For random effects, this occurred at the replication level, for each individual; thus, the absolute value was taken at the replication level in Equation 38. For random effects, change in generating

parameters also occurred at the condition level, across replications; in this case, the differing generating parameter is the mean $\bar{u}_r$ (Equation 37 can be reduced to estimate the difference $\bar{\hat{u}}_r - \bar{u}_r$). Thus, the mean absolute $Bias$ was also estimated as $\sum |Bias_r|/R$.

Model Fit

Finally, model fit was compared for models M0, M1, and M2, with the first considered to be a reduced form of the second, and the second a reduced form of the third. Deviance statistics were retained for each model at each replication, and likelihood ratio tests were used to test the appropriateness of the more complex models (M2 versus M1, M1 verses M0). Here, the more complex model is referred to as the alternative and the less complex nested model is referred to as the null.

The chi-square likelihood ratio (D) compares likelihoods for the null versus alternative models as:

$$D = -2 \log \left(\frac{likelihood_{null}}{likelihood_{alternative}} \right). \tag{40}$$

Distributing the $-2 \log$, this equation becomes a difference in deviance statistics:

$$D = deviance_{null} - deviance_{alternative}, \tag{41}$$

where $deviance = -2 \log(likelihood)$. By estimating more parameters the

likelihood of the alternative model will always be larger than that of the null model. As a result, the alternative deviance in Equation 41 will always be smaller, making D greater than 0. The size of the reduction in deviance D is used to assess whether or not the alternative model is more appropriate or fits better than the null. D is distributed approximately χ^2 with degrees of freedom df_{A-N} equal to the number of parameters estimated in the alternative model ($df_{alternative}$) minus the number of parameters estimated in the null (df_{null}).

Deviance statistics were also used to obtain the AIC and BIC for each model, where

$$\text{AIC} = 2 \times df + deviance, \tag{42}$$

and

$$\text{BIC} = \log(sample\ size) \times df + deviance. \tag{43}$$

With a multilevel model, different samples are present at each level. In the literature, the selection of BIC *sample size* varies by the particular field and the type of multilevel data that are modeled (McCoach & Black, 2008). The selection of *sample size* also differs by software package. For example, MPLUS and SAS PROC MIXED utilize the number of higher-level observations, which in this case would be J persons. *lme*4 utilizes the number of level-1 observations, which in this case would be $N \times J$ for complete data. BIC were obtained in this

study using the number of level-1 observations, a more conservative approach.

Model fit results were summarized across replications within each of the 12 conditions using the proportion of statistically significant likelihood ratio tests (favoring the alternative model) and the proportion of more complex models having lower AIC and lower BIC (also favoring the alternative model).

Chapter IV: Results

This chapter presents the results of the real data study and simulation study analyses described in Chapter 3. Results are first reported for the two real data studies, the first grade reading assessment and the GRE item sets. Results are then reported for the simulation study. A discussion of the results is presented in Chapter 5.

Real Data Studies

First Grade Reading Assessment

A base model, position effect model, and item-position effect model, M5, were fit to the reading test data. Model M5 was compared to the model without random item-position slopes using a likelihood ratio test. Results from the likelihood ratio test are included in Table 9. The $\chi^2_2 = 15.64$ was statistically significant at $\alpha = .05$ ($p = .0004$), supporting the inclusion of the random effect for item-position slopes in the model. Z-statistics for the position main effects were also statistically significant at $\alpha = .05$, in both the position effect model ($\gamma_{10} = -0.02$, $z = -18.82$, $p < .001$) and M5 ($\gamma_{10} = -0.02$, $z = -13.32$, $p < .001$). The AIC was smaller for M5, but the M5 BIC was slightly larger.

Model M5 was retained as the final model.

Table 10 contains the intercept and variance component estimates for each effect, for all three models. As discussed in Chapter 3, the intercept γ_{00} typically represents the average log-odds of correct response on the reference item. In the reading test data, without item indicator variables, γ_{00} was interpreted as the mean log-odds of correct response across all items, ignoring or averaging across item position. In the base model, this value was 0.0802, which can be converted to the probability metric using the inverse logistic function:

$$P(x) = logit^{-1}(x) = \frac{1}{1 + e^{-x}}. \tag{44}$$

The probability conversion of γ_{00} in the base model was .5200, indicating that the predicted mean percentage correct across all items and positions was 52%.

Table 9: RT Likelihood Ratio Test of Random Position Effects for Items

Model	*df*	AIC	BIC	logLik	χ^2	$\chi^2 df$	*p*
Position	3	123172	123200	-61583			
M5	5	123160	123207	-61575	15.64	2	.0004

The variance component for the intercept remained essentially unchanged from the base to the position effect model. However, the main effect γ_{00} increased to 0.3956. In this model, γ_{00} is interpreted as the mean log-odds at position zero, which was 0.3154 logits higher than γ_{00} in the base model. In the

Table 10: RT Intercepts and Variance Components

Model	Effect	Intercept	Variance	SD	Covariance
Base	γ_{00}	0.0802	0.2335	0.4832	
Position	γ_{00}	0.3956	0.2316	0.4813	
	γ_{10}	-0.0204			
M5	γ_{00}	0.3880	0.2932	0.5415	
	γ_{10}	-0.0199	0.0001	0.0072	-0.0024

Note: SD denotes standard deviation. The covariance is available only for M5, where random effects for intercepts and position slopes are estimated.

probability metric, γ_{00} converts to $logit^{-1}(0.3956) = .5976$, nearly 60%, or 8% greater than that of the base model. In other words, the predicted probability of correct response was 8% higher at the beginning of the form than in the middle.

The mean change in log-odds across position is described by the slope $\gamma_{10} = -0.0204$. For a one unit change in position the log-odds were estimated to decrease by 0.0204. The intercept and slope can be combined to determine the predicted mean log-odds across all items at the mean position. The main effect for position is multiplied by the mean item position, $\bar{p} = 12.3585$, to obtain -0.2525. Combining, this results in $\gamma_{00} + \gamma_{10}\bar{p} = 0.1431$, or a probability of .5357, slightly higher than the base model predicted probability of .5200.

In M5, γ_{00} and γ_{10} changed little from the previous models (see Table 10). The variance component for γ_{00} increased slightly to 0.2932. The variance component for the item-position slopes was small, 0.0001, with a standard

deviation of 0.0072.

Figure 3 depicts the variability of the M5 random effects around the fixed effect estimates. Included in Figure 3 are: (a) observed proportion correct plotted for each item across all positions, with light shaded lines representing items and the solid dark line representing the average across items; and (b) fitted proportion correct for each item across positions, again with light shaded lines for individual items and the dark line the intercept and slope main effects.

The variability in observed proportion correct makes it difficult to follow a given line across item positions; the first plot in Figure 3 is included to show this variability across items and position, and to show the general negative trend in observed proportion correct. Observed mean proportion correct decrease from .6264 at position 1 to .5005 at position 22. Mean fitted values in the second plot decrease from .5908 at position 1 to .4862 at position 22. This negative trend is noted in the majority of item-position effect slopes t_{1i} as well. In the plot of fitted values, items vary widely from the mean in terms of intercept but have similar decreasing slopes.

Figure 4 contains a plot of the M5 random slopes in the logit metric as $\gamma_{10} + t_{1i}$, against the random intercepts as $\gamma_{00} + t_{0i}$. All of the item slopes, on the y-axis, are negative, though some are relatively close to zero. A negative trend is evident in the relation between the slopes and intercepts, with a

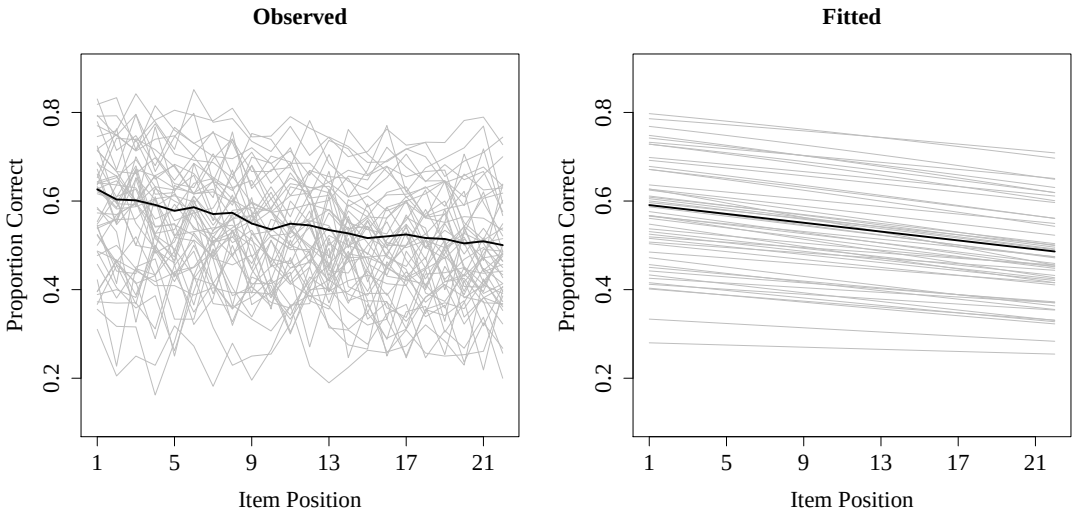

Figure 3: RT observed and fitted M5 proportion correct for each item across all positions. Individual items are represented by gray lines and observed means and fitted main effects by black lines.

correlation of $-.6116$. Items with higher initial logits tended to have stronger negative slopes; that is, easier items tended to increase in difficulty slightly more rapidly across the form than did more difficult items.

The relationship between the item effects in the three models is shown in Figure 5. The first plot contains estimates for the base and position effect models, where values for the latter are shown to shift upward, essentially by a constant. The solid line represents a 1 to 1 correspondence. The second plot contains estimates for the position effect model and M5, where each item fluctuates slightly from the solid line. The third plot contains estimates for the

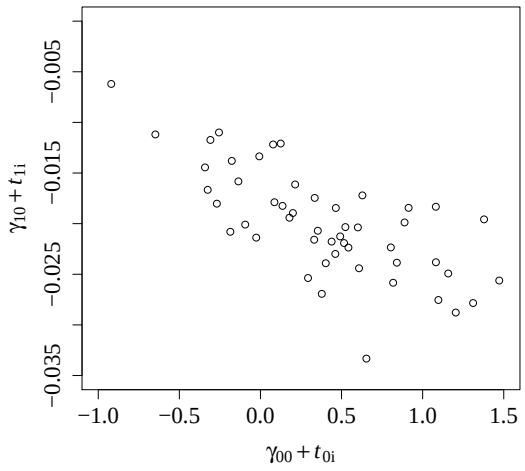

Figure 4: RT scatter plot of M5 random effects.

base model and M5, where values are shifted upward and again deviate slightly from a linear relationship.

Finally, Table 11 contains the combined fixed and random effects for a subset of 20 items under the three models, where $\beta_{0i} = \gamma_{00} + t_{0i}$ for the item effects and $\beta_{1i} = \gamma_{10} + t_{1i}$ for position effects in M5. Each column is described in the caption for the table. These values were used to create the plots in Figure 5. The plots in Figure 5 give a general description of how estimates change from one model to the next. Table 11 is useful as a reference for the Figure, showing the magnitude of these changes for a subset of items. Item effects for the position model (in the third column) were all higher than in the base model

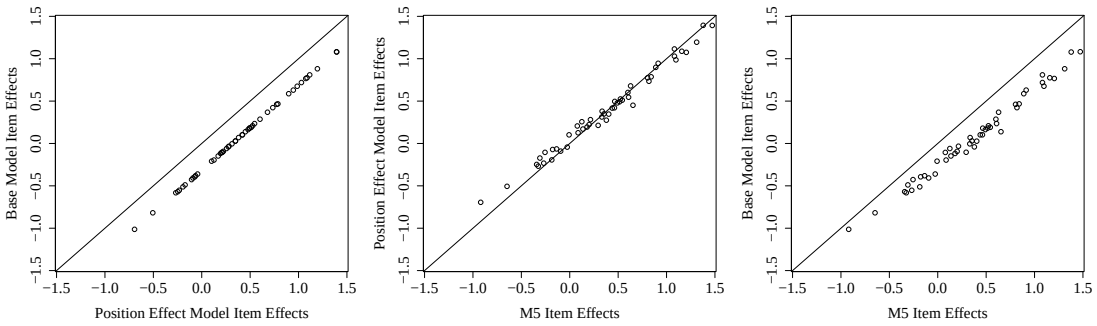

Figure 5: RT scatter plots of item effects for the base versus the position effect model, the position effect model versus M5, and the base model versus M5.

(second column) because of the negative position effect slope, $-.0204$. To obtain estimated item effects at $\bar{p} = 12.3585$, the constant $-.0204\bar{p} = -.2521$ would be subtracted from each of the values for the position effect model. This would shift the item effects from the position model to be nearly equivalent to corresponding values from the base model, shifting the points in the first plot of Figure 5 to the left. The item and position effects for M5 (columns 4 and 5), were also estimated at $p = 0$. The last column contains the M5 position effects at $\bar{p}$. Columns 4 and 6 would be combined to obtain M5 item effects which correspond to values for the base model; this would shift the points in the third plot of Figure 5 to the left.

Table 11: RT Item and Position Effects for a Subset of Items

Item	Base: β_{0i}	Pos: β_{0i}	M5: β_{0i}	M5: β_{1i}	M5: $\beta_{1i}\bar{p}$
1	0.59	0.90	0.89	−0.02	−0.25
2	1.08	1.39	1.38	−0.02	−0.24
3	0.81	1.12	1.08	−0.02	−0.23
4	0.63	0.95	0.91	−0.02	−0.23
5	1.08	1.39	1.47	−0.03	−0.32
6	0.78	1.09	1.16	−0.02	−0.31
7	0.77	1.08	1.20	−0.03	−0.36
8	0.72	1.03	1.08	−0.02	−0.29
9	0.37	0.68	0.63	−0.02	−0.21
10	0.47	0.79	0.84	−0.02	−0.29
11	−0.58	−0.27	−0.33	−0.02	−0.21
12	0.18	0.50	0.46	−0.02	−0.23
13	−0.00	0.31	0.33	−0.02	−0.27
14	−0.43	−0.10	−0.26	−0.01	−0.14
15	−1.01	−0.69	−0.92	−0.01	−0.08
16	−0.06	0.26	0.12	−0.01	−0.15
17	−0.11	0.21	0.08	−0.01	−0.15
18	−0.04	0.27	0.38	−0.03	−0.33
19	−0.41	−0.09	−0.09	−0.02	−0.25
20	0.18	0.49	0.52	−0.02	−0.27

Note: In each row are the combined fixed and random effects for a given item under the three models, where $\beta_{0i} = \gamma_{00} + t_{0i}$ for item effects and $\beta_{1i} = \gamma_{10} + t_{1i}$ for position effects. The first column contains item effects for the base model. The second column, Pos, contains item effects for the position effects model, which are all estimated at $p = 0$. The next two columns contain the item and position effects for M5, again as estimated at $p = 0$. The last column contains the M5 position effects at $\bar{p} = 12.3585$. The Pos main effect for position is $\beta_{1i} = \gamma_{10} = -.0204$ and the corresponding value at $\bar{p}$ is $-.2521$.

GRE

The base model M0, position effect model M1, and item-position effect model M2, were each fit to the three quantitative data sets GRE1, GRE2, and GRE3. Table 12 contains the model fit results by data set.

Model M2 was first compared to M1. The χ^2 were all statistically significant at $\alpha = .01$, supporting the inclusion of the interaction terms in each case. The AIC and BIC produced conflicting results. For GRE1, the AIC was smaller for M2 and the BIC was smaller for M1; for GRE2, both the AIC and BIC were smaller for M1; and for GRE3, the AIC was smaller for M2 and the BIC was smaller for M1. BIC was re-estimated using level-2 units as the *sample size*. Although the discrepancies between M2 and M1 values were reduced, the BIC still favored M1 over M2 for all three GRE data sets.

Next, under the assumption that model M1 was more appropriate than M2, M1 was compared to the base model M0. In this comparison, the likelihood ratio tests, AIC, and BIC all favored the more complex model M1, which included the main effect for position. Based on these results, M1 may have been more appropriate for GRE2 and perhaps for GRE1 and GRE3 as well. However, for demonstration purposes, M2 was considered further for all three data sets.

The model fit comparisons presented in Table 12 served as omnibus tests

for the entire set of item-position interaction terms included in models M2. The next step was to examine individual effects for practical and statistical significance. In Appendix A, Tables 20 through 23 contain the M2 item and item-position interaction effect estimates for each item set, along with standard errors, z-statistics, and p-values for the interaction effects. The z-statistics can be used to test the statistical significance of the item-position effects as differing from zero.

An unadjusted $\alpha = .05$ was used to identify interaction effects as statistically significant. This value was large, given the fact that 28 statistical tests were considered for each item set. However, since the purpose of this procedure was to identify items more susceptible to position bias, inclusiveness at the risk of increasing false positives seemed more acceptable than excluding potentially problematic items.

Table 21 contains interaction effects for GRE1, where all but the BIC favored the interaction effect model M2. Using $\alpha = .05$ as a cutoff, five interaction effects were considered statistically significant. These are $\gamma_{310} = 0.026$ ($z = 4.085$, $p < .001$), $\gamma_{380} = -0.030$ ($z = -4.058$, $p < .001$), $\gamma_{450} = -0.015$ ($z = -2.054$, $p = .040$), $\gamma_{470} = -0.016$ ($z = -2.210$, $p = .027$), and $\gamma_{500} = -0.024$ ($z = -2.425$, $p = .015$). Note that because of the effect coding used, these are interpreted as deviations from the main effect for

Table 12: GRE Likelihood Ratio Tests of Item-Position Interaction Effects

Item Set	Model	*df*	AIC	BIC	logLik	χ^2	$\chi^2 df$	p
GRE1	M0	29	52796	53052	-26369			
	M1	30	52758	53022	-26349	40.20	1	0.0000
	M1	30	52758	53022	-26349			
	M2	57	52744	53247	-26315	67.70	27	$< .0001$
GRE2	M0	29	55963	56220	-27953			
	M1	30	55856	56121	-27898	108.89	1	0.0000
	M1	30	55856	56121	-27898			
	M2	57	55862	56366	-27874	47.84	27	0.0080
GRE3	M0	29	51475	51732	-25709			
	M1	30	51409	51674	-25674	68.58	1	0.0000
	M1	30	51409	51674	-25674			
	M2	57	51377	51881	-25632	85.29	27	< 0.0001

position, the mean of the item-position interaction effects, $\gamma_{280} = -0.010$ ($z = -6.939$, $p < .001$). The main effect and interaction effects can be combined as $\gamma_{280} + \gamma_{(28+q)0}$ to obtain the total effect for position for a given item. The total effects for these five items were 0.017, -0.039, -0.025, -0.026, and -0.033. Thus, on average, items tended to have lower logits as position increased. Five items differed notably from this mean trend; the first (γ_{310}) was estimated to instead have a higher logit at higher item positions, and the four others were estimated to have even steeper declining slopes.

Table 22 contains interaction effects for GRE2, where both the AIC and BIC favored the position effect model M1. If model M2 were instead chosen as more appropriate, four interaction effects would be considered statistically

significant. These are $\gamma_{300} = -0.027$ ($z = -2.879$, $p = .004$), $\gamma_{390} = 0.017$ ($z = 2.539$, $p = .011$), $\gamma_{430} = -0.018$ ($z = -2.025$, $p = .043$), and $\gamma_{510} = 0.017$ ($z = 2.574$, $p = .010$). The main effect for position in GRE2 was $\gamma_{280} = -0.015$ ($z = -11.105$, $p < .001$), indicating that item logits again tended to decrease, with items becoming more difficult at higher positions. The total effects for position for the four items listed above were -0.042, 0.001, -0.033, and 0.002.

Table 23 contains interaction effects for GRE3, where all but the BIC again favored the interaction effect model M2. The main effect for position was $\gamma_{280} = -0.013$ ($z = -8.864$, $p < .001$), again indicating an overall decrease in item logit by position. Ten interaction effects were considered statistically significant. These are $\gamma_{330} = -0.019$ ($z = -2.603$, $p = .009$), $\gamma_{370} = 0.016$ ($z = 2.340$, $p = .019$), $\gamma_{400} = -0.023$ ($z = -2.965$, $p = .003$), $\gamma_{410} = -0.019$ ($z = -2.219$, $p = .026$), $\gamma_{420} = 0.016$ ($z = 2.498$, $p = .012$), $\gamma_{430} = 0.014$ ($z = 2.185$, $p = .029$), $\gamma_{440} = 0.018$ ($z = 2.523$, $p = .012$), $\gamma_{520} = 0.033$ ($z = 4.872$, $p < .001$), and $\gamma_{540} = -0.015$ ($z = -2.029$, $p = .042$). Five of these ten interactions resulted in total position slopes which were flatter than the mean, with values of $\gamma_{280} + \gamma_{(28+q)0}$ smaller than 0.01. Four of the ten had steeper decreasing slopes, and one item had a total positive slope of 0.020.

Figures 6, 7, and 8 contain plots of (a) observed proportion correct, as in the first plot of Figure 3; and (b) the lines $\gamma_{00} + \gamma_{q0} + \gamma_{N0}p + \gamma_{(q+N)0}p$ across

item position. Thus, these plots show the intercept and total position slope for each item. They demonstrate how item logits were estimated to change across the test forms.

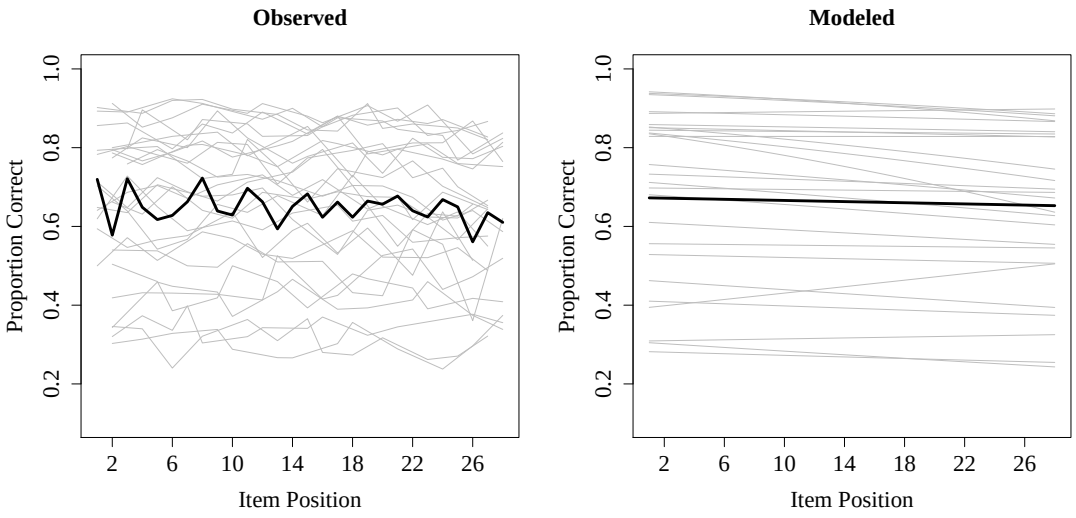

Figure 6: GRE1 observed and modeled M2 proportion correct for each item across all positions. Individual items are represented by gray lines and observed means and main effects by black lines.

Finally, Figure 9 contains plots of M2 ability u_{0j} for examinees taking each GRE item set. Ability was roughly normally distributed around 0 with a standard deviation of about 1 for each sample. Descriptive statistics are contained in Table 13. Ability estimates were essentially the same for M0, M1, and M2.

As described in Chapter 3, the results of the real data analysis were used

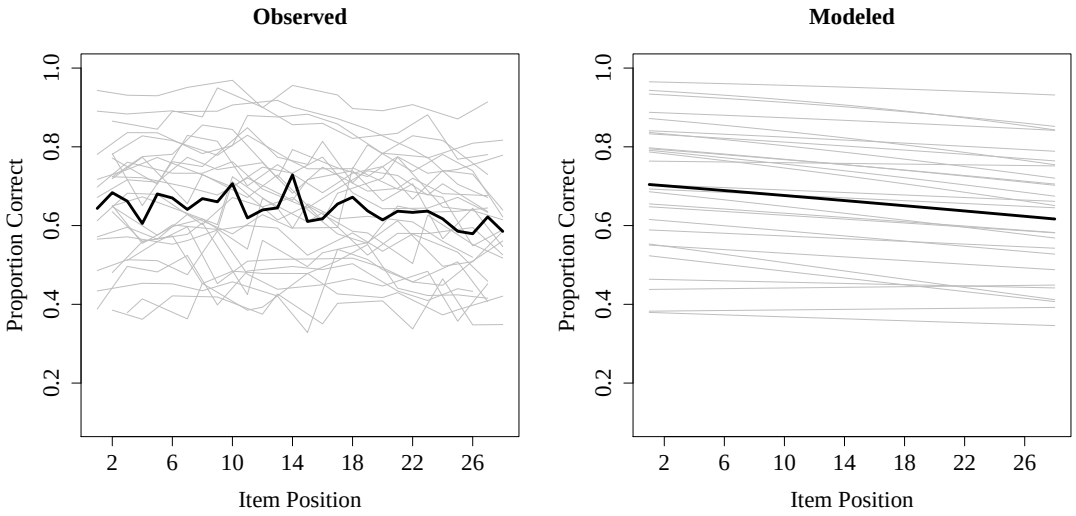

Figure 7: GRE2 observed and modeled M2 proportion correct for each item across all positions. Individual items are represented by gray lines and observed means and main effects by black lines.

