Detection of Speededness in Constructed Response Items Using Mixture IRT Models

by

Meereem Kim

(Under the Direction of Allan S. Cohen)

Abstract

Speededness effects tend to occur when tests have time limits (Lu & Sireci, 2007). Speededness is normally dealt with in psychometric models as a "nuisance" factor because it is a factor which is not the intended focus of the test. When speededness occurs, therefore, its effects intrude on the construct being measured and can seriously degrade the validity of the test results. A number of different approaches have been used to try to detect which examinees exhibit speededness effects. Speededness in constructed response (CR) items, however, has only recently been studied (Kim et al., 2016), although CR items are becoming increasingly prominent in standardized assessments as a means of getting students to produce a response rather than select a choice (Scalise, 2014). In this dissertation, we investigate test speededness in the context of CR items.

The first study examined a statistical model for detection of speededness effects in CR items using a two-class mixture graded response model (GRM; Samejima, 1969) for testlets. Traditional IRT models, unfortunately, cannot detect speededness, as the effects of speededness violate such models. In this first study, therefore, we considered an alternative model for estimating person and item parameters, when speededness effects are present. This approach

uses a mixture IRT model (Rost, 1990) and operates, in part, to classify examinees into one of two latent groups, a speeded group and a nonspeeded group.

In the second study, the model in the first study was extended to consider model parameters for both person and item as random effects. In particular, we investigated the performance of a random item mixture GRM for testlets with item covariates. The random item model considers both persons and items to be randomly sampled from a population (De Boeck, 2008). Treating items as random enables inclusion of item covariates directly in the model, which allows simultaneous detection of speededness effects and examination of the relationship between speededness effects in CR items and the item covariates.

In the third study, we described another possible way to characterize a latent group membership from a mixture IRT model. In general, a mixture IRT model does not readily provide a qualitative explanation of the latent dimension(s). In this dissertation, we investigated a statistical method for detecting latent themes or topics in the actual text that examinees used in giving their answers to CR items. This method is latent Dirichlet allocation (LDA; Blei, Ng, & Jordan, 2003), which is used to detect latent topics in text corpora. We investigated the use of LDA for usefulness in providing information about the qualitative differences in textual responses from the speeded and nonspeeded examinees.

INDEX WORDS: Speededness, mixture item response theory, graded response model, testlet effect, latent Dirichlet allocation, constructed response items

Detection of Speededness in Constructed Response Items Using Mixture IRT Models

by

Meereem Kim

B.A., Seoul National University, Korea, 2008

M.A., Seoul National University, Korea, 2010

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2017

Detection of Speededness in Constructed Response Items Using Mixture IRT Models

by

Meereem Kim

Major Professor: Allan S. Cohen

Committee: Seock-Ho Kim
Gary J. Lautenschlager
Zhenqiu (Laura) Lu

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2017

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speededness occurs when a test is administered within a limited time (Bejar, 1985). Under a dichotomy of power or speed tests, speededness is a construct of interest, when a test is designed to measure an examinee's speed of answering questions. Speededness is normally a "nuisance" factor, however, when time limits are implemented for test administration (Lu & Sireci, 2007). Most standardized assessments, for example, enforce time restrictions to ensure that all test takers are given the same amount of time so that a test is fair for every examinee (Powers & Fowles, 1996). In other words, such assessments may include some element of speededness even though standardized assessments are most often categorized as power tests.

Item response theory (IRT) models have been widely applied for measuring latent characteristics of items as well as individuals. The advantages of IRT over classical test theory are as follows: (a) person parameter estimates are assumed to be invariant with respect to the set of items that fits the IRT scale, (b) the precision of the estimates is available at the individual level, (c) item parameter estimates obtained from different samples can be in a common scale up to a linear transformation, and (d) it is possible to predict examinees' performance as well as the plausibility of the model (de Ayala, 2013). It should be noted that these advantages can be achieved when the underlying assumptions, such as unidimensionality and local independence, are satisfied.

These important assumptions for IRT models, however, may be violated under time restrictions. For given time limits, some examinees may have insufficient time to finish the test, which may make them resort to guessing at or skipping some items (Bejar, 1985). This

can result in responses to items (usually at the end of a test) being governed by an additional construct, for example, test speededness (Lord & Novick, 1968), besides the target latent ability the test was intended to measure. As a result, the assumption of unidimensionality may be violated due to the intrusion of this additional construct. This, in turn, violates the local independence assumption of IRT, because an examinee's responses to end-of-test items are dependent both on speededness effects and the target latent ability. In this case, item and person parameter estimates would be biased if the data are analyzed using traditional IRT models (Oshima, 1994).

The effects of speededness have been studied largely in the context of multiple-choice items (e.g., Bolt, Cohen, & Wollack, 2002; Goegebeur, De Boeck, Wollack, & Cohen, 2008; Yamamoto & Everson, 1997) as this item type tends to predominate on standardized tests. Speededness in constructed response (CR) items, however, has only recently been studied (e.g., Kim et al., 2016). CR items are increasingly being used in standardized assessments as a means of getting students to produce a response rather than select a choice (Scalise, 2014). The intent is to tap higher order skills more directly than may be possible with selection items. In this dissertation, we investigated test speededness in the context of CR items.

One concern with the use of CR items is that they tend to take longer to respond to than multiple-choice or other selected response items and are often graded with a single score. One way to increase the amount of information provided by CR items is to score the responses for multiple kinds of information (Ercikan, 2002). In this way, the CR item can be considered as a testlet, since the individual scores are all based on the same stimulus. This requires an implementation of testlet IRT models to analyze the data from CR items using multiple rubrics.

The remainder of this dissertation is organized as follows. The second chapter pertains to the detection of speededness effects in CR items using a two-class mixture graded response model (GRM; Samejima, 1969) for testlets (MixGRM-t). Generally, the item parameter estimates in IRT are considered invariant for all examinees, since they are assumed to belong to

the same population. This assumption, however, can be violated if more than one population exists in the sample of examinees (Rost, 1990). As will be discussed in the sequel, mixture IRT (MixIRT; Mislevy & Verhelst, 1990; Rost, 1990) models, including the MixGRM-t, allow the classification of examinees into several latent groups based on their responses to the items on the test. In MixIRT models, different item parameters are estimated between the latent classes, and examinees in the same class can have different ability estimates. Thus, MixIRT models classify examinees into latent classes formed along a categorical latent variable and have an ability estimate along a continuous latent variable.

This feature of MixIRT models has been shown to be useful in detecting speededness effects on latent groups of examinees in the sample taking the test. To be specific, MixIRT models can detect a latent group of students whose responses to items were affected by speededness effects. This effect has typically been assessed on items at the end of a test. Bolt et al. (2002) presented an example using a two-class mixture Rasch model to detect speededness effects. Bolt et al. assumed that students answered questions sequentially and that some of the students were influenced by speededness on items close to and at the end of the test. Results suggested this model detected two different latent groups of students based on their responses to end-of-test items. One group, the nonspeeded group, was not affected by speededness effects and one group, the speeded group, was affected by speededness effects. In this study, using the same assumptions as in Bolt et al. (2002), a two-class MixGRM-t is developed and investigated for use in detecting speededness effects in polytomously scored CR items. Further, since item responses were scored for multiple characteristics, a testlet structure is considered.

In the third chapter, we proposed a random item mixture (RIM; De Boeck, 2008) GRM for testlets (RIMGRM-t) with item covariates that helps to examine which item characteristic affects speededness in CR items. IRT algorithms for estimation of model parameters typically treat items as fixed and persons as random. This approach is also common in MixIRT models. A more theoretically appealing approach, however, is a model in which items are treated as

random as it is generally assumed that items are also randomly sampled from a population (De Boeck, 2008). The random item model considers both persons and items to be randomly sampled from a population.

Treating items as random in the MixIRT model also enables inclusion of item covariates. Inclusion of a covariate in the model allows simultaneous detection of speededness effects and examination of the relationship between item parameter estimates and the covariate or covariates. In this study, as noted above, we consider a testlet version of the RIMGRM (i.e., RIMGRM-t). A simulation study is conducted to investigate the performance of the RIMGRM-t with and without a covariate. To illustrate the application of the RIMGRM-t, data from the same source as that used in the first study are analyzed. Similar to the simulation study, results of the RIMGRM-t with and without a covariate are compared.

The characteristics of examinees in each latent class of a MixIRT model are typically determined based on differences in item performance (e.g., Cho, Cohen, Kim, & Bottge, 2010). In the fourth chapter, we describe another possible way to characterize latent class membership. Although a MixIRT model assigns class memberships based on an examinees' response patterns, the model does not readily provide a qualitative explanation about the latent dimension along which some examinees are classified into a given latent class. One way to identify the characteristic of each latent class is to determine the association between class membership and manifest information such as gender or ethnicity (e.g., Bolt et al., 2002). This approach requires additional information about an examinee, even though it is convenient to use. Another way to explain differences between latent groups is to interview examinees after a test (e.g., Izsák, Orrill, Cohen, & Brown, 2010). Whereas an interview with examinees is an abundant source of information, only a handful of examinees can be subjects due to the limitation of time and resources.

In this dissertation, we investigated a third method, latent Dirichlet allocation (LDA; Blei et al., 2003) which is a generative probabilistic model used to detect latent profiles in text corpora. Using LDA, we analyze the text of students' responses to the CR items. The

advantages of the LDA analysis are as follows. First, this method does not require additional demographic information. Second, it includes all of the examinees in the analysis. Finally, it uses examinees' written responses to the CR items, thereby helping to explain cognitive differences between latent classes that are reflected in these responses. An actual data set from a middle and high grades test of science inquiry skills was analyzed. In this study, a two-class MixGRM-t for testlets is used to classify examinees into speeded and nonspeeded groups and then topics, which are obtained from the LDA analysis, of each group are compared.

Chapter 2

Mixture Testlet Graded Response Model for Test Speededness

Tests that are administered under time constraints will often produce speededness effects (Evans & Reilly, 1972). For examinees who do not have sufficient time to complete a test, speededness effects will be manifested as an unwanted component to the construct being measured (Lord & Novick, 1968). This may cause poor estimation of ability for speeded examinees and poor estimation of item parameters, particularly for those items located at the end of the test (Douglas, Kim, Habing, & Gao, 1998; Oshima, 1994). Items at the end of speeded tests often appear harder than they would be on a nonspeeded test, because, as examinees run out of time, they often tend to hurry through or even completely omit the items at the end of the test (Bejar, 1985; Bolt et al., 2002; Oshima, 1994).

Recent evidence has suggested the usefulness of mixture item response theory (IRT) models in accounting for speededness effects on item parameter estimates (Bolt et al., 2002; Yamamoto, 1989; Yamamoto & Everson, 1997). Bolt et al. (2002) extended the mixture Rasch model by Rost (1990) to classify examinees into latent speeded or nonspeeded groups, based upon the difference in performance on items at the beginning and end of speeded tests. Parameter estimates for end-of-test items based only on the nonspeeded group were very similar to estimates for those same items when they were administered in nonspeeded locations on a different form of the test. Wollack, Cohen, and Wells (2003) applied the Bolt, et al. model to eleven years' worth of data on a college-level English Placement Test that reserved the item locations at the end of the test to pilot new items. Items that performed well became candidates to use on a future form of the test. Wollack et al. demonstrated that calibrating the item pool using only examinees in the nonspeeded class produced a more

stable and more unidimensional score scale than was produced by including all examinees in the calibration.

Wollack et al. scored all items with a dichotomous Rasch model, even though the reading comprehension section consisted of several reading passages with 6 to 8 associated questions. Items of this type are better treated as testlets to control for the local dependencies that exist among items for a common passage (Thissen, Steinberg, & Mooney, 1989; Wainer & Kiely, 1987; Wainer & Lewis, 1990; Yen, 1993). Cho, Cohen, and Kim (2014) presented an example using a mixture IRT testlet model to detect speededness effects in a reading comprehension test. The results from Cho et al. suggest that the mixture testlet model provided a markedly different solution than that provided by the regular dichotomous mixture IRT model.

Constructed response (CR) items are increasingly used in standardized assessments as a means of getting students to produce a response rather than select a choice (Scalise, 2014). CR items are often graded with a single score. When CR items are scored for multiple kinds of information, each item can be considered as a testlet, since the individual scores are all based on the same stimulus. The effects of speededness on CR items have not yet been studied. In this study, we investigate test speededness in the context of CR items that are part of testlets.

Since CR items can be scored in multiple ordered categories, in this paper, we use the graded response model (GRM; Samejima, 1969) version of the two-class mixture IRT model (Bolt et al., 2002) for detection of speededness effects in CR tests with testlet item structures. To illustrate the use of the graded response testlet model for detecting speededness, we present an empirical example using data from a middle grades assessment of science inquiry skills.

2.1 IRT Testlet Models for Speeded Test Data

2.1.1 Testlet Model

A testlet refers to a group of items based on a common stimulus (Wainer & Kiely, 1987). As it is possible that the items in a testlet are not locally independent, the testlet structure can result in biased parameter estimates (Wang & Wilson, 2005). In a testlet model, this dependency between responses is accounted for by adding a random effect to the IRT model. Bradlow, Wainer, and Wang (1999) modified a two-parameter logistic (2PL) model by adding a person-specific testlet effect. The 2PL testlet model, 2PL-t, can be expressed as

$$P_{ij} = P(u_{ij} = 1|\theta_j, \alpha_i, \beta_i, \gamma_{jt(i)}) = \frac{\exp[\alpha_i(\theta_j - \beta_i - \gamma_{jt(i)})]}{1 + \exp[\alpha_i(\theta_j - \beta_i - \gamma_{jt(i)})]},$$

where u_{ij} is scored as a 1 or 0 response of examinee j to item i, θ_j is the ability parameter of examinee j, α_i is the discrimination parameter of item i, β_i is the difficulty parameter of item i, and $\gamma_{jt(i)}$ is a person-specific testlet effect representing the interaction of examinee j with testlet $t(i)$ (i.e., testlet t includes item i).

The variance of the testlet effect parameter, $\gamma_{jt(i)}$, is interpreted as the amount of the testlet effect. The size of variance indicates the amount of local independence. If the variance of $\gamma_{jt(i)}$ is zero, it indicates that there is no dependency between items in one testlet, and the model becomes the standard 2PL model.

2.1.2 Two-Class Mixture Testlet Model for Speeded Data

In mixture IRT models, it is assumed that a population of examinees consists of discrete latent classes (Rost, 1990). In the two-class speededness model by Bolt et al. (2002), it is assumed that a population consists of two latent groups: a speeded group, which is affected by the time limits, and a nonspeeded group, which is not.

In the two-class mixture model, the separation of latent classes can be achieved using ordinal constraints on item difficulty parameters. At the beginning of the test, both speeded and nonspeeded groups are assumed to have the same item difficulty parameters. At the

end of the test, the items are assumed to be harder for the speeded group. For these items, ordinal constraints are imposed such that the item difficulties of the speeded group are larger than those of the nonspeeded group.

As CR items are often scored polytomously based on a rubric with ordered categories, it is appropriate to consider use of a polytomous IRT model for calibration. The graded response model (GRM; Samejima, 1969) is an IRT model for ordered polytomous data. It can be seen as a generalization of the 2PL model. This model defines the probability of getting a score of k as the difference between two cumulative probabilities of adjacent categories k and $k-1$. The boundary characteristic curve P^*_{ijk} represents the cumulative probability of responding with category score k or higher. The boundary characteristic curve for examinee j of obtaining a score of k or higher on item i can be described as

$$P^*_{ijk} = P(u_{ij} = k|\theta_j, \alpha_i, \beta_{ik}) = \frac{\exp[\alpha_i(\theta_j - \beta_{ik})]}{1 + \exp[\alpha_i(\theta_j - \beta_{ik})]],}$$

where θ_j is the ability of examinee j, α_i is the item discrimination parameter of item i, and β_{ik} is the category boundary location parameter for category k of item i. Then, the probability of getting a category score k can be expressed as

$$P_{ijk} = P^*_{ij,k-1} - P^*_{ijk},$$

where $k = 1, \ldots, m_i + 1$ when m_i is the maximum category score of item i. By definition, $P^*_{ij1} = 1$ and $P^*_{ij,m_i+1} = 0$.

The probability of examinee j in latent group g obtaining category k or higher under the two-class mixture GRM is expressed as

$$P^*_{ijgk} = P(u_{ijg} = k|\theta_{jg}, \alpha_{ig}, \beta_{igk}, g) = \frac{\exp[\alpha_{ig}(\theta_{jg} - \beta_{igk})]}{1 + \exp[\alpha_{ig}(\theta_{jg} - \beta_{igk})]},$$

where g is an index for the latent group and $g = 1, 2$.

Finally, the two-class mixture testlet GRM considers an additional random effect which comes from the testlet structure. The probability for examinee j in latent group g of obtaining

category k or higher under the two-class mixture testlet GRM is given by

$$P^*_{ijgk} = P(u_{ijg} = k|\theta_{jg}, \alpha_{ig}, \beta_{igk}, \gamma_{jt(i)}, g) = \frac{\exp[\alpha_{ig}(\theta_{jg} - \beta_{igk} - \gamma_{jt(i)})]}{1 + \exp[\alpha_{ig}(\theta_{jg} - \beta_{igk} - \gamma_{jt(i)})]}.$$

2.2 Example: Speededness in a Science Inquiry Practice Test

An example is presented here using data from a middle grades test of science inquiry skills to illustrate the effects of speededness in a constructed response test. Science education is currently undergoing its first major reform in two decades, with a new emphasis on a model of three-dimensional learning meant to support a college and career ready STEM workforce (National Research Council, 2013). In this model, science learning is viewed as a combination of science practices, core conceptual ideas, and crosscutting concepts (National Research Council, 2012). For students to succeed in this new vision of science education, they must learn to communicate their thinking about complex phenomena both orally and in writing (Lee, Quinn, & Valdés, 2013). Writing about actual science investigations or about science investigation scenarios can accomplish one part of this purpose (Gunel, Hand, & Prain, 2007). To this end, the next generation of science assessments, requiring substantial amounts of student writing on constructed response assessments, is currently being developed in an effort to measure student learning of these complex science practices (Scalise, 2014).

2.2.1 Methods

Data Sources

The sample consisted of 1,612 middle and high school students' responses to a test of science inquiry practices. The test was one of two pre-test forms used as part of a larger host study in which students were provided with instruction on science inquiry practices. Science inquiry practices in this context consisted of hypothesis testing, determining cause and effect, and explaining standard physical phenomena (e.g., the impact of heat on different colors of

clothing, the effect of adding yeast to flour, salt, and water). Students were given a single 50 minute class period to complete the test.

There were seven CR items on the test. Each of the items was scored from two to four ordered categories over two to six different characteristics. These characteristics were treated as items within a testlet. The seven CR items were treated as seven testlets with 7, 5, 5, 6, 6, 5, and 6 items, respectively. The first three testlets measured 'Cause and Effect Relationships,' the next two testlets measured 'Controlling Variables,' and the last two testlets measured 'Hypothesis, Observation, and Evidence.'

Model

In this study, two models were used: a testlet GRM (GRM-t) and a two-class mixture testlet GRM (MixGRM-t). The first model was used to estimate item parameters of the first 17 items (those for Testlet 1 to Testlet 3). It was assumed that the speededness effect was not present in these items. The second model was used to account for speededness effects using ordinal constraints on the last 17 items (those for Testlet 5 to Testlet 7). The items in Testlet 4 were not constrained and were estimated.

The mean ability of the speeded group in the MixGRM-t was constrained to be zero to resolve the identification problem so that item boundary location parameter estimates could be compared across latent groups. This solution is different from the norming condition which Bolt et al. implemented.

Modeling Speededness

For the two-class MixGRM-t, it was assumed that the end-of-test items included effects of speededness. This assumption was reflected in the model such that the item boundary location parameter estimates and item discriminations for the first 17 items (those for Testlet 1 to Testlet 3) were constrained to be equal for both the speeded and nonspeeded group. The item boundary location parameters of the last 17 items (those for Testlet 5 to Testlet

7) were constrained to be higher for the speeded group than the nonspeeded group. The boundary location parameter estimates for the items in Testlet 4 were unconstrained. The item discrimination parameters for items in Testlets 4 to 6 were freely estimated.

Estimation

A Markov chain Monte Carlo (MCMC) estimation algorithm employing Gibbs sampling as implemented in the OpenBUGS software (Thomas, O'Hara, Ligges & Sturtz, 2006) was used to estimate the MixGRM-t. MCMC algorithms have been used for estimation of mixture distributions (Robert, 1996), including for mixture IRT models (Bolt et al., 2001; Cho & Cohen, 2010, Cho et al., 2014).

The MCMC algorithm first samples a Markov chain in which values for parameters in the model are sampled from their full conditional posterior distributions over a typically large number of iterations. The algorithm begins by sampling a class membership for each examinee at each stage of the chain and then sampling values for class parameters conditional on those class memberships (Bolt et al., 2002).

Mixing proportions and class ability parameters were based on the frequencies with which examinees were sampled into each latent class. The frequencies with which an examinee was sampled in each latent class over the course of the Markov chain defined the posterior probability of latent class membership in that class. Mixing proportions (π_g) and class mean ability parameters (μ_g) were estimated along with the class memberships. Ability was assumed to be normally distributed with a variance of 1 in each latent class.

To derive the posterior distributions for each parameter, it is first necessary to specify their prior distributions. The following priors were used to estimate the parameters of the

MixGRM-t in this study:

$$
\begin{aligned}
a_i &\sim Normal(\mu_a, \sigma_a^2)I(0,) \\
b_i &\sim Normal(\mu_b, \sigma_b^2) \\
\gamma_{gt(i)} &\sim Normal(\mu_g, \sigma^2_{\gamma t(i)}) \\
\theta_{gi} &\sim Normal(\mu_g, 1), \qquad i = 1, \ldots, N, \\
\mu_g &\sim Normal(0, 1), \qquad g = 1, 2, \\
(\pi_1, \pi_2) &\sim Dirichlet(0.5, 0.5),
\end{aligned}
$$

where N is the total number of examinees and $I(0,)$ indicates that observations of a were sampled above zero. Hyperparameters used in this analysis were selected to be noninformative: $\mu_a \sim N(0,1),\ \mu_b \sim N(0,1),\ \sigma_a^2 \sim \chi^{-2}_{\nu_a},\ \sigma_b^2 \sim \chi^{-2}_{\nu_b}, \sigma^2_{\gamma g} \sim \gamma^{-1}_{(2.5,0.25)}$.

Starting values are provided for each parameter being sampled to define the first state of the Markov chain. For the latent class mixing proportions, π_1 and π_2, starting values were set at .5. Starting values for the remaining model parameters were randomly generated using the OpenBUGS software.

In an MCMC analysis, information from initial iterations is discarded. These iterations, called *burn-in* iterations, are discarded, because initial sampled values tend to be dependent on the starting values. The subsequent iterations are based on a chain that is assumed to have converged to its stationary distribution. Estimates of sampled parameters were calculated from these post-burn-in iterations. In this study, the Heidelberger and Welch (1983) convergence diagnostic, as implemented in the CODA package using R (Plummer, Best, Cowles, & Vines, 2006), was used to determine the number of burn-in iterations. This diagnostic consists of two tests: the stationary and halfwidth tests. The null hypothesis of the first test is that the estimates for each variable have reached their stationary state. The second test, which is done only for those variables that have passed the stationary test, is used to estimate the standard error of the variable.

Table 2.1: Descriptive Statistics for the Female and Male Students

		Testlet 4		Testlet 5		Testlet 6		Testlet 7		Total	
Gender	N	M	SD	M	SD	M	SD	M	SD	M	SD
Female	852	9.24	2.98	9.76	3.42	8.64	3.09	10.19	3.23	37.83	10.54
Male	754	8.76	2.96	9.13	3.26	7.94	2.92	9.81	3.32	35.64	10.49

2.2.2 Results

Descriptive Statistics for the Test Items

Descriptive statistics of the last four testlets for female and male students are presented in Table 2.1. The means of female students were higher than those of male students for all four testlets. Independent t-tests suggested that the differences between females and males were significant at $\alpha = .05$ in all four testlets.

Monitoring Convergence

The Heidelberger and Welch (1983) convergence diagnostic suggested a burn-in length of 10,000 iterations for the GRM-t and 11,000 iterations for the MixGRM-t. Also, the result proposed a post burn-in length of 5,000 iterations for the GRM-t and 4,000 iterations for the MixGRM-t. Estimates of model parameters were based on the means of the sampled values from the post burn-in iterations.

Item Parameter Estimates

The item discrimination estimates for items in Testlet 4 to Testlet 7 for both the speeded and nonspeeded groups are given in Figure 2.1. It is interesting that all of the estimates for the speeded group were larger than those for the nonspeeded group except for one item even though there was no constraint for estimation of the item discrimination parameters.

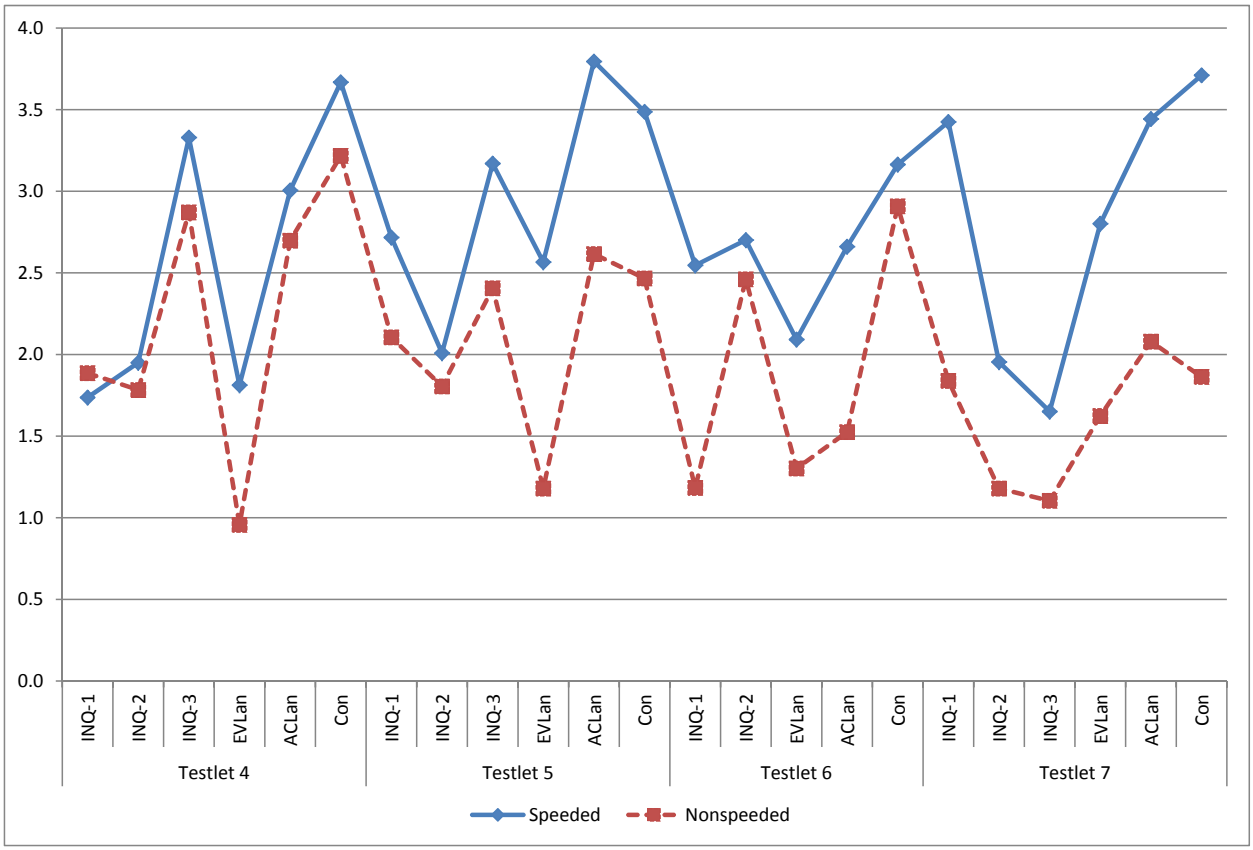

Figure 2.1: Item discrimination parameter estimates

Figure 2.2 plots the means of the boundary curve location estimates for items in Testlets 4 to Testlet 7 for both the speeded and nonspeeded groups. The average of boundary location parameter estimates for each item, as suggested by Masters (1982), was used to provide an index of the relationship between the item boundary location parameter estimates for speeded and nonspeeded groups. The item boundary location parameter estimates along with the item discrimination parameter estimates for the speeded and nonspeeded groups are given in Tables 2.2 and 2.3, respectively. In accordance with previous studies (Bolt et al., 2002; Kim et al., 2016), the distance between the speeded and nonspeeded groups appeared to be similar in Testlets 4 and 5 but increased in Testlet 6 and Testlet 7. This finding suggests that the speededness effect was largest at the end of the test.

