A Mixture Cross-Classification IRT Model for Test Speededness

by

Aijun Wang

(Under the direction of Allan S. Cohen)

Abstract

Previous research has shown that under time limits, items near the end of the test appear harder for the speeded examinees than for the non-speeded examinees (Bolt et al.,2002; Oshima, 1994). Moreover, speeded examinees tend to omit more items near the end of the test (Bolt et al., 2002; Cohen et al., 2002). Therefore, certain person and item characteristics related to test speededness may help to explain the differences in estimates of examinees' abilities and item difficulties. The test speededness models proposed thus far have focused on modeling speededness effects rather on attempting to explain them. In addition, the investigation of differential speededness is typically implemented in a two-step procedure. First, the measurement model is used to identify speeded groups, generally as latent classes. Next, a statistical analysis is done to examine characteristics of members of speeded and non-speeded groups. The purpose of this dissertation was to propose a mixture multilevel IRT model with person and item covariates that could be used to detect test speededness effect in paper-and-pencil test. Unlike the regular IRT models which treat persons as random and items as fixed, however, this dissertation treated both as random, making it possible to add item covariates into the model. De Boeck (2008) has shown that treating items as random not only makes sense theoretically, but also is promising for identifying DIF items and for explaining differential item difficulties. A multilevel mixture IRT model was developed in this dissertation for use in detection of speeded and non-speeded latent classes. Covariates are

illustrated as being incorporated into the model for use in helping to characterize members of each latent class.

INDEX WORDS: Test speededness, cross-classified item response model, mixture IRT, multilevel IRT

A Mixture Cross-Classification IRT Model for Test Speededness

by

Aijun Wang

B.A., Yantai Teachers University, China, 1997

M.A., Shanghai Institute of Foreign Trade, China, 2000

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

2011

A Mixture Cross-Classification IRT Model for Test Speededness

by

Aijun Wang

Approved:

Major Professor: Allan S. Cohen

Committee: Seock-Ho Kim
Gary J. Lautenschlager
Jonathan L. Templin

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2011

Dedication

To Steven and Cassie,

For being a constant source of joy, pride and inspiration.

Acknowledgments

This dissertation represents the combined efforts of a great many people. It is my honor and pleasure to thank those who made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever. I am deeply indebted to my advisor, Dr. Allan Cohen, for his continuous guidance, encouragement, caring, research and financial support, and providing me with an excellent atmosphere for doing research during my graduate study. I truly appreciate his commitment to students - both through his accessibility and availability as I frequently ran to him with a quick question or two. He guided me step by step growing from a naive graduate student to a mature researcher. Despite his busy schedule, Dr. Cohen always finds the time to read and revise my papers and to teach me how to write technically. With his generous financial support, I don't need to worry about my financial problems and was able to present my research at both national and international conferences. I am so grateful to him for providing me with the academic freedom and opportunity to explore topics of my interest. He sets the role example for me not only as a scholar but also as a person for being generous, considerate, tolerant, patient and humorous. Thanks to him, my graduate experiences are enjoyable and memorable. I couldn't find a mentor better than him.

I would like to express my sincere gratitude to my committee members, Dr. Seock-Ho Kim, Dr. Gary Lautenschlager and Dr. Jonathan Templin. They have always been a steady and generous source of guidance, feedback, comments and help. I would like to acknowledge Dr. Kim and Dr. Templin for their valuable discussions and constructive feedbacks that helped me to understand my research area better. They are always willing to share their thoughts and give me suggestions on my dissertation. I am also thankful to Dr. Lautenschlager for serving as my committee member and for his thought-provoking prelim questions that helped

guide me to improve my knowledge in the area of multilevel modeling, a general framework which this dissertation is based on.

I want to thank the other REMS faculty and students for their support and friendship. The family-like atmosphere in the REMS program makes me never feel alone. The care and friendship from the REMS friends make my study at UGA a pleasant memory for the rest of my life. I owe a special debt to Dr. Feiming Li for her help when I first got to Georgia and for her suggestions at the initial stage of this dissertation.

Finally, it is impossible to have my research career and pursue my Ph.D degree without the support and love of my family. I am grateful to my parents for their understanding on the value of education and for their unselfish love and support. My husband deserves my heartfelt thanks for standing behind me all the time and cheering me on when I was stressed beyond measure.

I truly thank all of you who have helped me to make my dream come true.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Statement of the Problem

Time limits on tests usually serve two purposes. Time limits are necessary, when the speed of responding is a construct that the test is designed to measure, but when the purpose of setting time limits is administrative, the speed of responding is not usually the construct of interest. Anastasi (1988) notes that "A pure speed test is one in which individual differences depend entirely on speed of performance. Such a test is constructed from items of uniformly low difficulty, all of which are well within the ability level of the persons for whom the test is designed. The time limit is made so short that no one can finish all the items. Under these conditions, each person's score reflects only the speed with which he or she worked. A pure power test, on the other hand, has a time limit long enough to permit everyone to attempt all items. The difficulty of the items is steeply graded, and the test includes some items too difficult for anyone to solve, so that no one can get a perfect score" (pp. 127-128). Anastasi's definitions are for pure speed and pure power tests. These are somewhat ideal definitions as most educational tests are administered under some type of time limits.