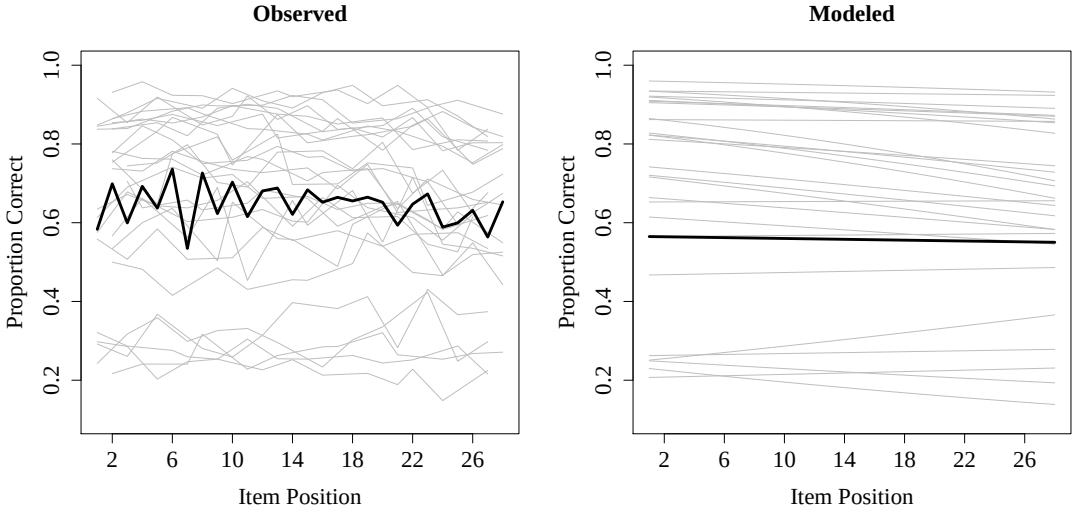

Figure 8: GRE3 observed and modeled M2 proportion correct for each item across all positions. Individual items are represented by gray lines and observed means and main effects by black lines.

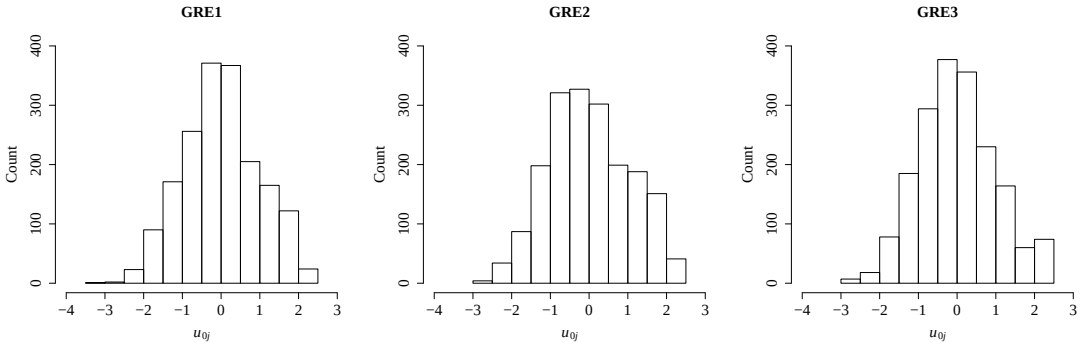

Figure 9: Distributions of ability u_{0j} for the three GRE samples.

Table 13: GRE Descriptive Statistics for u_{0j}

	$Mean$	SD	$Skew$	$Kurt$	Min	Max	N
GRE1	0.000	0.953	0.107	−0.320	−3.097	2.264	1797
GRE2	0.000	1.039	0.144	−0.519	−2.624	2.320	1852
GRE3	0.000	1.006	0.157	−0.354	−2.804	2.480	1843

primarily to inform the design of the simulation study. Fixed and random effect estimates from the GRE study, in particular, guided the selection of parameters for the item, item-position, and person effect simulation conditions. In the GRE study, item effects tended to be distributed slightly higher on the logit scale than persons, with means of 0.993, 1.04, and 1.06, for GRE1, GRE2, and GRE3, respectively. Item effects ranged from −1.35 to 3.35 across all three item sets. Main effects for position were all negative, and the majority of item-position interaction effects were negative as well, though a handful of items were estimated to increase in easiness across the test form.

Simulation Study

Results from the simulation study are presented below. Parameter recovery results are presented for the item parameters, first in terms of $Bias$, $AbsBias$, SE and $RMSE$, and then in terms of these same indices adjusted to the beginning and end of the test form. Next, recovery results for the M2 position parameters are presented in terms of $Bias$, $AbsBias$, SE and $RMSE$,

as are recovery indices for the random person parameters for all models. Parameter recovery at the item level, as opposed to the condition level across items, is then discussed. Finally, model fit results are presented. Conditions again represented by S1 and S2, for the small and medium sample sizes; T1 and T2, for the 20-item and 40-item tests; and P1, P2, and P3, for the negligible, positive, and negative position effect distributions.

Mean Item Parameter Recovery

Table 14 contains the item parameter $Bias$, $AbsBias$, SE, and $RMSE$ averaged across items by condition, at the center of the test form, i.e., $\bar{p}$. The 12 conditions are represented by columns and the parameter recovery indices, for models M0, M1, and M2, by rows. Figures 12, 13, 14, and 15 in Appendix B contain plots of the mean $Bias$, $AbsBias$, SE, and $RMSE$ by condition and model.

As shown in rows 1 through 3 of Table 14 and Figure 12, signed $Bias$ was highest for M2 in every condition except S2-T2-P2, where M1 was slightly higher. M0 $Bias$ was always lowest, and M1 tended to have signed $Bias$ slightly higher than M0, with the exception of S2-T2-P2. Differences in $Bias$ between models for a given condition were always less than 0.01 logit, with the largest difference being 0.006 between M0 and M2 at S1-T2-P3.

In terms of signed *Bias*, values closer to zero indicate higher estimation accuracy. Although M0 had the lowest *Bias* across all conditions, in many conditions (i.e., S1-T1-P2, S1-T1-P3, S1-T2-P1, S1-T2-P2, S2-T1-P3, and S2-T2-P2) this meant that M0 values were furthest from zero. Because *AbsBias* ignores sign, lower values always correspond to higher accuracy. M0 *AbsBias* tended to be lowest, but was higher than M1 and M2 at both sample sizes S for T2-P2 and at S2-T1-P3; otherwise, *AbsBias* was highest for M2 (see rows 4 through 6 of Table 14 and Figure 13). As with *Bias*, differences in *AbsBias* between models for a given condition were always less than 0.01 logit. The largest difference was 0.003, between M0 and M2 at S2-T2-P3.

SE was always smallest for M0, across all conditions (see rows 7 through 9 of Table 14 and Figure 14). *SE* for M1 were either the same as M0 or slightly higher. *SE* was always largest for M2. *SE* decreased from the smaller to the larger sample size, S1 to S2, for both test lengths, as well as from the shorter to longer test length, T1 to T2. Otherwise, there did not appear to be any trend across position conditions P. In S1-T1, *SE* decrease from P1 to P2, and then increase slightly at P3. In S1-T2, values decreased slightly from P1 to P2 and then decreased further at P3. In S2-T1, *SE* increased from P1 to P3, and at S2-T2 values were relatively consistent.

RMSE corresponded closely to *SE* (see rows 10 through 12 of Table 14

and Figure 15). In each set of conditions, M0 and M1 $RMSE$ were close to one another and values for M2 were slightly larger. The largest $RMSE$ across all conditions was 0.122, for M2 at S1-T1-P1. The largest difference was 0.001.

Mean Item Parameter Recovery at Positions 1 and N

Table 15 is structured similarly to Table 14, but contains recovery indices $Bias\, p_1$, $Bias\, p_N$, $AbsBias\, p_1$, and $AbsBias\, p_N$, which are the $Bias$ and absolute $Bias$ averaged across items at positions 1 and N ($N = 20$ for T1 and 40 for T2). These are depicted in Figures 16, 17, 18, 19 in Appendix B. Magnitudes of average $Bias$ at the beginning and end of the test form did not correspond to magnitudes found at the center. Instead, M2 tended to produce the least biased estimates, with M1 showing slightly more bias, and M0 the most.

$Bias\, p_1$ at P1 across all other conditions were roughly the same for M0, M1, and M2 (see row 1 of Table 15 and Figure 16). At P2 and P3, again for all other conditions, M0 differed noticeably from M1 and M2, increasing for P2 to as much as 0.186 at S2-T2-P2, and decreasing as low as -0.210 at S1-T2-P3. M0 $AbsBias\, p_1$ always increased from P1 to P3, across S and T, with the largest values at P3. M1 and M2 $AbsBias\, p_1$ were relatively stable across P, but values for M1 were substantially higher than for M2 (see row 2 of Table 15 and Figure 17).

$Bias\, p_N$ for M0 were essentially the opposite of $Bias\, p_1$, with M0 decreasing as low as -0.194 at P2 and then increasing to as much as 0.217 at P3. M1 and M2 were again relatively stable, remaining close to zero (see row 3 of Table 15 and Figure 18). Values for M0, M1, and M2 were again roughly the same at P1 across all other conditions. $AbsBias\, p_N$ also followed the same patterns, with similar magnitudes, as found with $AbsBias\, p_1$ (see row 4 of Table 15 and Figure 19).

Whereas parameter recovery at the center of the test tended to favor the simpler models M0 and M1, as in Table 14, $Bias$ and $AbsBias$ at the tails of the test favored model M2. The largest $AbsBias$ at the center of the test, from Table 14, was 0.013 logits, for M2 at S1-T1-P1. The largest $AbsBias$ at the beginning and end of the test, from Table 15, were 0.237 and 0.240, for M0 at S1-T2-P3 and S2-T2-P3, respectively.

Table 16 contains mean SE and $RMSE$ at the beginning and end of the test form. These indices reveal that M2 estimates were again significantly more variable than estimates from M0 or M1 (see Figures 20 and 21). For M0, $SE\, p_1$ and $SE\, p_N$ were equivalent to SE, i.e., $SE\, \bar{p}$, since M0 item difficulty was fixed. $SE\, p_1$ and $SE\, p_N$ for M1 were nearly the same as M1 SE at the center position. However, $SE\, p_1$ and $SE\, p_N$ for M2 were larger than M2 SE at $\bar{p}$, ranging from 0.148 to 0.221 compared to a range from 0.082 to 0.121.

$RMSE\,p_1$ and $RMSE\,p_N$ were essentially a combination of the corresponding $Bias$ and SE at the beginning and end of the test form. In most conditions, SE was substantially larger than $Bias$, which meant that $RMSE$ favored models M0 and M1 (see Figures 22 and 23). This is evident in S1-T1 and S2-T1, where $SE\,p_1$, $SE\,p_N$, $RMSE\,p_1$ and $RMSE\,p_N$ were highest for M2. When $Bias$ became larger for M0 than M1 or M2, $RMSE$ also increased. This is evident in P2 and P3 in S1-T2 and S2-T2, where values for $AbsBias\,p_1$, $AbsBias\,p_N$, $RMSE\,p_1$ and $RMSE\,p_N$ were highest for M0.

Mean Position Parameter Recovery

M2 recovery of item-position parameters was also summarized using mean $Bias$, $AbsBias$, SE, and $RMSE$ (see Table 17). Values for each were considerably smaller than corresponding item parameter recovery indices, likely because item-position parameters and estimates were considerably smaller. $Bias$ ranged from -0.0007 at S1-T1-P2 to 0.0004 at S2-T1-P2. $AbsBias$ ranged from 0.0005 at P2 and P3 in S2-T2 to 0.0019 at S1-T1-P1. SE and $RMSE$ were the same at all but one condition (S1-T1-P3) and ranged from 0.0066 at S2-T2-P2 to 0.0192 at S1-T1-P3. Condition P seemed to have little impact on recovery, with all values in a given condition S and T differing across P only in the thousandths decimal place.

Table 14: Mean Item Parameter Recovery Indices by Model and Condition

		S1						S2					
		T1			T2			T1			T2		
	Model	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3
Bias	M0	0.004	−0.003	−0.005	−0.005	−0.007	0.003	0.003	0.000	−0.003	−0.001	−0.002	0.004
	M1	0.004	−0.003	−0.005	−0.004	−0.005	0.005	0.003	0.000	−0.002	−0.001	−0.001	0.005
	M2	0.006	−0.001	−0.003	−0.002	−0.005	0.009	0.004	0.001	−0.001	0.001	−0.001	0.008
AbsBias	M0	0.011	0.009	0.008	0.009	0.011	0.009	0.007	0.006	0.006	0.005	0.007	0.006
	M1	0.011	0.009	0.009	0.009	0.009	0.009	0.007	0.007	0.005	0.005	0.006	0.007
	M2	0.013	0.009	0.009	0.010	0.009	0.011	0.008	0.008	0.005	0.006	0.006	0.009
SE	M0	0.121	0.115	0.118	0.118	0.117	0.113	0.082	0.083	0.084	0.081	0.082	0.081
	M1	0.121	0.116	0.118	0.118	0.117	0.113	0.082	0.083	0.084	0.081	0.082	0.082
	M2	0.121	0.116	0.119	0.118	0.118	0.114	0.082	0.083	0.084	0.082	0.082	0.082
RMSE	M0	0.121	0.115	0.118	0.118	0.117	0.113	0.082	0.083	0.083	0.081	0.082	0.081
	M1	0.121	0.116	0.118	0.118	0.117	0.113	0.082	0.083	0.083	0.081	0.082	0.082
	M2	0.122	0.116	0.118	0.118	0.118	0.114	0.082	0.083	0.084	0.082	0.082	0.082

Note: *Bias* is the mean of signed *Bias* estimates across items within a condition; *AbsBias* is the mean of absolute *Bias* estimates across items within a condition (see Equation 30).

Table 15: Mean Item Parameter Position *Bias* by Model and Condition

		S1						S2					
		T1			T2			T1			T2		
	Model	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3
Bias p_1	M0	0.003	0.089	−0.109	−0.008	0.181	−0.210	0.002	0.091	−0.107	−0.004	0.186	−0.210
	M1	0.006	−0.002	−0.009	0.000	−0.021	−0.003	0.008	−0.009	−0.006	0.001	−0.016	−0.001
	M2	0.009	0.006	−0.002	0.002	−0.010	0.009	0.009	−0.003	−0.001	0.002	−0.005	0.010
AbsBias p_1	M0	0.064	0.104	0.120	0.139	0.215	0.237	0.066	0.105	0.118	0.139	0.217	0.236
	M1	0.064	0.074	0.074	0.138	0.149	0.146	0.067	0.069	0.073	0.139	0.149	0.145
	M2	0.021	0.017	0.020	0.017	0.019	0.021	0.015	0.014	0.010	0.011	0.011	0.015
Bias p_N	M0	0.006	−0.094	0.099	−0.001	−0.194	0.217	0.005	−0.092	0.102	0.003	−0.190	0.217
	M1	0.002	−0.003	−0.001	−0.009	0.010	0.013	−0.001	0.009	0.002	−0.002	0.014	0.011
	M2	0.003	−0.008	−0.003	−0.007	0.000	0.008	0.000	0.005	0.000	0.000	0.003	0.006
AbsBias p_N	M0	0.072	0.108	0.112	0.139	0.219	0.239	0.069	0.107	0.114	0.138	0.216	0.240
	M1	0.072	0.070	0.069	0.138	0.146	0.148	0.069	0.075	0.070	0.137	0.146	0.148
	M2	0.020	0.016	0.011	0.020	0.016	0.020	0.009	0.014	0.010	0.013	0.012	0.010

Note: *Bias* p_1 is the mean of signed *Bias* estimates across items within a condition at position 1; *Bias* p_N is the mean of signed *Bias* estimates across items within a condition at position N, which is 20 for T1 and 40 for T2. *AbsBias* p_1 and *AbsBias* p_N are the corresponding averages of absolute *Bias* estimates at the beginning and end of the test.

Table 16: Mean Item Parameter Position SE and $RMSE$ by Model and Condition

		S1						S2					
		T1			T2			T1			T2		
	Model	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3
$SE\, p_1$	M0	0.121	0.115	0.118	0.118	0.117	0.113	0.082	0.083	0.084	0.081	0.082	0.081
	M1	0.128	0.122	0.124	0.119	0.119	0.117	0.086	0.087	0.087	0.083	0.085	0.082
	M2	0.217	0.214	0.221	0.219	0.216	0.221	0.148	0.154	0.153	0.153	0.155	0.158
$SE\, p_N$	M0	0.121	0.115	0.118	0.118	0.117	0.113	0.082	0.083	0.084	0.081	0.082	0.081
	M1	0.127	0.122	0.125	0.121	0.124	0.117	0.086	0.088	0.089	0.084	0.084	0.085
	M2	0.219	0.212	0.213	0.220	0.219	0.212	0.149	0.155	0.151	0.157	0.151	0.152
$RMSE\, p_1$	M0	0.142	0.165	0.175	0.193	0.257	0.270	0.113	0.141	0.150	0.170	0.242	0.256
	M1	0.148	0.148	0.147	0.194	0.202	0.195	0.116	0.117	0.118	0.171	0.180	0.174
	M2	0.218	0.214	0.222	0.219	0.217	0.222	0.149	0.154	0.153	0.153	0.155	0.158
$RMSE\, p_N$	M0	0.146	0.166	0.169	0.194	0.262	0.273	0.112	0.144	0.147	0.169	0.241	0.259
	M1	0.152	0.146	0.148	0.196	0.202	0.196	0.115	0.122	0.118	0.170	0.176	0.177
	M2	0.219	0.212	0.212	0.220	0.219	0.212	0.149	0.156	0.151	0.157	0.151	0.152

Note: $SE\, p_1$ and $RMSE\, p_1$ are the mean SE and $RMSE$ across items within a condition at position 1; $SE\, p_N$ and $RMSE\, p_N$ are the mean estimates across items within a condition at position N, which is 20 for T1 and 40 for T2.

Table 17: M2 Mean Position Parameter Recovery Indices by Condition

	S1						S2					
	T1			T2			T1			T2		
	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3
Bias	−0.0003	−0.0007	0.0000	−0.0002	0.0002	0.0000	−0.0005	0.0004	0.0001	0.0000	0.0002	−0.0001
AbsBias	0.0019	0.0015	0.0014	0.0007	0.0007	0.0008	0.0010	0.0010	0.0010	0.0006	0.0005	0.0005
SE	0.0191	0.0188	0.0192	0.0095	0.0094	0.0094	0.0131	0.0137	0.0134	0.0068	0.0066	0.0067
RMSE	0.0191	0.0188	0.0191	0.0095	0.0094	0.0095	0.0131	0.0137	0.0134	0.0068	0.0066	0.0067

Note: Only model M2 included position effect estimates for each item. Recovery indices are presented here as *Bias*, absolute *Bias*, *SE* and *RMSE* averaged across items within a condition.

Mean Person Parameter Recovery

Table 18 contains mean random effect *Bias*, *AbsBias*, *BiasAbs*, and *RMSE*. Within a given condition, mean parameter recovery indices were nearly all identical across models M0, M1, and M2. The largest change for any recovery index across models in a given condition was 0.001.

Recovery differed minimally across conditions as well. For example, within S1-T1 *Bias* increased from −0.005 at P1 to 0.004 at P2 and 0.008 at P3. Changes in *Bias* across P for other conditions were similarly small. *Bias* were all smaller than 0.01. *AbsBias* ranged from 0.022 at S2-T1-P3 to 0.041 at S1-T2-P2. *BiasAbs* ranged from 0.283 at S1-T2-P3 to 0.378 at S1-T1-P3. *RMSE* ranged from 0.358 at S1-T2-P3 and S2-T2-P2 to 0.475 at S1-T1-P3 and S2-T1-P2.

Table 18: Mean Person Parameter Recovery Indices by Model and Condition

		S1						S2					
		T1			T2			T1			T2		
	Model	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3
Bias	M0	−0.005	0.004	0.008	0.004	0.006	−0.003	−0.004	0.000	0.000	0.001	0.002	−0.006
	M1	−0.005	0.004	0.008	0.005	0.007	−0.003	−0.004	0.000	0.000	0.001	0.002	−0.006
	M2	−0.005	0.004	0.008	0.004	0.007	−0.003	−0.004	0.000	0.000	0.001	0.002	−0.006
AbsBias	M0	0.031	0.037	0.035	0.037	0.041	0.032	0.025	0.023	0.022	0.025	0.027	0.028
	M1	0.032	0.037	0.035	0.037	0.041	0.032	0.025	0.023	0.022	0.025	0.027	0.028
	M2	0.032	0.037	0.035	0.037	0.041	0.032	0.025	0.023	0.022	0.025	0.027	0.028
BiasAbs	M0	0.376	0.376	0.377	0.293	0.285	0.283	0.377	0.378	0.376	0.284	0.284	0.285
	M1	0.376	0.376	0.377	0.293	0.285	0.283	0.377	0.378	0.376	0.284	0.284	0.285
	M2	0.377	0.377	0.378	0.293	0.285	0.283	0.377	0.378	0.376	0.284	0.284	0.285
RMSE	M0	0.474	0.474	0.475	0.369	0.360	0.358	0.474	0.475	0.474	0.358	0.358	0.359
	M1	0.474	0.474	0.475	0.369	0.359	0.358	0.474	0.475	0.474	0.358	0.358	0.359
	M2	0.475	0.475	0.475	0.369	0.360	0.358	0.475	0.475	0.474	0.358	0.358	0.359

Note: *Bias* is the mean of signed *Bias* estimates across replications within a condition; *AbsBias* is the mean of absolute *Bias* estimates across replications within a condition (see Equation 37). *BiasAbs* is the mean of *Bias* estimates across replications, with the absolute value taken before averaging across persons (see Equation 38).

Item Level Parameter Recovery

The item parameter recovery indices presented above represent values averaged across items within a condition. Consideration was also given to parameter recovery at the individual item level, where an item parameter estimate is taken as the mean for that item across replications. These results help to clarify the impact of item position on the estimation of item difficulty across the test form.

Appendix C includes 18 figures, each containing 20 plots of true and estimated item difficulty across item position for M0, M1, and M2. The true item difficulty parameter is plotted across position as a black solid line; the M0 estimate of item difficulty, constant across position, is plotted as a horizontal grey dotted line; the M1 estimate of item difficulty, which varies by position but is constant across items within a condition, is shown as a grey dotted/dashed line; and the M2 estimate of item difficulty, which varies across position and by item, is shown as a black dashed line.

In each figure, the y-axis is the same across each row of plots. For example, in Figure 24, plots for items 1 through 4 in row 1 all share the same y-axis, which ranges from -1.2 to 0.2 logits. In the next row, plots for items 5 through 8 share a y-axis which ranges from -0.6 to 0.4 logits. Though the range

differs by row, and by figure, the span of the y-axis is fixed at 1 logit for all plots and all item difficulty figures. This facilitates comparisons of item difficulty slopes across items within a condition.

These figures depict *Bias* at the item level. Deviation of the estimates, the dotted and dashed lines, from the true item difficulty parameters, the black solid lines, indicates *Bias*. At the center of the x-axis in each plot, estimates were nearly the same for models M0, M1, and M2, and were all close to the true item difficulty. This point represents the item difficulty at the mean position, where *Bias* tended to be minimized. At the beginning and end of the test, the item difficulty estimates tended to depart from one another and from the true item difficulty.

Overall, estimates from M2 deviated the least from the true parameters, across all conditions. Estimates from M0 deviated the most, as they were constant across position. The *Bias* of M0 was most evident in the P2 and P3 conditions, shown in Figures 25 (S1-T1-P2), 26 (S1-T1-P3), 29 (S1-T2-P2 items 1 to 20), 30 (S1-T2-P2 items 21 to 40), 31 (S1-T2-P3 items 1 to 20), and 32 (S1-T2-P3 items 21 to 40). M0 *Bias* was smallest in the P1 conditions, where position effects were simulated to be small to negligible.

Estimates from M1 deviated noticeably, especially for items which increased or decreased in difficulty as the remaining items in the test changed in

the opposite direction. For example, in condition P2 item-position generating parameters tended to be positive, with a positive mean slope of 0.01; in Figure 25 the majority of lines are sloping upward. However, the generating slope for item 10 was −0.009, shown as a downward sloping solid black line in Figure 25, plot 10. In this case, the mean main effect slope estimate from M1, 0.010, resulted in a *Bias* of −0.201 at position 1 and 0.151 at position 20.