It should be noted that the item discrimination and item boundary location parameter estimates may not be clearly separated because the estimation of those parameters is related to each other. More specifically, it is highly likely that the items which are extremely easy or difficult tend to have lower item discrimination parameters than the items which are moderately difficult.

Testlet Effects

Table 2.4 illustrates the testlet effects for both speeded and nonspeeded groups. As noted earlier, the testlet effect indicates the magnitude of the local dependence among items which belong to the same testlet. Table 2.4 suggests that the larger testlet effect was detected for the speeded group for all of the testlets. For both groups, the testlet effects were larger at first (from Testlet 1 to Testlet 3), decreased in the middle (Testlet 4 and Testlet 5), and increased again at the end of the test (Testlet 6 and Testlet 7). For the speeded group, the testlet effect of the last two testlets was much larger compared to the nonspeeded group.

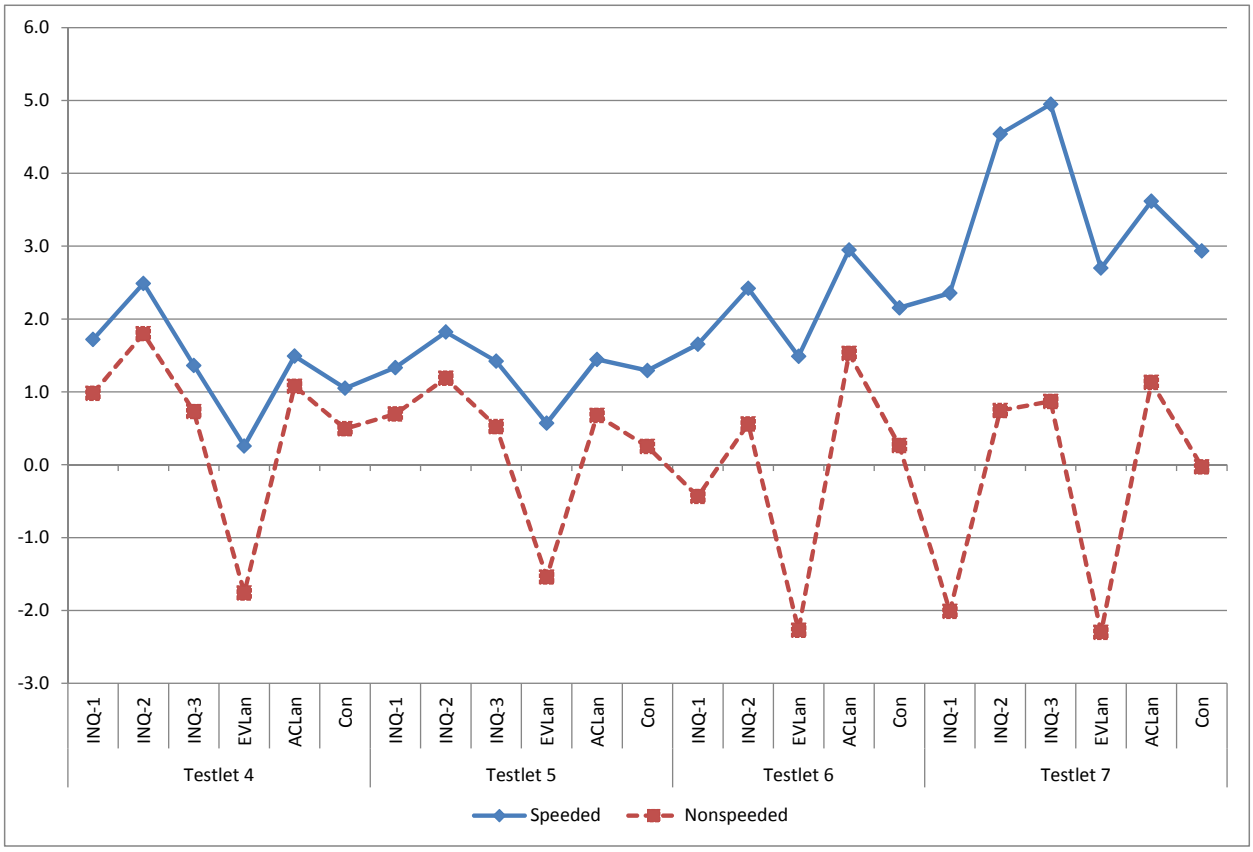

Figure 2.2: Means of boundary location parameter estimates

Table 2.2: Item Parameter Estimates for the Speeded Group

		α		β_1		β_2		β_3		β_4		
Testlet	Item	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD	$\bar{\beta}$
4	18	1.74	0.26	1.72	0.20							1.72
	19	1.95	0.37	2.49	0.30							2.49
	20	3.33	0.43	1.12	0.11	1.60	0.14					1.36
	21	1.81	0.18	-0.58	0.08	1.10	0.12					0.26
	22	3.01	0.33	0.59	0.09	1.22	0.12	1.82	0.16	2.34	0.20	1.49
	23	3.67	0.44	0.66	0.09	1.44	0.12					1.05
5	24	2.72	0.38	1.33	0.13							1.33
	25	2.01	0.31	1.82	0.20							1.82
	26	3.17	0.41	1.17	0.11	1.68	0.15					1.42
	27	2.57	0.28	-0.19	0.07	1.33	0.13					0.57
	28	3.80	0.42	0.71	0.09	1.17	0.11	1.63	0.14	2.28	0.19	1.45
	29	3.49	0.41	0.86	0.10	1.73	0.14					1.29
6	30	2.55	0.40	1.65	0.17							1.65
	31	2.70	0.46	2.37	0.17	2.47	0.17					2.42
	32	2.09	0.27	0.71	0.13	2.27	0.21					1.49
	33	2.66	0.38	1.99	0.18	2.46	0.20	3.36	0.28	3.98	0.36	2.95
	34	3.16	0.46	1.82	0.17	2.49	0.21					2.15
7	35	3.43	0.53	2.35	0.18							2.35
	36	1.95	0.39	4.19	0.30	4.90	0.40					4.54
	37	1.65	0.36	4.95	0.43							4.95
	38	2.80	0.40	2.03	0.16	3.36	0.23					2.70
	39	3.44	0.50	2.64	0.19	3.36	0.22	4.00	0.26	4.46	0.32	3.62
	40	3.71	0.53	2.33	0.18	3.54	0.22					2.94

Table 2.3: Item Parameter Estimates for the Nonspeeded Group

		α		β_1		β_2		β_3		β_4		
Testlet	Item	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD	$\bar{\beta}$
4	18	1.89	0.16	0.98	0.08							0.98
	19	1.78	0.18	1.80	0.13							1.80
	20	2.87	0.21	0.44	0.05	1.02	0.07					0.73
	21	0.96	0.08	-4.22	0.30	0.71	0.10					-1.76
	22	2.70	0.18	-0.02	0.04	0.74	0.06	1.44	0.08	2.14	0.12	1.08
	23	3.22	0.23	-0.03	0.04	1.02	0.07					0.49
5	24	2.11	0.18	0.70	0.06							0.70
	25	1.81	0.15	1.19	0.09							1.19
	26	2.41	0.17	0.21	0.05	0.84	0.06					0.52
	27	1.18	0.09	-3.97	0.26	0.89	0.09					-1.54
	28	2.62	0.17	-0.42	0.04	0.38	0.05	1.05	0.07	1.70	0.09	0.68
	29	2.47	0.17	-0.47	0.05	0.97	0.07					0.25
6	30	1.18	0.10	-0.44	0.07							-0.44
	31	2.46	0.20	0.36	0.05	0.76	0.06					0.56
	32	1.30	0.10	-3.96	0.24	-0.59	0.07					-2.27
	33	1.53	0.11	0.05	0.06	1.04	0.08	2.15	0.13	2.87	0.17	1.53
	34	2.91	0.24	-0.39	0.05	0.91	0.06					0.26
7	35	1.84	0.19	-2.01	0.12							-2.01
	36	1.18	0.10	0.46	0.08	1.02	0.10					0.74
	37	1.10	0.10	0.87	0.10							0.87
	38	1.62	0.14	-3.40	0.21	-1.20	0.08					-2.30
	39	2.08	0.16	-0.41	0.06	0.69	0.07	1.66	0.10	2.59	0.15	1.13
	40	1.86	0.15	-1.99	0.11	1.93	0.12					-0.03

Table 2.4: Testlet Effects

	Testlet						
Group	1	2	3	4	5	6	7
Speeded	3.38	2.30	2.79	0.30	0.29	1.55	3.52
Nonspeeded	1.98	1.15	1.11	0.26	0.36	0.61	0.73

Mean Ability and Proportion of Latent Classes

The mean ability of the speeded group was constrained as zero to resolve the identification problem and the mean ability estimate of the nonspeeded group was -1.01. This result suggests that the speeded group was more capable than the nonspeeded group. Also, this finding is in accordance with previous studies on speededness (Bolt et al., 2002; Cho et al., 2014). The estimates of the mixing proportions suggested that 30% of students were assigned to the speeded group and 70% to the nonspeeded group.

Classification of Examinees

Latent group memberships can be estimated based on the frequencies with which examinees are assigned into each latent group over the post-burn-in iterations. To understand the characteristics of latent classes, it helps to compare the class membership with manifest demographic information such as gender and ethnicity. The sample consisted of 852 female and 754 male students; six students did not report their gender. Also, the sample contained 690 Hispanic and 911 non-Hispanic students and 11 students did not report their ethnicity.

Table 2.5 shows the proportion of speeded and nonspeeded groups for gender and ethnicity, respectively. A two-way contingency table analysis was conducted to evaluate whether gender or ethnicity has a statistical association with speededness. First, gender and speededness were found to be significantly related ($\chi^2 = 8.800, df = 1, p = .003$). The proportion of female students who were classified as the speeded group (26.8%) was lower

Table 2.5: Relationship between Speededness Effects and Gender and Ethnicity

Group	N	%Speeded	%Nonspeeded	χ^2	df	p-value
Female	852	26.8	73.2			
Male	754	33.6	66.4	8.800	1	.003
Hispanic	690	30.4	69.6			
Non-Hispanic	911	29.7	70.3	0.088	1	.766

than that of male students (33.6%). Second, ethnicity and speededness were found to be insignificantly related ($\chi^2 = 0.001, df = 1, p = .980$). The proportion of Hispanic students who were classified as the speeded group (30.4%) was higher than that of non-Hispanic students (29.7%), but this difference was not significant.

2.3 Summary

Speededness effects have been shown to have an impact on the accuracy of item parameter estimates in IRT models. Previous work on speededness has focused on dichotomous models. In this study, speededness in CR items scored in multiple ordered categories was investigated. Speededness effects in testlets were investigated using the two-class mixture IRT model approach described by (Bolt et al., 2002). CR items are becoming increasingly important in educational research as researchers seek to expand the kinds of knowledge that can be measured by tests. These items are typically scored in multiple ordered categories. In the context of IRT, this type of scoring can usually be handled by using polytomous models.

In the present study, we examined the utility of one approach in the detection of speededness in the context of locally dependent item structures and also for items scored in multiple ordered categories. An example using a CR test with a testlet structure was presented. Results indicated that the gap in item boundary location parameter estimates between the

speeded and nonspeeded groups increased monotonically as the test proceeded. Also, the magnitude of the testlet effect was larger in the speeded group than in the nonspeeded group. The pattern of the testlet effect was similar in both groups such that the testlet effect was larger from Testlet 1 through Testlet 3, decreased from Testlet 4 and Testlet 5, and increased from Testlet 5 to Testlet 7.

Chapter 3

A Random Item Mixture IRT Model for a Speeded Test

Item response theory (IRT) algorithms for estimation of model parameters typically treat the items as fixed and persons as random. This is the way marginalized maximum likelihood estimation (MMLE; Bock & Aitkin, 1981) usually treats items when estimating parameters in the presence of the unobserved random latent variable(s). Under the MMLE procedure, the ability distribution is integrated over and thus removed from the likelihood function. It is then possible to estimate item parameters in the marginalized distribution. These estimates are independent from the estimation of ability parameters (Baker & Kim, 2004). It follows that we can estimate ability parameters using known item parameter estimates. In this way, both ability and item parameters are estimated.

As De Boeck (2008) has suggested, a more theoretically appealing approach is a model in which items are treated as random. Often, items are considered as a random sample from a domain but are treated as fixed when estimating item parameters. This is the case for software packages such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003), MULTILOG (Thissen, Chen, & Bock, 2003) and PARSCALE (Muraki & Bock, 2003). De Boeck suggests it may be more reasonable, however, to consider items as randomly sampled from a domain and, therefore, to treat items as random when estimating parameters. For example, assume an item bank, e.g., a pre-existing sample of calibrated items, needs to be constructed. The domain of the items, from which the bank is to be constructed, can be considered as the population from which each item in the domain is sampled.

In addition, a random item model can incorporate covariates that can help explain the variance of item difficulties. The linear logistic test model (LLTM; Fischer, 1973) has been

used to explain the components of item difficulty using a Q-matrix-like structure to account for the variability in the item parameter. The LLTM, however, has often been regarded as too stringent because it assumes that the item difficulties are explained completely by the components used in the model. Mislevy (1988) and Rijmen and De Boeck (2002) have extended the LLTM to relax its strong assumptions. Mislevy (1988) embedded the LLTM model in Bayesian estimation, thereby including the variances of item parameters. Rijmen and De Boeck (2002) proposed a random weights LLTM, which includes interactions between persons and item properties to permit different person-dependent item effects.

In addition, Frederickx, Tuerlinckx, De Boeck, and Magis (2010) has shown how differential item functioning (DIF) can be detected under a random item model. In a fixed item model, a set of anchor items is required in order to detect the DIF items. The quality of the anchor items will have an effect on the results of the DIF analysis (Kolen & Brennan, 2004). It is not necessary, however, in the context of a random item model to specify a priori a set of anchor items. Frederickx et al. (2010) suggested an alternative method for detecting DIF items using a random item mixture (RIM; De Boeck, 2008) model.

In this approach, an item is assumed to belong to one of two latent classes, a DIF class or a non-DIF class. The RIM model is then used to assign items into one of these two classes. In this approach, "the person groups are always manifest, whereas the item classes are latent" (p. 435). This is clearly different from the previous work on detection of DIF with mixture IRT models (e.g., Cohen & Bolt, 2005) which considers detecting DIF by classifying examinees into different latent groups.

There does not appear to be any research published yet on a RIM model for polytomous data. Considering that the implementation of constructed response (CR) items has increased in standardized assessment (Scalise, 2014), research on polytomous IRT models to analyze typically polytomous data from CR items would seem to be helpful. Kim et al. (2016) has recently reported speededness effects in a test composed of CR items using a mixture GRM. In this study, we extend the model by Kim et al. to include random item effects. Specifically,

we describe a RIM testlet graded response model (RIMGRM-t) and investigate its utility for the detection of speededness effects in the context of polytomously scored CR items. It should be noted that persons are assumed to belong to one of two latent class, a speeded group or a nonspeeded group, which is somewhat different from De Boeck's approach.

3.1 A Random Item Mixture Testlet Graded Response Model

The RIMGRM-t is an IRT model which incorporates a random component for item parameters (the random item part of the model), assumes that a population consists of several latent subpopulations (the mixture IRT part of the model), and can handle ordered polytomous categories nested in testlets (the graded response testlet part of the model). In this section, each component of the RIMGRM-t is introduced.

3.1.1 A Random Item Model

In the context of multilevel modeling, an IRT model treating both persons and items as random can be regarded as a special case of a cross-classified random effects model (Van den Noortgate, De Boeck, & Meulders, 2003). Typically, multilevel modeling assumes that the lower-level observation belongs to one and only one higher-level unit (Raudenbush & Bryk, 2002). Occasionally, however, the lower-level observation may belong to more than one higher-level unit. In such a case, it is necessary to include the effects of additional higher-level units on a dependent variable in the model. When this is the case, it can be parameterized as a cross-classified random effects model (Raudenbush, 1993).

Van den Noortgate et al. (2003) proposed a cross-classified multilevel logistic model in which the Rasch model is incorporated into a cross-classified random effects model. In this model, both persons and items can be regarded as higher-levels in which an examinee's response to an item is nested. It is possible, thus, for both levels to have random effects.

First, the Rasch model can be written as

$$\eta_{ij} = \log\left[\frac{P_i(\theta_j)}{1 - P_i(\theta_j)}\right] = \theta_j - \beta_i,$$

where θ_j is the ability parameter of examinee j; and β_i is the item difficulty parameter of item i. This relationship between θ_j and β_i can be reformulated as a level-1 model in the context of a cross-classified random effects model. The level-1 model can be described as

$$\eta_{ij} = \beta_{0j} + \beta_i,$$

where β_{0j} is the ability parameter of examinee j, and β_i is the item easiness parameter of item i. Note that the item parameter β_i has a different sign following the convention of multilevel models. This changes the interpretation of the parameter from item difficulty to item easiness. It should also be noted that there is no within-cell random effect because each cell has only one observation. That is, an examinee produces only one item response to an item.

The level-2 model includes random effects for person and item parameters and can be described as

$$\beta_{0j} = u_j,$$

$$\beta_i = u_i,$$

$$u_j \sim N(\mu_1, \tau_1^2)$$

$$u_i \sim N(\mu_2, \tau_2^2),$$

where u_j is the random effect of the person parameter, and u_i is the random effect of the item parameter.

The combined model under the assumption that both items and persons are treated as random (i.e., the random item Rasch model) can be expressed as

$$\eta_{ij} = \beta_0 + u_i + u_j,$$

where β_0 is the estimated logit for the probability of a correct response of an average person on an average item, u_i is the random effect of the item parameter, and u_j is the random effect of the person parameter. Van den Noortgate et al. (2003) suggested this model may not be useful because it does not provide the estimates for either ability or item difficulty parameters. However, this model "opens up the perspective of an error term for the items when the issue is to explain the item difficulties from item features" (p. 373).

Wang (2011) extended the cross-classified multilevel logistic model into a multilevel mixture IRT model (MMixIRTM) and applied it to the detection of test speededness. The MMixIRTM can be regarded as "a combination of the regular multilevel IRT model and the mixture IRT model" (p. 22). Under the multilevel mixture Rasch model, the level-1 model is expressed as

$$\eta_{ijg} = \theta_j + \beta_{ig},$$

where θ_j is the ability estimate of examinee j, and β_{ig} is the item easiness estimate of item i in latent group g.

Similar to the cross-classified multilevel logistic model, the random effects of the person and item parameters are specified in the level-2 model. The level-2 model related to the person parameter is defined as

$$\begin{aligned}
\theta_j &= \gamma_{0j} + u_{1j}, \\
\gamma_{0j} &\sim N(\mu_g, 1), \\
u_{1j} &\sim N(0, \sigma_1^2),
\end{aligned}$$

where γ_{0j} is the fixed person effect, μ_g is the mean of the ability estimates in latent group g, and u_{1j} is the random person effect. The mean ability of the first latent group is constrained to be 1 to solve the identification problem (i.e., $\mu_1 = 1$).

The level-2 model related to the item parameter is given by

$$\beta_{ig} = \gamma_{ig} + u_{2i},$$
$$\gamma_{ig} \sim N(\overline{\beta}_g, 1),$$
$$u_{2i} \sim N(0, \sigma_2^2),$$

where γ_{ig} is the fixed item effect for item i in latent group g, $\overline{\beta}_g$ is the mean of the item easiness estimates in latent group g, and u_{2i} is the random item effect.

The combined model with item and person covariates can be written as

$$\begin{aligned}\eta_{ijg} &= \theta_j + \beta_{ig} \\ &= b_{0g} + \lambda_{0g} + \sum_{a=1}^{A} b_{ag} N_j + \sum_{b=1}^{B} \lambda_{bg} M_i + u_{1j} + u_{2i},\end{aligned}$$

where b_{0j} is the mean of the ability estimates in latent group g when there is no person effect on the ability estimate, b_{ag} is the regression coefficient for person covariate N_j in latent group g, λ_{0g} is the mean of the item easiness estimates in latent class g when there is no item effect on the item easiness parameter, and λ_{bg} is the regression coefficient for item covariate M_i in latent group g.

An important aspect of these random item IRT models combined with multilevel modeling lies in their capability for including covariates to explain the variance of a parameter (i.e., the random effect of a parameter). For instance, Wang (2011) provided evidence showing that including covariates to explain the random components improved the accuracy of parameter estimates. Also, it is possible to estimate the item parameters of individual items, when using the IRT models under the cross-classified framework, whereas the LLTM provides the regression coefficients only for item properties. This study, therefore, used the same approach of Van den Noortgate et al. (2003) and Wang (2011) such that the model allows the item variance at the item-level. In the context of Bayesian estimation, this item variance can be specified as a hyperprior of the item parameters.

3.1.2 A Mixture IRT Model

A mixture IRT (MixIRT) model can be regarded as a combination of an IRT model and a latent class model (Rost, 1990). MixIRT models assume that a person population consists of discrete latent classes. One thing this means is that a single set of parameters cannot fully explain the relationship between a person's ability and item characteristics as there are subpopulations for which the item parameters may differ. MixIRT models resolve this problem by applying different item parameters to individual subpopulations. That is, a unique set of item parameters is estimated for each latent group under a MixIRT models. This reflects another important assumption of MixIRT models—an IRT model still holds within an individual latent class. The probability of getting an item correct in the mixture Rasch model, for example, is defined as

$$P(u_{ijg} = 1|\theta_{jg}, \beta_{ig}, g) = \frac{\exp(\theta_{jg} - \beta_{ig})}{1 + \exp(\theta_{jg} - \beta_{ig})},$$

where g is an index for the latent class, u_{ijg} is a response of examinee j in class g to item i, θ_{jg} is the ability of examinee j in class g, and β_{ig} is the item difficulty parameter of item i for class g.

As mentioned earlier, different sets of item parameters are defined within individual latent groups under MixIRT models. It is necessary, therefore, to solve the identification problems. This issue is one which takes into account that item parameter estimates from one latent group are not comparable to those from another latent group. There are two different solutions for this identification problem. First, a norming condition such that $\sum_i \beta_{ig} = 0$ can be implemented (Rost, 1990). Second, the mean ability of the first latent group can be constrained to be zero (i.e., $\mu_1 = 0$). Either of these techniques allows comparison of item and person parameter estimates across the latent classes.

3.1.3 Testlet Graded Response Model

The graded response model (GRM; Samejima, 1969) is an IRT model for ordered polytomous categories. Under this model, the probability of receiving a score of k is defined as the difference between two cumulative probabilities of adjacent categories k and $k-1$. The cumulative probability, also called the boundary characteristic curve, with which examinee j obtains a score of k or higher on item i is expressed as

$$P^*_{ijk} = P(u_{ij} = k|\theta_j, \alpha_i, \beta_{ik}) = \frac{\exp[\alpha_i(\theta_j - \beta_{ik})]}{1 + \exp[\alpha_i(\theta_j - \beta_{ik})]],}$$

where $k = 0, \ldots, m_i + 1$ when m_i is the maximum category score of item i, θ_j is the ability of examinee j, α_i is the item discrimination parameter of item i, and β_{ik} is the category boundary location parameter for category k of item i. By definition, $P^*_{ij0} = 1$ and $P^*_{ij,m_i+1} = 0$.

The probability of examinee j getting a score of k is the difference between adjacent cumulative probabilities and is defined as

$$P_{ijk} = P^*_{ij,k-1} - P^*_{ijk}.$$

The testlet graded response model (GRM-t) is an extension of the GRM for a test which consists of one or more testlets. When a set of items shares a common stimulus, the testlet effect can occur in item parameters (Wainer & Kiely, 1987). This follows, because of the violation of the local independence assumption, that biased estimates may be produced of person and item parameters. The testlet model tries to account for the dependency between answers by including additional random effects in the IRT model. The probability of examinee j obtaining category k or higher under the GRM-t is given by

$$P^*_{ijk} = P(u_{ijg} = k|\theta_j, \alpha_i, \beta_{ik}, \gamma_{jt(i)}) = \frac{\exp[\alpha_i(\theta_j - \beta_{ik} - \gamma_{jt(i)})]}{1 + \exp[\alpha_i(\theta_j - \beta_{ik} - \gamma_{jt(i)})]},$$

where $\gamma_{jt(i)}$ is a person-specific testlet effect which depicts the interaction of examinee j and testlet $t(i)$ (i.e., item i belongs to testlet t). The variance of the testlet effect is considered as the amount of the testlet effect.

3.1.4 A Random Item Mixture Testlet Graded Response Model

The RIMGRM-t expresses the probability of examinee j in latent group g of obtaining category k or higher as

$$P^*_{ijgk} = P(u_{ijg} = k|\theta_{jg}, \alpha_{ig}, \beta_{igk}, \gamma_{jt(i)}, g) = \frac{\exp[\alpha_{ig}(\theta_{jg} - \beta_{igk} - \gamma_{jt(i)})]}{1 + \exp[\alpha_{ig}(\theta_{jg} - \beta_{igk} - \gamma_{jt(i)})]}, \tag{3.1}$$

where g is an index for the latent group and $g = 1, 2$ because this study assumes that a person population consists of two subpopulations—the speeded and nonspeeded groups. Note that an unconditional model refers to the RIMGRM-t without covariates and a conditional model refers to the RIMGRM-t with covariates in this study. Both the unconditional and conditional models have the equivalent mathematical representation as Equation 3.1, since their difference comes from their prior distributions. The next part covers this difference between the unconditional and conditional models.