The item level parameter recovery results demonstrate that item level *Bias* for the less complex models M0 and M1 was greatest in the tails, where item position varied from the center of the form. The amount of *Bias* corresponded to the parameters excluded from models M0 and M1. Overall, *Bias* in M0 increased as the total position slope for an item, the main effect plus the interaction effect, differed from 0. For M1, *Bias* increased as the item-position slopes differed from the main position effect slope.

Model Fit

A total of 1200 data sets were generated in the simulation study, with 100 replications at each of 12 conditions. Models M0, M1, and M2 were fit to each data set, resulting in 3600 total model fits. Of these 3600, seven did not converge (0.002%). These seven were all for model M0, and were found in conditions T1-P1 (1 replication), T1-P2 (1 replication), and T2-P2 (4

replications) under S1, and T2-P3 (1 replication) under S2. All seven of these M0 models were successfully estimated in *lme4* with convergence. The missing *HLM6* parameter estimates for these models were imputed with the *lme4* results, and were thus incorporated into the results presented above. However, *HLM6* model fit statistics were left as missing for these seven replications. Thus, for the affected conditions the proportions reported here correspond to slightly fewer than 100 replications, the smallest of which was 96 for S1-T2-P2.

Table 19 contains proportions of fit statistics AIC, BIC, and χ^2 likelihood ratio favoring models M1 compared to M0 (rows 1 through 3), and M2 compared to M1 (rows 4 through 6). Proportions are also depicted in Figures 10, for M1, and 11, for M2.

Overall, model M1 tended to fit better than M0. At P1, across all conditions of S and T, fewer than 20% of replications favored M1 over M0; the largest AIC proportion at P1 was .17, the largest BIC was .01 and the largest χ^2 was .09. However, at P2 and P3, AIC and χ^2 were greater than .50 across S and T. And at T2-P2 and T2-P3, for both S, all three fit statistics favored M1 across all replications. Given the small increase in complexity (1 parameter) from M0 to M1, AIC was the most liberal of the three statistics, and BIC the most conservative. This point is discussed further in Chapter 5.

Model fit results favored M2 over M1 only at T2, and only in terms of

AIC and χ^2. At S1-T2 and S2-T2, across P, proportions of χ^2 were above .90. At S2-T2, for all P, all χ^2 and over 95% of AIC favored M2. At T1, the fit statistics tended to favor M1, with all proportions below .50. Proportions of BIC favoring M2 were all 0, across all conditions, suggesting that the *sample size* in Equation 43 may have been better defined by J as opposed to $N \times J$. This point is also discussed further in Chapter 5.

Table 19: Proportions of Fit Statistics Favoring M1 vs M0 and M2 vs M1 by Condition

		S1						S2					
		T1			T2			T1			T2		
	Fit	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3
M1	AIC	0.17	0.82	0.84	0.14	1.00	1.00	0.10	0.98	0.99	0.13	1.00	1.00
	BIC	0.01	0.25	0.34	0.00	1.00	1.00	0.00	0.67	0.72	0.00	1.00	1.00
	χ^2	0.09	0.61	0.68	0.05	1.00	1.00	0.04	0.97	0.95	0.02	1.00	1.00
M2	AIC	0.03	0.06	0.04	0.39	0.50	0.39	0.13	0.16	0.15	0.97	0.97	0.98
	BIC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	χ^2	0.17	0.20	0.19	0.97	0.91	0.96	0.30	0.45	0.39	1.00	1.00	1.00

Note: Rows labeled AIC and BIC contain the proportion of replications in a given condition where the AIC and BIC were smaller for M1 than M0 (in rows 1 and 2), and smaller for M2 than M1 (in rows 4 and 5). Rows labeled χ^2 contain the proportion of likelihood ratio tests with p-values less than .05 for M1 versus M0 (in row 3) and M2 versus M1 (in row 6).

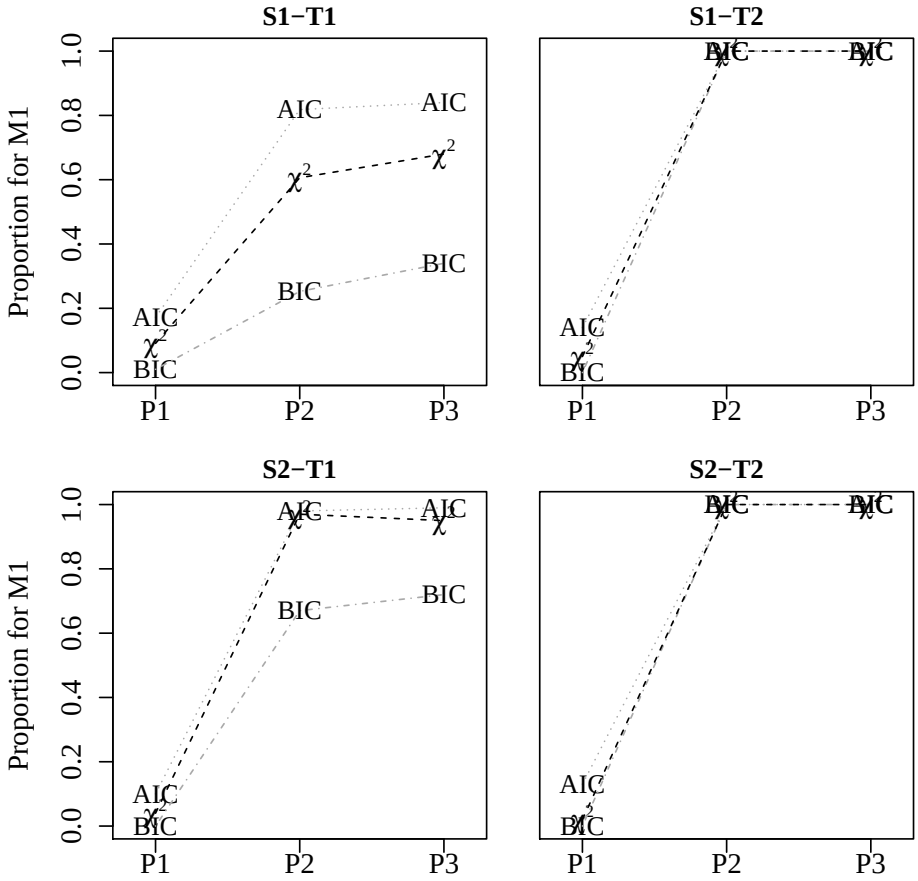

Figure 10: Proportion of AIC, BIC, and χ^2 likelihood ratios favoring M1 over M0. Each plot depicts a combination of sample size (S) and test length (T) conditions, with fit statistic proportions plotted across position effect condition (P).

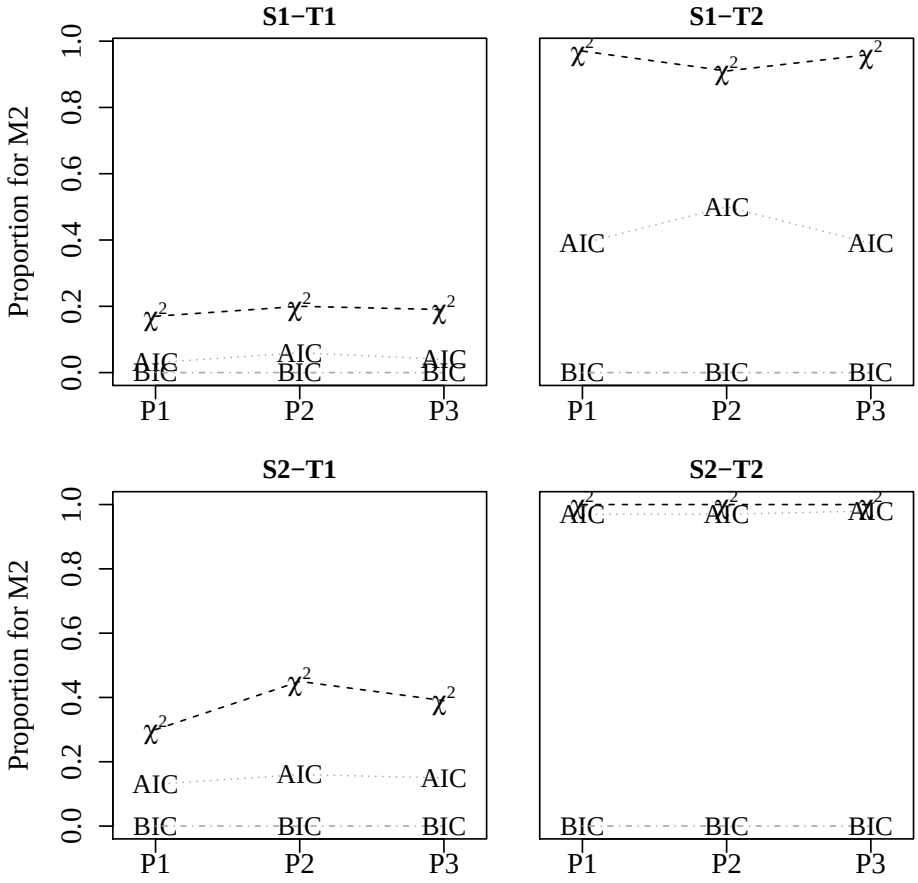

Figure 11: Proportion of AIC, BIC, and χ^2 likelihood ratios favoring M2 over M1. Each plot depicts a combination of sample size (S) and test length (T) conditions, with fit statistic proportions plotted across position effect condition (P).

Chapter V: Conclusion

As discussed in Chapters 1 and 2, testing programs often utilize IRT measurement scales which are based on multiple test forms. Context and position effects, introduced by common items which appear across these multiple forms, present a potential threat to the IRT assumption of local independence and thus also threaten the stability of such multi-form measurement scales. Position effects can be especially problematic in scrambled and computerized adaptive test administrations, where an item can essentially appear in any position, from the beginning to the end of the test.

Substantial research has addressed the practical implications of position effects. A wide range of studies were reviewed in Chapter 2, some addressing position effects for rearranged item sets, and more recent studies addressing position effects for common-item equating designs. Most often, position effects have been conceptualized as shifts in item difficulty attributed to change in item position. However, research on such item difficulty shifts has varied widely in study design, test design, and method of analysis. Although position effects have been uncovered for certain item types and test content, and the general

consensus is that item position can cause problems for the estimation of item difficulty and person ability, previous work has not presented a unified approach to examining these problems.

This lack of a unified approach was noted by Leary and Dorans (1985), Brennan (1992), and Davey and Lee (2011), each of whom referred to the necessity of estimating item, position, and person effects within a single model. In Chapter 2, four studies were reviewed which utilized different variations of such a model. The main purpose of the present study was to build upon these previous studies, and other earlier work, by formulating and demonstrating a hierarchical generalized linear model for examining the effect of item position on the estimation of item difficulty.

The HGLM is a flexible model which accommodates complex data structures and which allows for the simultaneous estimation of item and person effects along with effects for other covariates. As discussed in Chapter 3, the HGLM can be designed to examine the effects of item position from a variety of perspectives, depending on the test and study design and on the purpose of the study. Given the prevalence of operational test designs which allow an item to appear in any position, a series of models were presented in Chapter 3 which incorporate item position as a continuous covariate (M1, M2, M3, M4, M5). These models estimate position effects as slopes, changes in the difficulty of an

item based on changes in item position. Three of these models (M1, M2, M5) were demonstrated with real data. Models M1 and M2 were then compared to a base Rasch model using simulated data.

Real Data Study

The main purpose of the real data study was to compare position effect models to one another in terms of fit and examine the resulting item difficulty and position effect estimates. The results were then used to inform the design of the simulation study.

The real data study confirmed the usefulness of the item-position effect models. In the first grade reading test, model M5 was found to have slightly better fit than the position main effect model, supporting the inclusion of random item-position slopes. In the GRE data sets, the item-position effect model M2 was found to have slightly better fit in two of the three data sets, again supporting the inclusion of individual item-position slopes.

The real data study also revealed that item position had a significant impact on item difficulty. Main effects for position were negative in both the reading test and GRE, indicating an overall increase in difficulty across the test forms. The literature has tended to label increasing difficulty as a fatigue effect, where performance decreases and item difficulty thus increases as examinees are

assumed to decrease in performance-related variables such as motivation, concentration, and energy level (Davis & Ferdous, 2005). Decreases in performance may also result from increased pressure to provide a response as testing time becomes limited at higher item positions. As noted in Chapter 2, numerous studies have linked fatigue effects to the speededness of the test. Although the reading test and GRE administrations were not designed to be speeded, time limits may have impacted performance, contributing to the position effect findings.

Individual item-position effects were found to vary around the main effects in each real data study. In the reading test, all of the random slopes from M5 were also negative. Of the 84 GRE items, 19 (23%) were found to have statistically significant item-position interaction effects. Although the majority of these were negative, total position effects ranged in magnitude from -0.030 to 0.033, indicating for one item an estimated decrease of -0.828 logits across the 28-item form, and for another item an estimated increase of 0.930 logits. The first of these corresponds to a decrease in observed proportion correct of 0.17 from position 3 to position 28. The second corresponds to an increase in observed proportion correct of 0.082 from positions 1 to 27.

Were item level information available, the next step would be to examine the content and other characteristics of the items from each of these data sets

which were identified as being impacted by item-position bias. As noted above, these items may have been more susceptible than others to the speededness of the test; they may have involved complex content or wording which made them more difficult under the pressures of limited testing time. Some of the items in the GRE data sets were estimated to have positive slopes, indicating that these items decreased in difficulty as they appeared later in the form. These items may have involved novel or complex item types which were difficult early on but which became easier to understand as examinees were exposed to items of a similar type.

Without additional information, the causes of these effects cannot be identified. However, the item position effect models utilized with this data serve as tools for examining potential bias in item difficulty estimation due to item position. They are useful as a first step in the process of refining a test or item bank.

Simulation Study

The purpose of the simulation study was to examine the functionality of the HGLM as a tool for identifying position bias. Models M0, M1, and M2 were compared in terms of parameter recovery and model fit across three simulation conditions. M2 was the most complex model, including item and person effects,

a main effect for position, and item-position interaction effects. M1 was nested within M2 and included only the item and person effects and a main effect for position. The base model M0 was the least complex, including only item and person effects.

The results of the simulation study suggest that each of the models is more appropriate than the others under certain conditions. These conditions were sample size (S), test length (T), and position effect distribution (P). Results are discussed below for each model, with an emphasis on mean $RMSE$ as a comprehensive index of parameter recovery which captures both $Bias$ and SE.

Parameter Recovery

M0 was found to be the most appropriate model for the smaller sample size S1, with 500 examinees, and the shorter test length T1, with 20 items, when position effects were simulated to be negligible, as P1. Under these specific conditions, M0 resulted in the lowest mean $RMSE$, $RMSE\,p_1$, and $RMSE\,p_N$. Thus, with a small sample and short test, where position is expected to have a small influence on item difficulty, the estimation of either a main effect for position with M1 or interaction effects with M2 is not recommended. The same conclusion could be made for the short test length and negligible position effects with a larger sample size, such as S2. In this case, mean $RMSE$, $RMSE\,p_1$, and

$RMSE\,p_N$ were again lowest for M0. Although M0 was found to have higher mean $Bias\,p_1$ and $Bias\,p_N$ under each of these conditions, SE for the more complex models made them less appropriate.

M1 was found to be the most appropriate model for the smaller sample size S1, with 500 examinees, and the shorter test length T1, with 20 items, when position effects were simulated to be both positive and negative, as P1 and P2. Although M1 $RMSE$ at S1-T1-P1 and S2-T1-P1 were slightly higher than those for M0, for positive and negative position effects with a shorter test they were lowest for M1. M1 was also found to be most appropriate for positive and negative position effects under the smaller sample size and the longer test length, where $RMSE\,p_1$, and $RMSE\,p_N$ were again lowest for M1. Thus, when position effects are assumed to be present, a main effect for position may be supported, even with a somewhat small sample size and shorter test.

A combination of sample size and test length were needed for M2 to produce the lowest $RMSE$ of the three models. M2 was found to be most appropriate with the larger sample size S2, and the longer test length T2. These conditions resulted in M2 having the lowest $RMSE\,p_1$ and $RMSE\,p_N$ for the negligible, positive, and negative position effect conditions. At the center of the form, $RMSE$ was lowest for M2 only in the positive position effect condition.

Larger sample sizes will result in more accurate estimation as standard

errors are reduced. Thus, although M2 had the highest SE across all conditions, with a sample size of 1000, M2 SE more closely approximated SE for M0 and M1. On the other hand, a longer test will allow position effects, e.g., fatigue or practice effects, to have a greater net impact on item difficulty estimates. Thus, models which ignore the effect of position will tend to be more biased for longer tests. As sample size and test length increase, in particular beyond $N = 1000$ and $J = 40$, the item-position effect model M2 can be expected to outperform the less complex models.

Model Fit

Parameter recovery provided only one perspective of model appropriateness. Model fit, based on AIC, BIC, and χ_2 likelihood ratio indices, was used as an additional indication of how the models compared under the different simulation conditions. Overall, model fit results corresponded closely to the parameter recovery findings discussed above.

AIC, BIC, and χ_2 indices all confirmed that M0 was the most appropriate model under the small sample, short test, negligible position effect condition. They also suggested that M0 was most appropriate for the negligible position effect conditions under the remaining large sample and long test conditions. Thus, at P1, model M0 always resulted in the best fit.

For positive and negative position effect conditions, model M1 tended to fit better than M0. With a smaller sample and shorter test, model fit results were less clear, but with a larger sample and longer test M1 resulted in better fit than M0. M1 also fit better than M2, on average, when positive and negative position effects were present in the data. This was the case for the shorter test under both the small and large sample sizes.

With positive and negative position effects and a longer test length, model fit results were somewhat unclear. According to the χ_2 results, M2 fit the best in this condition for the smaller sample size. According to both the AIC and χ^2 results, M2 fit best in this condition for the larger sample size.

The conflicting results for AIC, BIC and χ_2 likelihood ratios may have been due to the differing penalties for model complexity imposed by each. The likelihood ratio test will always be the most liberal of the three indices, favoring the more complex model, when the increase in model complexity df_{A-M} exceeds 6 parameters and $\alpha = .05$; when $df_{A-M} > 6$, the χ^2 critical value for selecting the more complex model is always lower than the thresholds (i.e., the deviance penalty for the alternative model minus that of the null) imposed by the AIC and BIC. At $df_{A-M} = 7$ and $\alpha = .05$, the χ^2 critical value is 14.07, whereas the decrease in deviance must be 15 or larger for the AIC, and even larger for the BIC, depending on N and J.

The AIC and BIC are designed to be more conservative than the likelihood ratio test, with each index penalizing a model based on the number of parameters estimated. As shown in Equation 42, the AIC included a penalty of $2 \times df$, where df represents the number of parameters in the model; thus, the number of parameters in the model were added twice to the deviance statistic to produce the AIC. As in Equation 43, the BIC included a penalty of $\log(N \times J) \times df$; thus, the deviance statistic was increased by the number of parameters in the model multiplied by the log of the number of level-1 observations. The combination of sample size and test length resulted in M2 AIC penalties ranging from 82 to 162, and BIC penalties ranging from 378 to 858, for conditions S1-T1 and S2-T2 respectively. These penalty sizes help explain the differences in model selection by fit index.

Defining *sample size* in Equation 43 as the number of level-2 units J rather than the number of level-1 units $N \times J$ may have lead to different model comparison results. Model selection is a topic which deserves attention in future research with explanatory IRT models and HGLM such as M1 and M2. Model fit and selection involve a balance between model complexity and parsimony. When fit indices produce conflicting model comparison results, the choice of index (and model) may be made based on which of these two features, complexity or parsimony, is deemed to be more important. In the present study,

model M2 was demonstrated as a screening tool, one which could be used, for example, in the item analysis process where the number of items and item-position effects is fixed. In this case, complexity may be preferable to parsimony and the AIC may be a more appropriate index. For detailed discussions of these issues, see Burnham and Anderson (2004) and Kuha (2004).

Applications

Items affected by position bias can be problematic for scale maintenance. In the GRE data sets, the largest negative item-position effect was -0.030, which corresponded to a predicted change of -0.828 logits from the beginning to the end of the 28-item test form. This item was predicted to become significantly more difficult as it appeared later in the test. Were it utilized to anchor forms to a base scale, this bias could impact the calibration of new items, depending on the change in position across forms. In an extreme case, the item may appear in position 1 at its initial calibration, where it is estimated to have a difficulty of, say, 2, making it a relatively easy item. In a later form the item may appear in position 28, where it is estimated to have a difficulty of 1.72. In this case the remaining items in the new form would seem easier, relative to the biased item, and the bias would thus be propagated.

As shown in the simulation study, person parameter recovery did not

appear to depend on model. That is, controlling for position effects in M2 did not seem to improve recovery of ability parameters, and ignoring them in M0 did not seem to degrade it. Mean $BiasAbs$ and $RMSE$ were large, over one third of a standard deviation; however, they remained essentially unchanged across models M0, M1 and M2. Other study designs, discussed below, would allow for a more direct examination of the impact of position effects on person ability estimates.

As suggested above, the item-position HGLM M2 and variations of it seem most useful as tools for item and test development. Prior to operational administration in a scrambled test form or CAT, items could be administered non-operationally in randomized sequences as part of a pilot study. M2 could then be used to examine the susceptibility of items to position effects. Such a design seems most feasible for larger testing programs utilizing a pre-equating design, wherein items are piloted regularly so as to maintain an item bank.

In practice, item-position effects should be identified as significant based on both statistical tests and effect size results. In the GRE study, omnibus tests of the full set of M2 interaction effects were conducted by comparing model fit indices for M2 versus M1. Next, potentially biased items were selected using statistical significance tests. Effects with the smallest p-values, in this case values below .05, were those which deviated the most from the main effect for

position in each item set. Many of these effects were substantial, resulting in a predicted change of greater than 0.50 logits from the beginning to the end of the test form. Were total effect sizes smaller, for example, within the standard errors associated with the item effects, they may not warrant attention.

Future Studies

Although a substantial amount of research has examined position and context effects in testing, much remains unknown. The present study examined three basic conditions under which model fit and parameter recovery could be affected. The results suggest that longer test and larger samples justify the estimation of item-position effects. Because the simulation study was designed based on data and results from the GRE (Davey & Lee, 2011), generalizations to other testing contexts and administration designs may be limited.

Given that tests with more than 40 items and samples with over 1000 examinees are commonly found in practice, the performance of these models should be examined with greater numbers of items and examinees. Additional real data studies may uncover distinct nonlinear relations between item difficulty and position as the number of items increases. Furthermore, the simulation study examined only three position effect distributions, where all items exhibited at least some change in difficulty by position. In some situations, subsets of items

may be more susceptible to position bias than others, as suggested by previous research. Given that model performance depended on test length, the number of items affected by position bias is likely a key factor and deserves attention.

A simulation study which incorporates an equating design across multiple forms could shed light on the impact of position bias on person ability estimates. The position of common items could be manipulated across forms and multiple equating methods applied, for example, the fixed common item parameter method or test characteristic curve methods (Kolen & Brennan, 2004). Varying factors such as test length and number of biased items would provide a detailed description of how item position bias can be expected to affect ability estimation.

Future research should also consider variables related to test performance and how they interact with item position to bias estimation. As discussed above and in Chapter 2, such variables include item type, item content and complexity, item difficulty, item latency and test time. These variables and others could be included within the HGLM to estimate relationships among them and their shared impact on performance.

Final Thoughts

This study, along with previous research, has shown that the commonly held but infrequently tested assumption of local independence in IRT may not

be tenable under certain testing conditions. Changes in item position can be especially problematic for item parameter estimation. Multilevel models, such as the HGLM demonstrated here, offer researchers and practitioners a flexible tool for examining the effects of position change and other variables on estimation, thereby improving item and test development.

References

Alexandrowicz, R., & Matschinger, H. (2008). Estimation of item location effects by means of the generalized logistic regression model: A simulation study and an application. *Psychology Science Quarterly*, *50*(1), 64–74.

Bejar, I. I., & Wingersky, M. S. (1982). A study of pre-equating based on item response theory. *Applied Psychological Measurement*, *6*, 309–325.

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.

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

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

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–304.

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 Rep. No. 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.

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

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, *20*(2), 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.

Flaugher, R. L., Melton, R. S., & Myers, C. T. (1968). Item rearrangement under typical test conditions. *Educational and Psychological Measurement*, *28*, 813–824.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and*

multilevel/hierarchical models. New York, NY: Cambridge University Press.

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

Holland, P. W., & Thayer, D. T. (1985). Section pre-equating in the presence of practice effects. *Journal of Educational and Behavioral Statistics*, *10*, 109–120.

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.

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

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8*, 147–154.

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

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly*, *50*, 311–327.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research*, *33*, 188–229.

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.

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

Meyers, J. L., Miller, G. E., & Way, W. D. (2008). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education*, *22*, 38–60.

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

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

Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement*, *30*, 100–120.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Sax, G., & Carr, A. (1962). An investigation of response sets on altered parallel forms. *Educational and Psychological Measurement*, *22*, 371–376.

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

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.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational*

Measurement, *27*, 361–370.

Swinton, S. S., Wild, C. W., & Wallmark, M. (1983). *Investigation of practice effects on item types in the Graduate Record Examination Aptitude Test* (ETS Research Rep. No. RR-82-56). Princeton, NJ: ETS.

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.

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

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

Wightman, L. E. (1981). *GMAT within-test practice effects studies.* Paper presented at the meeting of the National Council on Measurement in Education, Los Angeles, CA.

Wightman, L. E., & Leary, L. F. (1981). *Dealing with practice effects in section pre-equating.* Paper presented at the meeting of the American Educational Research Association, Montreal, Canada.

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

Measurement, *17*, 297–311.

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

Appendix A: GRE Item and Item-Position Effects

This appendix includes 4 tables containing fixed effect results from fitting model M2 to the GRE data sets using effects coding. The first, Table 20, contains the M2 item effects for data sets GRE1, GRE2, and GRE3. The next three, Tables 21, 22, and 23, contain the M2 item-position interaction effects and significance test results for GRE1, GRE2, and GRE3, respectively.

Table 20: GRE Item Effects and Standard Errors

Effect	GRE1	SE	GRE2	SE	GRE3	SE
γ_{00}	0.993	0.034	1.043	0.035	1.061	0.036
γ_{10}	0.755	0.135	0.289	0.120	−0.368	0.114
γ_{20}	1.747	0.161	1.813	0.160	1.413	0.147
γ_{30}	−1.438	0.111	−1.529	0.114	1.204	0.150
γ_{40}	−0.765	0.104	0.346	0.116	1.602	0.160
γ_{50}	0.817	0.136	0.905	0.135	0.503	0.122
γ_{60}	−0.224	0.107	2.305	0.199	0.772	0.136
γ_{70}	−0.536	0.105	0.312	0.120	−0.113	0.105
γ_{80}	−1.806	0.118	0.133	0.112	−0.097	0.115
γ_{90}	0.583	0.129	−0.243	0.113	−2.097	0.109
γ_{100}	0.670	0.130	−1.185	0.102	−0.584	0.110
γ_{110}	0.023	0.110	−1.523	0.112	−0.431	0.107
γ_{120}	−1.133	0.112	−0.932	0.102	0.830	0.132
γ_{130}	0.160	0.112	−0.808	0.110	1.403	0.148
γ_{140}	1.122	0.148	−0.831	0.116	−1.195	0.108
γ_{150}	−0.274	0.105	1.640	0.158	−0.801	0.102
γ_{160}	−1.350	0.113	−0.390	0.104	−2.409	0.123
γ_{170}	0.787	0.130	0.580	0.124	−2.248	0.117
γ_{180}	−1.799	0.108	−0.559	0.104	0.414	0.122
γ_{190}	0.672	0.120	−0.676	0.107	−2.151	0.114
γ_{200}	−1.923	0.116	1.039	0.131	1.620	0.157
γ_{210}	1.694	0.165	−0.169	0.111	2.136	0.197
γ_{220}	1.827	0.167	−0.423	0.110	0.011	0.113
γ_{230}	0.702	0.133	−1.294	0.106	0.494	0.130
γ_{240}	−0.154	0.114	0.634	0.125	−2.175	0.114
γ_{250}	1.064	0.142	0.330	0.116	1.281	0.155
γ_{260}	−0.074	0.107	−0.221	0.108	0.536	0.124
γ_{270}	−0.875	0.113	0.616	0.129	1.247	0.150
γ_{qref0}	−0.270	0.106	−0.160	0.107	−0.798	0.105

Note: Although the same row labels are used, items differ across GRE1, GRE2, and GRE3. γ_{00} is the mean of item effects. γ_{qref0} is the effect for the reference item.

Table 21: GRE1 Item-Position Interaction Effects

Effect	Estimate	SE	z	p
γ_{280}	−0.010	0.001	−6.939	0.000
γ_{290}	0.003	0.008	0.434	0.664
γ_{300}	−0.015	0.010	−1.525	0.127
γ_{310}	0.026	0.006	4.085	0.000
γ_{320}	0.008	0.007	1.258	0.208
γ_{330}	0.005	0.008	0.578	0.563
γ_{340}	−0.003	0.007	−0.377	0.706
γ_{350}	0.001	0.007	0.201	0.841
γ_{360}	−0.002	0.007	−0.222	0.824
γ_{370}	0.010	0.008	1.278	0.201
γ_{380}	−0.030	0.007	−4.058	0.000
γ_{390}	0.003	0.007	0.426	0.670
γ_{400}	−0.000	0.007	−0.068	0.946
γ_{410}	−0.006	0.007	−0.820	0.412
γ_{420}	0.001	0.009	0.124	0.901
γ_{430}	0.007	0.007	1.038	0.299
γ_{440}	0.004	0.007	0.635	0.525
γ_{450}	−0.015	0.007	−2.054	0.040
γ_{460}	0.013	0.007	1.820	0.069
γ_{470}	−0.016	0.007	−2.210	0.027
γ_{480}	0.005	0.007	0.670	0.503
γ_{490}	−0.015	0.010	−1.529	0.126
γ_{500}	−0.024	0.010	−2.425	0.015
γ_{510}	0.007	0.008	0.901	0.367
γ_{520}	0.008	0.007	1.140	0.254
γ_{530}	0.014	0.009	1.619	0.105
γ_{540}	−0.004	0.007	−0.660	0.509
γ_{550}	0.007	0.007	1.001	0.317
$\gamma_{(N+qref)0}$	0.007	0.007	1.000	0.318

Note: γ_{280} is the main effect for position, the mean item-position interaction effect. $\gamma_{(N+qref)0}$ is the interaction effect for the reference item.

Table 22: GRE2 Item-Position Interaction Effects

Effect	Estimate	SE	z	p
γ_{280}	−0.015	0.001	−11.105	0.000
γ_{290}	−0.010	0.007	−1.431	0.152
γ_{300}	−0.027	0.009	−2.879	0.004
γ_{310}	0.010	0.007	1.497	0.134
γ_{320}	−0.004	0.007	−0.504	0.614
γ_{330}	−0.014	0.008	−1.855	0.064
γ_{340}	−0.011	0.011	−0.948	0.343
γ_{350}	−0.007	0.007	−1.003	0.316
γ_{360}	0.013	0.007	1.857	0.063
γ_{370}	−0.003	0.007	−0.509	0.610
γ_{380}	0.012	0.006	1.864	0.062
γ_{390}	0.017	0.007	2.539	0.011
γ_{400}	−0.002	0.006	−0.310	0.756
γ_{410}	−0.006	0.007	−0.853	0.394
γ_{420}	0.006	0.007	0.904	0.366
γ_{430}	−0.018	0.009	−2.025	0.043
γ_{440}	0.004	0.007	0.584	0.559
γ_{450}	−0.001	0.008	−0.075	0.940
γ_{460}	0.002	0.007	0.310	0.756
γ_{470}	0.008	0.007	1.287	0.198
γ_{480}	0.001	0.008	0.094	0.925
γ_{490}	0.008	0.007	1.199	0.231
γ_{500}	0.005	0.007	0.753	0.452
γ_{510}	0.017	0.007	2.574	0.010
γ_{520}	0.003	0.008	0.334	0.738
γ_{530}	−0.002	0.007	−0.347	0.729
γ_{540}	0.007	0.007	1.124	0.261
γ_{550}	−0.010	0.007	−1.373	0.170
$\gamma_{(N+qref)0}$	0.001	0.007	0.130	0.897

Note: γ_{280} is the main effect for position, the mean item-position interaction effect. $\gamma_{(N+qref)0}$ is the interaction effect for the reference item.

Table 23: GRE3 Item-Position Interaction Effects

Effect	Estimate	SE	z	p
γ_{280}	−0.013	0.001	−8.864	0.000
γ_{290}	0.000	0.007	0.040	0.968
γ_{300}	−0.000	0.009	−0.055	0.956
γ_{310}	0.000	0.008	0.015	0.988
γ_{320}	0.007	0.010	0.692	0.489
γ_{330}	−0.019	0.007	−2.603	0.009
γ_{340}	0.012	0.008	1.435	0.151
γ_{350}	−0.009	0.007	−1.346	0.178
γ_{360}	−0.004	0.007	−0.617	0.537
γ_{370}	0.016	0.007	2.340	0.019
γ_{380}	0.003	0.007	0.389	0.697
γ_{390}	0.014	0.007	1.960	0.050
γ_{400}	−0.023	0.008	−2.965	0.003
γ_{410}	−0.019	0.009	−2.219	0.026
γ_{420}	0.016	0.006	2.498	0.012
γ_{430}	0.014	0.007	2.185	0.029
γ_{440}	0.018	0.007	2.523	0.012
γ_{450}	−0.010	0.008	−1.293	0.196
γ_{460}	−0.001	0.007	−0.190	0.849
γ_{470}	0.001	0.007	0.133	0.895
γ_{480}	−0.017	0.010	−1.775	0.076
γ_{490}	−0.008	0.011	−0.687	0.492
γ_{500}	−0.004	0.007	−0.600	0.549
γ_{510}	−0.007	0.007	−0.984	0.325
γ_{520}	0.033	0.007	4.872	0.000
γ_{530}	−0.008	0.009	−0.871	0.384
γ_{540}	−0.015	0.007	−2.029	0.042
γ_{550}	−0.001	0.009	−0.071	0.943
$\gamma_{(N+qref)0}$	0.011	0.006	1.698	0.090

Note: γ_{280} is the main effect for position, the mean item-position interaction effect. $\gamma_{(N+qref)0}$ is the interaction effect for the reference item.

Appendix B: Mean Parameter Recovery Plots

This appendix includes 12 figures, each containing 4 plots for a particular parameter recovery index, averaged over items, by condition and model. Figures 12, 13, 14, and 15 contain the mean $Bias$, $AbsBias$, SE, and $RMSE$ at the center of the test form, i.e., $\bar{p}$. Figures 16, 17, 18, and 19 contain the mean $Bias$ and $AbsBias$ at positions p_1 and p_N. Figures 20, 21, 22, and 23 contain the mean SE and $RMSE$ at positions p_1 and p_N.

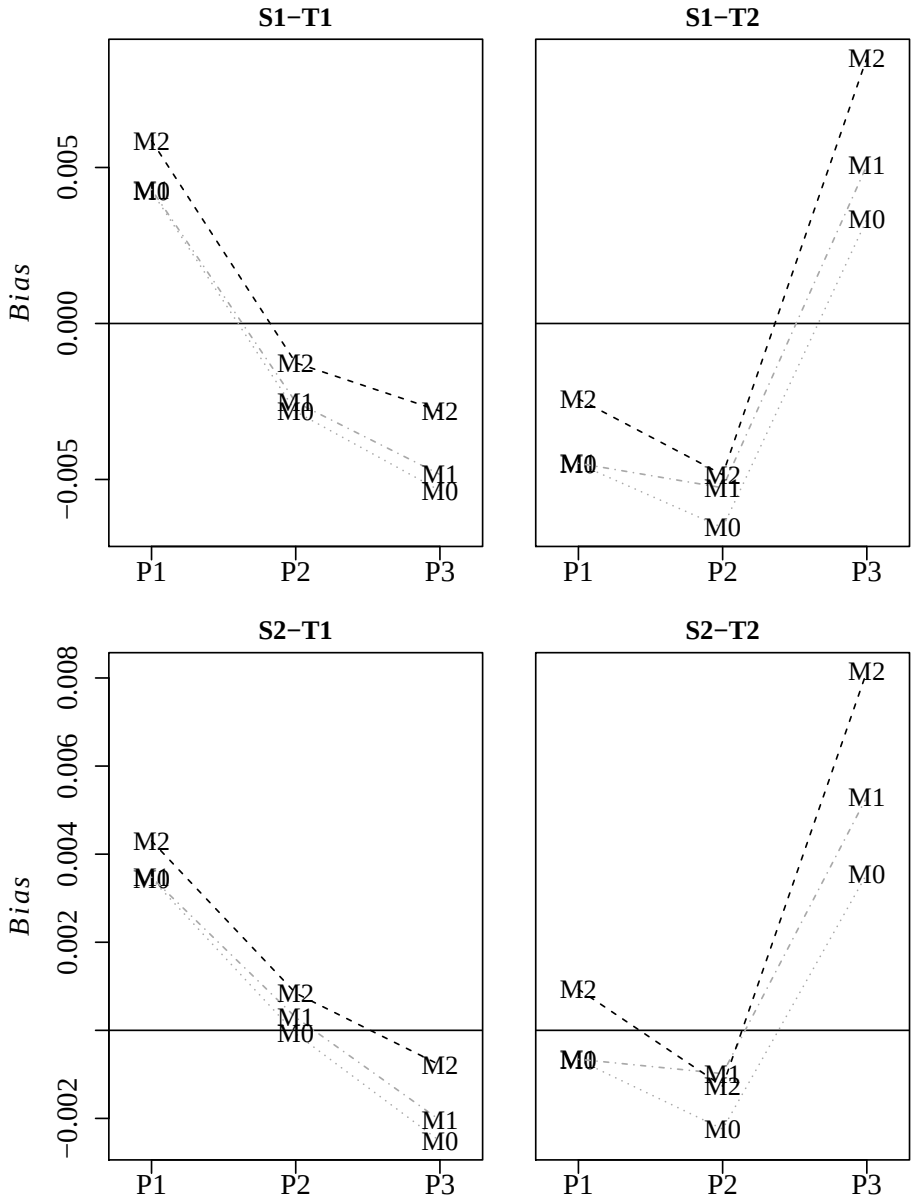

Figure 12: Average *Bias* for models M0, M1, and M2. Each plot depicts a combination of sample size (S) and test length (T) conditions, with *Bias* by model across position effect condition (P).

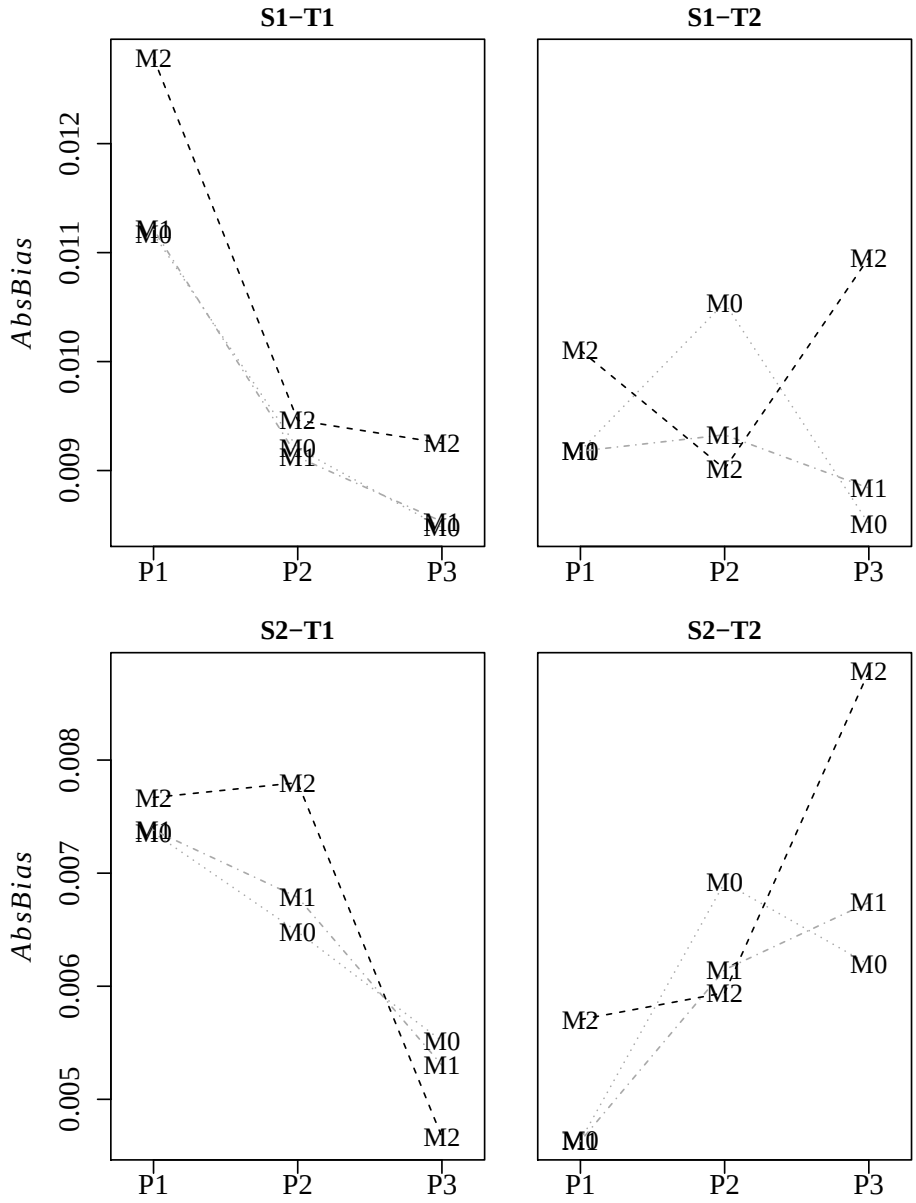

Figure 13: Average *AbsBias* for models M0, M1, and M2. Each plot depicts a combination of sample size (S) and test length (T) conditions, with *AbsBias* by model across position effect condition (P).

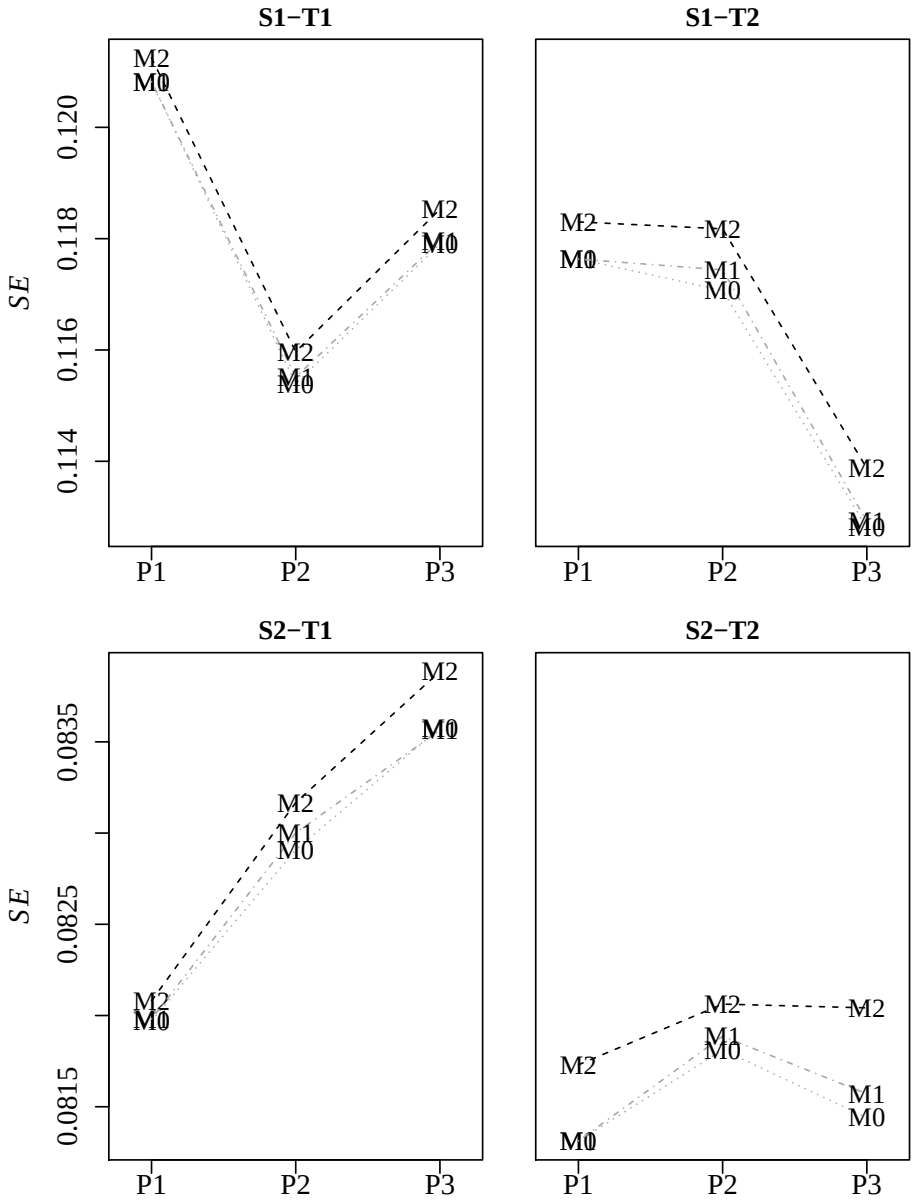

Figure 14: Average SE for models M0, M1, and M2. Each plot depicts a combination of sample size (S) and test length (T) conditions, with SE by model across position effect condition (P).

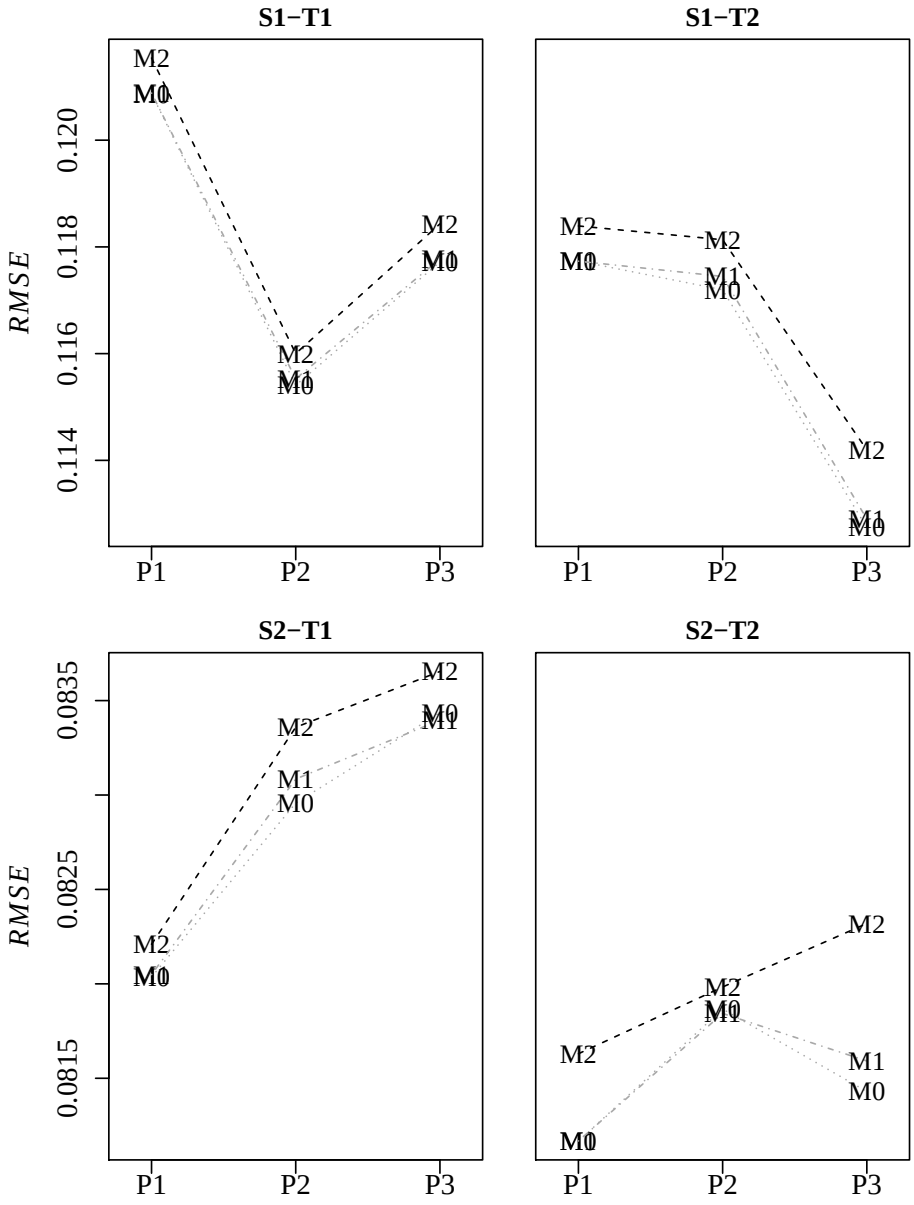

Figure 15: Average $RMSE$ for models M0, M1, and M2. Each plot depicts a combination of sample size (S) and test length (T) conditions, with $RMSE$ by model across position effect condition (P).