3.2 Parameter Estimation

3.2.1 Bayesian Estimation Using Markov Chain Monte Carlo

The RIMGRM-t parameters were estimated using a Markov chain Monte Carlo (MCMC) method with Gibbs sampling as implemented in the OpenBUGS software (Thomas, O'Hara, Ligges, & Sturtz, 2006).

For models with higher complexity, including mixture IRT models, integration over high-dimensional distributions is required to make inferences about parameters. In such cases, it may not always be possible to implement numerical integration. MCMC methods provide an alternative way to solve this problem by employing Monte Carlo integration using Markov chains (Gilks, Richardson, & Spiegelhalter, 1996). To be more precise, Monte Carlo estimates the expectation of model parameters using a mean of the samples drawn by using a Markov chain. The Markov chain is a sequence of estimates of random variables whose distribution depends only on the current state of the chain. The Markov chain gradually forgets its initial state and, when the chain converges, becomes a stationary distribution. The expectation of

the parameters can be calculated following discarding of the burn-in iterations by using the remaining iterations.

Gibbs sampling, which is one of the MCMC sampling algorithms, iteratively draws samples from the full conditional distribution of model parameters and observed responses (Spiegelhalter, Best, Gilks, & Inskip, 1996). This full conditional distribution, which is also called the posterior distribution, is proportional to the product of the prior distribution and the likelihood function. For example, the prior distribution of the unconditional model can be written as

$$
\begin{aligned}
&P(\mu_g, \bar{\beta}_g, \sigma^2_\beta, \sigma^2_\gamma, \theta_{jg}, \alpha_{ig}, \beta_{igk}, \gamma_{jt(i)}, g, \pi_g | Y) \\
&\quad \propto P(\mu_g, \bar{\beta}_g, \sigma^2_\beta, \sigma^2_\gamma, \theta_{jg}, \alpha_{ig}, \beta_{igk}, \gamma_{jt(i)}, g, \pi_g) P(Y | \mu_g, \bar{\beta}_g, \sigma^2_\beta, \sigma^2_\gamma, \theta_{jg}, \alpha_{ig}, \beta_{igk}, \gamma_{jt(i)}, g, \pi_g) \\
&\quad = P(\mu_g, \bar{\beta}_g, \sigma^2_\beta, \sigma^2_\gamma) P(\theta_{jg}, \alpha_{ig}, \beta_{igk}, \gamma_{jt(i)}, g, \pi_g | \mu_g, \bar{\beta}_g, \sigma^2_\beta, \sigma^2_\gamma) P(Y | \theta_{jg}, \alpha_{ig}, \beta_{igk}, \gamma_{jt(i)}, g, \pi_g),
\end{aligned}
$$

where $P(Y|\theta_{jg}, \alpha_{ig}, \beta_{igk}, \gamma_{jt(i)}, g, \pi_g)$ is the probability of getting a score of k to item i under the unconditional RIMGRM-t.

As described in Chapter 2, the MCMC algorithm samples a class membership for each examinee at each stage of the chain. Next, it samples values for the parameters of each class conditional on class membership. First a class membership, $g_j = 1, \ldots, G$, and an examniee's ability, θ_{jg}, are sampled for each examinee at each stage of the Markov chain proportional to the probability of membership in that class and conditional upon all other class parameters.

The mixing proportions, π_g, and class ability distribution, μ_g, parameters are defined by the frequency each of the examinees was sampled into each latent class. The frequency examinees are sampled in each latent class defines the posterior membership probability for the examinee in that class. The mixing proportions and class mean ability parameters are then estimated along with the class memberships of the respondents. Ability, in this study, was assumed to be normally distributed with mean 0 and variance 1 in each latent class.

3.2.2 Priors

The following priors and hyperpriors were used for both unconditional and conditional models:

$$
\begin{aligned}
\theta_{jg} &\sim N(\mu_g, 1), && j = 1, \ldots, N,\ g = 1, 2, \\
\alpha_{ig} &\sim N(0, 1)I(0,), && i = 1, \ldots, n,\ g = 1, 2, \\
\gamma_{jt(i)} &\sim N(\mu_g, \sigma^2_\gamma), && j = 1, \ldots, N,\ t = 1, \ldots, T, \\
(\pi_1, \pi_2) &\sim Dirichlet(0.5, 0.5), \\
\mu_g &\sim N(0, 1), && g = 1, 2, \\
\sigma^2_\gamma &\sim Inverse\text{-}Gamma(2.5, 0.25),
\end{aligned}
$$

where $I(0,)$ indicates that the observations of α_{ig} will be sampled above zero. The mean ability of the first group was fixed as zero (i.e., $\mu_1 = 0$) to resolve the identification problem. Note that the variance of the discrimination parameter is fixed as one. In other words, this study focused on the random effect of item boundary location parameters.

The difference between the unconditional and conditional model is specified as how the mean of the item boundary location parameter is defined. For the unconditional model, the prior distribution of the item boundary location parameter is given by

$$
\begin{aligned}
\beta_{igk} &\sim N(\bar{\beta}_g, \sigma^2_\beta), && i = 1, \ldots, n,\ g = 1, 2,\ k = 1, \ldots, m_i, \\
\bar{\beta}_g &\sim N(0, 1), && g = 1, 2, \\
\sigma^2_\beta &\sim Uniform(0, 2),
\end{aligned}
\tag{3.2}
$$

where $\bar{\beta}_g$ is the mean of the item boundary location parameters for latent group g which implies that mean values can be different for speeded and nonspeeded groups and σ^2_β is the variance of the item boundary location parameter (i.e., the random item effect).

For the conditional model, the equivalent distribution of the item boundary location parameter is still used. The mean of the item boundary location parameter, however, is

expressed rather differently by using item covariates and can be written as

$$\begin{aligned} \bar{\beta}_g &= \bar{\beta}_{0g} + \textstyle\sum_{a=1}^{A} c_{ag} IC_a, \quad g = 1, 2, \\ \bar{\beta}_{0g} &\sim N(0, 1), \qquad\qquad\quad g = 1, 2, \end{aligned} \tag{3.3}$$

where $\bar{\beta}_{0g}$ is the mean of the item boundary location parameter for latent group g when the item covariates have no impact on the item boundary location parameter, c_{ag} is the effect of the item covariate for latent group g, and IC_a is the item covariate. It is important to note that the item covariate coefficient c_{gk} can be different for each latent group.

3.2.3 Checking Convergence

As mentioned earlier, it is expected that the posterior distribution converges to the stationary distribution after a sufficiently long burn-in. The burn-in iterations should be determined based on evidence of convergence, since the length of the burn-in can vary depending on a number of factors including the complexity of the model or the data. In this study, Heidelberger and Welch (1983) convergence diagnostics as implemented in the CODA package (Plummer et al., 2006) were used to check the convergence.

3.3 Example: Detection of Speededness on a Science Inquiry Test

In this section, an example is presented to illustrate the detection of speededness using the RIMGRM-t. Data for the example were taken from a middle grades and high school grades test of science inquiry knowledge. In the example, both the unconditional and conditional RIMGRM-t's were used to detect test speededness in CR items.

3.3.1 Methods

Data

The sample consisted of 1,245 middle (77.23%) and 367 high (22.77%) school students. The data were taken from the responses to a test designed to evaluate students' knowledge and use

of science inquiry practices. This test was composed of seven CR items. The responses were scored using four different rubrics depending on the type of information being scored for the item. This was done to increase the amount of information which comes from an examinee's written answer (Ercikan, 2002). In this case, the seven items were regarded as seven testlets and scores from various rubrics as ordered categories for items nested in testlets. The data were treated as responses nested in the seven testlets and consisted of 6, 4, 5, 6, 6, 5, and 6 items, respectively.

For the conditional model, the item type was used as an item covariate. As noted earlier, there are four different item types (i.e., Science Inquiry, Everyday Language, Academic Language, and Science Content), and an item type was coded as 0 and 1 (i.e., dummy coding).

Model

Three models were used in this study: a random item testlet GRM model (RIGRM-t) and a two two-class RIMGRM-t model, the unconditional and conditional models, respectively. The item discrimination and boundary location parameters for the first 15 items (those for Testlet 1 to Testlet 3) were estimated by the first model. Then, an equality constraint was imposed on the item parameters of these 15 items based on the assumption that those items were not affected by test speededness. The item parameters of the remaining 28 items (those for Testlet 4 to Testlet 7) were estimated by both the unconditional and conditional RIMGRM-t's.

The mean ability parameter of the speeded group was constrained to be zero so that the parameter estimates for the nonspeeded group could be compared with those for the speeded group.

Modeling Speededness

The speededness model suggested in Bolt et al. (2002) was used in this study. That model assumes that the effects of test speededness are most likely to be detected on items at the end of the test. The ordinal constraints for this assumption were mentioned in the previous chapter and were used to model speededness in this example.

3.3.2 Results

Monitoring Convergence

The Heidelberger and Welch (1983) convergence diagnostic was used to determine the burn-in and post burn-in iterations. The results suggested a burn-in length of 6,000 and a post burn-in length of 6,000 for the RIGRM-t. For the unconditional and conditional RIMGRM-t's, the diagnostic indicated that the unconditional model converged after 13,000 burn-in iterations and the conditional model converged after 10,000 burn-in iterations. Also, the result from the diagnostic proposed a post burn-in of 2,000 iterations for the unconditional model and 5,000 iterations for the conditional model.

Parameters Estimates

The item discrimination parameter estimates for the speeded and nonspeeded models from the unconditional and conditional models are illustrated in Figures 3.1 and 3.2. The differences between the unconditional and conditional models for the speeded group were more appreciable compared to those for the nonspeeded group. Given that the sample size for the speeded group was much smaller than that for the nonspeeded group, it is possible that the item discrimination parameter estimates are more dependent on the sample size than the item boundary location parameter estimates which are presented below.

Figure 3.3 shows plots of the differences between the unconditional and conditional models in the means of item boundary location parameters for each testlet for the speeded group. Similarly, Figure 3.4 shows the differences between the unconditional and conditional

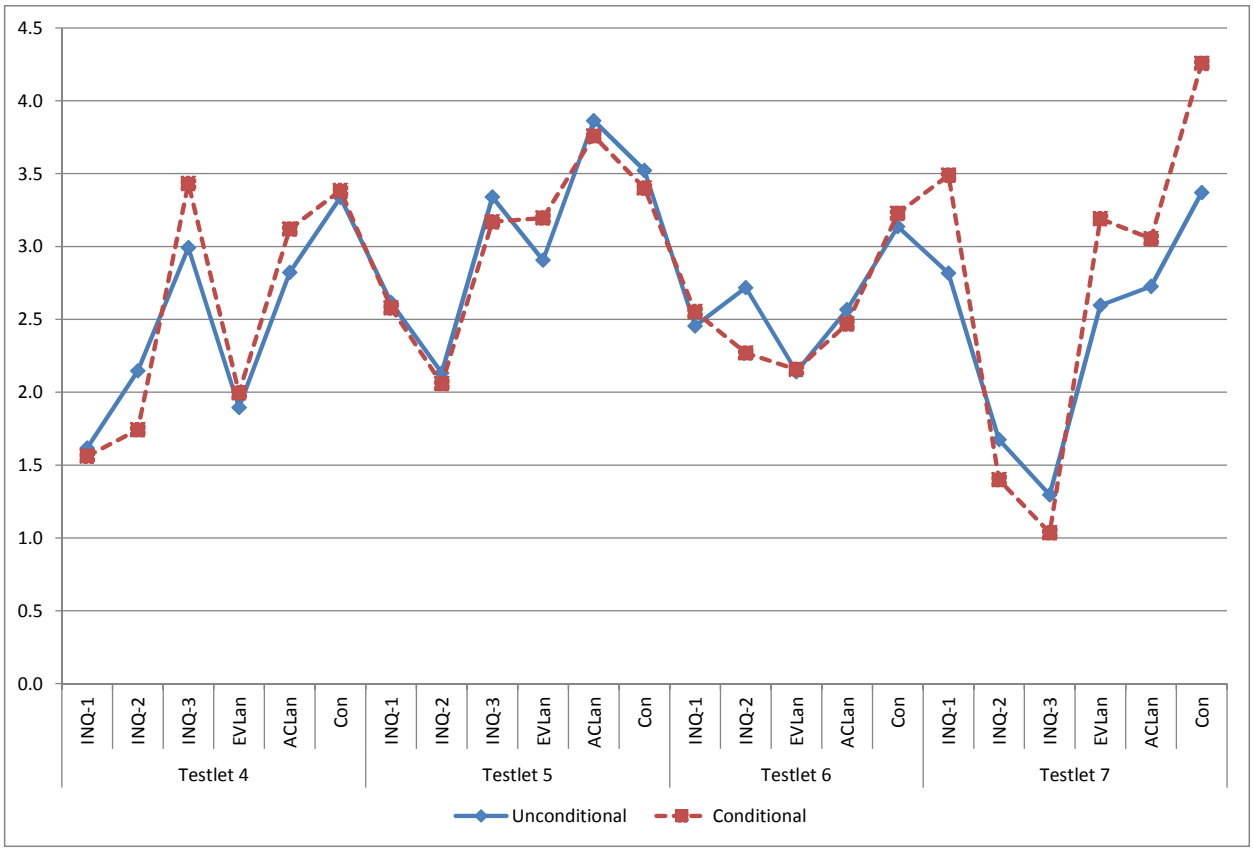

Figure 3.1: Item discrimination parameter estimates: The speeded group

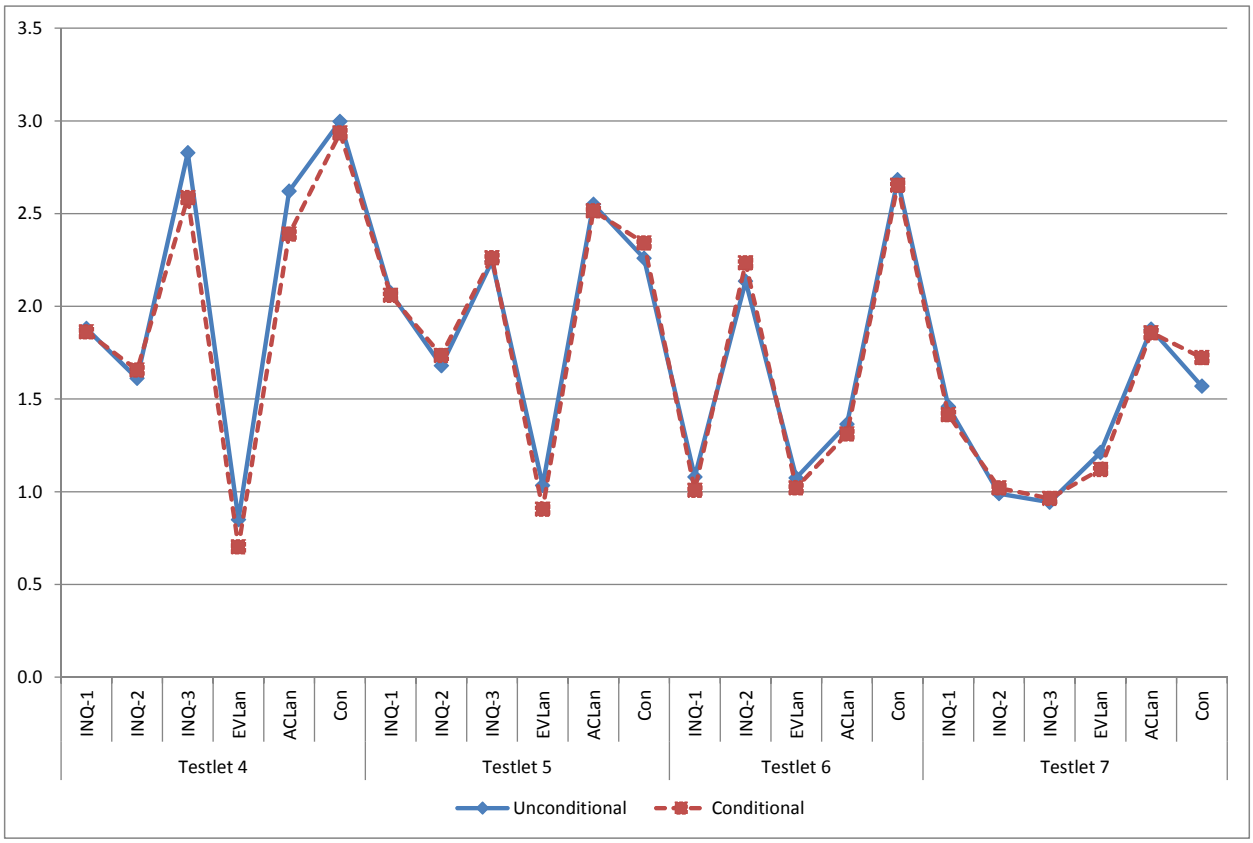

Figure 3.2: Item discrimination parameter estimates: The nonspeeded group

models in the means of item boundary location parameters for the nonspeeded group. Since the first 15 items on the test were constrained to be equal for the speeded and nonspeeded groups for both the conditional and unconditional models, they served as anchor items to link the metrics of the conditional and unconditional models. Therefore, the means of item boundary location parameters for the speeded and nonspeeded groups were on the same scale.

Both figures suggest that the item boundary location parameter estimates from the unconditional and conditional models were similar to each other. There were, however, a couple of differences between the estimates from these models. For the speeded group, the difference between the two models became larger at the end of the test (i.e., Testlet 6 and Testlet 7). For the nonspeeded group, the boundary location parameter estimates of Everyday Language were smaller for the conditional model. This might possibly be explained by the effects of item covariates (explained below). The item discrimination parameter estimates as well as the item boundary location parameter estimates for the speeded and nonspeeded groups obtained from the unconditional model are provided in Tables 3.1 and 3.2, and those for the nonspeeded group obtained from the conditional model are provided in Tables 3.3 and 3.4.

The item variance of the unconditional model was 1.926 ($SD = 0.068$), whereas the item variance of the conditional model was 1.861 ($SD = 0.114$). These results suggest that the inclusion of the item covariate helped explain the random variance in items. Results in Table 3.5 show the effects of the item covariates on item boundary location parameter estimates. The credibility interval from 2.5% to 97.5% in Table 3.5 indicates whether the parameter was significantly different from zero. The result suggests that the effects of all item covariates did not affect the item boundary location parameters of the speeded group, but the effects of Everyday Language and Academic Language covariates were contained in the interval at the .05 level for the nonspeeded group. These significant coefficients helped explain the differences between the unconditional and conditional models in item boundary location

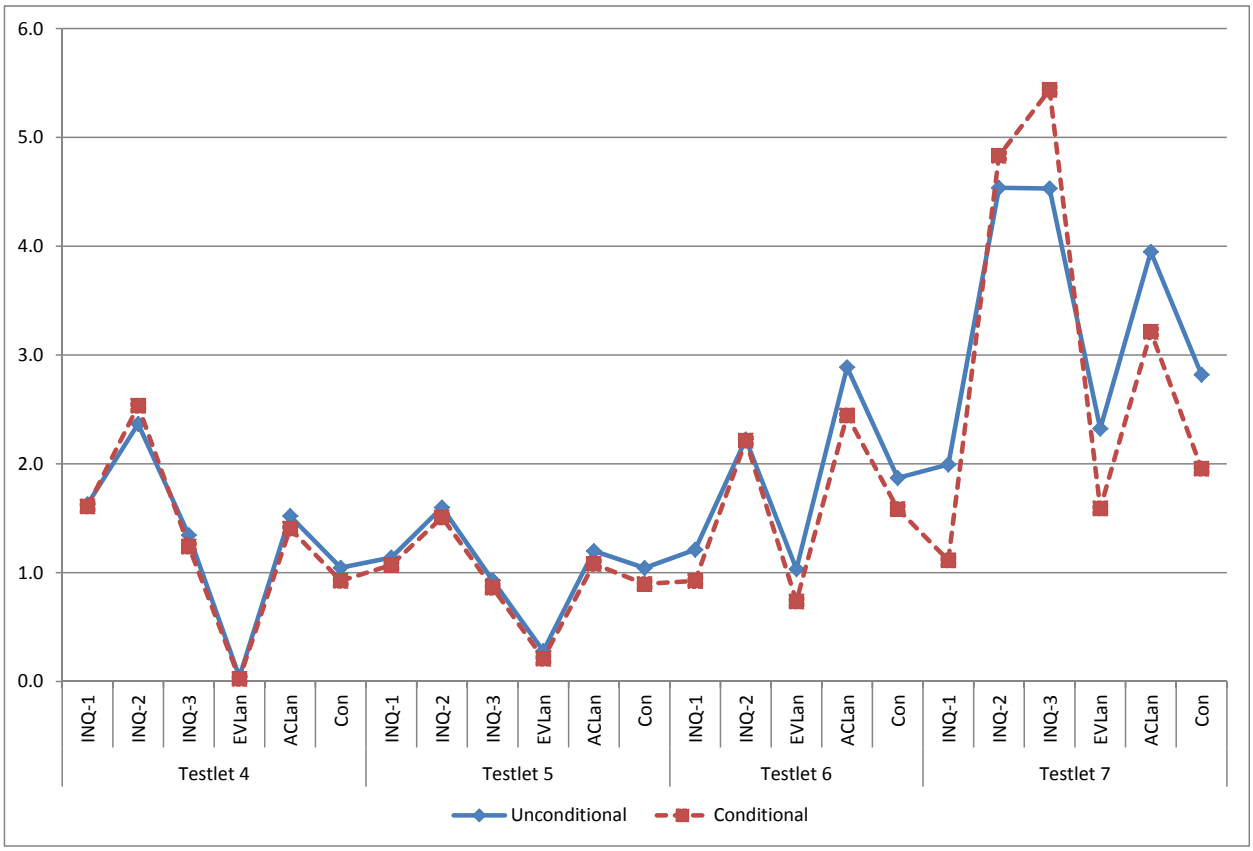

Figure 3.3: Means of boundary location parameter estimates: The speeded group

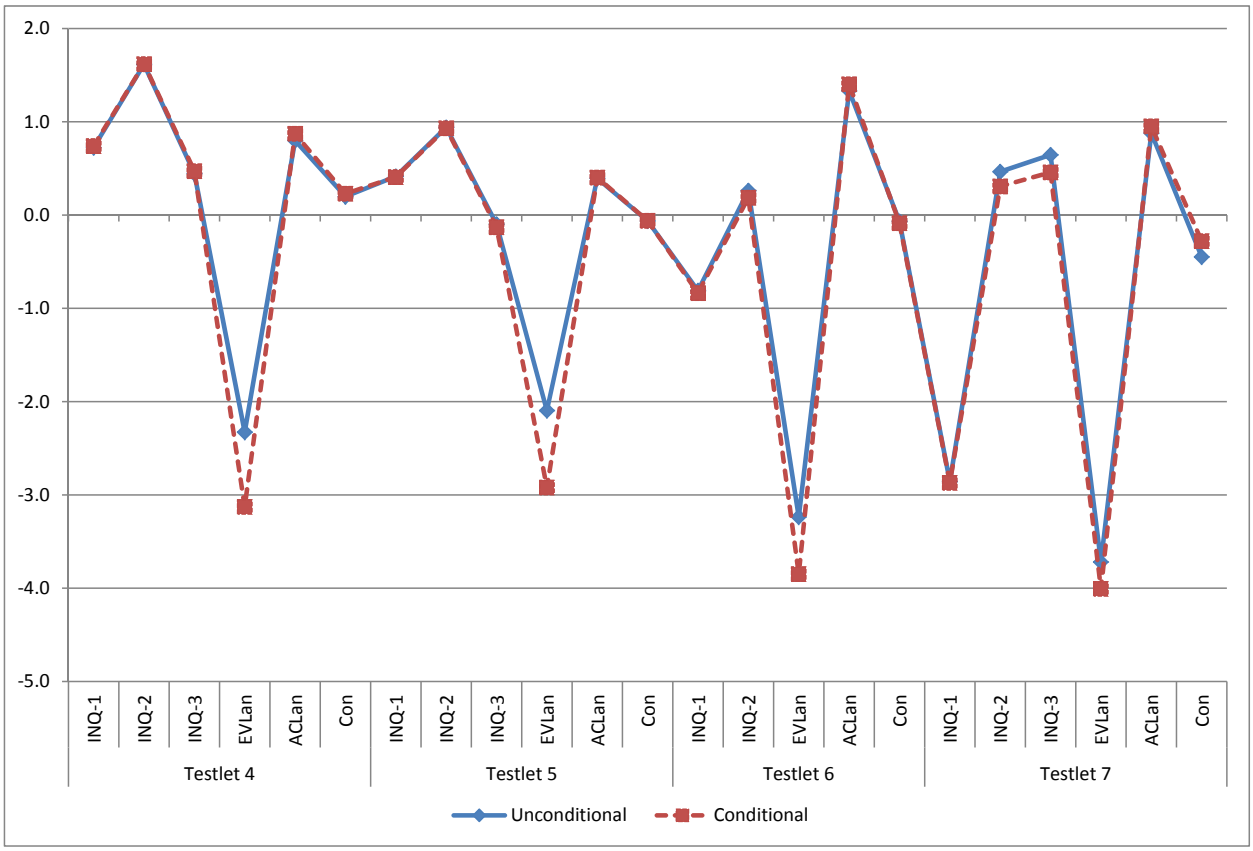

Figure 3.4: Means of boundary location parameter estimates: The nonspeeded group

Table 3.1: Item Parameter Estimates for the Speeded Group: The Unconditional Model

		α		β_1		β_2		β_3		β_4		
Testlet	Item	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD	$\bar{\beta}$
4	16	1.62	0.22	1.63	0.17							1.63
	17	2.15	0.36	2.37	0.22							2.37
	18	2.99	0.39	1.07	0.09	1.62	0.13					1.34
	19	1.90	0.19	-0.90	0.09	0.99	0.11					0.04
	20	2.82	0.30	0.50	0.07	1.20	0.10	1.91	0.14	2.47	0.20	1.52
	21	3.34	0.38	0.63	0.07	1.47	0.11					1.05
5	22	2.62	0.33	1.14	0.09							1.14
	23	2.13	0.31	1.60	0.15							1.60
	24	3.34	0.41	0.93	0.08	1.44	0.12					1.18
	25	2.91	0.29	-0.54	0.07	1.10	0.09					0.28
	26	3.86	0.37	0.45	0.06	0.92	0.07	1.40	0.09	2.03	0.14	1.20
	27	3.52	0.38	0.60	0.07	1.48	0.11					1.04
6	28	2.45	0.32	1.21	0.13							1.21
	29	2.72	0.44	2.11	0.22	2.34	0.24					2.22
	30	2.14	0.25	0.11	0.10	1.96	0.20					1.03
	31	2.57	0.34	1.63	0.16	2.26	0.21	3.28	0.30	4.39	0.46	2.89
	32	3.14	0.40	1.49	0.14	2.25	0.19					1.87
7	33	2.82	0.39	1.99	0.18							1.99
	34	1.67	0.32	4.18	0.31	4.90	0.38					4.54
	35	1.30	0.24	4.53	0.38							4.53
	36	2.60	0.35	1.40	0.16	3.24	0.25					2.32
	37	2.73	0.40	2.76	0.21	3.46	0.23	4.30	0.27	5.28	0.39	3.95
	38	3.37	0.47	2.06	0.18	3.58	0.24					2.82

Table 3.2: Item Parameter Estimates for the Nonspeeded Group: The Unconditional Model

		α		β_1		β_2		β_3		β_4		
Testlet	Item	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD	$\bar{\beta}$
4	16	1.88	0.16	0.72	0.08							0.72
	17	1.61	0.15	1.62	0.12							1.62
	18	2.83	0.22	0.15	0.05	0.75	0.06					0.45
	19	0.85	0.07	-5.10	0.35	0.44	0.11					-2.33
	20	2.62	0.17	-0.32	0.05	0.46	0.06	1.16	0.08	1.88	0.11	0.80
	21	3.00	0.22	-0.36	0.05	0.75	0.06					0.20
5	22	2.07	0.18	0.41	0.06							0.41
	23	1.68	0.14	0.94	0.09							0.94
	24	2.24	0.15	-0.10	0.05	0.57	0.06					0.24
	25	1.03	0.08	-4.84	0.33	0.65	0.10					-2.10
	26	2.55	0.17	-0.75	0.05	0.09	0.05	0.78	0.06	1.46	0.09	0.39
	27	2.26	0.15	-0.84	0.05	0.71	0.06					-0.06
6	28	1.08	0.10	-0.81	0.08							-0.81
	29	2.14	0.20	0.05	0.06	0.48	0.07					0.26
	30	1.08	0.09	-5.33	0.37	-1.14	0.09					-3.24
	31	1.36	0.10	-0.28	0.07	0.80	0.10	2.02	0.16	2.79	0.21	1.33
	32	2.68	0.22	-0.79	0.06	0.65	0.07					-0.07
7	33	1.46	0.16	-2.86	0.17							-2.86
	34	0.99	0.10	0.14	0.09	0.79	0.12					0.46
	35	0.94	0.09	0.65	0.11							0.65
	36	1.21	0.12	-5.46	0.45	-1.98	0.13					-3.72
	37	1.88	0.18	-0.85	0.07	0.41	0.07	1.49	0.11	2.47	0.18	0.88
	38	1.57	0.14	-2.81	0.15	1.91	0.14					-0.45

Table 3.3: Item Parameter Estimates for the Speeded Group: The Conditional Model

		α		β_1		β_2		β_3		β_4		
Testlet	Item	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD	$\bar{\beta}$
4	16	1.56	0.20	1.61	0.17							1.61
	17	1.74	0.30	2.53	0.31							2.53
	18	3.43	0.40	0.98	0.10	1.50	0.12					1.24
	19	2.00	0.17	-0.99	0.08	1.04	0.12					0.02
	20	3.12	0.32	0.45	0.09	1.11	0.11	1.80	0.16	2.27	0.18	1.40
	21	3.39	0.37	0.51	0.09	1.34	0.12					0.93
5	22	2.58	0.31	1.07	0.10							1.07
	23	2.06	0.28	1.51	0.16							1.51
	24	3.17	0.36	0.86	0.08	1.30	0.11					1.08
	25	3.20	0.29	-0.65	0.06	1.07	0.10					0.21
	26	3.76	0.34	0.32	0.06	0.88	0.07	1.30	0.09	1.82	0.13	1.08
	27	3.40	0.34	0.41	0.06	1.39	0.11					0.90
6	28	2.55	0.32	0.92	0.10							0.92
	29	2.27	0.36	2.11	0.15	2.32	0.17					2.21
	30	2.16	0.21	-0.17	0.08	1.64	0.13					0.74
	31	2.47	0.30	1.32	0.11	1.95	0.14	2.85	0.21	3.66	0.33	2.44
	32	3.23	0.42	1.20	0.10	1.97	0.14					1.58
7	33	3.49	0.45	1.11	0.15							1.11
	34	1.40	0.24	4.31	0.37	5.35	0.51					4.83
	35	1.04	0.21	5.44	0.54							5.44
	36	3.19	0.43	0.65	0.13	2.54	0.22					1.59
	37	3.05	0.41	1.88	0.17	2.68	0.22	3.69	0.27	4.61	0.39	3.21
	38	4.26	0.52	1.15	0.15	2.76	0.21					1.96

Table 3.4: Item Parameter Estimates for the Nonspeeded Group: The Conditional Model

		α		β_1		β_2		β_3		β_4		
Testlet	Item	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD	$\bar{\beta}$
4	16	1.86	0.16	0.74	0.08							0.74
	17	1.66	0.18	1.62	0.15							1.62
	18	2.59	0.20	0.15	0.06	0.79	0.07					0.47
	19	0.70	0.07	-6.69	0.57	0.44	0.13					-3.13
	20	2.39	0.16	-0.33	0.05	0.51	0.06	1.25	0.09	2.06	0.14	0.87
	21	2.94	0.23	-0.36	0.05	0.81	0.07					0.23
5	22	2.06	0.18	0.41	0.07							0.41
	23	1.74	0.16	0.93	0.09							0.93
	24	2.26	0.16	-0.13	0.05	0.58	0.07					0.22
	25	0.91	0.08	-6.51	0.56	0.67	0.11					-2.92
	26	2.52	0.17	-0.75	0.05	0.06	0.05	0.79	0.07	1.50	0.09	0.40
	27	2.34	0.17	-0.81	0.05	0.69	0.07					-0.06
6	28	1.01	0.10	-0.84	0.09							-0.84
	29	2.24	0.20	-0.04	0.06	0.41	0.07					0.19
	30	1.02	0.09	-6.39	0.53	-1.31	0.10					-3.85
	31	1.31	0.10	-0.28	0.07	0.85	0.09	2.12	0.15	2.92	0.20	1.40
	32	2.65	0.24	-0.84	0.06	0.67	0.07					-0.09
7	33	1.42	0.17	-2.87	0.18							-2.87
	34	1.02	0.10	-0.02	0.09	0.64	0.11					0.31
	35	0.96	0.10	0.46	0.11							0.46
	36	1.12	0.11	-5.73	0.49	-2.28	0.14					-4.01
	37	1.86	0.16	-0.81	0.07	0.48	0.08	1.56	0.13	2.58	0.18	0.95
	38	1.72	0.16	-2.70	0.14	2.15	0.17					-0.28

Table 3.5: Effects of the Item Covariates

Group	Covariate	Estimate	SD	2.5%	97.5%
Speeded	Inquiry	0.847	0.545	-0.222	1.928
	EVLan	-0.495	0.588	-1.674	0.637
	ACLan	0.697	0.542	-0.364	1.768
	Content	0.087	0.587	-1.092	1.240
Nonspeeded	Inquiry	0.542	0.549	-0.526	1.612
	EVLan	-2.485	0.603	-3.645	-1.280
	ACLan	1.199	0.531	0.163	2.252
	Content	0.306	0.581	-0.814	1.447

Note. Inquiry = Science Inquiry, EVLan = Everyday Language, ACLan = Academic Language, Content = Science Content.

parameter estimates of Everyday Language and Academic Language. More specifically, the differences in Everyday Language was larger than that in Academic Language.

Table 3.6 presents the estimated testlet effects for the unconditional and conditional models. Both the unconditional and conditional models showed a similar pattern in testlet effects. The testlet effects in the speeded group were large both at the beginning of the test (Testlet 1 to Testlet 3) and at the end of the test (Testlet 7) and smaller in the middle of the test (Testlet 4 to Testlet 6). The testlet effects in the nonspeeded group were large at the beginning of the test (Testlet 1 to Testlet 3) and small at the middle and end of the test (Testlet 4 to Testlet 7).

As can be seen in Table 3.7, the proportion of the speeded group obtained from the unconditional model was lower than the proportion obtained from the conditional model. That is, the mixture proportion of the speeded group was 35.6% in the unconditional model and 40.2% in the conditional model. The mixture proportion of the nonspeeded group was 64.4% in the unconditional model and 59.8% in the conditional model.

Table 3.6: Testlet Effects

Model	Group	Testlet						
		1	2	3	4	5	6	7
Unconditional	Speeded	11.330	8.155	9.184	0.329	0.152	1.390	4.366
	Nonspeeded	6.614	4.315	2.699	0.233	0.317	0.726	0.791
Conditional	Speeded	9.977	7.884	8.722	0.367	0.164	1.192	3.923
	Nonspeeded	6.947	4.305	2.579	0.262	0.274	0.772	0.831

Table 3.7: Proportions of Latent Classes

Model	Proportions (%)	
	Speeded	Nonspeeded
Unconditional	35.6	64.4
Conditional	40.2	59.8

Results in Table 3.8 show a cross-tabulation of group membership assigned by the unconditional and conditional models. Both models detected 86.0% of the total examinees in the same latent group. When the item covariates were used (i.e., under the conditional model), 4.7% of the total examinees were classified as the speeded group and 9.3% of the total examinees were classified as the nonspeeded group.

3.4 A Simulation Study

The purpose of this simulation study was to examine the performance of a RIMGRM-t for the detection of speededness effects under practical testing conditions. Results from the unconditional and conditional models were compared.

Table 3.8: Cross-Tabulation of Latent Group Membership

		Conditional Model		
		Speeded	Nonspeeded	Total
Unconditional Model	Speeded	498 (30.9%)	76 (4.7%)	574 (35.6%)
	Nonspeeded	101 (9.3%)	899 (55.1%)	1000 (64.4%)
Total		698 (40.2%)	983 (59.8%)	1612 (100.0%)

3.4.1 Simulation Conditions

Aside from the differences between the unconditional and conditional models, two factors were considered in the simulation study based on the results from Wang (2011). First, the sample size was manipulated. Wang used three different conditions (1,000, 2,000 and 3,000) for this factor. The results from Wang suggested that the improvements in the recovery of the parameters were negligible when the sample size increased from 2,000 to 3,000, suggesting that one of the conditions can be dropped. Considering that a RIMGRM-t is more complex than the model implemented in Wang, two sample sizes were simulated, 1,000 and 3,000 examinees.

Second, the membership proportions of the speeded group were manipulated. In previous work, the proportions of the speeded group range from 10% to 30% (e.g., Bolt et al., 2002; Kim et al., 2016; Wang, 2011). In this study, two proportions of test speededness were simulated: small and high speededness. The former includes 10% and the later 30% of the speeded examinees in the sample. These conditions were compared to the ones obtained from the analysis of the generated data to check the classification accuracy.

Twenty replications were simulated for each of the 8 different conditions considered (i.e., 2 models $\times$ 2 sample sizes $\times$ 2 proportions of the speeded group = 8 conditions).

3.4.2 Data Generation Procedure

The same test length and structure of the real data were used to generate the response data. That is, a 38-item test length was simulated with 7 testlets. The item parameter estimates obtained from the real data were used as generating parameters. Specifically, the estimates of the first 15 items obtained from the RIGRM-t were used as the generating values for both the unconditional and conditional models. As mentioned above, the parameters of these 15 items were assumed to be equal across the latent classes. The generating parameters of the items for the first three testlets are given in Table 3.9. The item parameter estimates of the other 23 items from the unconditional and conditional models were implemented as the generating parameters for each model. The generating parameters of the items for the last four testlets of the unconditional and conditional models are presented in Table 3.10 and Table 3.11, respectively.

For the conditional model, the item type was simulated as covariates on the item boundary location parameters. The item types of the real data, which included Science Inquiry, Everyday Language, Academic Language, and Science Content, were also used for the simulation. The types of items describe the major issue considered in scoring the response. Each response was scored as belonging to only one item type. Every testlet has one item of Science Content, Everyday Language, and Academic Language, whereas the number of Science Inquiry items ranges from 2 to 3. Dummy coding was used to indicate the item type, which resulted in adding four coefficients to the model. The effect of the item type on each latent group was fixed when the data was generated.

Again, the testlet effects obtained from the real data were used as the variance of the normal distribution for the person-specific testlet effect parameter.

Table 3.9: Generating Parameters for Testlet 1 to Testlet 3

Testlet	Item	α	β_1	β_2	β_3	β_4
1	1	0.621	-0.980			
	2	0.992	0.998			
	3	0.625	-2.737			
	4	0.964	0.603			
	5	0.624	-2.252	0.212	2.640	4.340
	6	2.192	-2.282	1.320		
2	7	0.821	-2.487			
	8	0.669	0.944			
	9	0.942	-1.510	0.937	3.531	5.194
	10	1.850	-3.533	1.807		
3	11	3.454	-1.080			
	12	3.877	-1.156			
	13	1.202	-5.642	-4.811		
	14	0.986	-0.642	0.488	3.293	4.814
	15	4.882	-1.309	0.172		

Table 3.10: Generating Parameters for Testlet 4 to Testlet 7: The Unconditional Model

		Speeded					Nonspeeded				
Testlet	Item	α	β_1	β_2	β_3	β_4	α	β_1	β_2	β_3	β_4
4	16	1.62	1.63				1.88	0.72			
	17	2.15	2.37				1.61	1.62			
	18	2.99	1.07	1.62			2.83	0.15	0.75		
	19	1.90	-0.90	0.99			0.85	-5.10	0.44		
	20	2.82	0.50	1.20	1.91	2.47	2.62	-0.32	0.46	1.16	1.88
	21	3.34	0.63	1.47			3.00	-0.36	0.75		
5	22	2.62	1.14				2.07	0.41			
	23	2.13	1.60				1.68	0.94			
	24	3.34	0.93	1.44			2.24	-0.10	0.57		
	25	2.91	-0.54	1.10			1.03	-4.84	0.65		
	26	3.86	0.45	0.92	1.40	2.03	2.55	-0.75	0.09	0.78	1.46
	27	3.52	0.60	1.48			2.26	-0.84	0.71		
6	28	2.45	1.21				1.08	-0.81			
	29	2.72	2.11	2.34			2.14	0.05	0.48		
	30	2.14	0.11	0.11			1.08	-5.33	-5.33		
	31	2.57	1.63	2.26	3.28	4.39	1.36	-0.28	0.80	2.02	2.79
	32	3.14	1.49	2.25			2.68	-0.79	0.65		
7	33	2.82	1.99				1.46	-2.86			
	34	1.67	4.18	4.90			0.99	0.14	0.79		
	35	1.30	4.53				0.94	0.65			
	36	2.60	1.40	3.24			1.21	-5.46	-1.98		
	37	2.73	2.76	3.46	4.30	5.28	1.88	-0.85	0.41	1.49	2.47
	38	3.37	2.06	3.58			1.57	-2.81	1.91		

Table 3.11: Generating Parameters for Testlet 4 to Testlet 7: The Conditional Model

		Speeded					Nonspeeded				
Testlet	Item	α	β_1	β_2	β_3	β_4	α	β_1	β_2	β_3	β_4
4	16	1.56	1.61				1.86	0.74			
	17	1.74	2.53				1.66	1.62			
	18	3.43	0.98	1.50			2.59	0.15	0.79		
	19	2.00	-0.99	1.04			0.70	-6.69	0.44		
	20	3.12	0.45	1.11	1.80	2.27	2.39	-0.33	0.51	1.25	2.06
	21	3.39	0.51	1.34			2.94	-0.36	0.81		
5	22	2.58	1.07				2.06	0.41			
	23	2.06	1.51				1.74	0.93			
	24	3.17	0.86	1.30			2.26	-0.13	0.58		
	25	3.20	-0.65	1.07			0.91	-6.51	0.67		
	26	3.76	0.32	0.88	1.30	1.82	2.52	-0.75	0.06	0.79	1.50
	27	3.40	0.41	1.39			2.34	-0.81	0.69		
6	28	2.55	0.92				1.01	-0.84			
	29	2.27	2.11	2.32			2.24	-0.04	0.41		
	30	2.16	-0.17	-0.17			1.02	-6.39	-6.39		
	31	2.47	1.32	1.95	2.85	3.66	1.31	-0.28	0.85	2.12	2.92
	32	3.23	1.20	1.97			2.65	-0.84	0.67		
7	33	3.49	1.11				1.42	-2.87			
	34	1.40	4.31	5.35			1.02	-0.02	0.64		
	35	1.04	5.44				0.96	0.46			
	36	3.19	0.65	2.54			1.12	-5.73	-2.28		
	37	3.05	1.88	2.68	3.69	4.61	1.86	-0.81	0.48	1.56	2.58
	38	4.26	1.15	2.76			1.72	-2.70	2.15		

The ability parameters of the speeded group for the unconditional and conditional models were generated from a normal distribution with mean 0 and variance 1, whereas the ability parameters for the nonspeeded group for both models were generated from a normal distribution with estimated mean ability parameters and variance 1.

Given those generating parameters, the response data of the unconditional and conditional models were generated based on equation 3.1 and distributions 3.2 and 3.3.

3.4.3 Recovery Analysis

A recovery analysis was conducted to evaluate the performance of the two RIMGRM-t's (i.e., the unconditional and conditional random item testlet graded response models) for the detection of test speededness. Results of the recovery analysis indicate the extent to which the generating parameters were recovered from the simulated data. For item discrimination, item boundary location parameters, and testlet effects, the bias, the root mean square error (RMSE), and the correlation between the generating item parameters and the estimated item parameters were computed. For the mean ability parameter of each latent class, only the mean ability of the nonspeeded group was recovered because the mean ability of the speeded group was constrained to be 0 to solve the identification problem. The predetermined latent group membership of the speeded and nonspeeded examinees were compared with the estimated latent group membership from the simulated data.

3.4.4 Results

In this section, the results of the recovery analysis of item parameters (i.e., item discrimination and item boundary location parameters), testlet effects, and percentages of correct detection of latent group memberships are given separately for the unconditional and conditional models. This is mainly because the number of examinees who were assigned to the speeded group was much smaller than that of examinees who were assigned to the nonspeeded group. The recovery of the mean ability of the nonspeeded group is also presented.

Simulation Results of the Unconditional Model

The bias, RMSEs, the correlations between the estimates and the true parameters, and the percentages of the correct classification of latent group memberships for the speeded and nonspeeded groups obtained from the unconditional model are presented in Table 3.12. In general, the absolute values for the bias and the RMSEs for the speeded group were larger than those for the nonspeeded group, and the correlations for the speeded group were lower than those of the nonspeeded group. The percentage of correct membership for the speeded group was slightly larger than those for the nonspeeded group.

Recovery of the Item Discrimination Parameters. The absolute values for bias and RMSEs for the item discrimination parameters for both the speeded and nonspeeded groups decreased as the sample size increased or the proportion of the speeded group increased. For example, the RMSE of the speeded group, when the sample size was 1,000, decreased from 0.974 to 0.634 as the proportion of the speeded group increased from 10% to 30%. The absolute values for bias of the item discrimination parameters for the speeded group, which ranged from 0.257 to 0.804, were larger than those for the nonspeeded group, which ranged from 0.061 to 0.222. The RMSEs for the item discrimination parameters for the speeded group ranged from 0.411 to 0.974 and were also larger than those for the nonspeeded group, which ranged from 0.177 to 0.256.

The correlations for both the speeded and nonspeeded groups increased as the sample size increased. When the proportion of the speeded group increased, however, the correlations for the speeded group increased, while the correlations for the nonspeeded group decreased. Considering that the proportion of the speeded group is also related to the sample size of each latent class, these results suggest that a sufficient sample size would have a positive impact on the recovery of the correlations of the item discrimination parameters. Note that the correlations for the speeded group were lower than .9 under all conditions, whereas those for the nonspeeded group were higher than .9 under all conditions. To be more precise, the

correlations for the speeded group, which ranged from .548 to .884, were smaller than those for the nonspeeded group, which ranged from .952 to .990.

Recovery of the Item Boundary Location Parameters. The absolute values for the bias and the RMSEs for the item boundary location parameters for both the speeded and nonspeeded groups decreased as either the sample size increased or as the proportion of the speeded group increased, except in a few conditions. When the sample size was 3,000, the absolute value of the bias and the RMSE for the speeded group increased as did the absolute value of bias for the nonspeeded group as the proportion of the speeded group increased from 10% to 30%. The absolute values for bias of the item boundary location parameters for the speeded group, which ranged from 0.055 to 0.181, were larger than those for the nonspeeded group, which ranged from 0.022 to 0.116. The RMSEs for the item boundary location parameters for the speeded group ranged from 0.530 to 0.898 and were also larger than those for the nonspeeded group, which ranged from 0.210 to 0.334.

The correlations of the item boundary location parameters for the speeded group increased as the total sample increased, only when the sample size was 1,000. Under the 3,000 examinee condition, the correlations for the speeded group remained the same as the proportion of the speeded group increased. Similar to the recovery of the item discrimination parameters, the correlations for the nonspeeded group increased as the sample size increased and the correlations decreased as the proportion of the speeded group increased. These differences, were negligible as all the correlations for the nonspeeded group were higher than .99. The correlations for the speeded group, which ranged from .809 to .936, were smaller than those for the nonspeeded group, which ranged from .993 to .999.

Recovery of the Testlet Effects. The absolute values for the bias of the testlet effects for both the speeded and nonspeeded groups increased as the proportion of the speeded group increased within each sample size. For example, the absolute value for the bias, when the sample size was 3,000, increased from 0.286 to 0.744 for the speeded group, and from 0.208

to 0.310 for the nonspeeded group. The absolute values for the bias of testlet effects for the speeded group, which ranged from 0.286 to 0.744 was larger than those for the nonspeeded group, which ranged from 0.208 to 0.310.

The RMSEs for the testlet effects for the speeded group, which ranged from 1.009 to 1.910, increased as the proportion of the speeded group increased only when the sample size was 3,000 and the RMSEs for the speeded group decreased as the sample size increased. The RMSEs for testlet effects for the nonspeeded group, which ranged from 0.401 to 0.648, increased as the proportion of the speeded group increased. This result seems reasonable because a smaller sample size was involved in the estimation of the testlet effects for the nonspeeded group when the proportion of the speeded group increased. Note that the RMSEs were larger than 1 only for the speeded group, which indicates a small portion of the total sample may be insufficient for recovering the true testlet effects.

The correlations of the testlet effects for the speeded group were lower than those for the nonspeeded group, when the sample size was 1,000, whereas the correlations for both groups were similar when the sample size was 3,000. Again, the correlations for both groups increased as the sample size increased. Also, the correlations for the speeded group increased and the correlations for the nonspeeded group decreased when the proportion of the speeded group increased. All of the correlations for both groups were higher than .9. Specifically, the correlations for the speeded group ranged from .983 to .992 and the correlations for the nonspeeded group ranged from .981 to .993.

Recovery of Latent Group Membership. The recovery of latent group membership was evaluated by the percentages of the correct classification of each examinee's latent group membership. As expected, the percentages of the correct detection for the speeded group increased and the percentages for the nonspeeded group decreased as the proportion of the speeded group increased. Similarly, the percentages for both groups increased when the sample size increased, except for the 3,000 examinee condition for the nonspeeded group. These results

imply that the proportion of the speeded group affects the recovery of the latent group membership for the nonspeeded group even when the sample size is sufficient.

Recovery of the Latent Group Mean for the Nonspeeded Group. The bias and the RMSEs for the latent group mean for the nonspeeded group is presented in Table 3.13. When the sample size was 1,000, both the absolute values for the bias and the RMSEs for the mean ability for the nonspeeded group increased as the proportion of the speeded group increased. When the sample size was 3,000, the absolute values for the bias increased, whereas the RMSEs decreased as the proportion of the speeded group increased.

Table 3.12: Results of Recovery Analysis: The Unconditional Model

			α			β			Testlet Effects			
Group	N	Proportion	BIAS	RMSE	Corr.	BIAS	RMSE	Corr.	BIAS	RMSE	Corr.	Correct%
Speeded	1000	10%	-0.804	0.974	.548	0.177	0.898	.809	-0.567	1.910	.947	84.90
		30%	-0.468	0.634	.763	-0.055	0.573	.917	-0.617	1.211	.983	92.85
	3000	10%	-0.506	0.682	.718	0.136	0.513	.936	-0.286	1.009	.983	87.57
		30%	-0.257	0.411	.884	-0.181	0.530	.936	-0.744	1.123	.992	93.52
Nonspeeded	1000	10%	0.156	0.246	.966	0.107	0.334	.995	-0.211	0.434	.992	94.36
		30%	0.083	0.244	.952	-0.022	0.306	.993	-0.288	0.644	.981	84.74
	3000	10%	0.222	0.256	.990	0.095	0.243	.999	-0.208	0.401	.993	96.84
		30%	0.061	0.177	.979	-0.116	0.210	.997	-0.310	0.648	.984	82.46
Speeded			-0.509	0.675	.728	0.019	0.629	.900	-0.554	1.313	.976	89.71
Nonspeeded			0.131	0.231	.972	0.016	0.273	.996	-0.254	0.532	.988	89.60
	1000		-0.258	0.525	.807	0.052	0.528	.929	-0.421	1.050	.976	89.21
	3000		-0.120	0.382	.893	-0.017	0.374	.967	-0.387	0.795	.988	90.09
		10%	-0.233	0.540	.806	0.129	0.497	.935	-0.318	0.939	.979	90.91
		30%	-0.145	0.367	.895	-0.094	0.405	.961	-0.490	0.907	.985	88.39

Table 3.13: Recovery Analysis of Latent Group Mean: The Unconditional Model

		Group Mean	
N	Proportion	BIAS	RMSE
1000	10%	-0.018	0.078
	30%	0.027	0.132
3000	10%	0.004	0.134
	30%	0.006	0.036
1000		-0.017	0.334
3000		0.005	0.085
	10%	-0.007	0.106
	30%	0.017	0.084

Simulation Results of the Conditional Model

The bias, RMSEs, the correlations between the estimates and the true parameters, and the percentages of the correct classification of the latent group memberships for the speeded and nonspeeded group obtained from the conditional model are presented in Table 3.14. In general, similar to the result from the unconditional model, the absolute values for bias and RMSEs for the speeded group were larger than those for the nonspeeded group, and the correlations for the speeded group were larger than those for the nonspeeded group. Contrary to the unconditional model, the percentages of correct membership for the speeded group were smaller than those for the nonspeeded group.

Recovery of the Item Discrimination Parameters. The absolute values for bias and RMSEs for the item discrimination parameters for the speeded group decreased as the sample size increased or the proportion of the speeded group increased. For the nonspeeded group, however, there was no specific pattern in the absolute values for bias and RMSEs. The absolute values for bias of the item discrimination parameters for the speeded group, which ranged from 0.215 to 1.035, were larger than those for the nonspeeded group, which ranged from 0.158 to 0.220. The RMSEs for the item discrimination parameters for the speeded

group, which ranged from 0.386 to 1.251, were also larger than those for the nonspeeded group, which ranged from 0.255 to 0.376.