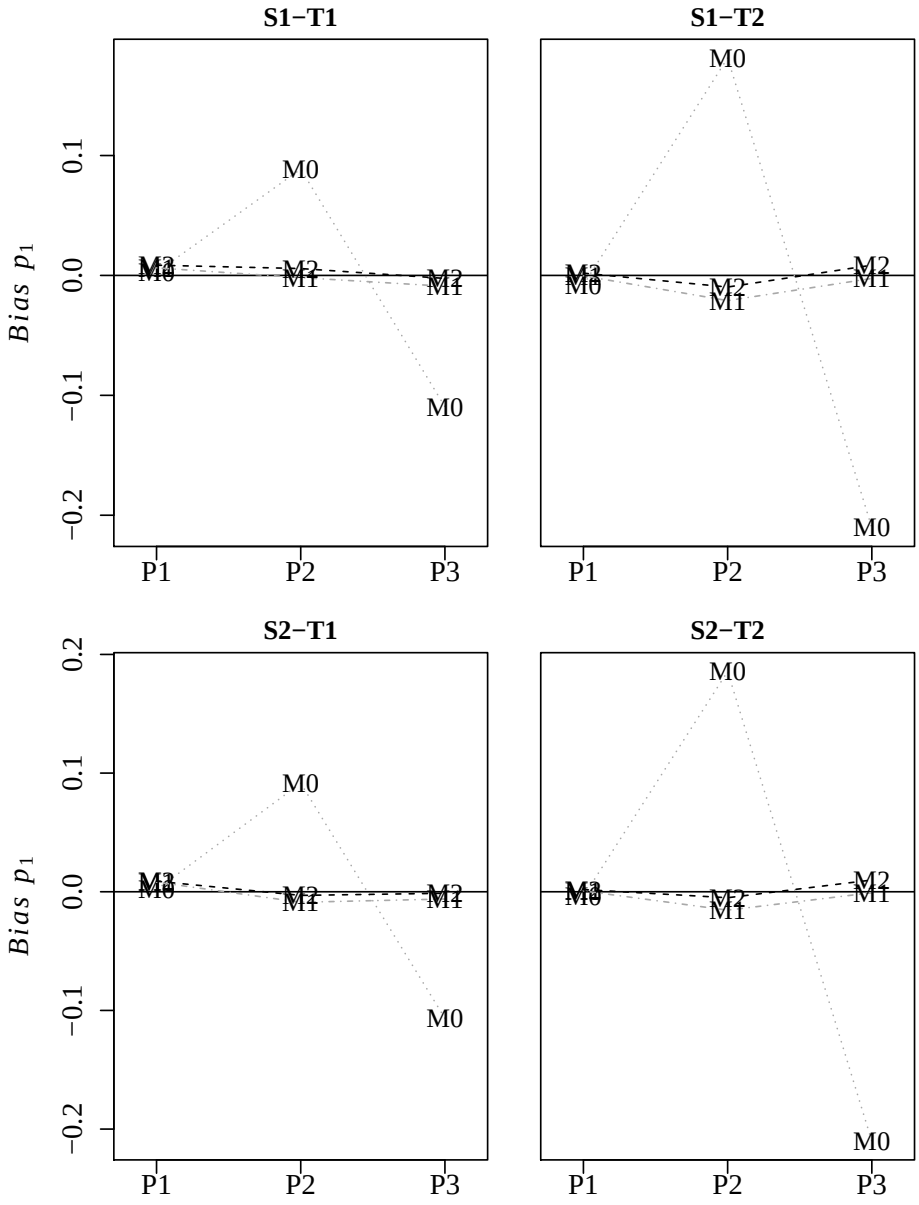

Figure 16: Average $Bias\ p_1$ for models M0, M1, and M2. Each plot depicts a combination of sample size (S) and test length (T) conditions, with $Bias\ p_1$ by model across position effect condition (P).

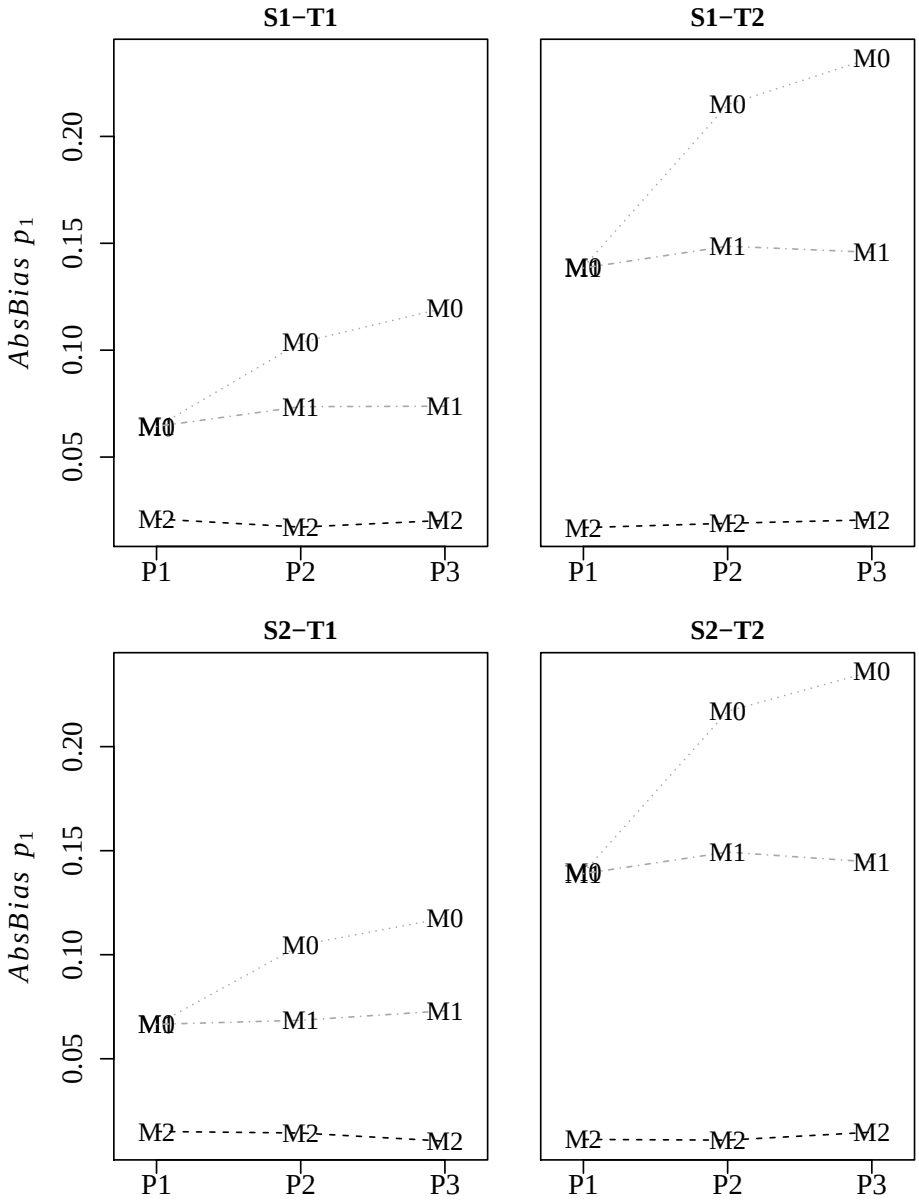

Figure 17: Average $AbsBias\ p_1$ for models M0, M1, and M2. Each plot depicts a combination of sample size (S) and test length (T) conditions, with $AbsBias\ p_1$ by model across position effect condition (P).

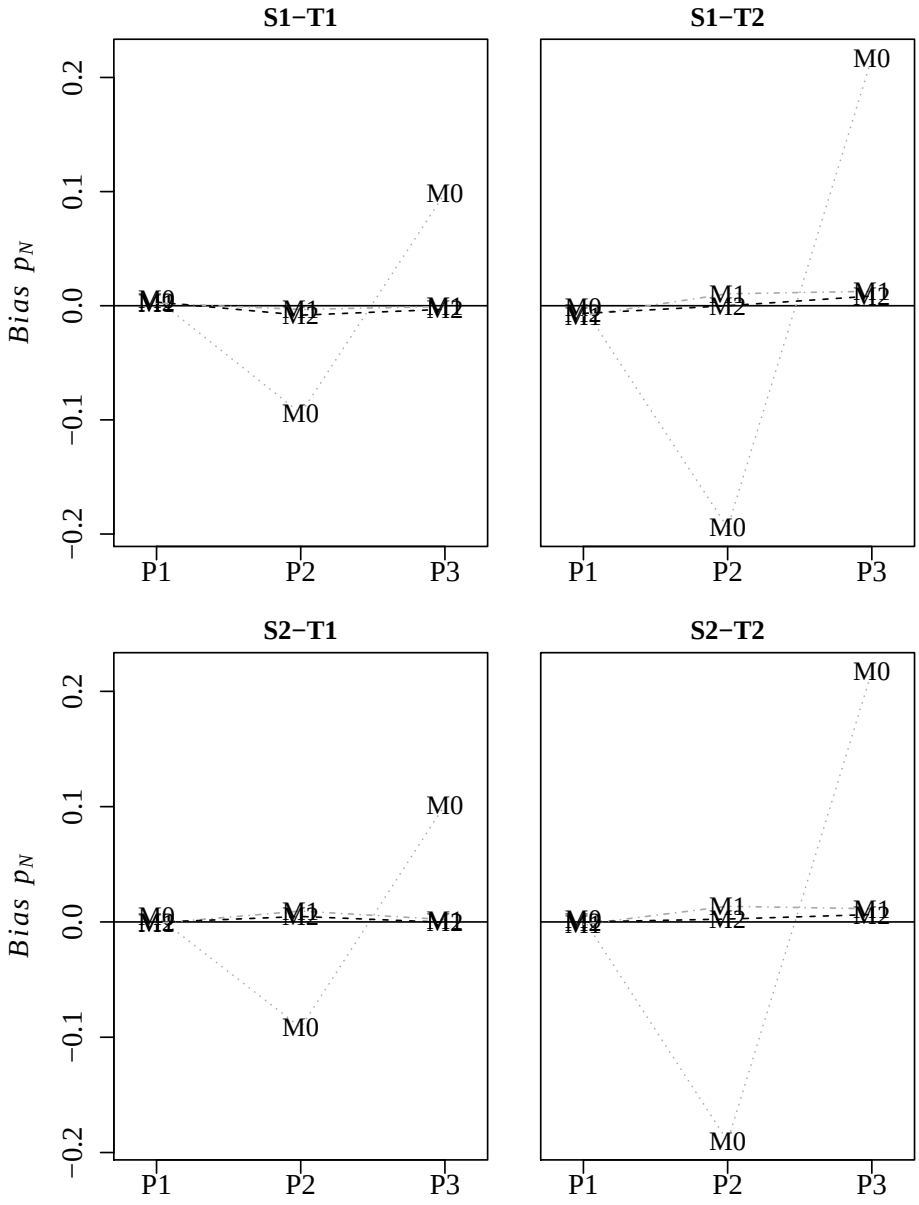

Figure 18: Average *Bias* p_N for models M0, M1, and M2. Each plot depicts a combination of sample size (S) and test length (T) conditions, with *Bias* p_N by model across position effect condition (P).

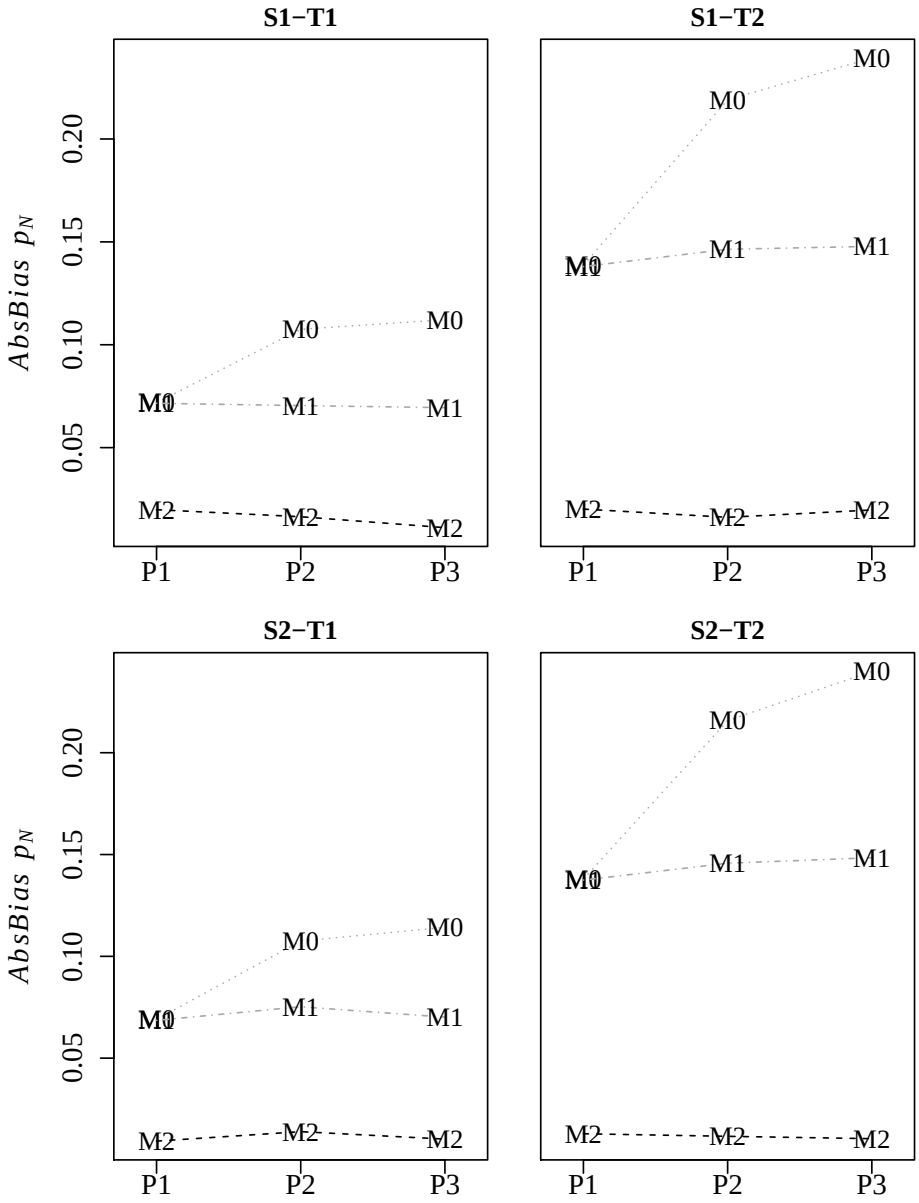

Figure 19: Average $AbsBias\, p_N$ for models M0, M1, and M2. Each plot depicts a combination of sample size (S) and test length (T) conditions, with $AbsBias\, p_N$ by model across position effect condition (P).

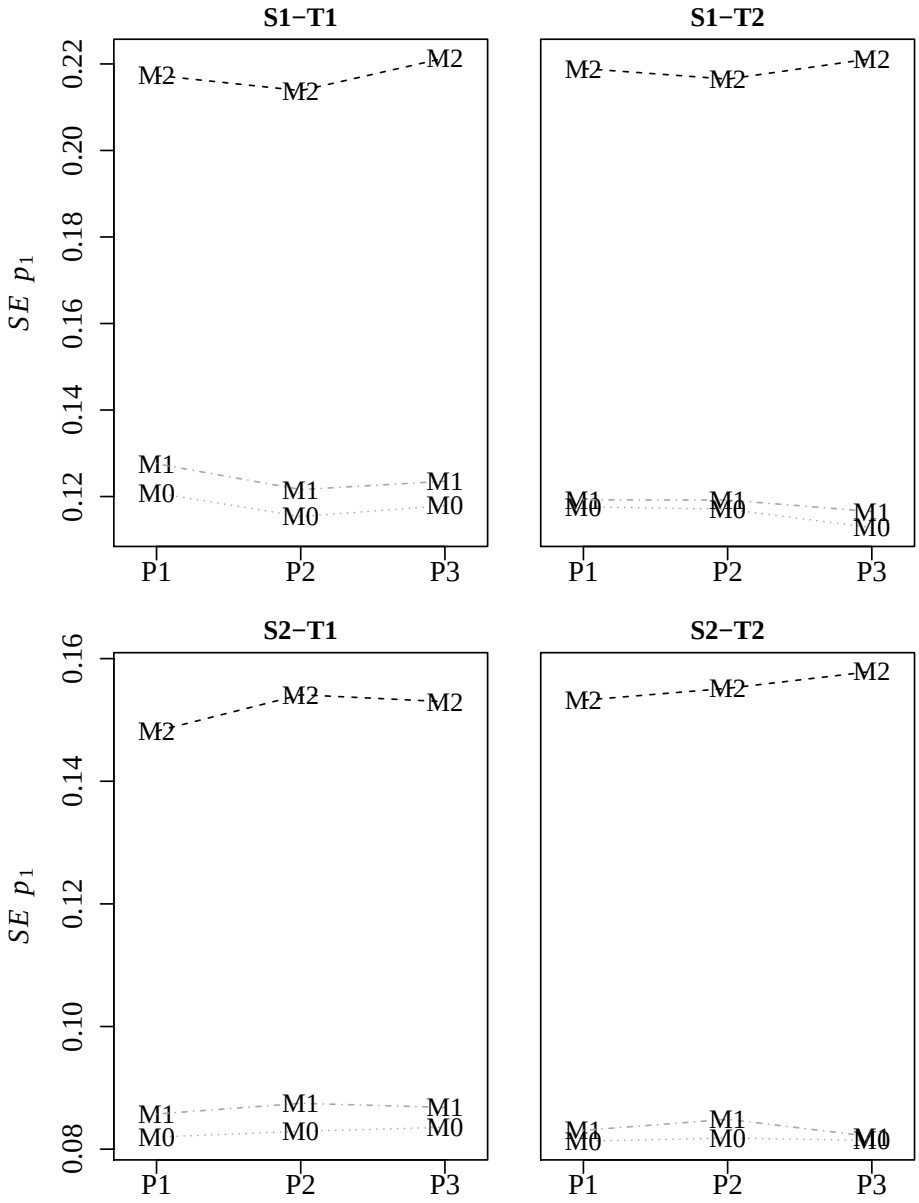

Figure 20: Average $SE\,p_1$ for models M0, M1, and M2. Each plot depicts a combination of sample size (S) and test length (T) conditions, with $SE\,p_1$ by model across position effect condition (P).

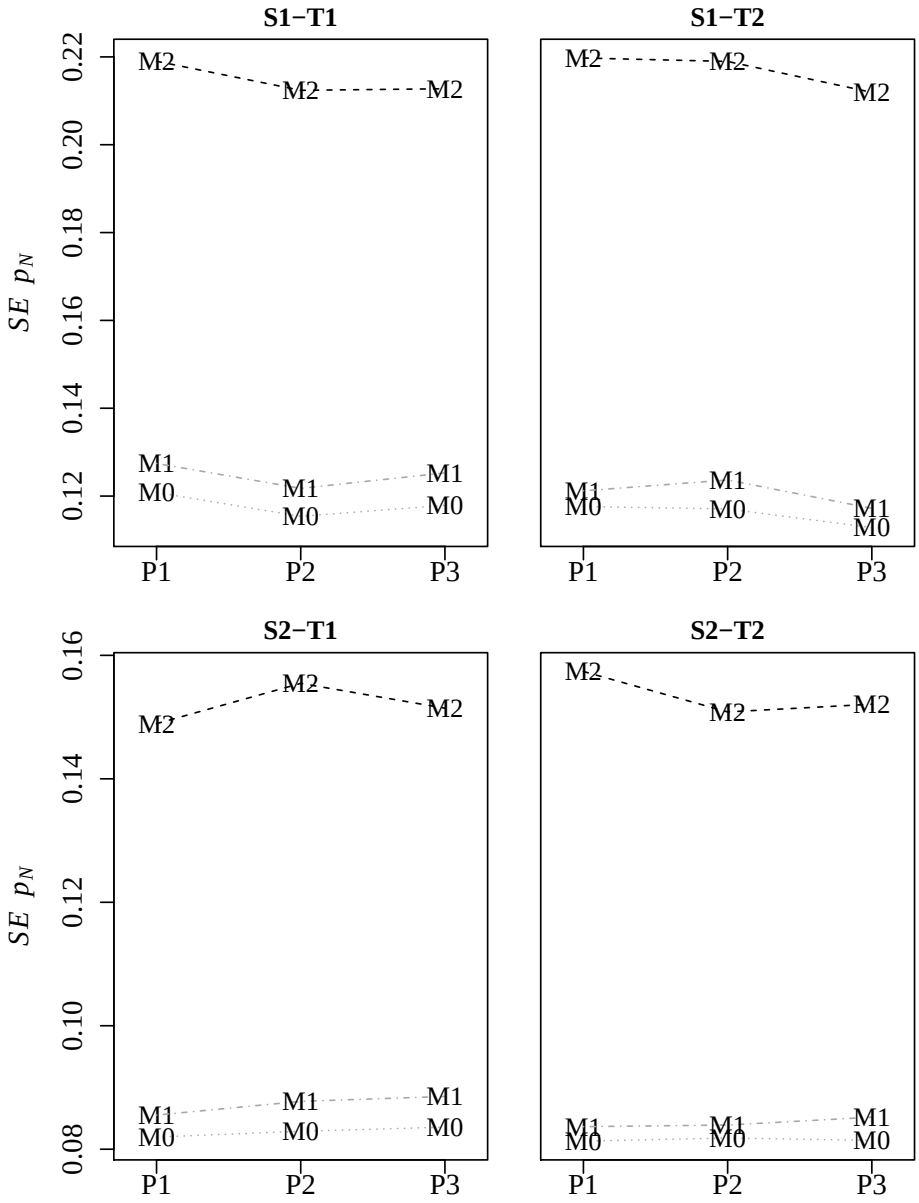

Figure 21: Average $SE\,p_N$ for models M0, M1, and M2. Each plot depicts a combination of sample size (S) and test length (T) conditions, with $SE\,p_N$ by model across position effect condition (P).

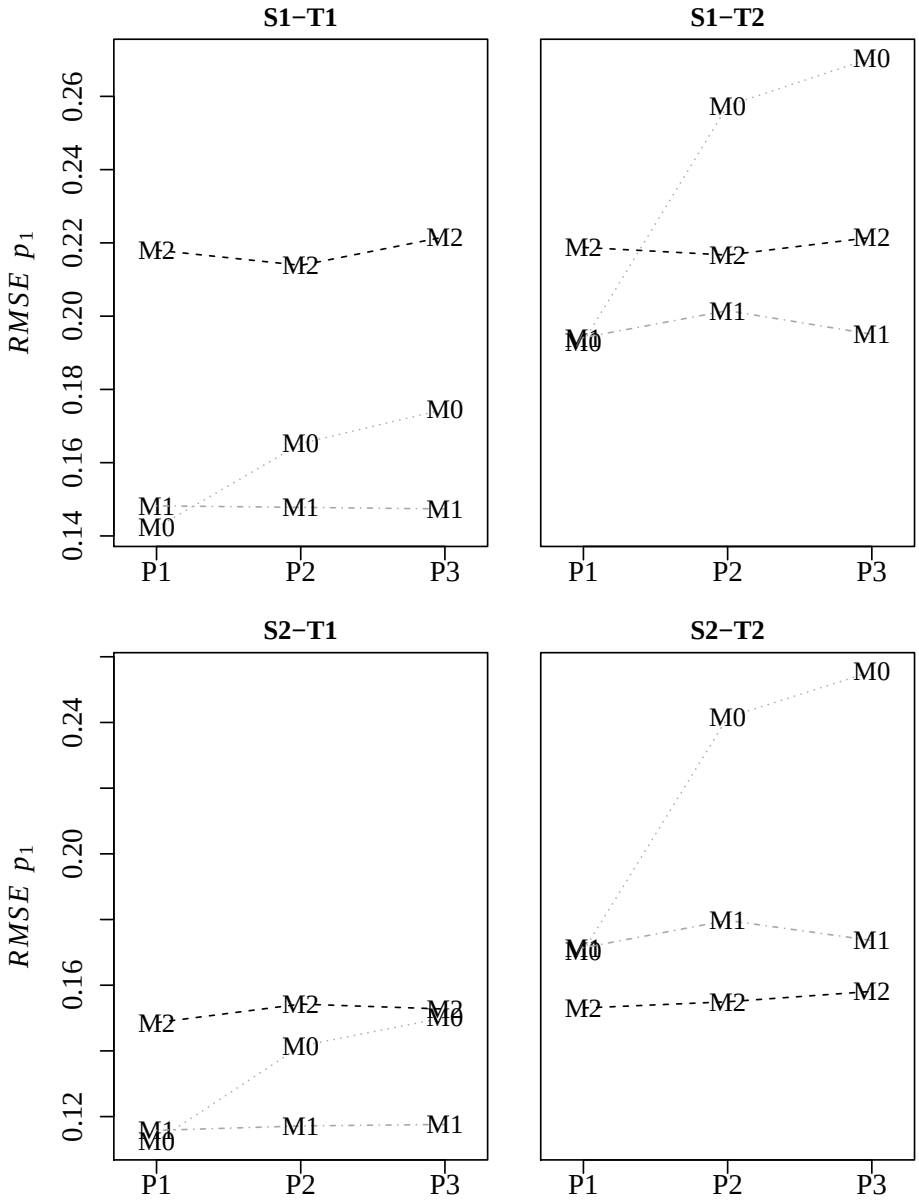

Figure 22: Average $RMSE\,p_1$ for models M0, M1, and M2. Each plot depicts a combination of sample size (S) and test length (T) conditions, with $RMSE\,p_1$ by model across position effect condition (P).