Again, the correlations for both the speeded and nonspeeded groups increased as the sample size increased. When the proportion of the speeded group increased, the correlations for the speeded group increased, whereas the correlations for the nonspeeded group decreased. As noted earlier, the proportion of the speeded group is related to the sample size of each group. To be more precise, the sample for the speeded group increased and the sample for the nonspeeded group decreased when the proportion of the speeded group increased. These results, therefore, indicate that the recovery of the correlations of the item discrimination parameters is affected by the sample size. It should be noted that the correlations for the speeded group were lower than .9 under all conditions except for one condition (i.e., 3,000 examinees and 30% speeded group), whereas the correlations for the nonspeeded group were higher than .9 under all conditions. The correlations for the speeded group, which ranged from .513 to .927, were smaller than those for the nonspeeded group, which ranged from .934 to .990.

Recovery of the Item Boundary Location Parameters. The absolute values for the bias and the RMSEs for the item boundary location parameters for both the speeded and nonspeeded groups decreased as either the sample size increased or as the proportion of the speeded group increased, except in one condition for the nonspeeded group (i.e., 3,000 examinees). When the sample size was 3,000, the absolute value and the RMSE for the nonspeeded group increased as the proportion of the speeded group increased from 10% to 30%, which was similar to the pattern identified from the bias and RMSE values for the item discrimination parameters. The absolute values for bias of the item boundary location parameters for the speeded group, which ranged from 0.038 to 0.749 were larger than those for the nonspeeded group, which ranged from 0.060 to 0.185, except in one condition (i.e., 3,000 examinees and 30% speeded group).

The correlations of the item boundary location parameters for the speeded group increased as the sample size increased or as the proportion of the speeded group increased. For example, when the proportion of the speeded group was 10%, the correlation for the speeded group increased from .854 to .965 as the sample size increased from 1,000 to 3,000. The correlations for the nonspeeded group increased as the proportion of the speeded group increased, only when the sample size was 1,000. The correlations for the nonspeeded group remained the same, when the sample size was 3,000, as the proportion of the speeded group increased. Again, these differences between the correlations for the nonspeeded group were trivial because all the correlations for the nonspeeded group were higher than .99 under all conditions. The correlations for the speeded group, which ranged from .854 to .990, were smaller than those for the nonspeeded group, which ranged from .990 to .998.

Recovery of the Testlet Effects. The absolute values for the bias of the testlet effects for the speeded group increased as the proportion of the speeded group increased, while those for the nonspeeded group decreased under a 1,000 examinee condition or increased under a 3,000 examinee condition as the proportion of the speeded group increased. The absolute values for the bias of the testlet effects for the speeded group increased under the 10% speeded group condition or decreased under the 30% speeded group condition as the sample size increased. The absolute values for the bias of testlet effects for the nonspeeded group increased as the sample size increased. The range of the absolute values for the bias of the speeded group was from 0.001 to 0.572 and the range for the nonspeeded group was from 0.148 to 0.215.

The RMSEs for the testlet effect for the speeded group increased as the proportion of the speeded group increased when the sample size was 1,000 and decreased when the sample size was 3,000. The RMSEs for the nonspeeded group increased as the proportion of the speeded group increased under both the 1,000 and 3,000 examinee conditions. The RMSEs for the speeded group decreased as the sample size increased, whereas those for the nonspeeded group increased under the 10% speeded group condition and decreased under the 30% speeded group condition as the sample size increased. The RMSEs for the speeded

group, which ranged from 0.510 to 2.056, were larger than those for the nonspeeded group, which ranged from 0.367 to 0.503. Similar to the result from the unconditional model, the RMSEs for the speeded group under the 1,000 examinee condition were larger than 1, which implies the estimates for testlet effects might be unreliable under this condition.

When the sample size was 1,000, the correlations of the testlet effects for the speeded group were lower than those for the nonspeeded group. When the sample size was 3,000, the correlations of both groups were similar to each other. The correlations for both groups increased as the proportion of the speeded group increased except in one condition for the nonspeeded group (i.e., 10% speeded group). The correlations for the speeded group ranged from .886 to .996 and those for the nonspeeded group ranged from .987 to .995.

Recovery of Latent Group Membership. Similar to the result from the unconditional model, the percentages of the correct classification for the speeded group increased and those for the nonspeeded group decreased as the proportion of the speeded group increased. The percentages for both groups increased as the sample size increased. The percentages for the speeded group were larger than .90 only when the proportion of the speeded group was 30%, whereas those for the speeded group were larger than .90 under all conditions. These results suggest that the successful recovery of the latent group membership requires a higher proportion of the speeded group even when the sample size is satisfactory.

Recovery of the Latent Group Mean for the Nonspeeded Group. The bias and the RMSEs for the latent group mean for the nonspeeded group are provided in Table 3.15. Both the absolute values for the bias and the RMSEs for the mean ability for the nonspeeded group increased as the proportion of the speeded group increased, when the sample size was 1,000. On the contrary, both the absolute values for the bias and the RMSEs for the nonspeeded group decreased as the proportion of the speeded group increased, when the sample size was 3,000.

Table 3.14: Results of Recovery Analysis: The Conditional Model

			α			β			Testlet Effects			
Group	N	Proportion	BIAS	RMSE	Corr.	BIAS	RMSE	Corr.	BIAS	RMSE	Corr.	Correct%
Speeded	1000	10%	-1.035	1.251	.513	0.749	1.139	.854	-0.001	1.635	.932	80.20
		30%	-0.673	0.931	.764	0.282	0.642	.924	-0.572	2.056	.886	90.85
	3000	10%	-0.572	0.728	.855	0.252	0.464	.965	-0.029	0.857	.984	85.32
		30%	-0.215	0.386	.927	0.038	0.215	.990	-0.216	0.510	.996	92.92
Nonspeeded	1000	10%	0.158	0.229	.975	0.152	0.383	.997	-0.180	0.367	.995	97.83
		30%	0.194	0.376	.934	0.185	0.620	.990	-0.148	0.503	.987	93.35
	3000	10%	0.220	0.256	.990	0.118	0.298	.998	-0.189	0.371	.994	97.87
		30%	0.180	0.233	.985	0.060	0.239	.998	-0.215	0.436	.992	93.74
Speeded			-0.624	0.824	.765	0.330	0.615	.933	-0.205	1.265	.950	87.32
Nonspeeded			0.188	0.274	.971	0.129	0.385	.996	-0.183	0.419	.992	95.70
	1000		-0.339	0.697	.797	0.342	0.696	.941	-0.225	1.140	.950	90.56
	3000		-0.097	0.401	.939	0.117	0.304	.988	-0.162	0.544	.992	92.46
		10%	-0.307	0.616	.833	0.318	0.571	.954	-0.100	0.808	.976	90.30
		30%	-0.129	0.482	.903	0.141	0.429	.976	-0.288	0.876	.965	92.71

Table 3.15: Recovery Analysis of Latent Group Mean: The Conditional Model

		Group Mean	
N	Proportion	BIAS	RMSE
1000	10%	-0.020	0.062
	30%	0.052	0.166
3000	10%	0.006	0.031
	30%	-0.001	0.027
1000		-0.214	0.429
3000		0.003	0.029
	10%	-0.007	0.047
	30%	0.026	0.097

Summary of Recovery Analysis

The RMSE values for the unconditional and conditional models are presented in Figures 3.5 and 3.6, respectively. For both the unconditional and conditional models, the RMSEs for the speeded group were larger than those for the nonspeeded group. For the unconditional model, the RMSEs for the item discrimination and item location boundary parameters are similar, whereas the RMSEs for the testlet effects were larger than the RMSEs for the item parameters. For the conditional model, a similar pattern is identified, whereas the RMSEs for the item discrimination, item location boundary parameters, and testlet effect were comparable only for the nonspeeded group. The RMSEs for the mean ability parameter for the nonspeeded group were much smaller compared to the other RMSEs in both models.

The correlations between the generating and estimated parameters are illustrated in Figures 3.7 and 3.8. Again, the pattern was similar in both the unconditional and conditional models. In both models, the correlations for the item discrimination parameters for the speeded group were relatively smaller than the others, whereas the correlations for the item

boundary location parameters and testlet effects were close to 1 regardless of the latent group membership.

Comparison of Unconditional and Conditional Models

The RMSE values for the item discrimination, item boundary location parameters, and testlet effects for the speeded and nonspeeded group are presented in Figures 3.9 to 3.11. For the item discrimination parameters, the RMSEs from both the unconditional and conditional models were similar except for those for the speeded group under the 1,000 examinee condition. For the item boundary location parameters, the RMSEs for the speeded group from the unconditional model were smaller than those from the conditional model when the sample size was 1,000, whereas the RMSEs for the speeded group from the unconditional model were larger when the sample size was 3,000. The RMSEs for the nonspeeded group from the unconditional model were smaller than those from the conditional model under all conditions. For the testlet effects, the RMSEs for both the speeded and nonspeeded groups were larger than those from the conditional model except in one condition (i.e., 1,000 examinees and 30% speeded group).

Also, the RMSE values for the mean ability for the nonspeeded group are plotted in Figure 3.12. The RMSEs for the mean ability for the nonspeeded group from the unconditional model were larger than those from the conditional model except in one condition (i.e., 1,000 examinees and 30% speeded group). The difference between two models, however, was negligible except for one condition (i.e., 3,000 examinees and 10% speeded group).

The percentages of correct detection of latent class memberships are presented in Figure 3.13. The percentages for the speeded group from the unconditional model were larger than those from the conditional model under all conditions, whereas the percentages for the nonspeeded group from the unconditional model were smaller than those from the conditional model.

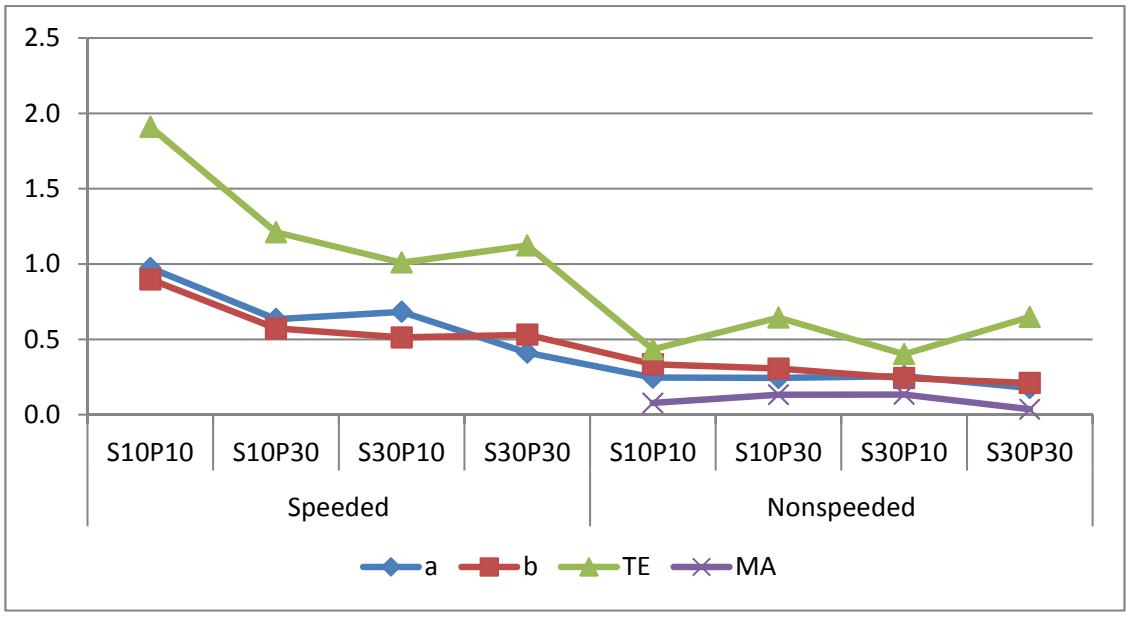

Figure 3.5: RMSE values for recovery analysis: The unconditional model

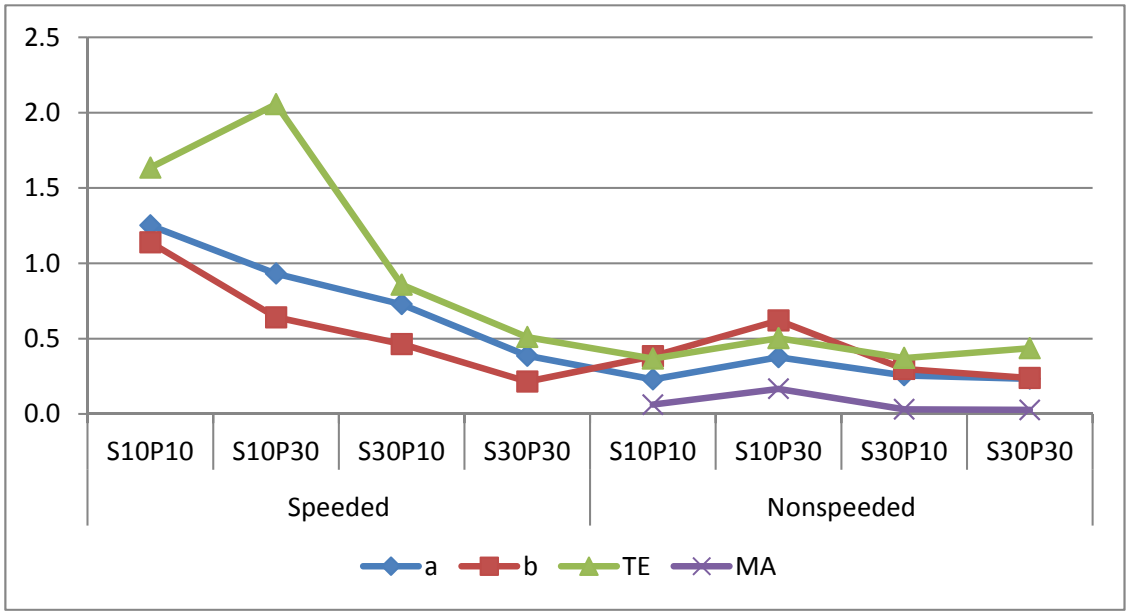

Figure 3.6: RMSE values for recovery analysis: The conditional model

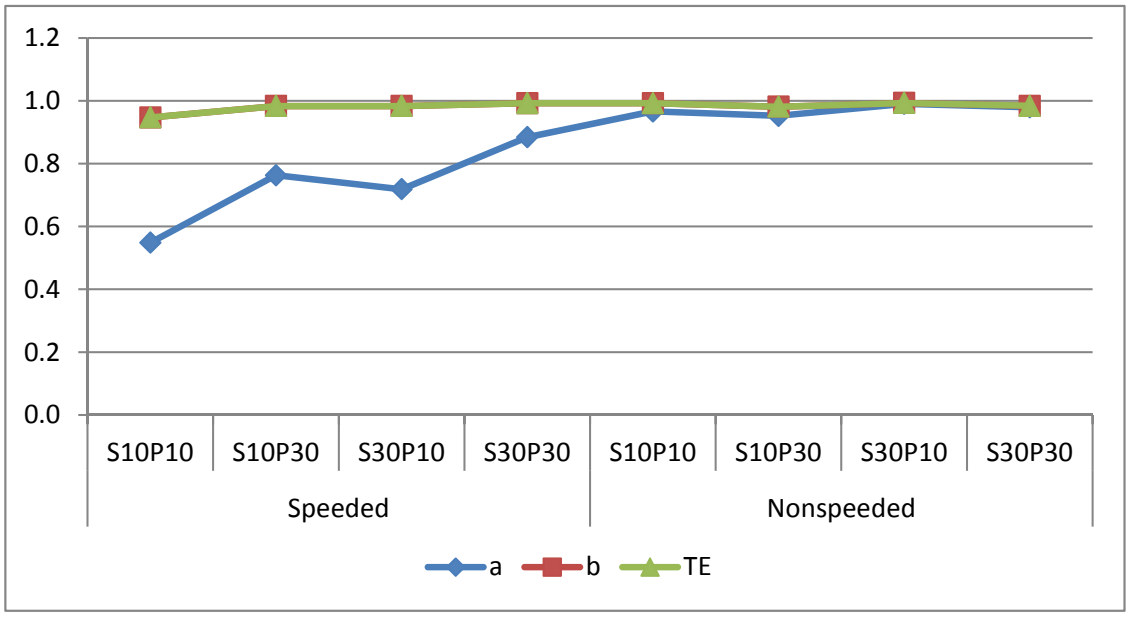

Figure 3.7: Correlations between true and estimated values: The unconditional model

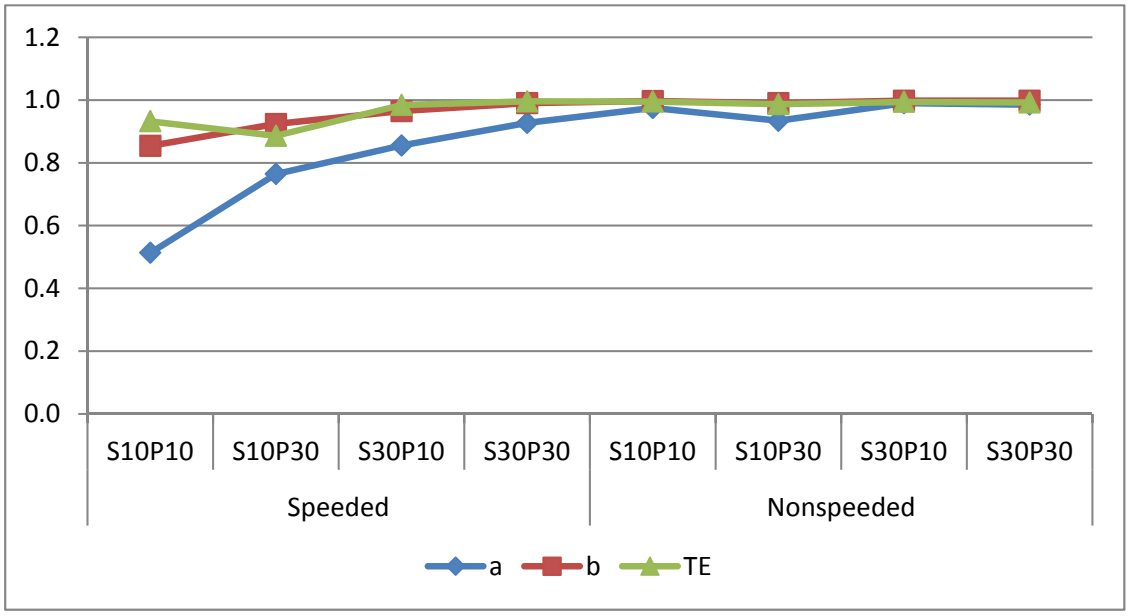

Figure 3.8: Correlations between true and estimated values: The conditional model

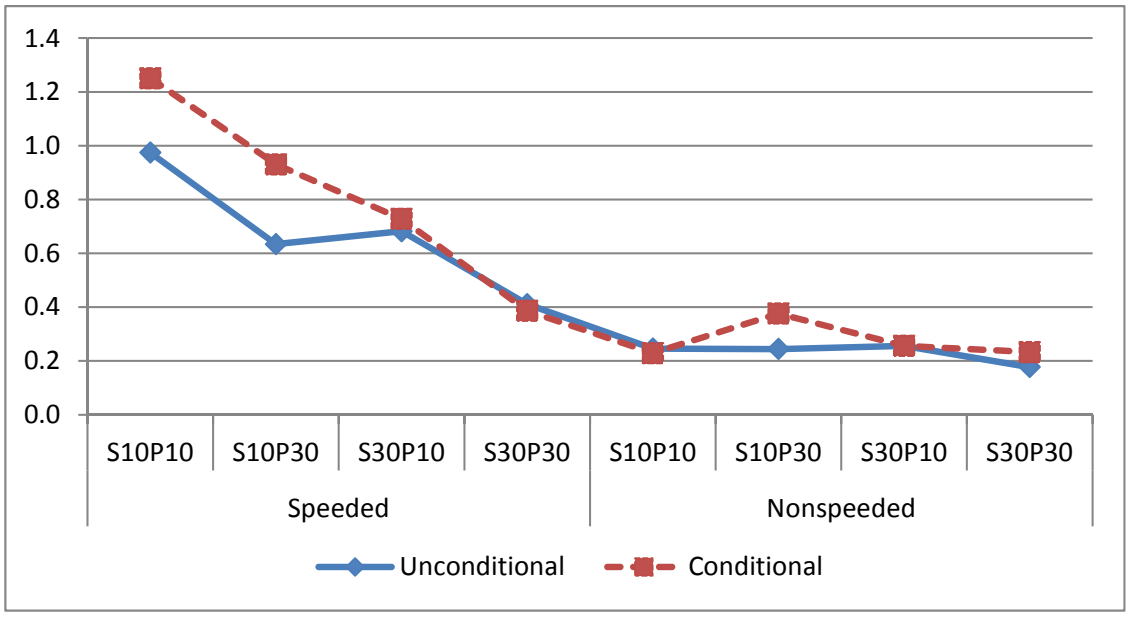

Figure 3.9: RMSE values for recovery analysis: Item discrimination parameters

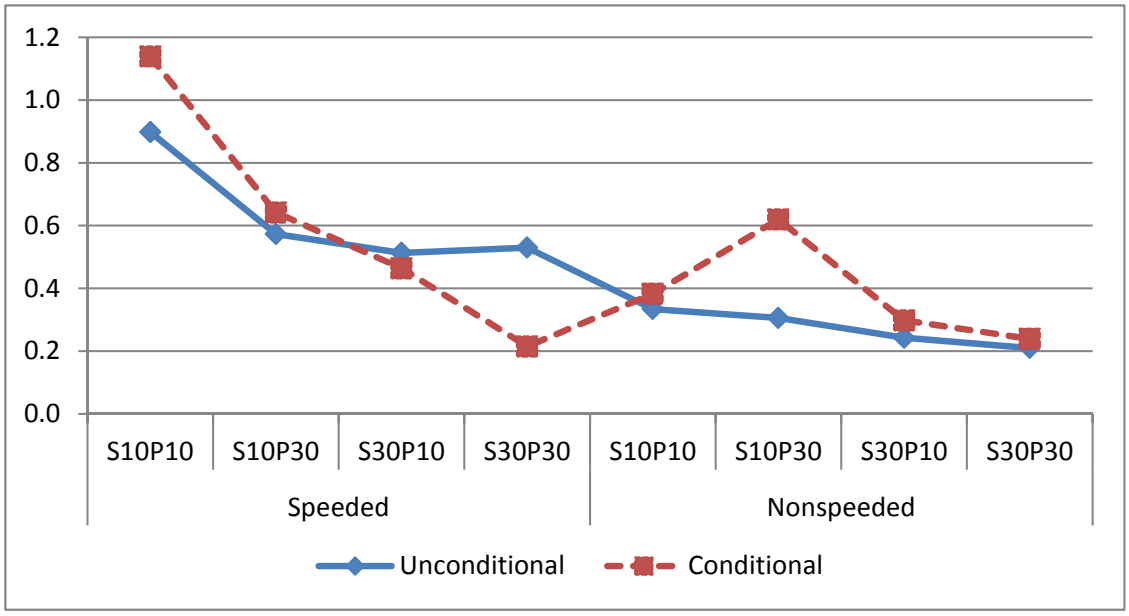

Figure 3.10: RMSE values for recovery analysis: Item boundary location parameters

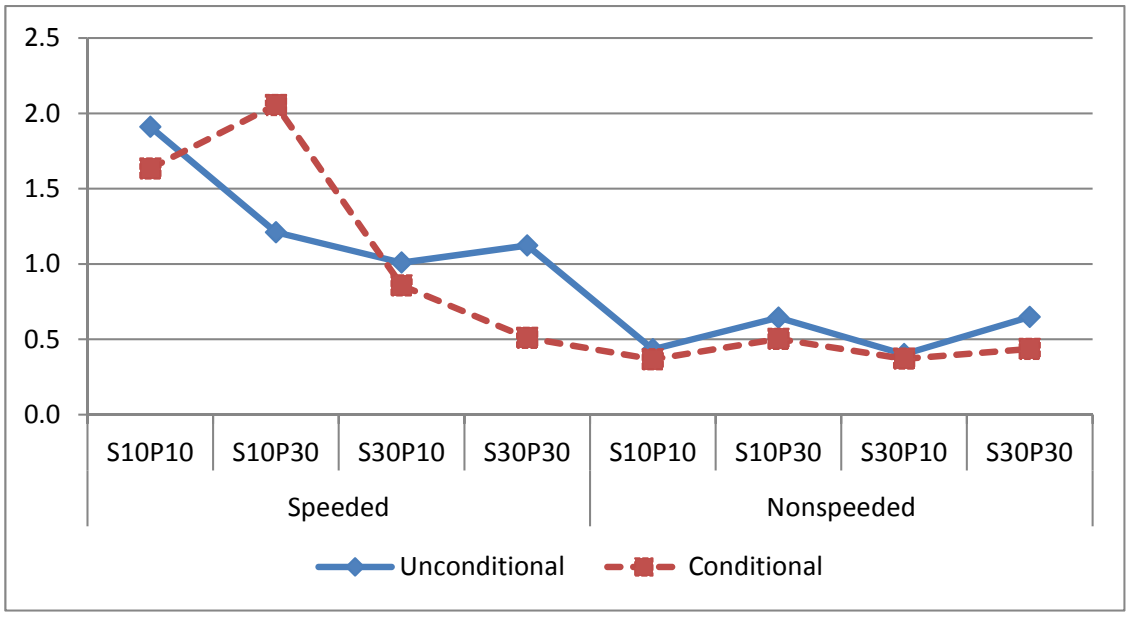

Figure 3.11: RMSE values for recovery analysis: Testlet effects

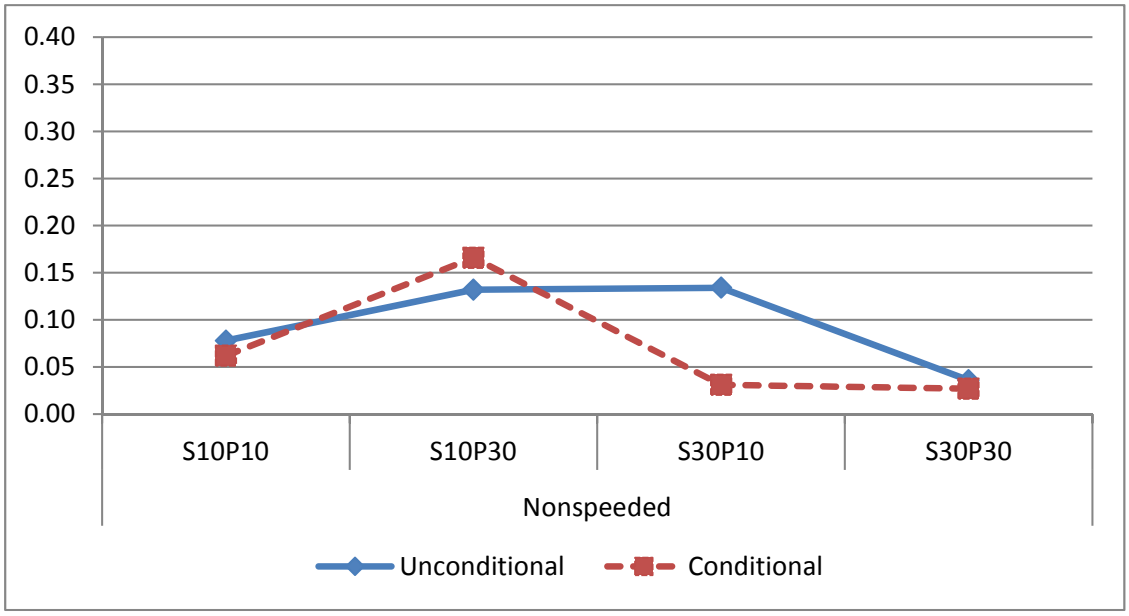

Figure 3.12: RMSE values for recovery analysis: Mean ability

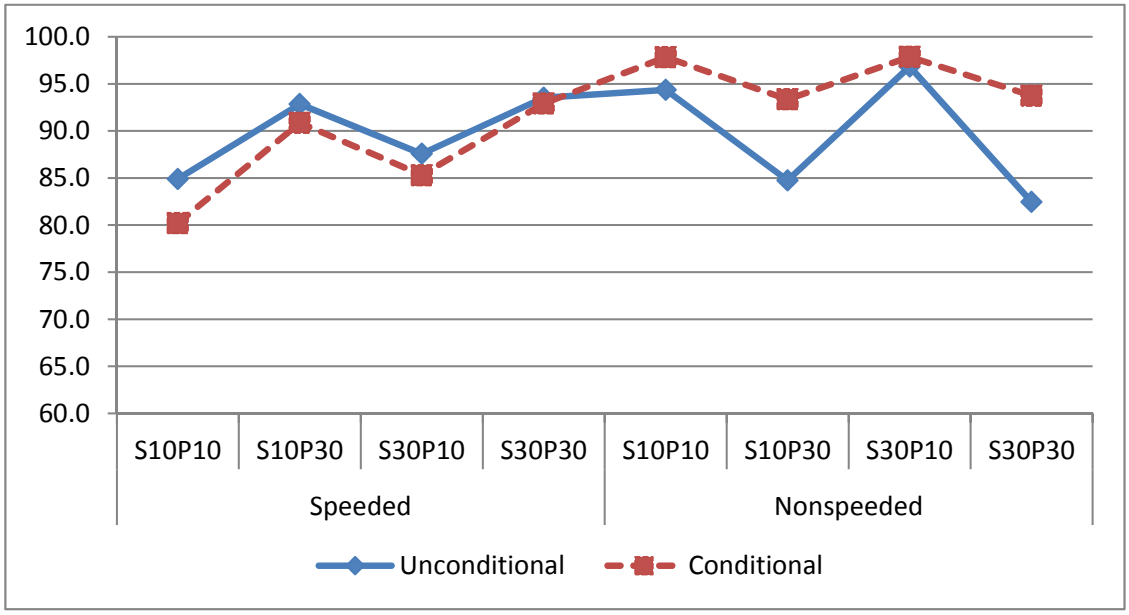

Figure 3.13: Percentages of correct detection of latent group membership

3.5 Summary

The purpose of this study was to examine the performance of the RIMGRM-t for the detection of test speededness. The RIMGRM-t can be considered as a random item mixture IRT model for testlets. Two RIMGRM-t's, the unconditional and conditional models, were compared to examine the effects of item covariates on item boundary location parameters. The data used in this study were obtained from CR items on a test designed to measure middle and high school students' knowledge of science inquiry practices. Each CR item was scored with multiple rubrics, and the multiple scores from one answer were treated as individual item responses in a testlet.

The results from the real data analysis suggested that item boundary location parameter estimates from the unconditional and conditional models were similar to each other. The mixing proportions (i.e., the latent class membership proportions) in the two latent classes

indicated that both unconditional and conditional models had similar results. Further, most examinees were assigned to the same latent group by both models. In addition, the testlet effect estimates from both models had similar patterns. These results suggested that the performance of the unconditional and conditional models in the detection of the speededness in CR items was comparable. The effects of the item covariates on item boundary location parameters, however, were different depending on the latent class.

A simulation study was also conducted to compare the performance of the unconditional and conditional RIMGRM-t's under various testing conditions. Two factors, the sample size and the proportion of the speeded group, were considered along with the difference between the unconditional and conditional models. The bias, RMSEs, and correlations were used to evaluate the recovery. The results from the recovery analysis indicated that the recovery from both the unconditional and conditional models was acceptable, and a sufficient amount of the sample size is required to recover model parameters, especially testlet effects. The percentages of the correct classification were also similar in both models.

Chapter 4

Exploring Characteristics of Speeded Examinees Using LDA

This chapter considers the application of latent Dirichlet allocation (LDA; Blei et al., 2003) for characterizing latent class membership to see the difference between the speeded and nonspeeded groups. A mixture IRT (MixIRT; Mislevy & Verhelst, 1990; Rost, 1990) model was implemented in the first two chapters because this model is one common way to detect speededness. It is not possible, however, to obtain a qualitative explanation about each latent class from this model, since the mixture IRT model classifies examinees based on different item performance (e.g., Cho et al., 2010). Additional analysis is required, therefore, to see the attribute of each latent class from MixIRT models.

One way to describe the characteristics of each latent class is to examine the association between latent class membership and manifest characteristics such as gender or ethnicity (e.g., Bolt et al., 2002). Bolt et al. used the mixture Rasch model to detect speededness in a college-level mathematics placement examination. The results suggested that approximately 24% of examinees were classified as speeded and 76% as nonspeeded. Differences in the gender and ethnicity characteristics of the speeded and nonspeeded groups were analyzed using Pearson chi-square tests for a randomly selected sample of 1,000 examinees. Bolt et al. found that the association between gender and speededness was not significant ($\chi^2 = 2.37, p = .124$), but the association between ethnicity and speededness was ($\chi^2 = 3.92, p = .048$).

Another way to characterize differences between latent groups is to interview examinees after a test (e.g., Izsák et al., 2010). Izsák et al. used a mixture Rasch model to analyze the responses on a multiple-choice test designed to measure the mathematical knowledge necessary for teaching arithmetic. The mixture Rasch model detected two latent groups.

Results from follow-up interviews were used to determine the different aspects of the latent groups. The authors interviewed 16 of the 201 teachers in the sample—seven belonged to the one latent group and nine to the other latent group. The follow-up interviews revealed important differences between the two groups. For example, the teachers in the first group identified "appropriate reference units for numbers and parts-of-parts of quantities more consistently" (p. 293) than the teachers in the second group. The number of interviewees, however, was very small relative to the size of the total sample, and represents a drawback of this approach. That is, this approach permits use of interview data from only a handful of examinees because of limited time and resources.

In this dissertation, a third method, the LDA model, was investigated. LDA is a generative probabilistic model and is used to uncover latent topics in text corpora. It is worth noting that the analysis unit of the LDA model is a word, which enables the analysis of the text from examinees' responses to the constructed response (CR) items. There are a number of advantages derived from using LDA to analyze text data. First, it is not necessary for this method to have additional demographic information, whereas the additional analysis in Bolt et al. (2002) required manifest information. Second, this approach includes all of the examinees in the analysis, whereas the number of examinees is limited when using interviews. Third, and possibly most importantly, this process analyzes examinees' written responses to the CR items, whereas the MixIRT analyzes only the item scores. In this study, a two-class mixture graded response testlet (MixGRM-t) model was used to classify examinees as the speeded or nonspeeded group as the item response data are polytomous. Then the text of each latent group's responses was characterized using the LDA model.

4.1 Theoretical Framework

4.1.1 Mixture Testlet Graded Response Model

The mixture testlet graded response model (MixGRM-t) can be considered as a combination of a mixture IRT model (Rost, 1990), a graded response model (GRM; Samejima, 1969), and

a testlet IRT model (Wainer & Kiely, 1987). The features of each model are connected to the characteristics of the data. First, a mixture IRT model separates examinees into latent classes based on their responses. In this study, examinees were assigned into two latent classes, the speeded and nonspeeded groups, with the assumption that some of the examinees' responses were affected by test speededness. Second, GRM can be seen as an extension of the two-parameter logistic model to analyze the ordered categories. In this study, the CR items were scored in two or more ordered categories, which suggests that it is appropriate to analyze the data using a polytomous IRT model. Lastly, the testlet IRT model accounts for the local dependency which comes from the testlet structure by adding a person-specific testlet effect. The responses to CR items are sometimes scored by multiple rubrics in order to make multiple inferences and interpretations (Ercikan, 2002). In this case, the individual scores from one response can be regarded as nested in a testlet.

GRM specifies the probability of obtaining a score of k using cumulative probability (i.e. the probability of obtaining a score of k or higher) which is also called a boundary probability. The probability for examinee j in latent group g of obtaining category k or higher under the mixture testlet GRM (GRM-t) is given by

$$P^*_{ijgk} = P(u_{ijg} = k|\theta_{jg}, \alpha_{ig}, \beta_{igk}, \gamma_{jt(i)}, g) = \frac{\exp[\alpha_{ig}(\theta_{jg} - \beta_{igk} - \gamma_{jt(i)})]}{1 + \exp[\alpha_{ig}(\theta_{jg} - \beta_{igk} - \gamma_{jt(i)})]},$$

where

- g: an index for latent group, $g = 1, 2$,
- u_{ijg}: the response of examinee j in class g to item i,
- θ_{jg}: the ability parameter of examinee j in class g,
- α_{ig}: the item discrimination parameter of item i for class g,
- β_{igk}: the item difficulty parameter of item i of category k for class g,
- $\gamma_{jt(i)}$: a person-specific testlet effect representing the interaction of examinee j with testlet $t(i)$ (i.e., testlet t includes item i).

By definition, the probability of getting a score of zero or higher is 1 (i.e., $P^*_{ijg0} = 1$) and the probability of getting a score of m_i+1 or higher is 0 ($P^*_{ijg,m_i+1} = 0$) where m_i is the maximum

category of item i. The probability of getting a score of k is defined as the difference between adjacent boundary probabilities, which is expressed as

$$P_{ijgk} = P^*_{ijg,k-1} - P^*_{ijgk},$$

where $k = 0, \ldots, m_i$. Since this study assumes that students belong to either the speeded or nonspeeded group, this model becomes a two-class MixGRM-t (i.e., $g = 1, 2$).

4.1.2 Latent Dirichlet Allocation

Several studies have been conducted to describe the characteristics of a collection of documents in an efficient way. The LDA model also comes from the same motivation. Compared with other methodologies for text corpora, however, this model involves mixture distributions for single words contained in the document (Blei et al., 2003). To be specific, the basic ideas of LDA are that each document is assumed to be a mixture of topics and also of words; the documents are assumed to be generated from these topics. The LDA model can be seen as a hidden variable model. That is, in an LDA analysis, "the observed data are the words of each document and the hidden variables represent the latent topical structure" (Blei & Lafferty, 2009, p. 73).

The probabilistic generative process of the LDA model can be expressed as follows (Blei & Lafferty, 2009).

1. For each topic $k = 1, \ldots, K$
   (a) Draw a distribution over words $\phi_k \sim \text{Dirichlet}(\beta)$.
2. For each document $d = 1, \ldots, D$
   (a) Draw a vector of topic proportions $\theta_d \sim \text{Dirichlet}(\alpha)$.
   (b) For each word $i = 1, \ldots, N_d$
      i. Draw a topic assignment $z_i \sim \text{Multinomial}(\theta_d)$, $z_i \in 1, \ldots, K$.
      ii. Draw a word $w_i \sim \text{Multinomial}(\phi_{z_i})$, $w_i \in 1, \ldots, V$.

where K is a specified number of topics, V is the number of words in the vocabulary, Dirichlet(β) is a K-dimensional Dirichlet, Dirichlet(α) is a V-dimensional Dirichlet, N_d is the individual document length, z_i is a topic index, and w_i is a word index.

The Dirichlet distribution which plays an important role in the LDA model has the following density:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1},$$

where Γ is the Gamma function which can be regarded as a real-valued extension of the factorial function. According to Blei et al. (2003), "the Dirichlet is a convenient distribution on the simplex—it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution" (p. 996).

Based on the probabilistic generative process, the joint distribution of the random variables (w, z, ϕ, θ) is expressed as

$$p(w, z, \phi, \theta|\alpha, \beta) \propto p(\phi|\beta)p(\theta|\alpha)p(w|z, \phi)p(z|\theta).$$

Then, the marginal probability of the corpus A (i.e., the D observed documents) given α and β is obtained by integrating over both ϕ and θ and summing over z; it is written as

$$p(A|\alpha, \beta) = \int_{\phi} \int_{\theta} \sum_{z} p(\phi|\beta)p(\theta|\alpha)p(w|z, \phi)p(z|\theta)d\theta d\phi.$$

Since this probability is not computationally intractable in general (Blei et al., 2003), some approximations of the marginal probability, including variational inference (Blei et al., 2003), collapsed variational inference (Teh, Newman, & Welling, 2006) and collapsed Gibbs sampling (Griffiths & Steyvers, 2004), have been developed.

4.2 Methods

4.2.1 Data

For the MixGRM-t, the data were taken from the responses of 876 middle school students to a constructed response test that was designed to examine middle school students' under-

standing of independent and dependent variables, cause and effect, and hypothesis testing. The test consisted of six CR items and was scored with four different rubrics: science inquiry practices, use of everyday language, use of academic language, and science content. This process produced 33 scores with scores for each item ranging from 0 to 4.

For the LDA model, the texts of the responses of 137 of the 876 students were analyzed to help sort examinees into speeded and nonspeeded latent classes. The corpus consisted of 137 answer documents and 631 words. Words that occurred in only a few documents or those that were considered to be "stop" words were removed by using TF-IDF (Salton & McGill, 1986), reducing the number of words to 572. Stop words are those that appear with high frequency but contribute little to the meaning. These words are context dependent but typically include words such as *if*, *and*, *a*, and *the*.

4.2.2 Modeling Speededness in Mixture Testlet GRM

To model test speededness, we assumed that the speeded examinees were not affected by speededness at the beginning of the test, whereas they were affected by speededness at the end of the test. Based on this assumption, the item discrimination and boundary location parameters of the first two testlets (Testlet 1 and Testlet 2) were constrained to be equal. Also, the item boundary location parameters for the speeded group in the last three testlets (Testlet 4 to Testlet 6) were constrained to be larger than those for the nonspeeded group. The model parameters of the third testlet were unconstrained and freely estimated.

4.2.3 Procedure

The constrained two-class MixGRM-t was applied to classify the examinees as the speeded or nonspeeded group using the total sample which included the sample of the LDA analysis, because the sample for the LDA analysis was insufficient to estimate the latent group membership using a MixGRM-t. The latent group membership for individual students was then

used to classify students who belonged to the sample for the LDA analysis. This made it possible to compare the proportions of topics used by the speeded and nonspeeded groups.

Then, the LDA model was used to extract latent topics used by examinees in their written responses, which were part of the total sample. Cluster analysis of the proportions of topic usage for each person and each group was used to characterize the latent groups. For cluster analysis, hierarchical clustering was used to explore the proportions for topic usages for each topic to set the number of clusters. K-means was then used to identify the patterns in the proportions of topic usage for the speeded and nonspeeded groups.

4.2.4 Estimation

Mixture Testlet GRM

A Markov chain Monte Carlo (MCMC) algorithm using Gibbs sampling, as implemented in the computer software OpenBUGS (Thomas et al., 2006), was used to estimate the parameters of the MixGRM-t. The MCMC algorithm samples a Markov chain repeated from the full conditional posterior distributions over a large number of iterations. After a sufficiently long burn-in, the Markov chain is assumed to reach its stationary distribution. The subsequent sampling becomes dependent on this stationary distribution. Then, the sample mean (or mode) obtained from the remaining iterations can be used as the parameter estimate from the posterior.

Under the MCMC algorithm, the mixture distribution is estimated by sampling a class membership parameter for each examinee at each stage of the chain (Bolt et al., 2002). Once a class membership parameter and an examinee's ability parameter are sampled, the class parameters, which include item parameters and the mean and variance of ability parameters for each class, are sampled conditional on the class membership and ability parameters.

It is necessary to specify prior distributions for parameters that are estimated. The following priors were used for MixGRM-t:

$$
\begin{aligned}
a_{ig} &\sim N(0,1)I(0,), && i = 1,\ldots,n,\ g = 1,2,\\
b_{igk} &\sim N(\bar{\beta}_g, 1), && i = 1,\ldots,n,\ g = 1,2,\ k = 1,\ldots,m_i,\\
\gamma_{jt(i)} &\sim N(\mu_g, \sigma^2_\gamma), && j = 1,\ldots,N,\ t = 1,\ldots,T,\\
\theta_{jg} &\sim N(\mu_g, 1), && j = 1,\ldots,N,\ g = 1,2,\\
(\pi_1, \pi_2) &\sim \mathit{Dirichlet}(0.5, 0.5),
\end{aligned}
$$

where n is the total number of items, N is the total number of examinees, and $I(0,)$ indicates that observations of a_{ig} are sampled above zero. Hyperparameters in this analysis were selected to be noninformative:

$$
\begin{aligned}
\mu_g &\sim N(0,1), && g = 1,2,\\
\bar{\beta}_g &\sim N(0,1), && g = 1,2,\\
\sigma^2_\gamma &\sim \mathit{Inv\text{-}Gamma}(2.5, 0.25).
\end{aligned}
$$

The convergence was checked by using Heidelberger and Welch (1983) convergence diagnostics as implemented in the CODA package (Plummer et al., 2006).

LDA

The Gibbs sampling (Griffiths & Steyvers, 2004), as implemented in the *topicmodels* package written in R (Grün & Hornik, 2011), was used to fit the LDA model. Under Gibbs sampling, all variables are sampled from their full conditional distributions conditioning on the current state of all other variables and data (Griffiths & Steyvers, 2004). When the data come from D documents involving K topics over W unique words, the full conditional distribution is expressed as

$$
P(z_i = j | \boldsymbol{z}_{-i}, \boldsymbol{w}) \propto \left[\frac{n^{(w_i)}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j} + W\beta} \right] \left[\frac{n^{(d_i)}_{-i,j} + \alpha}{n^{(d_i)}_{-i,\cdot} + K\alpha} \right],
$$

where z_i is the latent variable indicating the current topic membership of the ith word, $\boldsymbol{z}_{-i}$ is the current topic membership for all words except for the ith word, $\boldsymbol{w}$ is the corpus which consists of words w_i's (i.e., $\boldsymbol{w} = \{w_1, \ldots, w_n\}$), $n_{-i,j}^{(w_i)}$ is the frequency of that word w_i that has been classified as topic j without the current membership, $n_{-i,j}^{(\cdot)}$ is the sum of the frequencies of all the words assigned to topic j without the current membership, $n_{-i,j}^{(d_i)}$ is the frequency of that document d_i that has been classified as topic j without the current membership, and $n_{-i,\cdot}^{(d_i)}$ is the sum of the frequencies of that d_i that has been classified as any topics without the current membership. This expression, therefore, suggests that the full conditional distribution is proportional to the two probabilities: the probability of word w_i assigned to topic j under all words assigned to topic j and the probability of topic k under document d_i (Griffiths & Steyvers, 2004).

The Dirichlet priors of α and β are used in the estimation of the posterior distributions. There does not yet appear to be a consensus on these priors. Consequently, in this study the priors $50/K$ and $200/W$ were implemented for α and β, respectively, as suggested in Blei et al. (2003).

Removing stop words was done using the term frequency-inverse document frequency (TF-IDF) measures (Blei & Lafferty, 2009). A TF-IDF measure is defined as a multiplication of TF and IDF measures, and it is given by

$$\begin{aligned} \text{TF-IDF}_i &= \text{TF}_i \times \text{IDF}_i, \\ &= \frac{k_{ij}}{K_j} \times \log \frac{N}{n_i}, \end{aligned}$$

where k_{ij} is the frequency of vocabulary t_i in document j, K_j is the total number of vocabulary in document j, n_i is the frequency of documents that contain the term t_i, and N is the total number of documents. A TF-IDF measure can prune out words that have lower frequencies using the first term, a TF measure, as well as words which occurred in lots of documents (e.g., about, the, will), that can be called 'stop' words, using the second term, an IDF measure (Grün & Hornik, 2011).

Choosing a value for the TF-IDF index is somewhat context dependent. For example, Blei and Lafferty (2009) chose the top 10,000 terms by TF-IDF, whereas Grün and Hornik (2011) used an arbitrary number near the median TF-IDF measure as a cut-off score. Similar to Grün and Hornik, this study also removed words that had a TF-DIF measure lower than a cut-off score. The cut-off score of this study was 0.014 and was smaller than the median TF-IDF measure of 0.030. This followed the suggestion from Kwak, Kim, and Cohen (2017) which used a similar data set to this study. The justification of this cut-off score was that it was the minimum TF-IDF number that allowed important academic words (e.g., increase and decrease) to remain in the corpus.

Exploratory LDA analysis was used to determine the best fitting model for the corpus. This was done by fitting different LDA models, each with a pre-specified number of topics. Models were then compared based primarily on interpretability and information criterion indices. This study considered models with two to ten topics and used the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) to select the best fitting model, which seems to be a natural choice in that LDA implements the Bayesian estimation. DIC was computed with the following equation:

$$DIC = \overline{D(\theta)} + p_D,$$

where $\overline{D(\theta)}$ is the posterior mean of the deviance and p_D is the number of effective parameters. The latter term is given by

$$p_D = \overline{D(\theta)} - D(\hat{\theta}),$$

where $D(\hat{\theta})$ is the deviance of the posterior model. This criterion selects the model with the smallest value as the best fitting model.

Table 4.1: Descriptive Statistics for the Female and Male Students

		Testlet 3		Testlet 4		Testlet 5		Testlet 6		Total	
Gender	N	M	SD	M	SD	M	SD	M	SD	M	SD
Female	78	10.04	1.90	12.12	2.43	8.18	2.15	11.04	2.75	41.37	6.72
Male	58	9.52	1.94	11.31	2.99	7.57	1.85	12.17	3.66	40.57	8.75

4.3 Results

4.3.1 Mixture Testlet GRM

Descriptive Analysis

Table 4.1 presents descriptive statistics for the final four testlets for females and males. The means of female students were higher than those of male students for all testlets except for Testlet 7. The total scores of the four testlets were similar to one another. Independent t-tests suggested that the difference between female and male students in Testlet 6 was significant at $\alpha = .05$.

Monitoring Convergence

The Heidelberger and Welch (1983) convergence diagnostic proposed a burn-in length of 5,000 iterations and a post burn-in length of 11,000 iterations for the GRM-t and a burn-in length of 12,000 iterations and a post burn-in length of 2,000 iterations for the MixGRM-t.

Item Parameter Estimates

The discrimination parameter estimates for items in Testlet 3 to Testlet 6 for both the speeded and nonspeeded groups are plotted in Figure 4.1. In general, the discrimination estimates for the speeded group were larger for Testlet 3 and Testlet 4 and those for the nonspeeded group were larger for Testlet 5 and Testlet 6.

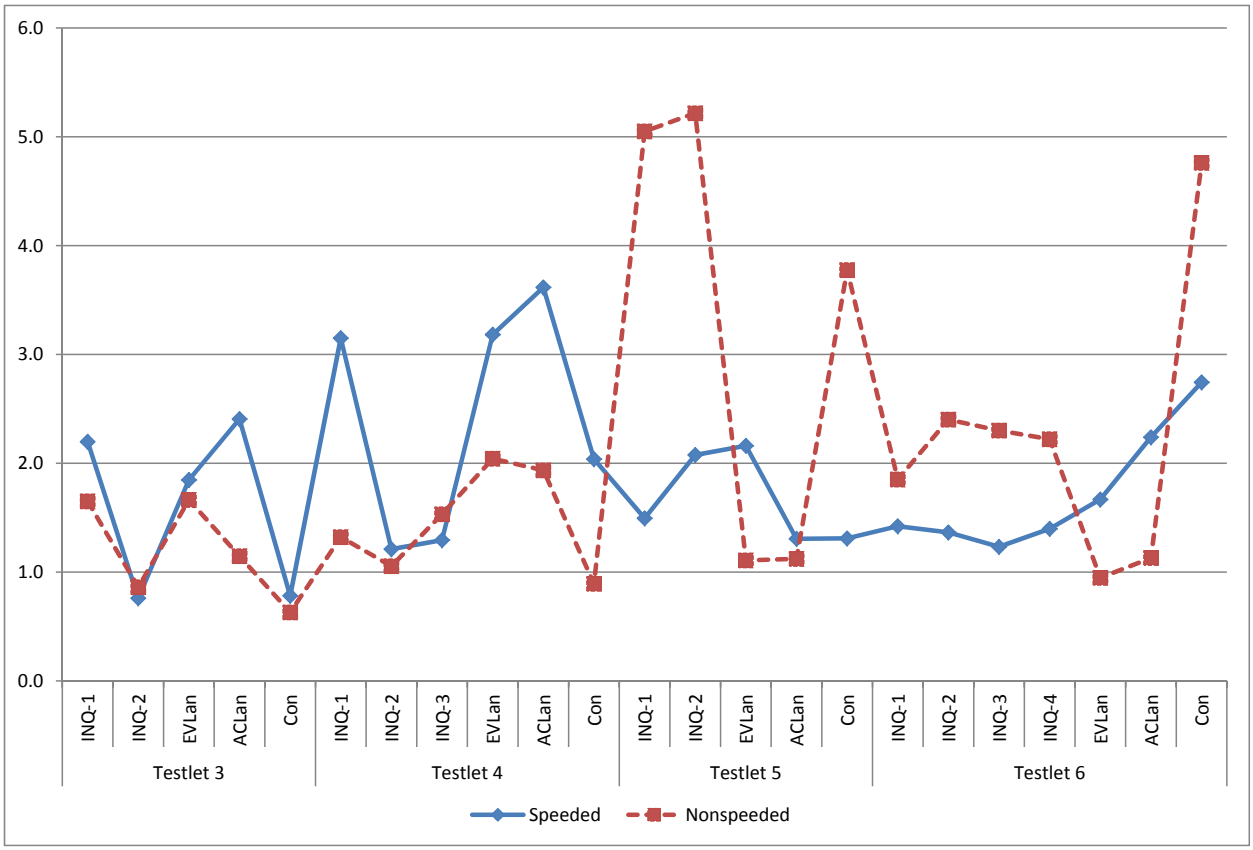

Figure 4.1: Item discrimination parameter estimates

The means of the boundary location parameter estimates for items in Testlet 3 to Testlet 6 for both the speeded and nonspeeded groups are presented in Figure 4.2. Both the item discrimination and boundary location parameter estimates for the speeded and nonspeeded group are presented in Tables 4.2 and 4.3, respectively. In Testlet 3, the means for the speeded group were larger than those for the nonspeeded group at items for Science Inquiry and Everyday Language, but smaller at Academic Language and Science Content. From Testlet 4 to Testlet 6, the means for the speeded group were always larger than those for the nonspeeded group due to the ordinal constraint.