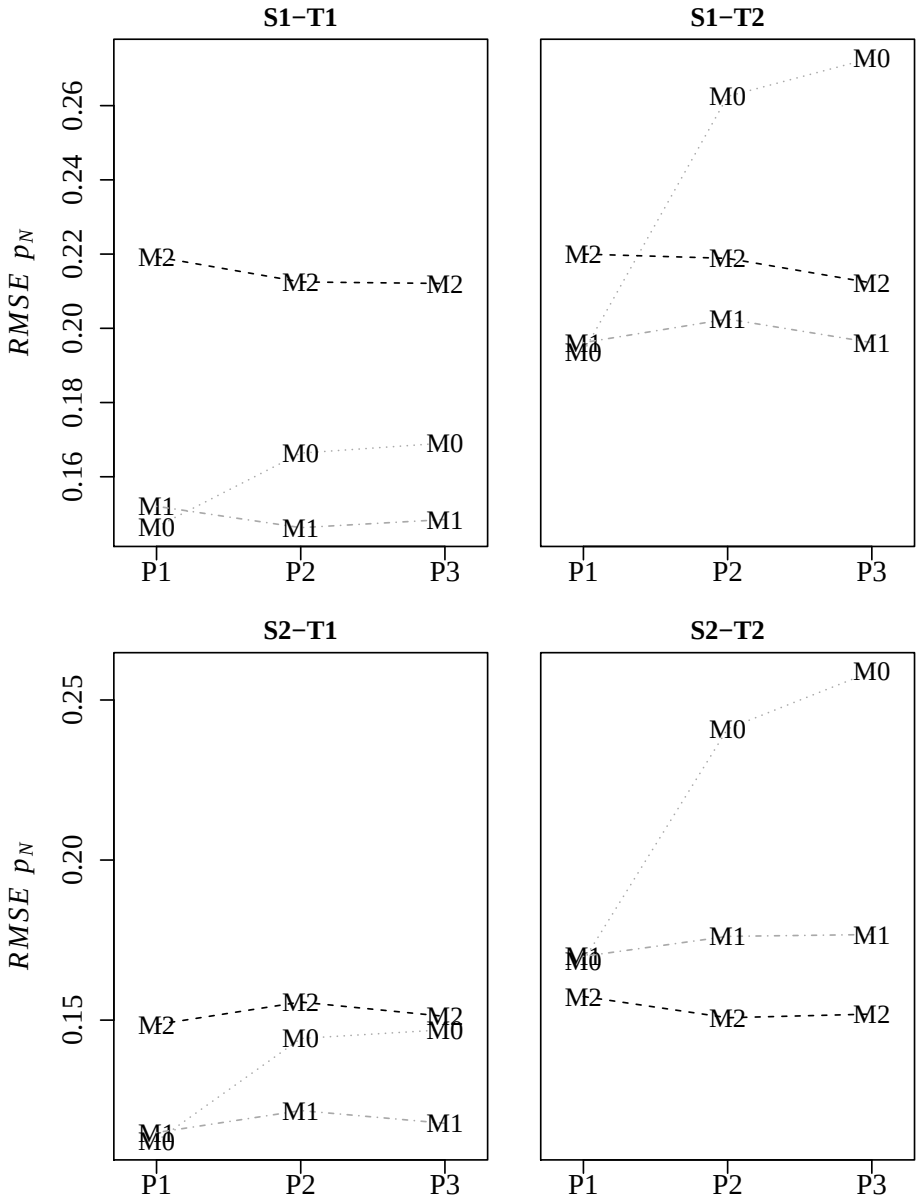

Figure 23: Average $RMSE\,p_N$ for models M0, M1, and M2. Each plot depicts a combination of sample size (S) and test length (T) conditions, with $RMSE\,p_N$ by model across position effect condition (P).

Appendix C: Item Difficulty Plots

This appendix includes 18 figures, each containing 20 plots of true and estimated item difficulty across item positions. Each plot contains the following: (1) the true item difficulty parameter at each position as a black solid line; (2) the M0 estimate of item difficulty, constant across position, as a horizontal grey dotted line; (3) the M1 estimate of item difficulty, constant across items within a condition, as a grey dotted/dashed line; and (4) the M2 estimate of item difficulty across position, as a black dashed line.

Each of the 20-item conditions, from T1, is contained in a single figure. The 40-item conditions, from T2, are divided across two figures. Figures 24, 25, and 26 contain plots for conditions S1-T1-P1, S1-T1-P2, and S1-T1-P3. Condition S1-T2-P1 is plotted in Figures 27 and 28; S1-T2-P2 in Figures 29 and 30; and S1-T2-P3 in Figures 31 and 32.

Figures 33, 34, and 35 contain plots for conditions S2-T1-P1, S2-T1-P2, and S2-T1-P3. Condition S2-T2-P1 is plotted in Figures 36 and 37; S2-T2-P2 in Figures 38 and 39; and S2-T2-P3 in Figures 40 and 41.

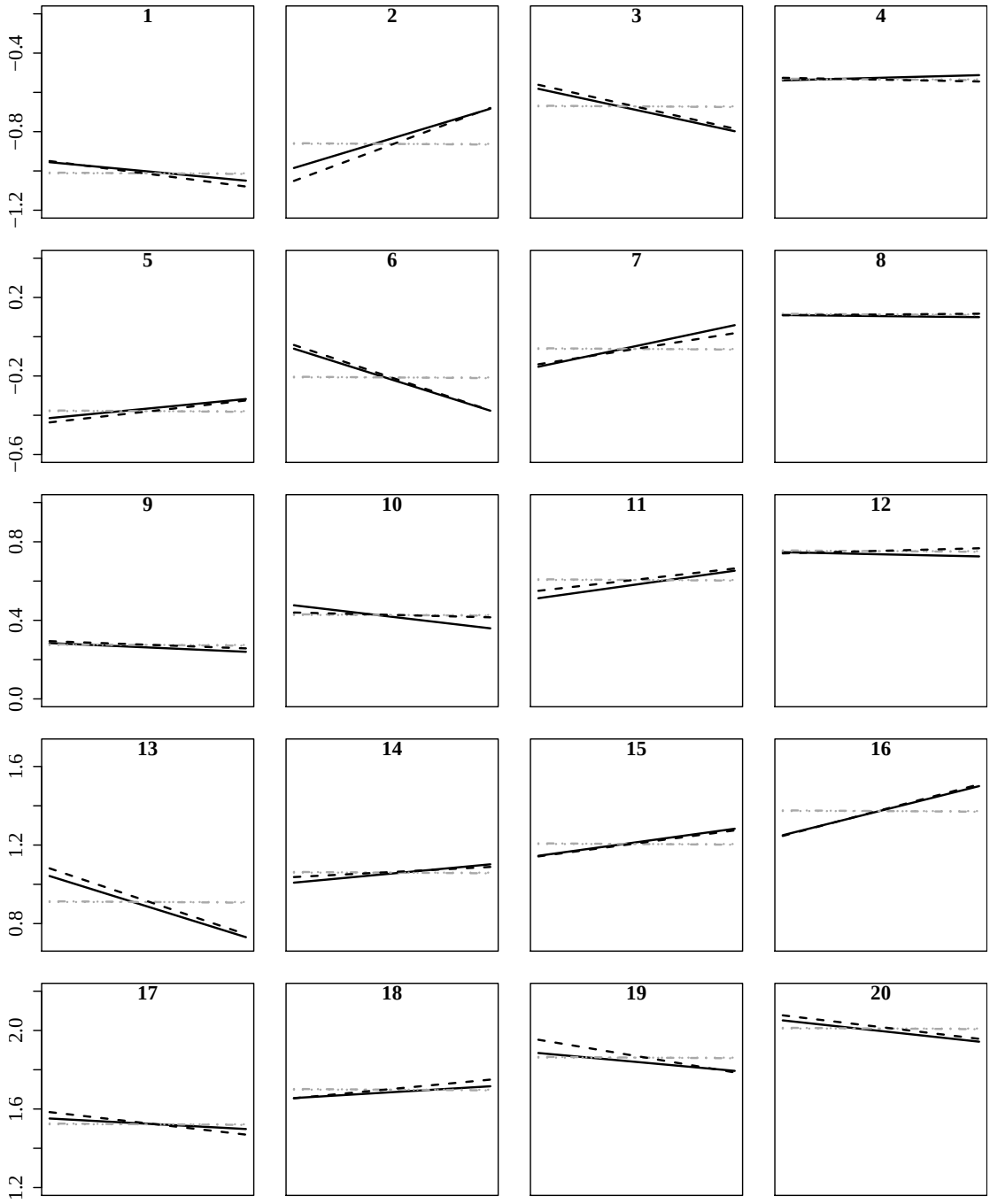

Figure 24: S1-T1-P1 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed) across position.

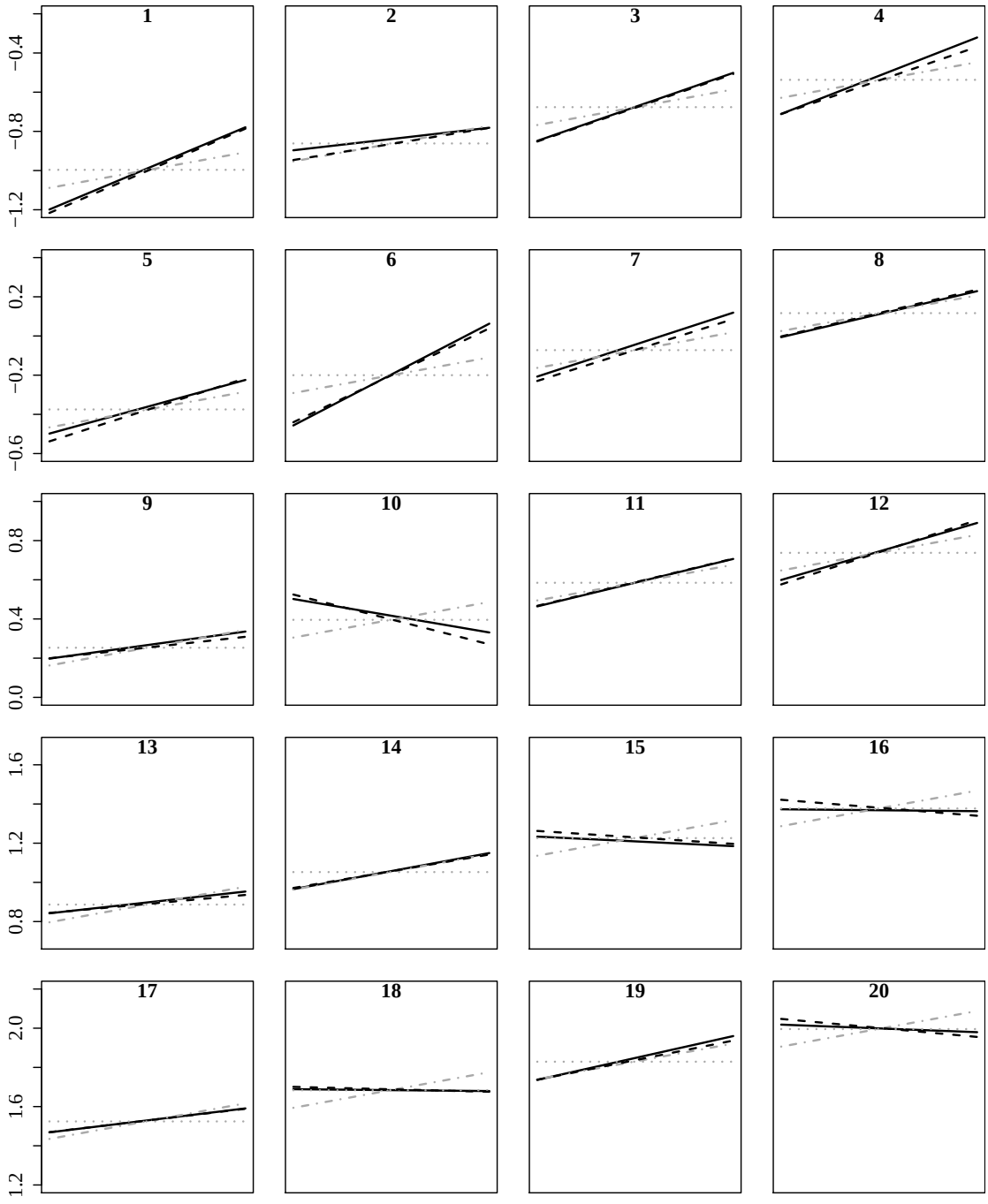

Figure 25: S1-T1-P2 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed) across position.

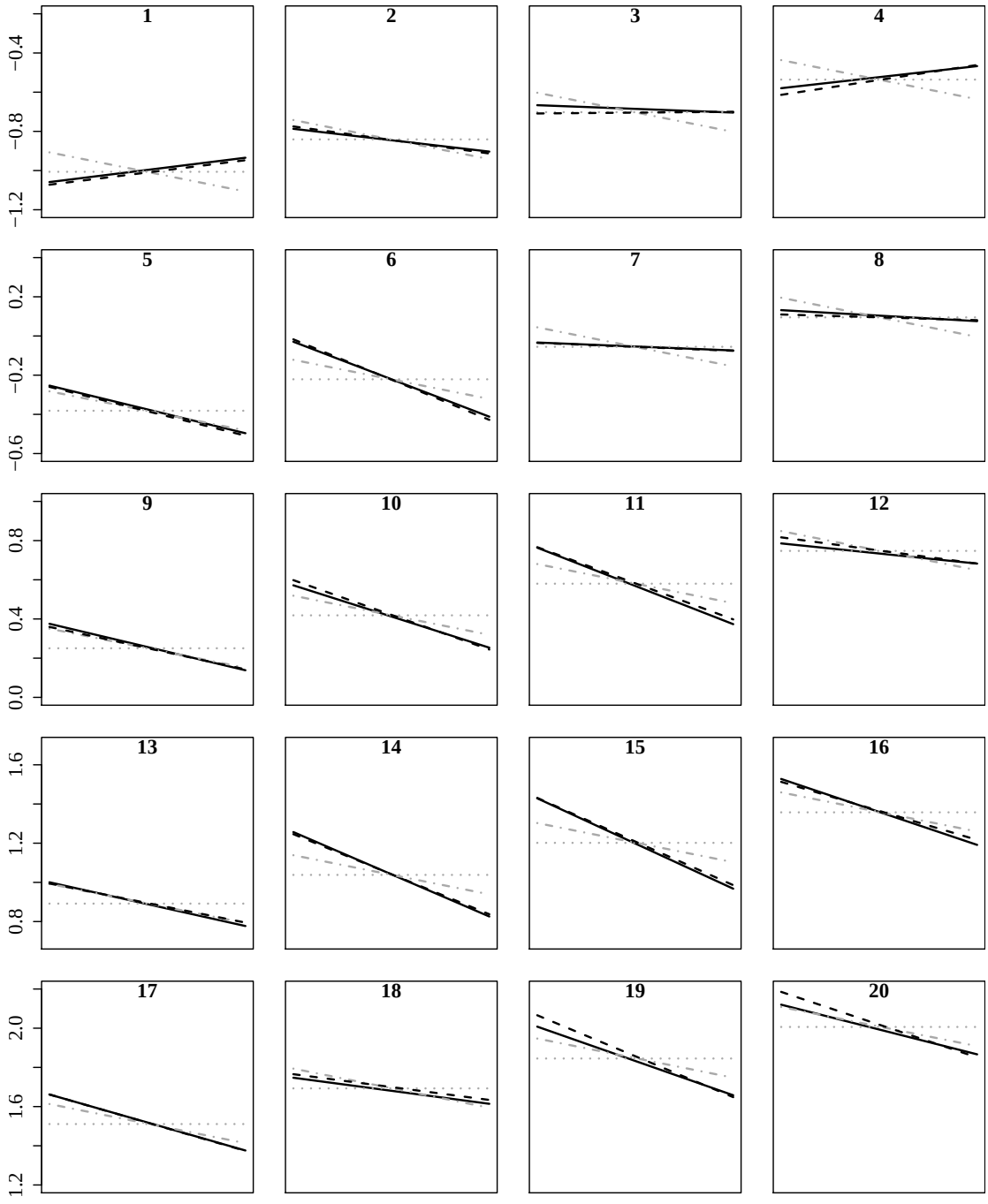

Figure 26: S1-T1-P3 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed) across position.

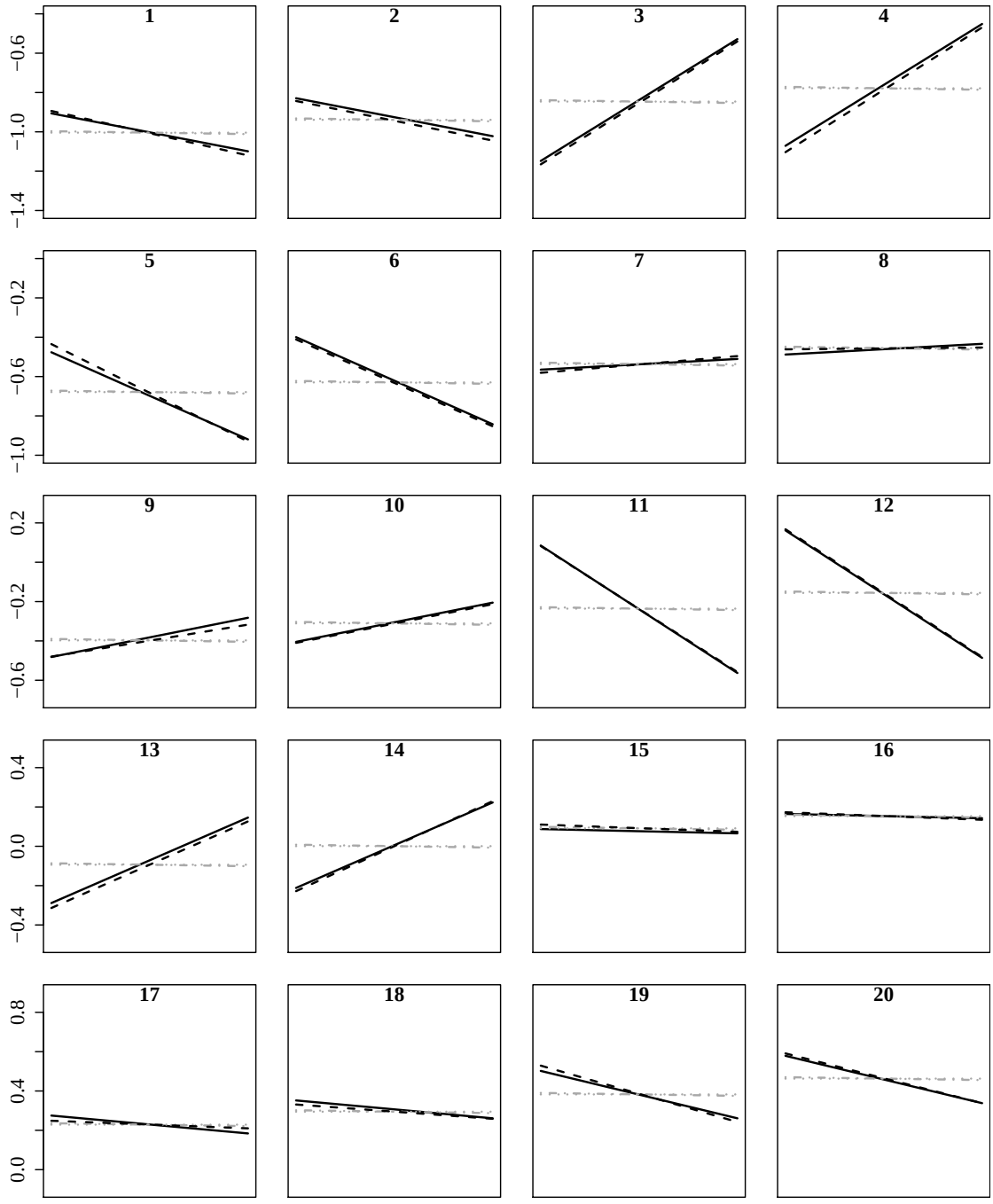

Figure 27: S1-T2-P1 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed), items 1 to 20.

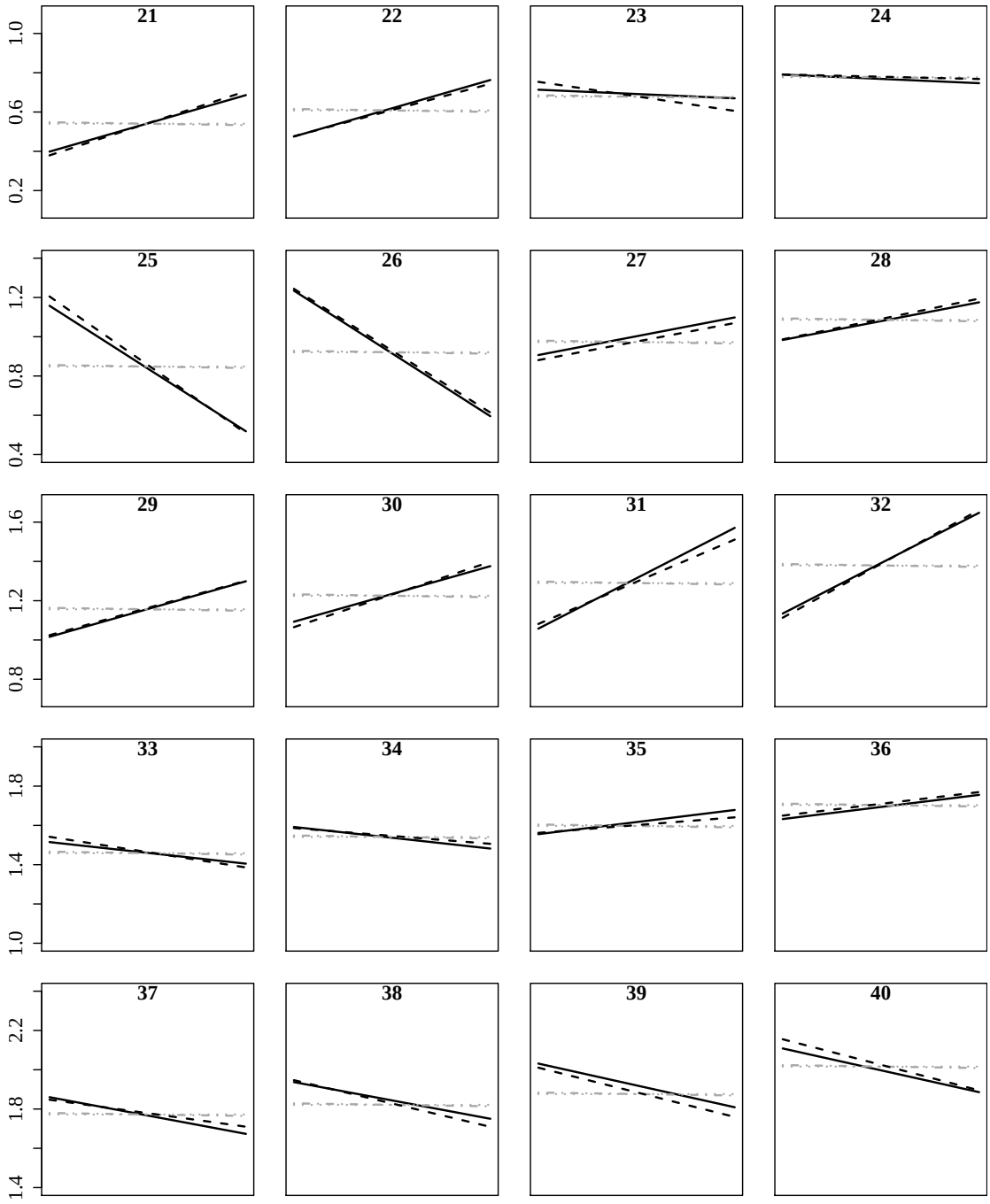

Figure 28: S1-T2-P1 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed), items 21 to 40.