Note that the pattern of the means of the boundary location parameter estimates is different from that seen in previous research (Bolt et al., 2002; Kim et al., 2016). More specifically, the pattern observed in Bolt et al. (2002) and Kim et al. (2016) was that the distance between the speeded and nonspeeded groups increased at the end of the test. In this study, however, the distance between the two groups was already large in the middle of the test (i.e., in Testlet 4) except for items for Academic Language. The distance between the two groups on items for Academic Language was rather small. This suggests that these items were not affected by speededness. Also, the distance between the two groups remained relatively consistent across the last three testlets (i.e., from Testlet 4 to Testlet 6). This suggests that the speededness was already pronounced in the middle of the test.

Testlet Effects

The estimated testlet effects in the speeded and nonspeeded groups are given in Table 4.4. Similar to the testlet effects detected in the previous two studies, the testlet effects in the speeded group were large both at the beginning of the test (Testlet 1 and Testlet 2) and at the end of the test (Testlet 6). The testlet effects in the nonspeeded group remained less than 1 across all the testlets except for the last testlet.

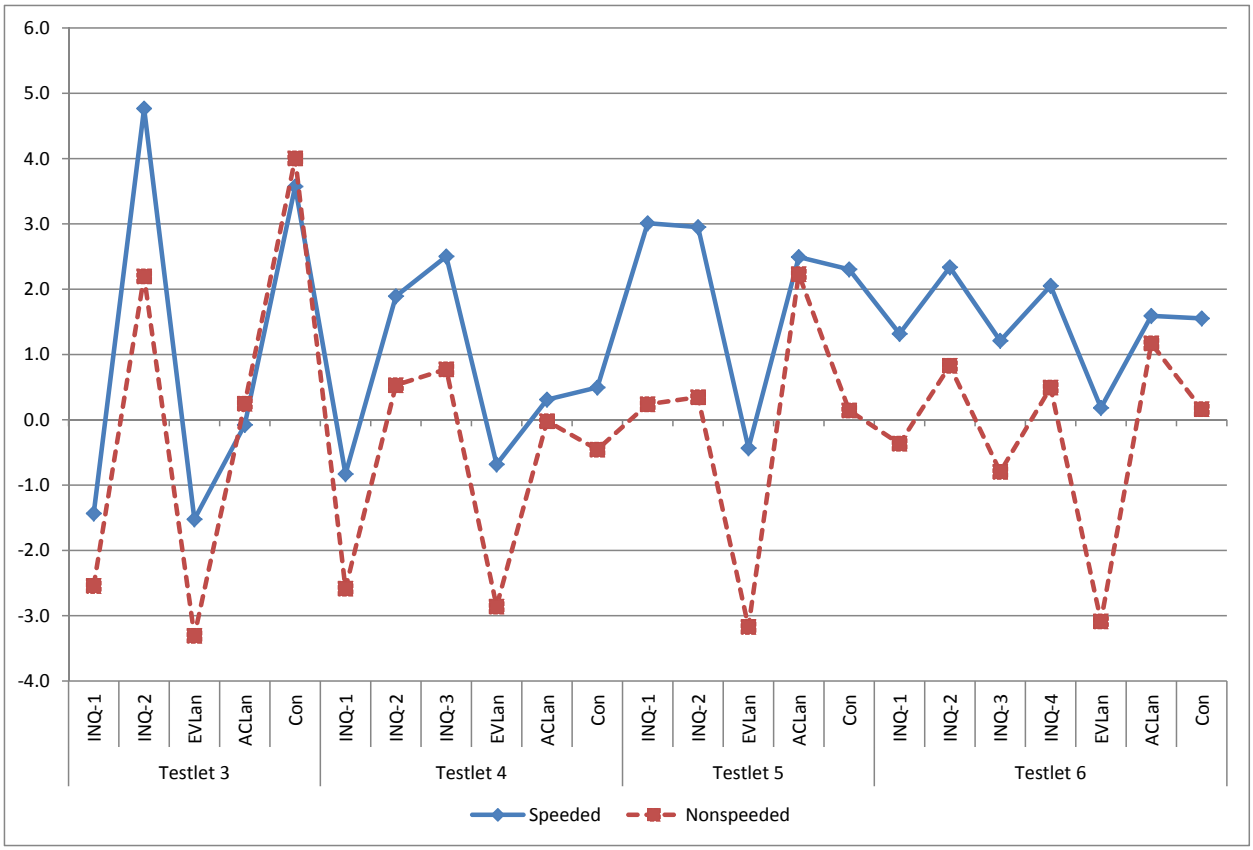

Figure 4.2: Means of the boundary location parameter estimates

Table 4.2: Item Parameter Estimates for the Speeded Group

		α		β_1		β_2		β_3		β_4		
Testlet	Item	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD	$\bar{\beta}$
3	11	1.86	0.16	-1.43	0.19							-1.43
	12	1.66	0.18	2.88	0.49	6.66	1.43					4.77
	13	2.59	0.20	-2.19	0.23	-0.86	0.15					-1.52
	14	0.70	0.07	-1.05	0.16	-0.40	0.15	0.31	0.18	0.83	0.21	-0.08
	15	2.39	0.16	1.86	0.40	5.29	1.10					3.57
4	16	2.94	0.23	-0.83	0.13							-0.83
	17	2.06	0.18	1.42	0.26	2.37	0.39					1.89
	18	1.74	0.16	2.50	0.49							2.50
	19	2.26	0.16	-1.36	0.13	-0.01	0.12					-0.68
	20	0.91	0.08	-0.60	0.11	0.17	0.10	0.58	0.09	1.08	0.10	0.31
	21	2.52	0.17	-0.67	0.12	1.66	0.18					0.50
5	22	2.34	0.17	3.01	0.52							3.01
	23	1.01	0.10	2.95	0.53							2.95
	24	2.24	0.20	-0.76	0.14	-0.11	0.13					-0.43
	25	1.02	0.09	1.21	0.16	1.70	0.17	3.21	0.25	3.86	0.34	2.49
	26	1.31	0.10	1.11	0.25	3.50	0.53					2.30
6	27	2.65	0.24	1.32	0.24							1.32
	28	1.42	0.17	2.34	0.31							2.34
	29	1.02	0.10	1.21	0.27							1.21
	30	0.96	0.10	2.05	0.29							2.05
	31	1.12	0.11	-0.93	0.21	1.29	0.25					0.18
	32	1.86	0.16	0.22	0.17	1.29	0.19	2.28	0.22	2.58	0.21	1.59
	33	1.72	0.16	0.77	0.18	2.33	0.31					1.55

Table 4.3: Item Parameter Estimates for the Nonspeeded Group

		α		β_1		β_2		β_3		β_4		
Testlet	Item	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD	$\bar{\beta}$
3	11	1.65	0.30	-2.54	0.24							-2.54
	12	0.86	0.13	1.03	0.23	3.37	0.57					2.20
	13	1.66	0.24	-4.08	0.37	-2.54	0.18					-3.31
	14	1.15	0.14	-1.94	0.17	-0.30	0.09	0.94	0.17	2.30	0.31	0.25
	15	0.63	0.09	1.82	0.30	6.19	0.96					4.01
4	16	1.32	0.21	-2.59	0.26							-2.59
	17	1.06	0.14	0.30	0.12	0.76	0.16					0.53
	18	1.53	0.21	0.77	0.13							0.77
	19	2.04	0.26	-3.74	0.35	-1.98	0.13					-2.86
	20	1.94	0.15	-1.12	0.07	-0.32	0.06	0.40	0.07	0.96	0.08	-0.02
	21	0.89	0.08	-2.48	0.18	1.57	0.17					-0.46
5	22	5.05	0.54	0.24	0.07							0.24
	23	5.22	0.53	0.35	0.07							0.35
	24	1.11	0.13	-3.50	0.26	-2.85	0.19					-3.17
	25	1.12	0.09	0.90	0.11	1.50	0.13	2.99	0.20	3.53	0.27	2.23
	26	3.78	0.44	-0.44	0.06	0.73	0.08					0.14
6	27	1.85	0.21	-0.37	0.08							-0.37
	28	2.40	0.31	0.83	0.11							0.83
	29	2.30	0.28	-0.80	0.09							-0.80
	30	2.22	0.28	0.50	0.10							0.50
	31	0.95	0.11	-4.39	0.36	-1.78	0.15					-3.09
	32	1.13	0.09	-0.43	0.10	0.80	0.12	1.84	0.16	2.47	0.19	1.17
	33	4.76	0.51	-0.56	0.07	0.88	0.09					0.16

Table 4.4: Testlet Effects

	Testlet					
Group	1	2	3	4	5	6
Speeded	1.51	2.36	0.65	0.13	0.09	2.19
Nonspeeded	0.87	0.38	0.20	0.11	0.97	1.53

Mean Ability and Proportion of Latent Classes

The mean ability of the speeded group was fixed as zero to remove indeterminacy of the scale. The estimated mean ability of the nonspeeded group was 0.602, which suggests that the nonspeeded group was more capable than the speeded group. The results indicated that 24% of the total students were assigned to the speeded group and 76% to the nonspeeded group. In the sample for the LDA analysis, similar proportions of 22% in the speeded group and 78% in the nonspeeded group were observed.

4.3.2 LDA

Model Selection

In this study, DIC was used to select the best fitting model. DIC was calculated based on hold-out likelihood, which requires one to separate the data set into two parts, the training and test sets. In this study, 60% of the documents were chosen as the training set and 40% as the test set. This selection resulted in 82 documents in the training set and 49 documents in the test set.

The calculated DIC values for nine different models are shown in Figure 4.3. The results indicated that a three-topic model was the best fitting model.

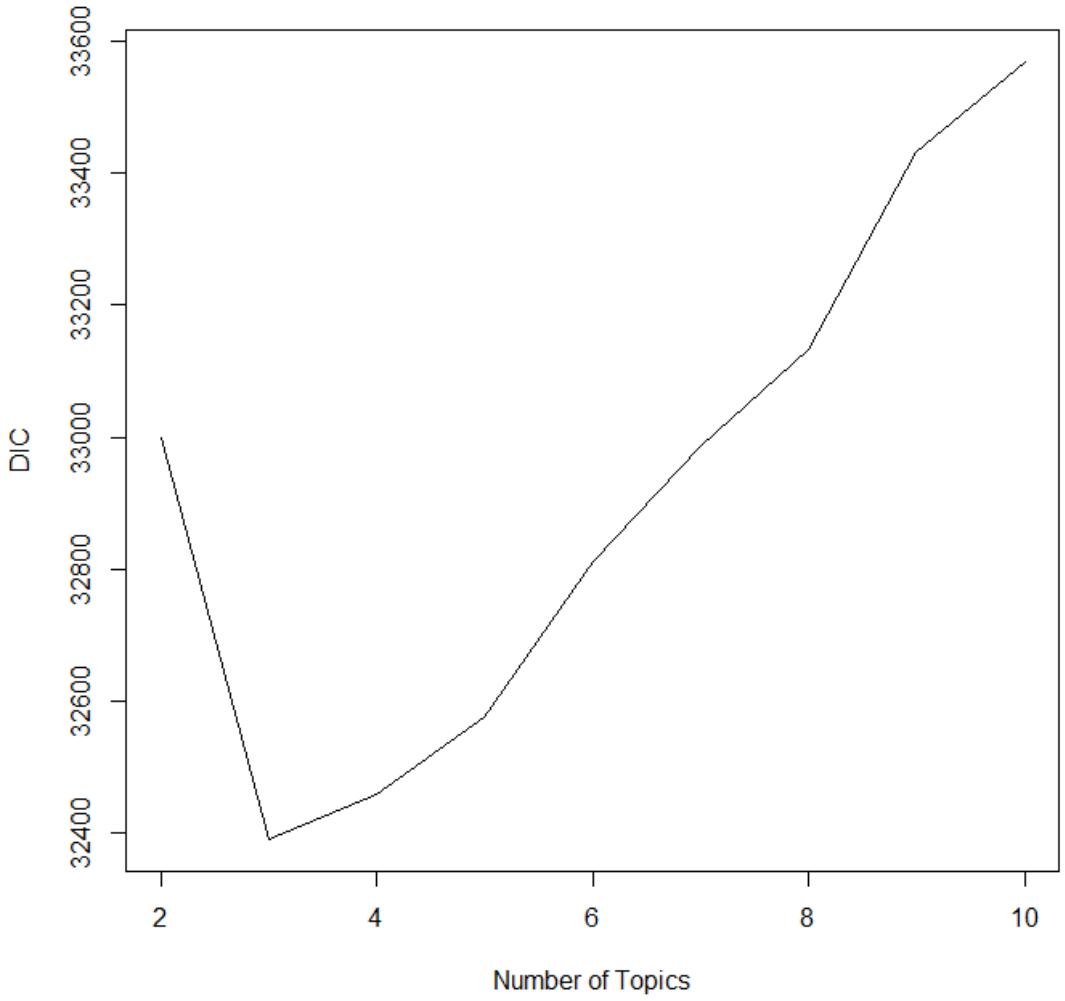

Figure 4.3: Results of the model comparison using DIC

Topic Specification

It is necessary to inspect the words in each topic to characterize the topic. The top 20 most frequently occurring words in each topic are listed in Table 4.5. Topic 1 includes words such as *need*, *when*, *think*, *can*, *white*, *big*, *small*, *slower*, *fast*, and *little*. These words suggest that this topic is related to words which can be used in everyday life. The words in Topic 1, therefore, would be considered to be 'everyday words.'

Topic 2 includes words such as *variable*, *circuit*, *hypothesize*, *experiment*, *independent*, *temperature*, and *dependent*, These words suggest that this topic is related to words which can be used to write scientific assumptions or statements. The words in Topic 2, therefore, would be considered to be 'general academic words.' The word *would* could possibly be considered a stop word, but the TF-IDF index failed to detect it as such.

Topic 3 includes words such as *hemisphere*, *northern*, *amount*, *metal*, *sunlight*, *Celsius*, and *conduct*. This suggests that this topic is related to words which would be taught in a science class. The words in Topic 3, therefore, would be considered to be 'discipline-specific words.' It should be noted that Topics 2 and 3 are both taught in science class, but Topic 2 is limited to words which would be used when crafting scientific statements such as a hypothesis.

Proportions of Topic Usage

The descriptive statistics for the proportions of usage of each topic for the speeded and nonspeeded groups are given in Table 4.6. Both means and standard deviations of the proportions for both groups were similar to one another. The proportions of topic usage in each topic for the speeded and nonspeeded groups are presented in Figure 4.4. Again, the median values for both groups were similar across the topics. The interquartile ranges of Topics 1 and 2 for the speeded group were larger than those for the nonspeeded group, and the interquartile range of Topic 3 for the speeded group was smaller than that for the nonspeeded group. These results suggest that the proportions for the speeded group were more variable than

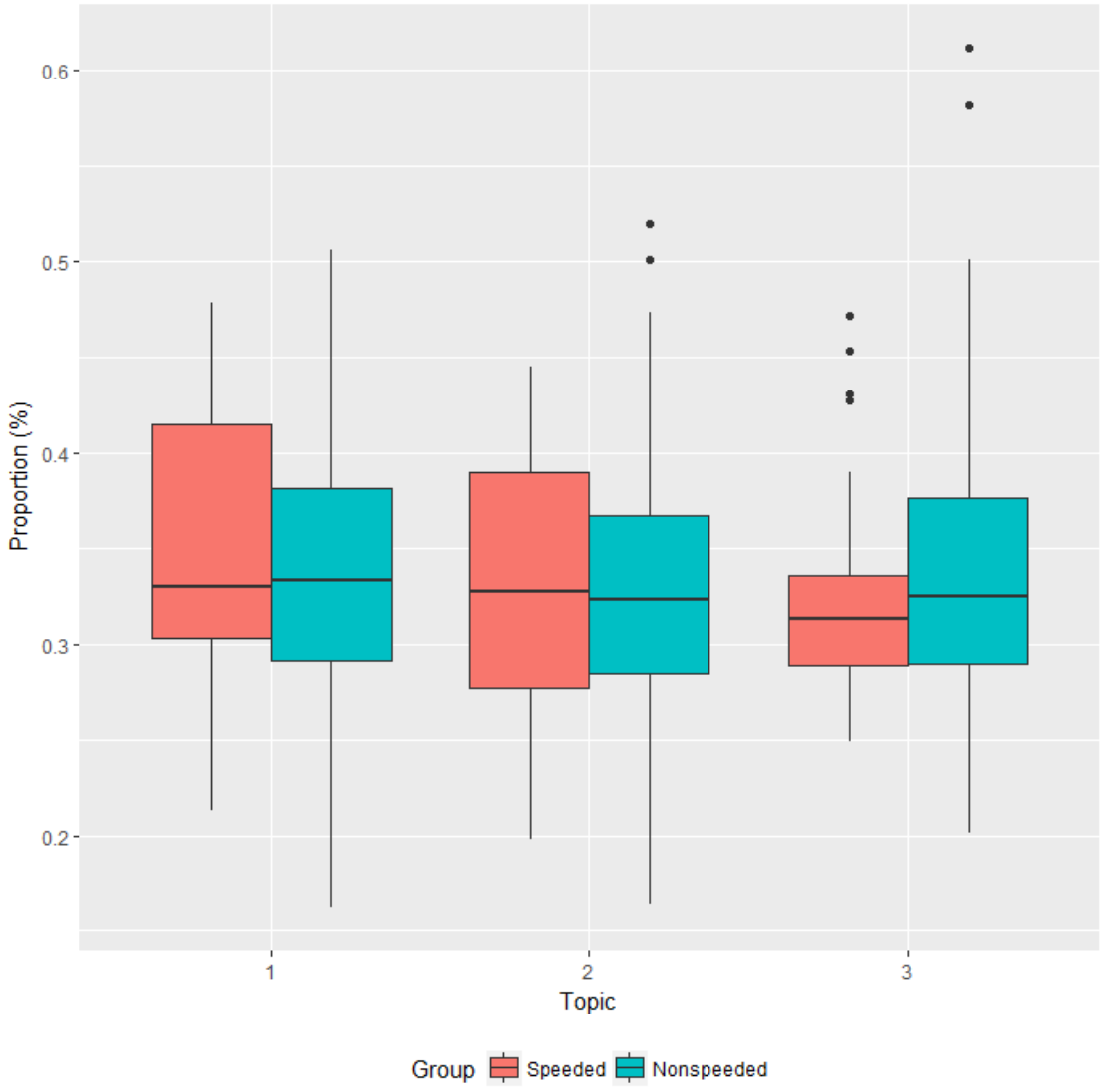

Figure 4.4: The proportions of the topic used by the speeded and nonspeeded groups

Table 4.5: Top 20 Frequent Words for Each Topic

Topic	Twenty Most Frequent Words
1	need, burn, steelball, when, think, connect, can, attract, white, big, small, weight, might, all, other, slower, turn, fast, little, earth
2	would, work, variable, circuit, hypothesize, out, experiment, independent, temperature, there, energy, two, change, dependent, what, stay, keep, less, nothing, smaller
3	same, hemisphere, cause, put, size, northern, type, which, amount, use, effect, give, through, metal, white, sunlight, Celsius, bucket, conduct, was

Table 4.6: Descriptive Statistics for the Proportions of Topic Usage

		Topic 1		Topic 2		Topic 3	
Group	N	M	SD	M	SD	M	SD
Speeded	30	0.35	0.08	0.33	0.07	0.32	0.06
Nonspeeded	170	0.34	0.07	0.33	0.07	0.34	0.08

those for the nonspeeded group for Topics 1 and 2 and less variable for the speeded group for Topic 3.

This study also used cluster analysis to explain the proportions of topic usage in each group. The results from hierarchical clustering suggested that there were two clusters in the proportions of topic usage for both groups. K-means with two clusters was used to identify these clusters. The clusters identified in the proportions for the speeded and nonspeeded groups are shown in Figures 4.5 and 4.6. The patterns in the proportions of topic usage as well as the proportion of the examinees in each cluster for the speeded and nonspeeded groups are given in Tables 4.7 and 4.8, respectively.

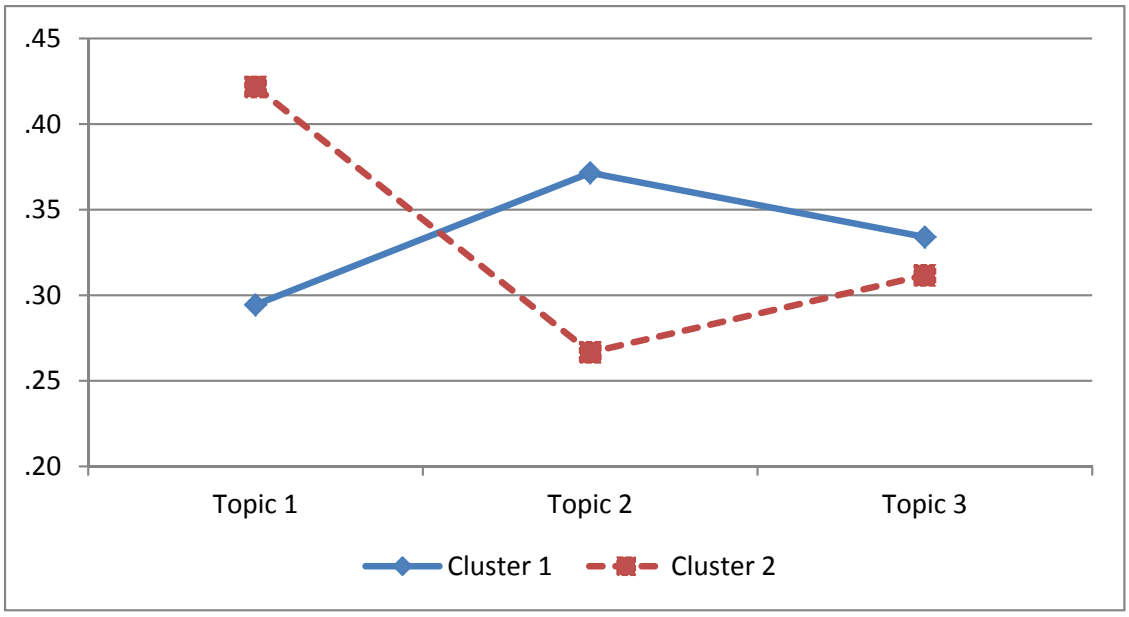

Figure 4.5: Patterns in proportions of topic usage identified by K-means: The speeded group

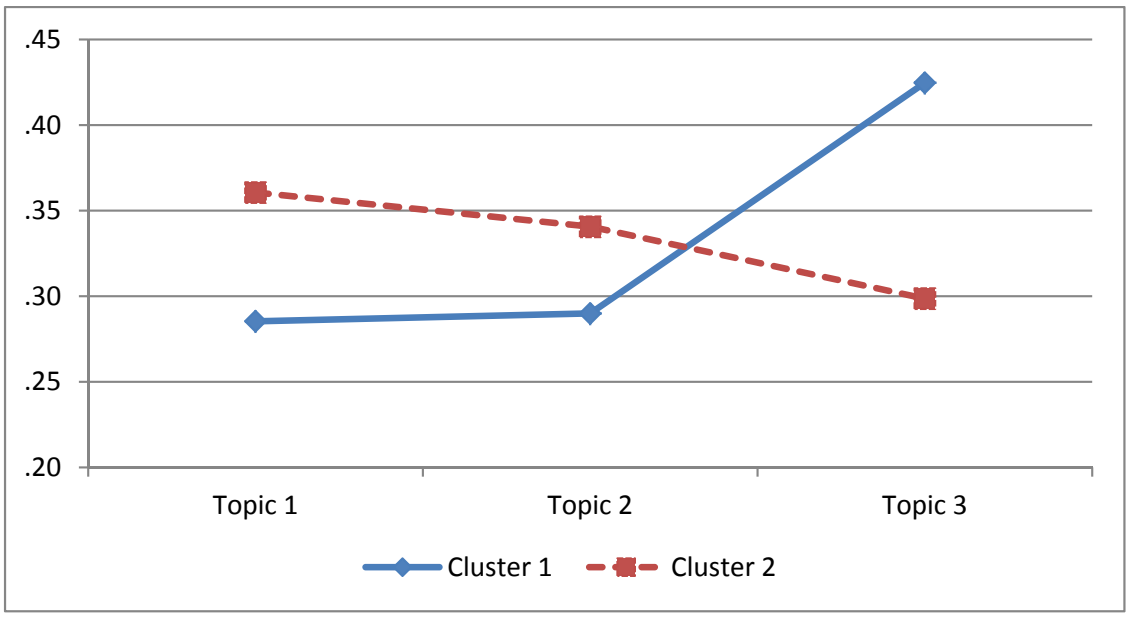

Figure 4.6: Patterns in proportions of topic usage identified by K-means: The nonspeeded group

Table 4.7: Patterns in the Proportions of Topic Usage: The Speeded Group

	Topic Usage (%)				
Cluster	Topic 1	Topic 2	Topic3	N	%
1	29.4	37.2	33.4	17	56.7
2	42.2	26.7	31.2	13	43.3

Table 4.8: Patterns in the Proportions of Topic Usage: The Nonspeeded Group

	Topic Usage (%)				
Cluster	Topic 1	Topic 2	Topic3	N	%
1	28.5	29.0	42.5	33	30.8
2	36.1	34.1	29.9	74	69.2

The results indicated that the patterns of the speeded and nonspeeded groups were different from one another. For the speeded group, 56.7% of the speeded examinees used general academic words (37.2%) most frequently followed by discipline-specific words (33.4%) and everyday words (29.4%) in their answers, whereas 43.3% of the speeded examinees used everyday words (42.2%) most frequently followed by discipline-specific words (31.2%) and everyday words (26.7%). For the nonspeeded group, 30.8% of the nonspeeded examinees used discipline-specific words (42.5%) most frequently followed by general academic words (29.0%) and everyday words (28.5%) in their answers, whereas 69.2% of the nonspeeded examinees used everyday words (36.1%) most frequently followed by general academic words (34.1%) and discipline-specific words (29.9%). It is worth pointing out that there was no common pattern shared by both groups.

4.4 Summary

In this study, the LDA model was implemented to characterize the latent groups, the speeded and nonspeeded groups, obtained from a two-class MixGRM-t. The LDA model was studied for its utility as a tool to help understand possible differences in the use of words between speeded and nonspeeded examinees. Two data sets were used in this study. The first data set was the rubric-based score of the total sample of 870 examinees who took one of the forms of a middle grades test on science inquiry knowledge. The second data set was a subset of the total set in which the written responses of 130 examinees were hand-entered into a text file. The first data set was used to classify the examinees into speeded and nonspeeded latent groups. The second data set was used to extract the latent topics of examinees' written responses and calculate the proportions of topic usage for the speeded and nonspeeded latent groups. The latent group memberships of the examinees in the second data set were taken from the results of the analysis for the total sample. The proportions of topic usage for each group were then compared to determine whether differences could be observed between the speeded and nonspeeded groups.