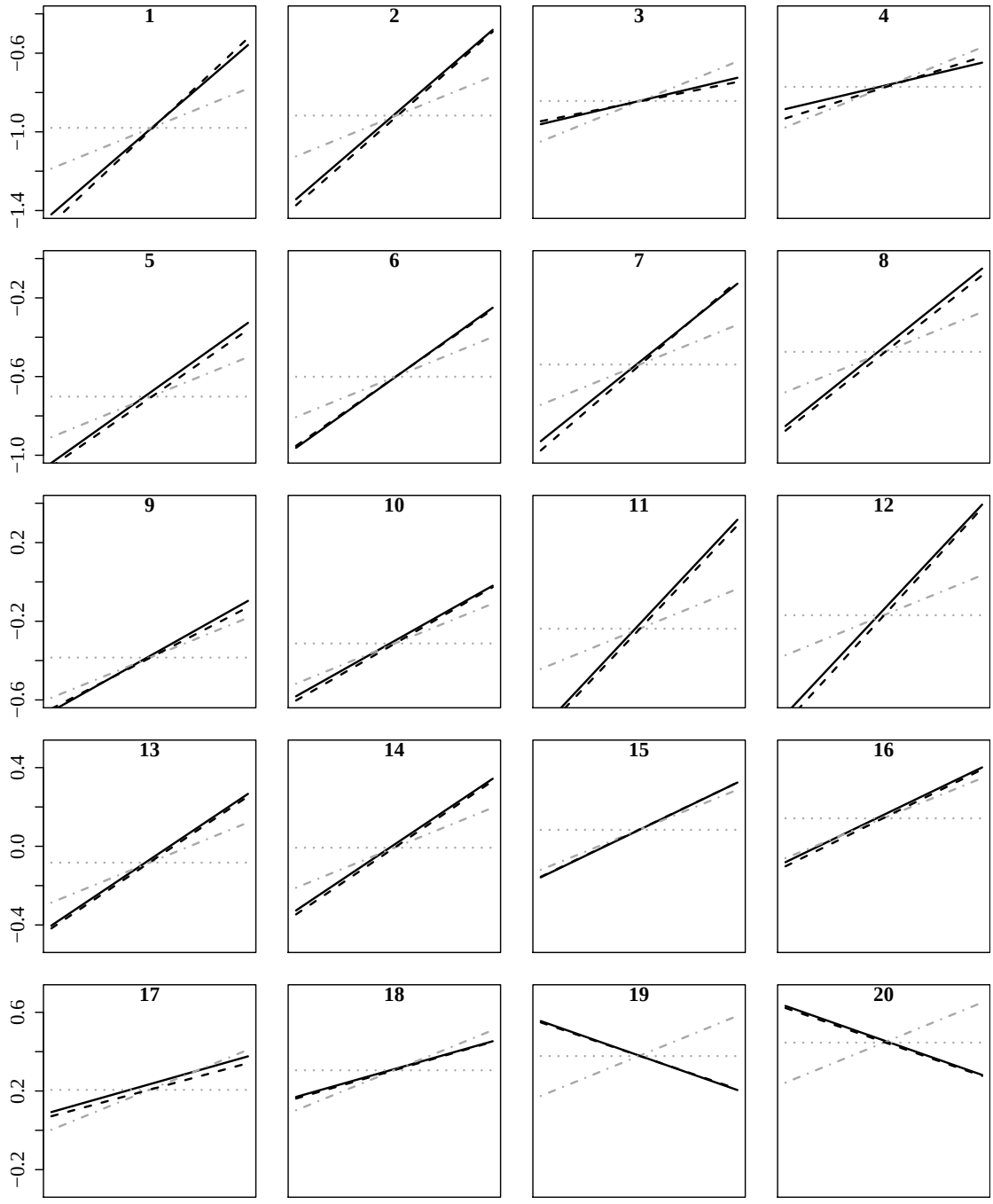

Figure 29: S1-T2-P2 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed), items 1 to 20.

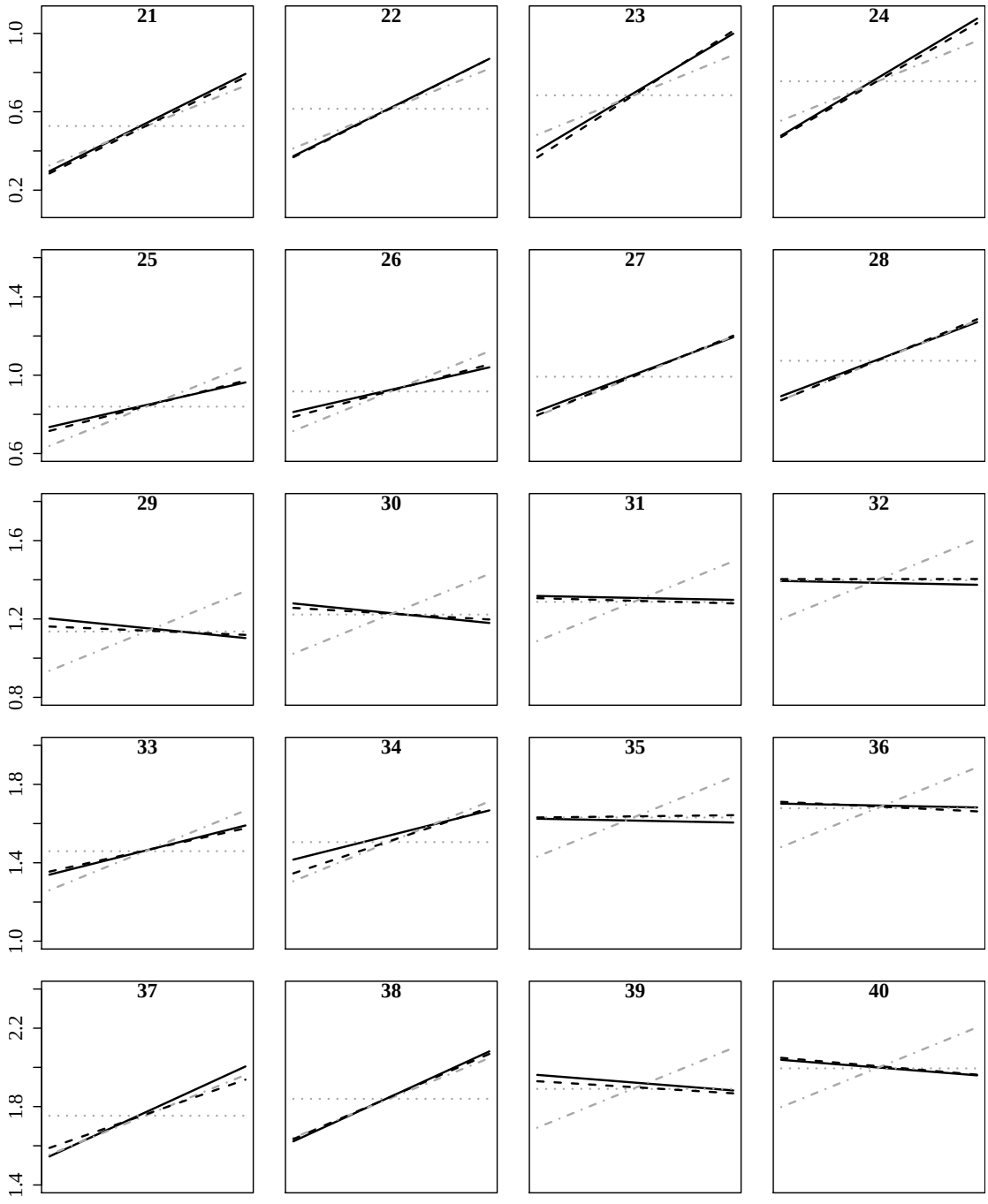

Figure 30: S1-T2-P2 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed), items 21 to 40.

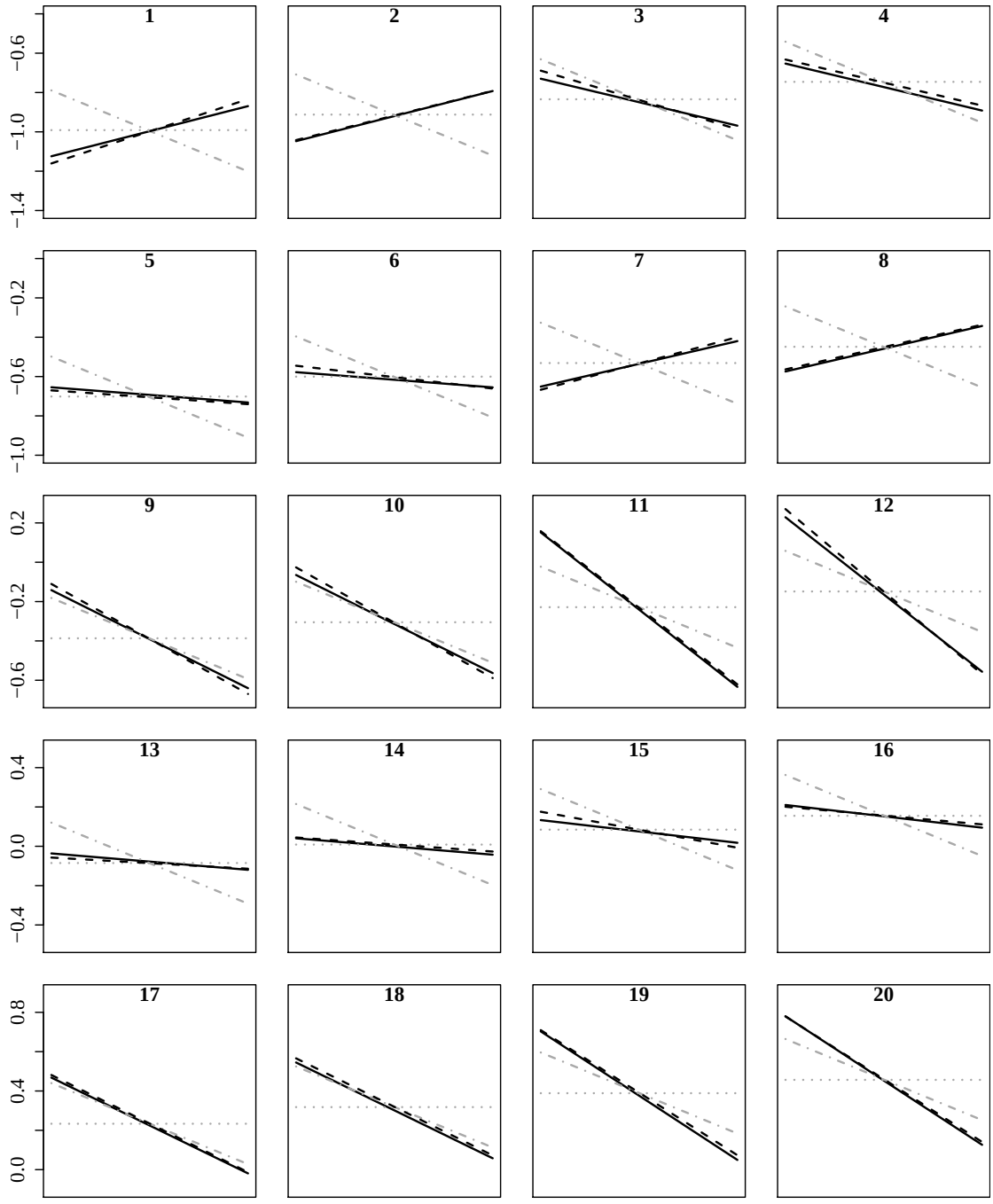

Figure 31: S1-T2-P3 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed), items 1 to 20.

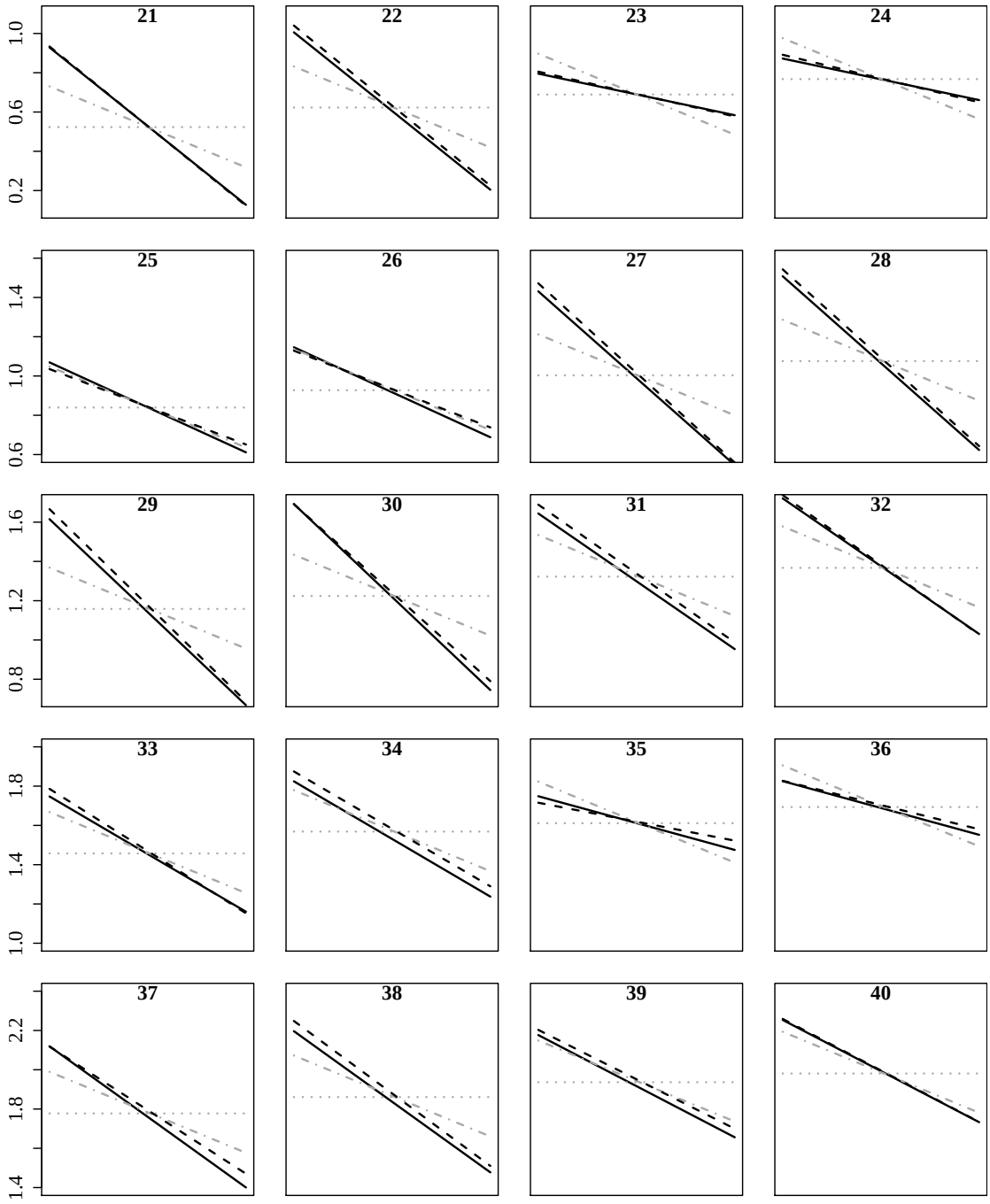

Figure 32: S1-T2-P3 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed), items 21 to 40.

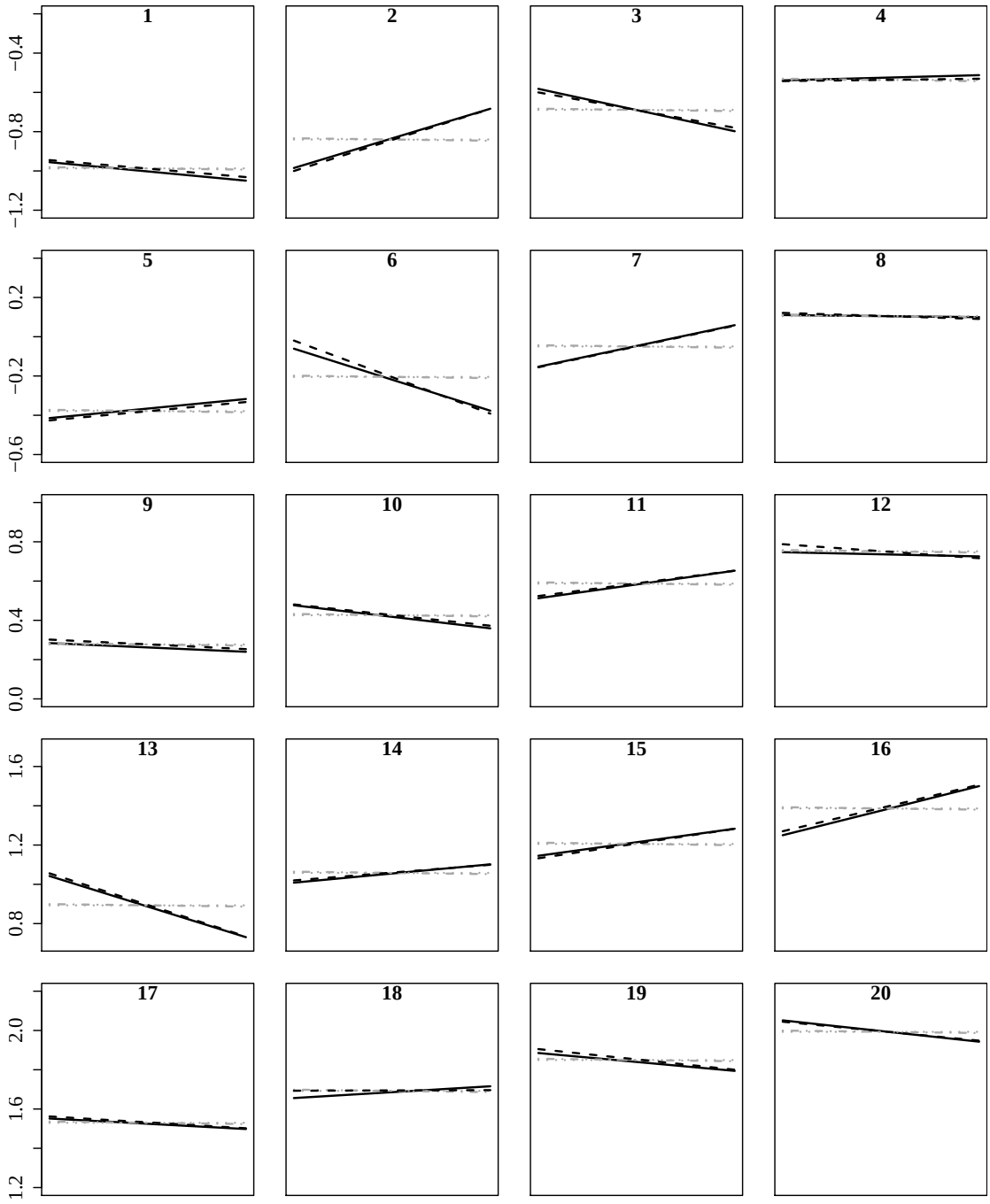

Figure 33: S2-T1-P1 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed) across position.

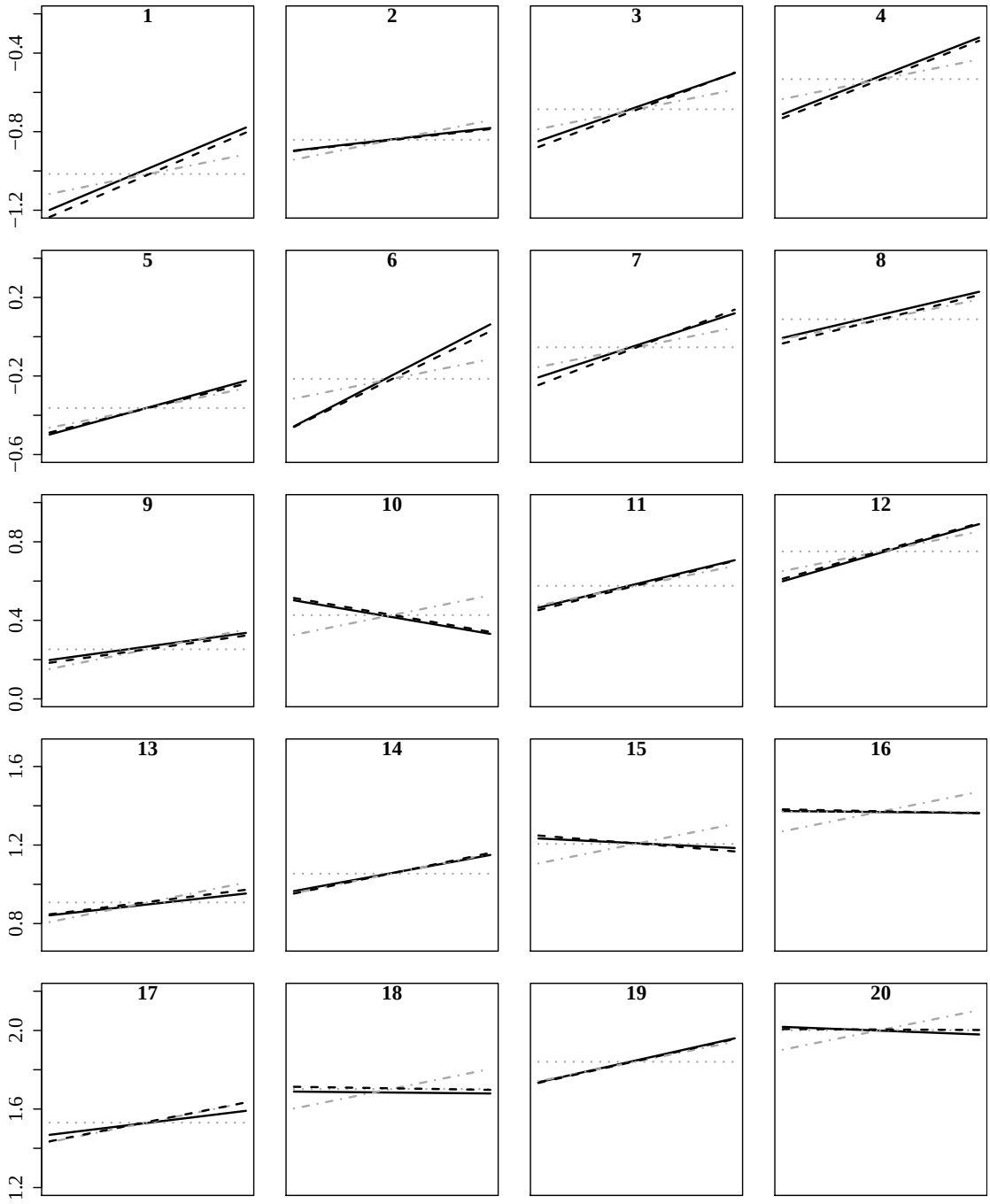

Figure 34: S2-T1-P2 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed) across position.

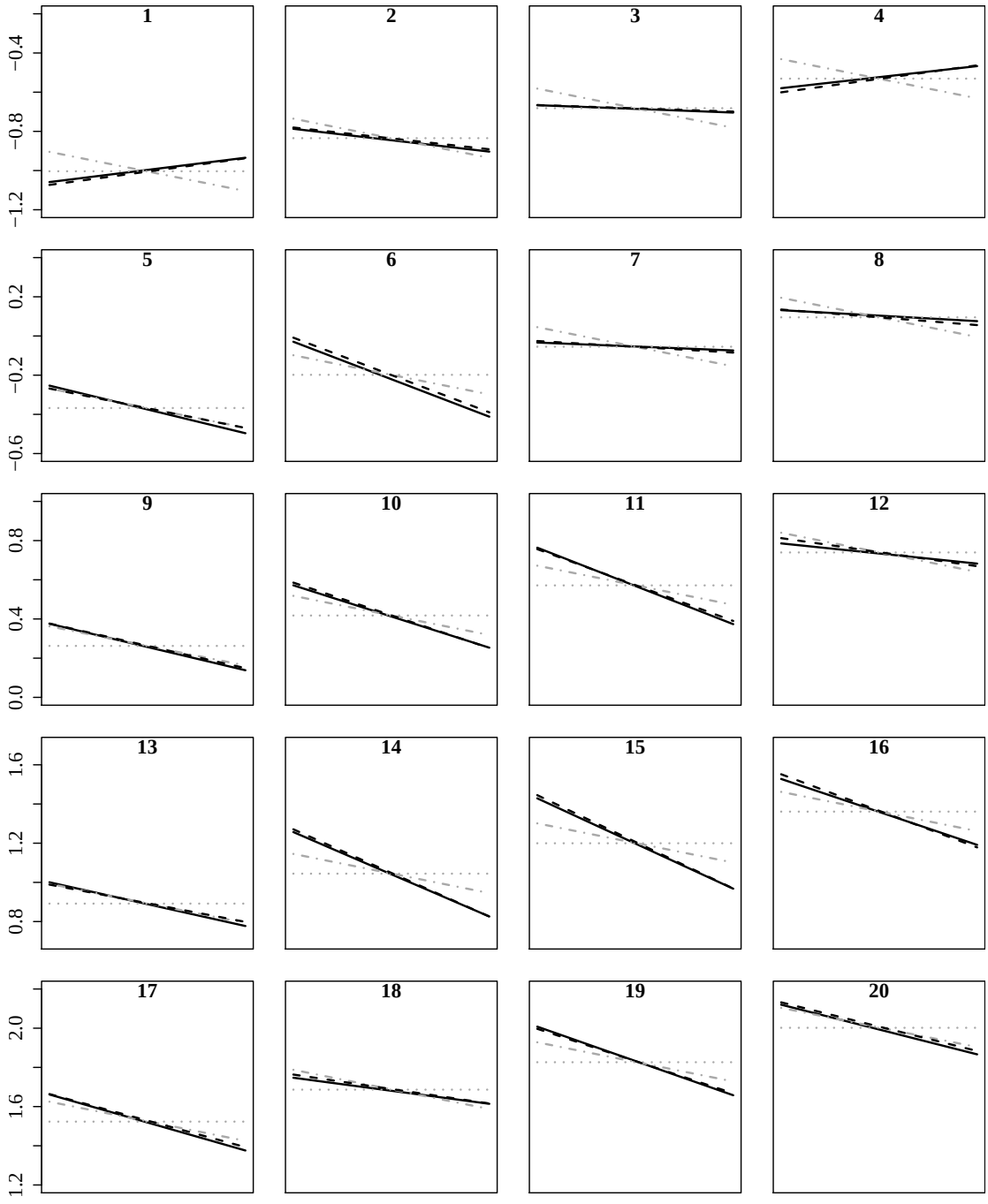

Figure 35: S2-T1-P3 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed) across position.