The results from the MixGRM-t for the total sample suggested that speededness effects were more pronounced in the middle of the test and that the nonspeeded group was more capable than the speeded group. In the sample used for the LDA, 22% of the 130 examinees were classified as the speeded group and 78% of the 130 examinees as the nonspeeded group. DIC suggested a three topic LDA model as the best fitting model. Based on the top 20 most frequent words for each topic, topics were characterized as everyday words, general academic words, and discipline-specific words. The descriptive statistics for the proportions of topic usage for the speeded and nonspeeded groups indicated that the means and standard deviations were similar to each other, but the distributions of the proportions were slightly different. A cluster analysis indicated that the patterns of the proportions of topic usage for the speeded and nonspeeded groups were also different.

Chapter 5

Discussion

The purpose of this dissertation was to investigate test speededness on constructed response (CR) items. To do this, we employed different types of mixture IRT models, each designed to treat the underlying traits somewhat differently. Test speededness can occur when time for test administration is limited. This implies that most standardized assessments, which usually enforce time limits for reasons of test fairness, can be affected by test speededness. When the responses from the speeded test are analyzed by IRT models, the model parameter estimates for items at the end of the test tend to be biased. This is because test speededness can undermine the important assumptions of IRT models such as unidimensionality and local independence.

Previous research has shown that mixture IRT models (Mislevy & Verhelst, 1990; Rost, 1990) can be effective separating examinees into two groups, the speeded and nonspeeded groups (Bolt et al., 2002; Goegebeur et al., 2008; Wang, 2011). The examinees in the speeded group were affected by speededness and, thus used different response strategies such as guessing due to insufficient time, primarily at the end of the test. The nonspeeded group consisted of students who were not affected by speededness and, thus used the same response strategies throughout the test. Mixture IRT models used in this way permitted unbiased estimates to be obtained from the nonspeeded group as well as to investigate each group's characteristics with estimated latent group membership for each examinee.

In Study 1 in this dissertation, a mixture graded response model for testlets (MixGRM-t) was used to detect speededness in CR items nested in testlets. Previously reported approaches to account for the impact of speededness effects on test performance using mixture IRT

models have focused on multiple-choice items. This is natural as these items have been prevalent in standardized assessments for many decades. CR items are increasingly employed in standardized assessments, however, as automated means of scoring these items improve and as measurement specialists seek to have students produce more of their answers rather select them. The impact of test speededness on CR items, although an important issue in education, has not been studied extensively. As the CR items in this study were scored with multiple rubrics, they were considered to be nested within testlets. Nesting in this way helped to expand the amount of information obtained from a student's response. Thus, a testlet IRT model was used to analyze the data.

In Study 2, a random item mixture graded response model for testlets (RIMGRM-t) was studied for detection of speededness effects in CR items nested in testlets. Generally, IRT models assume that persons are randomly sampled from a population, whereas items have been treated as fixed, when the model parameters are estimated. The random item IRT models, however, treat both person and items as random parameters, which allows one to estimates the variance terms for both the person or item parameters. One of the advantages of this approach is that it is possible to include covariates to explain the variance in item responses. Two models, an unconditional model, which was a model with no covariates, and a conditional model, which was a model with covariates, were compared to see the effects of the item covariates on the item parameter estimates.

In Study 3, a new approach to examining differences between speeded and nonspeeded examinees. Latent Dirichlet allocation (LDA) was first used to detect the latent topic structure of the texts of the CR responses. Next, the use of the latent topics was compared between the speeded and nonspeeded latent groups. This was done to help understand the characteristics of the answers of each latent group. Results were examined to determine whether examinees' written responses could be used to explore differences in examinees' responses.

5.1 Discussion for Chapter 2

An empirical example of using a MixGRM-t to detect speededness in a CR test was presented. The data were taken from the seven CR items from a secondary grades science inquiry test. The responses to these seven items were short answers on average of from one to three sentences.

Item Parameters. Results indicated that the item discrimination parameter estimates for the speeded group were larger than those for the nonspeeded group except for one item. This item was designed to measure understanding of science inquiry. It was the first item of the test, suggesting that speededness may not have affected the discrimination of this item because of its location.

The differences in the item boundary location parameter estimates between the speeded and nonspeeded groups increased toward the end of the test. The location parameter were higher for the speeded group than for the nonspeeded group and suggested that speededness effects were present in these parameter estimates. This was consistent with previous research with dichotomous items (Bolt et al., 2002). The differences between item location parameter estimates for the speeded and nonspeeded groups were smaller in the middle of the test and increased toward the end of the test. This pattern suggested that the effects of speededness tended to be greater the closer the items were to the end of the test.

Person Parameters. Consistent with previous research, the mean ability of the speeded group was higher than that of the nonspeeded group (Bolt et al., 2002; Cho et al., 2014). This finding suggested that the speeded examinees were more capable than the nonspeeded examinees. What may also be likely, however, is that the measurement of speeded examinees' ability was less accurate (Cohen, Wollack, Bolt, & Mroch, 2002). Cohen et al. found that speeded examinees had a grade point average that was almost higher than that of nonspeeded examinees. Semmes, Davison, and Close (2011) found that ACT performance was positively

correlated with speededness, suggesting that examinees with higher ability may be more affected by test speededness.

The estimated mixing proportion indicated that 30% of the examinees were classified in the speeded group and 70% in the nonspeeded group. An examination of manifest characteristics indicated gender was significantly related to speededness but ethnicity was not.

5.2 Discussion for Chapter 3

Two RIMGRM-t's, an unconditional model and a conditional model, were used to analyze a data set from the same study as that analyzed in Study 1. The unconditional model did not include item covariates and the conditional model did. A simulation study was conducted to examine the performance of both the unconditional and conditional models under practical testing conditions.

Real Data Example. In general, the item and person parameter estimates from both the unconditional and conditional models were comparable. The differences in item discrimination parameter estimates between the two models for the speeded group, however, were quite different. The estimates for the nonspeeded group, on the other hand, were almost the same for both models. This suggested that the item discrimination parameter estimates might be related to sample size as the sample size for the speeded group were smaller than for the nonspeeded group. More specifically, the numbers of speeded examinees were 574 (35.6%) and 698 (40.2%) out of 1,612 for the unconditional and conditional models, respectively.

The differences in item boundary location parameter estimates between the unconditional and conditional models were also clearly evident. This appeared to be related to the fact that two of four item covariates for the nonspeeded group in the conditional model had a significant effect on the item boundary location parameter estimates.

The item variance estimates from the unconditional model were larger than those from the conditional model. This suggests that inclusion of item covariates reduced item variances.

Testlet effects estimated from the two models also showed a similar pattern. For the mixing proportions, the proportion of the speeded group classified by the unconditional model was smaller than that by the conditional model. The cross-tabulation of latent group membership classified by the unconditional and conditional models indicated that 86.0% of the examinees in the sample were classified into the same latent group by both models.

Simulation Study. The simulation conditions included two sample sizes (1,000 and 3,000 examinees), two proportions for the speeded group (10% and 30%), and two estimation models (an unconditional model and a conditional model). This yielded eight different simulation conditions. Twenty replications were simulated for each condition. The recovery of the item parameters, including item discrimination, boundary location, and testlet effects, as well as mean ability were evaluated using bias, RMSEs, and correlations between the true and estimated values. Recovery of latent group membership was evaluated by the percentages of the correct classification.

In general, both models showed similar levels of recovery of the model parameters. As expected, the recovery of the parameters for the nonspeeded group was better than that for the speeded group. One likely reason for this pattern may be that the sample size of the nonspeeded group was much larger than that of the speeded group. As an example, when the sample size was 1,000 and the proportion of the speeded group was 10%, there were 100 members in the speeded group but 900 members in the nonspeeded group.

Both the item discrimination and boundary location parameters were recovered moderately well for the nonspeeded group in both the unconditional and conditional models. For both the item discrimination and boundary location parameters, the reduction of the RMSEs in the speeded group was much larger than that in the nonspeeded group. The estimates for the nonspeeded group from the conditional model seemed to be related to the increase of the proportions of the speeded group. For example, when the sample size was 1,000, the RMSEs of the item boundary location parameters for the nonspeeded group from the conditional model increased as the proportion of the speeded group increased from 10% to 30%. This

increase did not occur, however, when the sample size was 3,000. In fact, the RMSEs for the nonspeeded group decreased as the proportion of the speeded group increased from 10% to 30% when the sample size was 3,000.

The recovery of the testlet effects for the speeded group was poor in both models when compared with recovery of either the item discrimination or boundary location parameters. The conditional model had lower RMSEs for the testlet effect than the unconditional model except in the 1,000 examinees and 30% speeded examinees condition.

The recovery of the mean ability parameter for the nonspeeded group was generally better for both models compared to the recovery of either the item discrimination or boundary location parameters. When the sample size was 1,000, the RMSEs from both models increased as the proportion of the speeded group increased. When the sample size was 3,000, however, the RMSEs for ability from the unconditional model decreased whereas those from the conditional model did not change.

The recovery of latent group membership for the speeded group was better with the unconditional model and better for the nonspeeded group with the conditional model. This tendency may explain why some of the examinees were assigned to different latent groups in the real data analysis depending on the estimation model. In addition, when the proportion of the speeded group was 30%, the difference in proportions between the two models for the nonspeeded group was large.

5.3 Discussion for Chapter 4

In this study, LDA was used to analyze the actual words of the responses students in the speeded and nonspeeded groups used in their answers to the CR items. Examinees were first classified into the speeded and nonspeeded groups using a MixGRM-t on the total sample, which were scores of 870 students. Next, LDA was used done on a subsample of the total sample. This subsample consisted of the written responses of 137 students, that had been transcribed into machine-readable form. Finally, the latent topic structure was compared for

examinees in the speeded and nonspeeded groups. Data used in this study were from the same NSF-funded project data used in the first two studies but from a different academic year.

Results from the Mixture testlet GRM. In general, the item discrimination parameter estimates for the speeded group were larger in the middle of the test (Testlet 3 to Testlet 4), whereas those for the nonspeeded group were larger at the end of the test (Testlet 5 and Testlet 6). For the item boundary location parameters, the distance between the speeded and nonspeeded groups did not depend on the location of the item but instead was a function of the item type like Everyday Language or Academic Language. For example, the difference between two groups was smaller for items dealing with Academic Language and larger for items dealing with Everyday Language. From these results, it appears that test speededness varied depending on the item type rather than on the location of item. This result is somewhat different from previous studies (Bolt et al., 2002; Kim et al., 2016) which showed that the difference between the speeded and nonspeeded groups was largest at the end of the test. The testlet effects for the speeded group were larger both at the beginning and at the end of the test, whereas those for the nonspeeded group were larger only at the end of the test. The mixing proportions for the samples for the LDA analysis indicated that 22% of students were assigned to the speeded group and 78% to the nonspeeded group.

Results from LDA. The DIC indices suggested that a two-topic model was the best-fitting of the models considered. The top 20 most frequently occurring words in each topic were examined to help characterize the topic. As a result, the first topic was classified as everyday words, the second topic as general academic words, and the third topic as discipline-specific words.

An LDA analysis also provides the proportions of usage for each topic. Differences between these proportions between the speeded and nonspeeded groups are available at the individual examinee level. This allows one to investigate the topic usage characteristics of speeded and

nonspeeded examinees. It is possible, therefore, to compare these proportions for the speeded and nonspeeded groups using various methods. For example, cluster analysis was used in this dissertation to determine the differences in latent topic use between the two latent groups. The descriptive statistics for the proportions of topic usage suggested that the distribution of the proportions for two groups did differ, but the differences did not appear to be meaningful as the means and standard deviations for the speeded and nonspeeded groups were almost the same. However, the results from clustering analysis on the proportions for each group suggested that the patterns in the proportions of topic usage for the speeded group were different from those for the nonspeeded group. More specifically, the majority of the speeded examinees used general academic words most frequently followed by the discipline-specific words and then by everyday words, whereas the majority of the nonspeeded examinees used the everyday words most frequently followed by the general academic words and the discipline-specific words. Moreover, the speeded and nonspeeded groups did not share a common pattern in topic usage.

5.4 Suggestions for Future Study

This dissertation used ordinal constraints only on the item boundary location parameters to model speededness. It is possible, however, that changes in response patterns of speeded examinees due to time limits could also affect the estimation of the item discrimination parameters. For dichotomous data, Oshima (1994) suggested that test speededness could distort the item discrimination, difficulty, and guessing parameter estimates. It might be useful to investigate the relation between speededness and estimation of the item discrimination parameters in the context of the CR items and to find useful constraints to model speededness using the item discrimination as well as the item boundary location parameters.

For the conditional RIMGRM-t in the second study, the effect of person covariates needs to be examined further. This dissertation used item covariates and confirmed that they

affected item boundary location parameter estimates. Wang (2011) proposed that the inclusion of a person covariate helped to classify the examinees into different latent classes. It would be useful to inspect the influence of person covariates in the context of the graded response model.

More testing conditions, especially ones involving item characteristics, might be useful for a simulation study. The simulation study presented in the second study considered two testing conditions, the sample size and the proportion of the speeded examinees. The item characteristics such as item difficulty might also be considered since test speededness might possibly vary with these characteristics. For example, the influence of time limits may differ on a test which has difficult items at the beginning of the test from a test which has difficult items at the end of the test.

The graded response model was implemented in this dissertation to deal with multiple scores of the data. It is likely that other models for analyzing polytomous data might also be potentially useful such as the rating scale model (Andrich, 1978b, 1978a), the partial credit model (Masters, 1982), or the generalized partial credit model (Muraki, 1992). Based on the taxonomy suggested by Thissen and Steinberg (1986), the graded response model belongs to the category of 'difference models,' whereas the rating scale model, the partial credit model, and the generalized partial credit model belong to the category of 'divided-by-total models.' Since this classification is based on how the probability of certain category score k is defined, the impact of test speededness on the estimation of the model parameters is likely to be different depending on which estimation model is implemented.

Bibliography

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied psychological measurement*, *2*, 581–594.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

Bejar, I. I. (1985). *Test speededness under number-right scoring: An analysis of the test of English as a foreign language* (ETS Research Report No. 85-11). Princeton, NJ: Educational Testing Service.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. N. Srivastava & M. Sahami (Eds.), *Text mining: Classification, clustering, and applications* (pp. 71–93). Boca Raton, FL: Chapman & Hall/CRC.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331–348.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2014). A mixture group bifactor model for binary responses. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 375–395.

Cho, S.-J., Cohen, A. S., Kim, S.-H., & Bottge, B. (2010). Latent transition analysis with a mixture item response theory measurement model. *Applied Psychological Measurement*, *34*, 483–504.

Cohen, A. S., Wollack, J. A., Bolt, D. M., & Mroch, A. A. (2002, April). *A mixture rasch model analysis of test speededness.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

de Ayala, R. J. (2013). *The theory and practice of item response theory.* New York, NY: Guilford Publications.

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

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics*, *23*, 129–151.

Ercikan, K. (2002). Scoring examinee responses for multiple inferences: Multiple scoring in assessments. *Educational Measurement: Issues and Practice*, *21*(2), 8–14.

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement*, *9*, 123–131.

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

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement*, *47*, 432–457.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 1–19). Boca Raton, FL: Chapman & Hall/CRC.

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*, 65–87.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*, 5228–5235.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30.

Gunel, M., Hand, B., & Prain, V. (2007). Writing for learning in science: A secondary analysis of six studies. *International Journal of Science and Mathematics Education*, *5*, 615–637.

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*, 1109–1144.

Izsák, A., Orrill, C. H., Cohen, A. S., & Brown, R. E. (2010). Measuring middle grades teachers' understanding of rational numbers with the mixture Rasch model. *The Elementary School Journal*, *110*, 279–300.

Kim, M., Cohen, A. S., Lu, Z., Kim, S., Buxton, C., & Allexsaht-Snider, M. (2016, April). *Speededness in a constructed response science test.* Paper presented at the annual meeting of National Council on Measurement in Education, Washington, DC.

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

Kwak, M., Kim, S., & Cohen, A. S. (2017). *Using latent dirichlet allocation to analyze constructed response answers.* (Manuscript submitted for publication)

Lee, O., Quinn, H., & Valdés, G. (2013). Science and language for English language learners in relation to Next Generation Science Standards and with implications for Common Core State Standards for English language arts and mathematics. *Educational Researcher*, *42*, 223–233.

Lord, F. M., & Novick, M. S. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*(4), 29–37.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement*, *12*, 281–296.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Muraki, E., & Bock, D. (2003). PARSCALE (version 4.1). [Computer software]. Mooresville, IN: Scientific Software International.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* National Academies Press.

National Research Council. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*, 200–219.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News*, *6*(1), 7–11.

Powers, D. E., & Fowles, M. E. (1996). *Effects of applying different time limits to a proposed GRE writing test* (ETS Research Report No. 96-28). Princeton, NJ: Educational Testing Service.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, *18*, 321–349.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement*, *26*, 271–285.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*.

Scalise, K. (2014, April). *Assessment system design options for the Next Generation Science Standards (NGSS) reflections on some possible design approaches.* Paper presented at the Invitational Research Symposium on Science Assessment, Princeton, NJ.

Semmes, R., Davison, M. L., & Close, C. (2011). Modeling individual differences in numerical reasoning speed as a random effect of response time limits. *Applied Psychological Measurement*, *35*, 433–446.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 583–639.

Spiegelhalter, D. J., Best, N. G., Gilks, W. R., & Inskip, H. (1996). Hepatitis B: a case study in MCMC methods. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 21–43). Boca Raton, FL: Chapman & Hall/CRC.

Teh, Y. W., Newman, D., & Welling, M. (2006). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Neural Information Processing Systems*, *6*, 1378–1385.

Thissen, D., Chen, W., & Bock, R. (2003). MULTILOG (version 7.0). [Computer software]. Mooresville, IN: Scientific Software International.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*, 247–260.

Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006). Making BUGS open. *R News*, *6*(1), 12–17.

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*, 369–386.

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

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, *27*, 1–14.

Wang, A. (2011). *Mixture cross-classification IRT model for test speededness.* Unpublished doctoral dissertation, University of Georgia, Athens, GA.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126–149.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement*, *40*, 307–330.

Yamamoto, K. (1989). *HYBRID model of IRT and latent class models* (ETS Research Report No. 89-41). Princeton, NJ: Educational Testing Service.

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). New York, NY: Waxman.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. (2003). BILOG-MG (version 3.0). [Computer software]. Mooresville, IN: Scientific Software International.

Appendix A

OpenBUGS Code for the Unconditional Random Item Mixture Graded Response Model for Test Speededness

```
# NE: the number of examinees
# NI: the number of items
# NT: the number of testlet
# mI: the maximum category of each item

model
{
   for (j in 1:NE)  {
      for (i in 1:NI)  {
         r[j, i]<- resp[j, i]
         r[j, i]~ dcat(pcat[j, i, 1:mI[i]])
      }
   }

   # GRM

   for (j in 1:NE)  {
      for (i in 1:NI)  {
         for (k in 1:(mI[i]-1)) {
            p[j, i, k]<- 1/ (1+ exp(-a2[gmem[j], i]*(theta[j]- beta[gmem[j],
    i, k]- gamma[j, d[i]])))
         }
         p[j, i, mI[i]]<- 0
      }
      for (k in 1:NT){
         gamma[j, k]~ dnorm(mut[gmem[j]], pr.gamma[gmem[j], k])
      }
      theta[j]~ dnorm(mut[gmem[j]], 1)
      gmem[j]~ dcat(pi[1:2])
   }

   for (j in 1:NE)  {
      for (i in 1:NI)  {
         pcat[j, i, 1]<- 1- p[j, i, 1]
         for (k in 2:mI[i]) {
```

```
         pcat[j, i, k]<- p[j, i, k-1]- p[j, i, k]
      }
   }
}
pi[1:G2]~ ddirch(alph[1:2])
mut[1]<- 0
mut[2]~ dnorm(0, 1)

# Item variance

var.i~ dunif(0, 2)
tau.i<- pow(var.i, -1)

# Priors

for (j in 1:G2){
   for (i in 16:NI) {
      a2[j, i]~ dnorm(0, 1) I(0,)
   }
}

for (i in 1:G2) {
   mbeta[i]~ dnorm(0, 1)
}

for (i in 1:G2){
   for (k in 1:NT){
      pr.gamma[i, k]~ dgamma(2.5, 0.25)
   }
   for (k in 1:NT){
      testlet[i, k]<- pow(pr.gamma[i, k], -1)
   }
}

# Constraint on Testlet 1 to Testlet 3: a1 = a2 & b1 = b2

a2[1,1]     <- 0.621
a2[1,2]     <- 0.992
...
a2[1,14]    <- 0.986
a2[1,15]    <- 4.882

a2[2,1]     <- 0.621
a2[2,2]     <- 0.992
...
a2[2,14]    <- 0.986
a2[2,15]    <- 4.882
```

```
   beta[1,1,1]    <- -0.980
   beta[1,2,1]    <- 0.998
   ...
   beta[1,15,1]   <- -1.309
   beta[1,15,2]   <- 0.172

   beta[2,1,1]    <- -0.980
   beta[2,2,1]    <- 0.998
   ...
   beta[2,15,1]   <- -1.309
   beta[2,15,2]   <- 0.172

   # No constraint on Testlet 4

   for (i in 16:21) {
      beta[1, i, 1]~ dnorm(mbeta[1, i], tau.i)
      beta[2, i, 1]~ dnorm(mbeta[2, i], tau.i)
      for (k in 2: (mI[i]-1)) {
         beta[1, i, k]~ dnorm(mbeta[1, i], tau.i) I(beta[1, i, k-1], )
         beta[2, i, k]~ dnorm(mbeta[2, i], tau.i) I(beta[2, i, k-1], )
      }
   }

   # Constraint on Testlet 5 to Testlet 7: b1 > b2

   for (i in 22:NI){
      beta[1, i, 1]~ dnorm(mbeta[1, i], tau.i)
      beta[2, i, 1]~ dnorm(mbeta[2, i], tau.i) I(,beta[1, i, 1])
      for (k in 2:(mI[i]-1)){
         beta[1, i, k]~ dnorm(mbeta[1, i], tau.i) I(beta[1, i, k-1], )
         beta[2, i, k]~ dnorm(mbeta[2, i], tau.i) I(beta[2, i, k-1], beta
    [1, i, k])
      }
   }
}
```

Appendix B

OpenBUGS Code for the Conditional Random Item Mixture Graded Response Model for Test Speededness

```
# NE: the number of examinees
# NI: the number of items
# NT: the number of testlet
# mI: the maximum category of each item

model
{
   for (j in 1:NE)  {
      for (i in 1:NI)  {
         r[j, i]<- resp[j, i]
         r[j, i]~ dcat(pcat[j, i, 1:mI[i]])
      }
   }

   # GRM

   for (j in 1:NE)  {
      for (i in 1:NI)  {
         for (k in 1:(mI[i]-1)) {
            p[j, i, k]<- 1/ (1+ exp(-a2[gmem[j], i]*(theta[j]- beta[gmem[j],
    i, k]- gamma[j, d[i]])))
         }
         p[j, i, mI[i]]<- 0
      }
      for (k in 1:NT) {
         gamma[j, k]~ dnorm(mut[gmem[j]], pr.gamma[gmem[j], k])
      }
      theta[j]~ dnorm(mut[gmem[j]], 1)
      gmem[j]~ dcat(pi[1:2])
   }

   for (j in 1:NE)  {
      for (i in 1:NI)  {
         pcat[j, i, 1]<- 1- p[j, i, 1]
         for (k in 2:mI[i]) {
```

```
         pcat[j, i, k]<- p[j, i, k-1]- p[j, i, k]
      }
   }
}
pi[1:G2]~ ddirch(alph[1:2])
mut[1]<- 0
mut[2]~ dnorm(0, 1)

# Item variance

var.i~ dunif(0, 2)
tau.i<- pow(var.i, -1)

# Priors

for (j in 1:G2){
   for (i in 16:NI) {
      a2[j, i]~ dnorm(0, 1) I(0,)
   }
}

for (i in 1:G2){
   for (k in 1:NT){
      pr.gamma[i, k]~ dgamma(2.5, 0.25)
   }
   for (k in 1:NT){
      testlet[i, k]<- pow(pr.gamma[i, k], -1)
   }
}

# Item covariates

for (j in 1:G2) {
   for (i in 1:NI) {
      mbeta[j, i]<- mb[j]+ coef.1[j]*item.cv1[i]+ coef.2[j]*item.cv2[i]+
coef.3[j]*item.cv3[i]
   }
   mb[j]~ dnorm(0, 1)
   coef.1[j]~ dnorm(0, 1)
   coef.2[j]~ dnorm(0, 1)
   coef.3[j]~ dnorm(0, 1)
}

# Constraint on Testlet 1 to Testlet 3: a1 = a2 & b1 = b2

a2[1,1]     <- 0.621
a2[1,2]     <- 0.992
```

```
   ...
   a2[1,14]    <- 0.986
   a2[1,15]    <- 4.882

   a2[2,1]     <- 0.621
   a2[2,2]     <- 0.992
   ...
   a2[2,14]    <- 0.986
   a2[2,15]    <- 4.882

   beta[1,1,1]    <- -0.980
   beta[1,2,1]    <- 0.998
   ...
   beta[1,15,1]   <- -1.309
   beta[1,15,2]   <- 0.172

   beta[2,1,1]    <- -0.980
   beta[2,2,1]    <- 0.998
   ...
   beta[2,15,1]   <- -1.309
   beta[2,15,2]   <- 0.172

   # No constraint on Testlet 4

   for (i in 16:21) {
      beta[1, i, 1]~ dnorm(mbeta[1, i], tau.i)
      beta[2, i, 1]~ dnorm(mbeta[2, i], tau.i)
      for (k in 2: (mI[i]-1)) {
         beta[1, i, k]~ dnorm(mbeta[1, i], tau.i) I(beta[1, i, k-1], )
         beta[2, i, k]~ dnorm(mbeta[2, i], tau.i) I(beta[2, i, k-1], )
      }
   }

   # Constraint on Testlet 5 to Testlet 7: b1 > b2

   for (i in 22:NI){
      beta[1, i, 1]~ dnorm(mbeta[1, i], tau.i)
      beta[2, i, 1]~ dnorm(mbeta[2, i], tau.i) I(,beta[1, i, 1])
      for (k in 2:(mI[i]-1)){
         beta[1, i, k]~ dnorm(mbeta[1, i], tau.i) I(beta[1, i, k-1], )
         beta[2, i, k]~ dnorm(mbeta[2, i], tau.i) I(beta[2, i, k-1], beta
   [1, i, k])
      }
   }
}
```