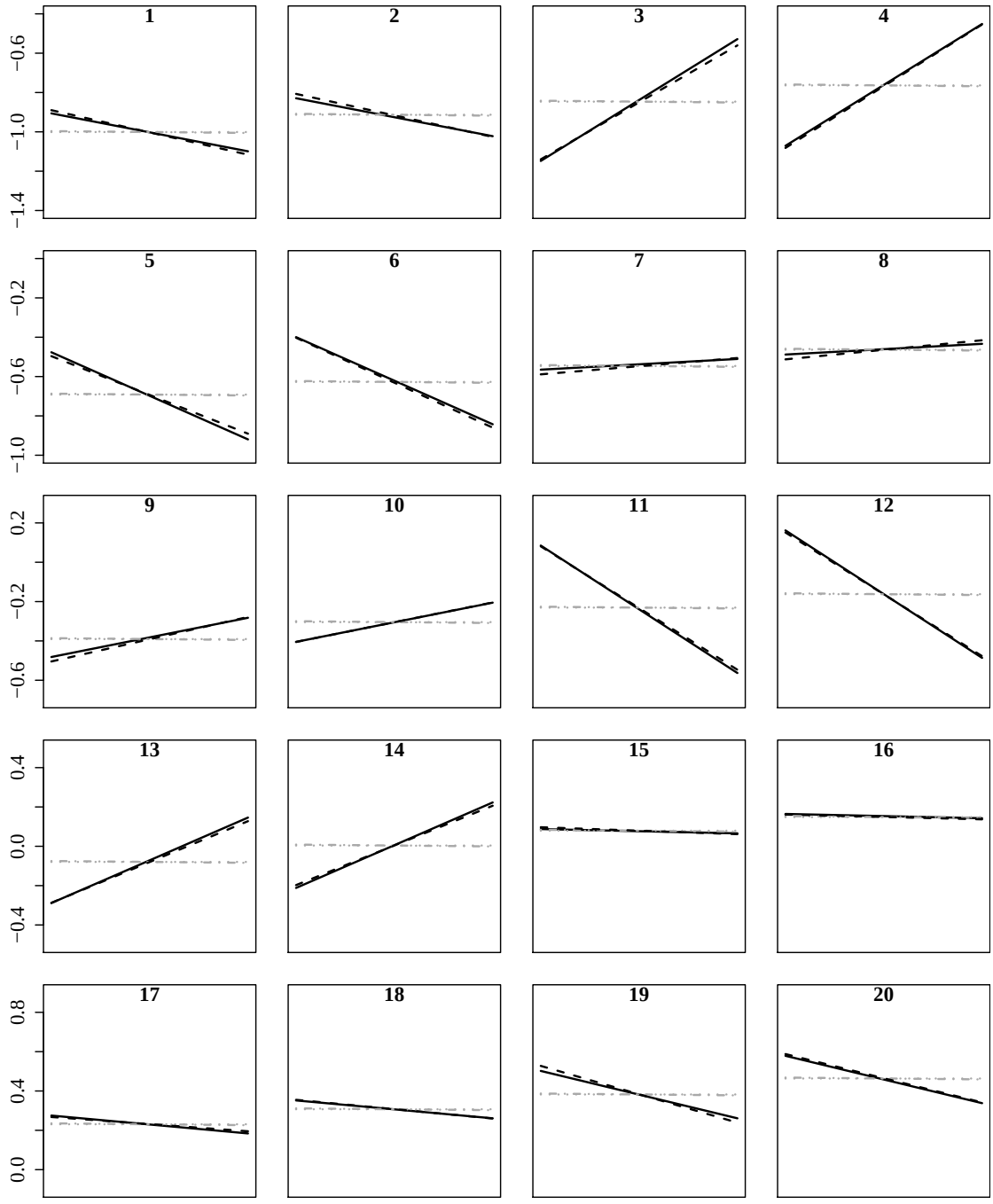

Figure 36: S2-T2-P1 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed), items 1 to 20.

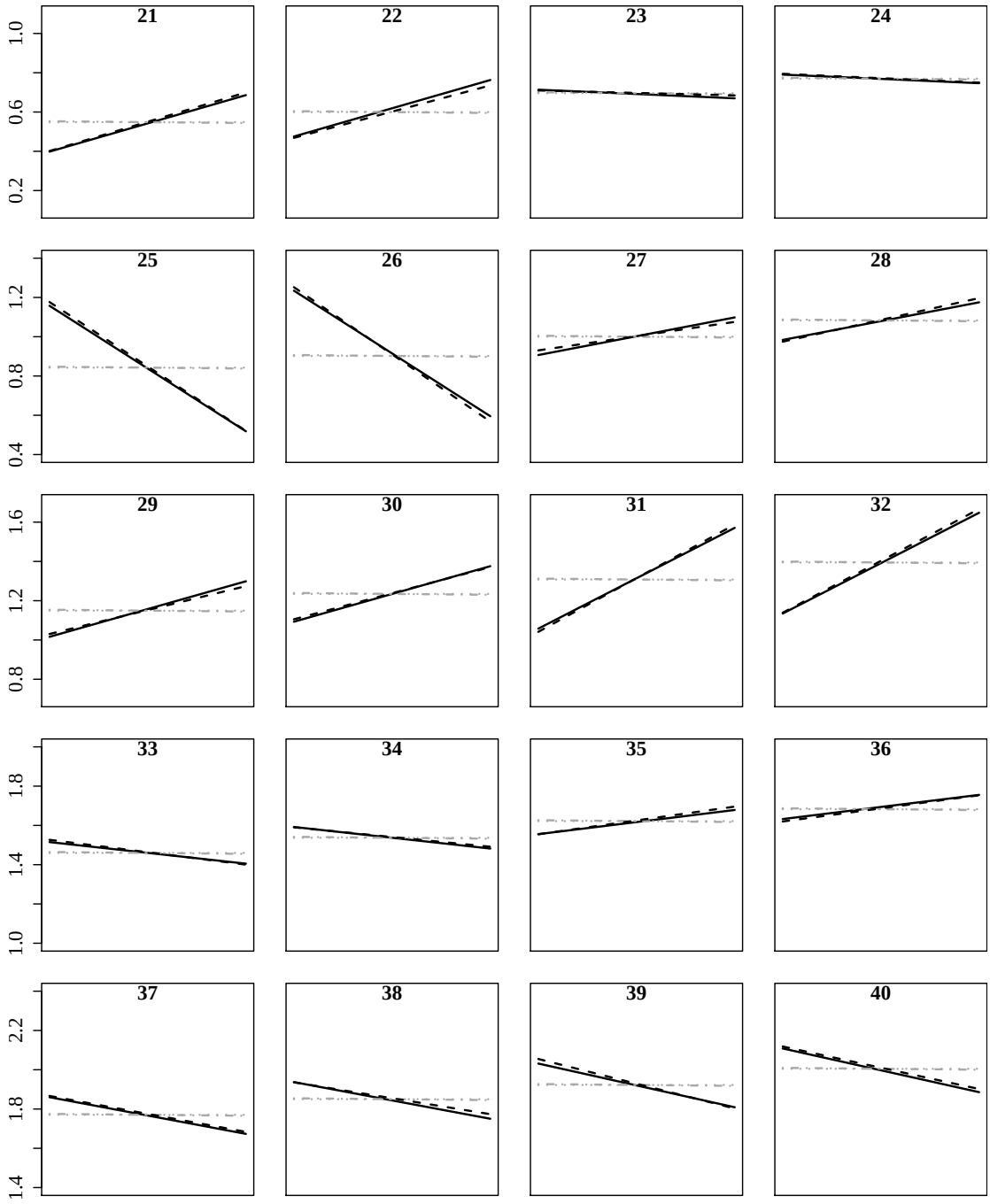

Figure 37: S2-T2-P1 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed), items 21 to 40.

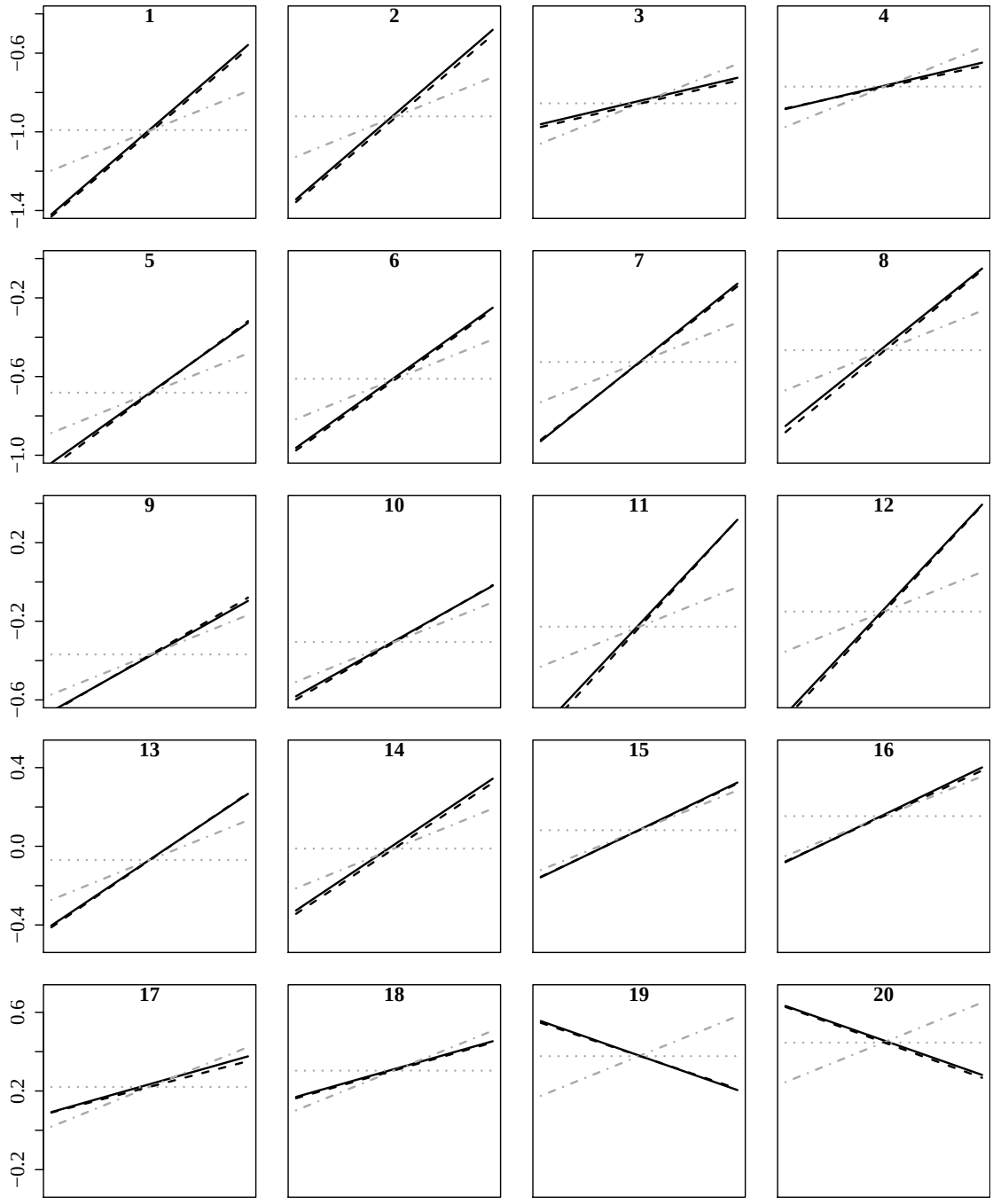

Figure 38: S2-T2-P2 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed), items 1 to 20.

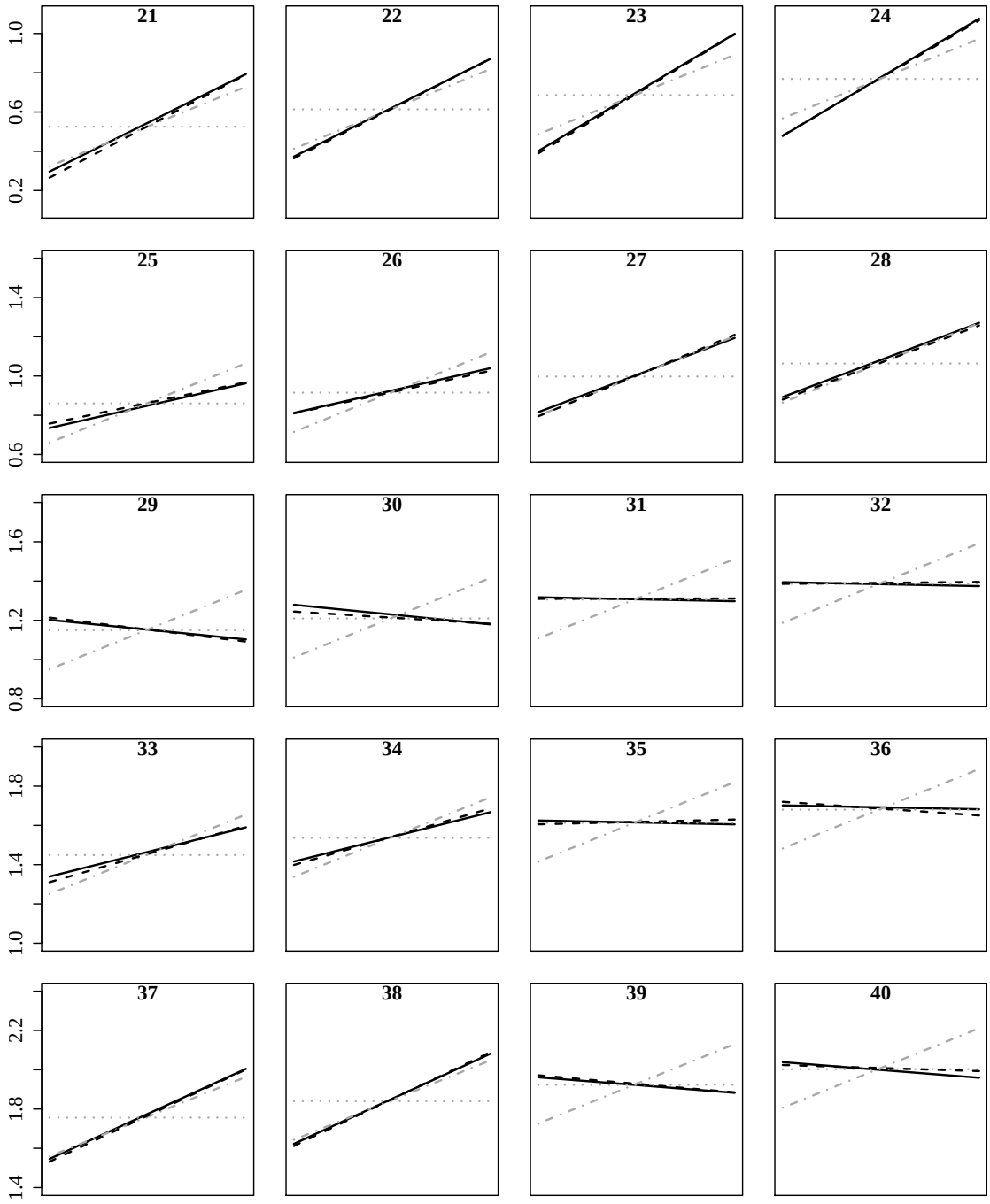

Figure 39: S2-T2-P2 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed), items 21 to 40.

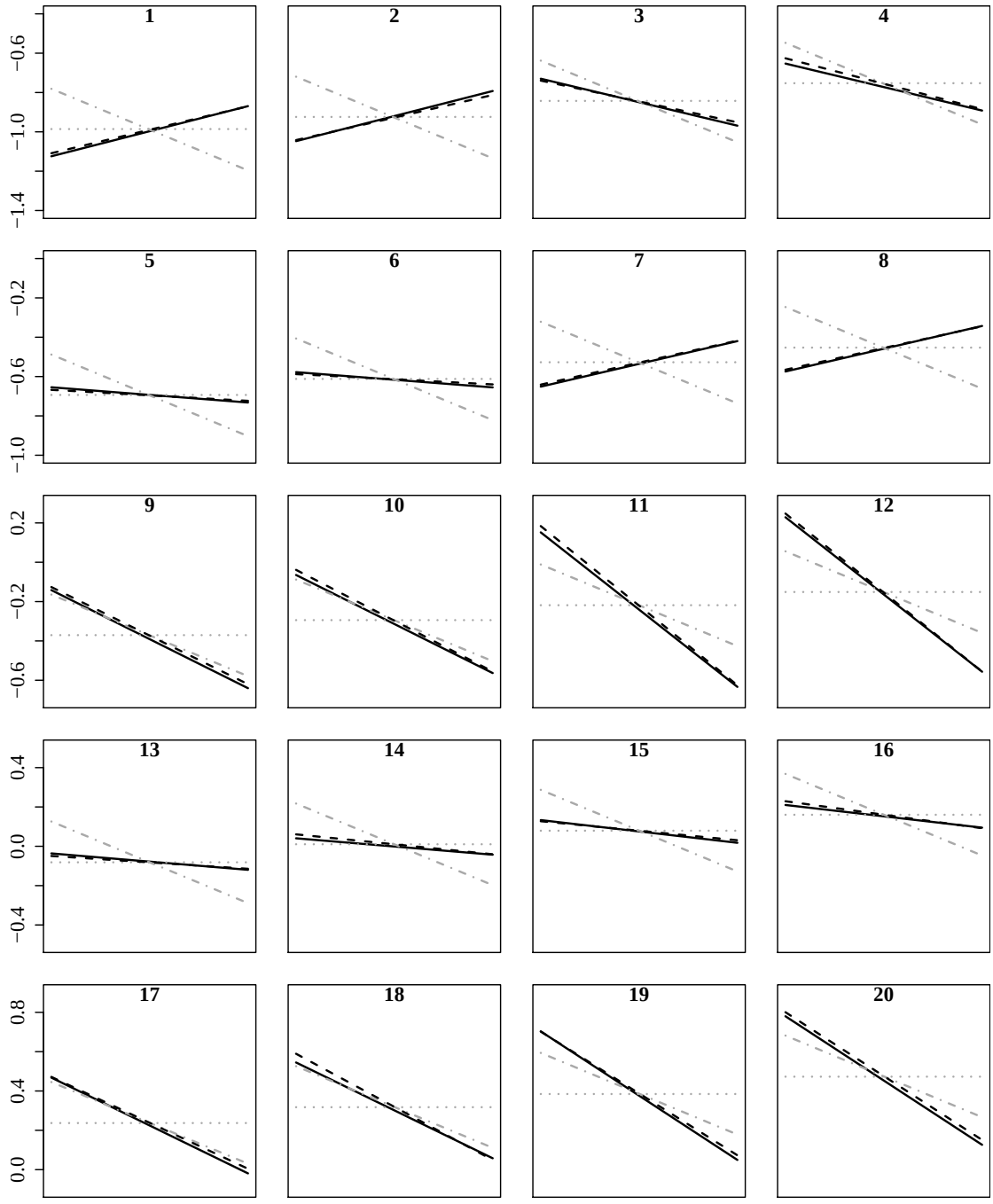

Figure 40: S2-T2-P3 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed), items 1 to 20.

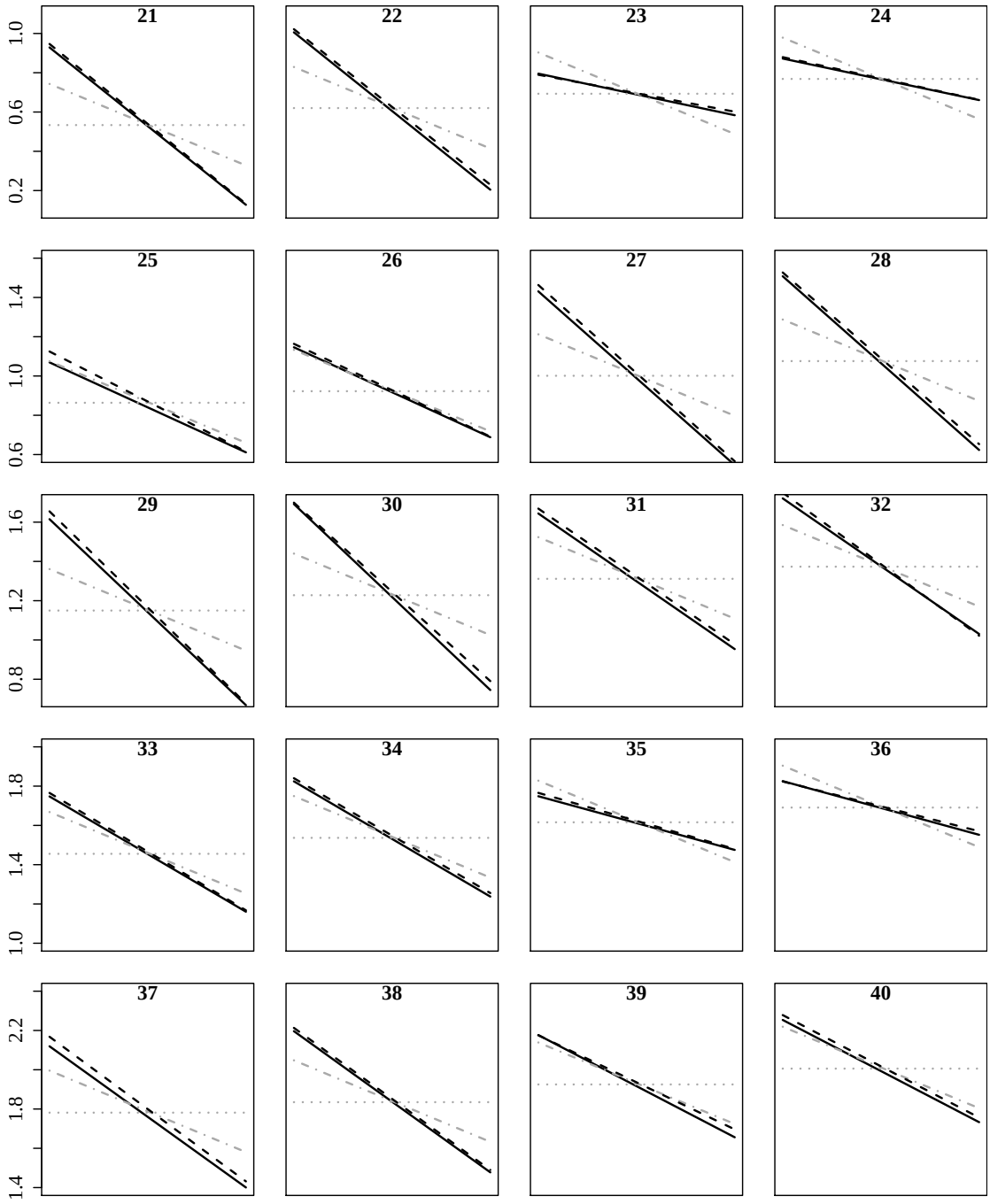

Figure 41: S2-T2-P3 true (solid black) and estimated item difficulty for M0 (grey dotted), M1 (grey dotted/dashed) and M2 (black dashed), items 21 to 40.