For a power test, the expectation (based on Anastasi's definition) is that all examinees should have sufficient time to at least consider, if not finish each test item. When some examinees don't have sufficient time to try all the items on the test, time limits may affect their performance at least to some degree. In such a case, the test would be regarded as speeded and the effects of time limits on examinee performance referred to as speededness effects (Evans & Reilly, 1972). When this is the case, then speededness should be taken into account in estimating item and examinee performance.

When speed of responding is not part of the construct being measured, then test speededness likely introduces construct irrelevant variance, thus potentially jeopardizing the construct validity of the test (Lu & Sireci, 2007). When a test is speeded, in other words, part of the variance in the test score is potentially related to speededness. If the construct of interest of the test does not include the speed of responding, then the presence of test speededness changes the interpretation of the test scores and the inferences based on the test scores. Briel, O'Neil, and Schueneman (1993) note that the Graduate Record Examinations (GRE) is primarily a measure of intellectual power rather than rate of responding, and differences in examines' response rates constitute an irrelevant source of difficulty in test performance. In this same regard, time limits have been shown to have an effect on the validity of intelligence test scores (Wilhelm & Schulze, 2002). Wilhelm and Schulze further note that "the simple manipulation of relaxing time constraints in the measurement of reasoning (indicates) that properties of measurement instruments are not stable when conditions of administration are altered. Speeded and nonspeeded tests of reasoning ability do not equally tap the same constructs. Removing the time constraints from reasoning measurement removes mental speed variance" (p. 551). Mellon, Daggett, MacManus, and Moritsch (1996) found speeded tests were susceptible to coaching effects, as examinees could be taught to fill in all the items at the end of the test (i.e., the ones most likely to show speededness effects) as time limits expired. For those examinees, responses at the end of the test would not reflect the same construct as responses at the beginning of the test. Sager, Peterson, and Oppler (1994) found speeded and non-speeded administrations of the General Aptitude Test Battery (GATB) measured similar but not identical constructs. Using a confirmatory factor analysis (CFA), Sager et al. showed a speed factor was present for the speeded administration. Lu and Sireci (2007) note that the presence of omissions at the end of the test poses a threat to the content validity of the test. If a large number of these items are left blank, in other words, the content domain represented by these items will not be complete.

Reliability has also been shown to be affected by test speededness. Evans and Reilly (1972) found increasing speededness on the LSAT produced a lower KR-20 coefficient for tests of fee-free examinees compared to regular-fee examinees. Likewise, Attali (2005) found the internal consistency reliability of multiple-choice tests was lowered by speededness. According to Attali, examinees who run out of time tend to guess on the remaining items instead of omitting them. This random guessing behavior introduces noise into examinee's responses or creates inconsistent responses, thus lowering the reliability of the test. Schnipke and Scrams (1997) found higher ability examinees tended to randomly guess more on items near the end of the test. The effect of random guessing behavior noted by Schnipke and Scrams was to reduce the variance among the examinee abilities, thereby reducing the reliability of the test.

Item response theory (IRT) models are currently widely used for scaling and scoring of a large proportion of educational tests. If the IRT model fits the data, the ability scores estimated by the IRT model are independent of the set of items used on the test. The ability of examinees who have taken different sets of items calibrated to the same scale, in other words, can be placed on that same scale. In addition, if the model fits the data, item parameters estimated by an IRT model are sample-independent. This means that item parameter estimates obtained from different samples of examinees can be transformed to the same scale. To achieve this, however, IRT models require strong assumptions about the data, such as unidimensionality and local independence. Such assumptions are sometimes difficult to meet with real test data. If a test is speeded, meaning the construct being measured includes extraneous variability due to test time limits, then item responses are not unidimensional and the IRT model likely is not appropriate. Lack of local independence is reflected as a non-zero correlation among the speeded items near the end of the test. Depending on the extent of speededness effects, sometimes, the speededness effect may form a secondary dimension (Lord, 1956; Sager, Peterson, & Oppler, 1994; Wilhelm & Schulze, 2002).

Some previous research has also found the use of IRT models in speeded tests resulted in inaccurate model parameters. Oshima (1994) found test speededness affected both item

parameters and ability parameters. Although relative standing of examinees was relatively stable under speeded conditions, speededness resulted in inaccurate estimates of item parameter estimates. Discrimination and difficulty parameters were overestimated and guessing parameters were underestimated for the speeded items near the end of the test. Yamamoto and Everson (1997) found similar results using the hybrid model. Not only were item difficulty and discrimination parameters overestimated, examinees' abilities were underestimated, particularly especially for the very able examinees.

1.2 Purpose of the Study

A number of attempts have been made to investigate the differential effects of speededness as a function of examinee characteristics such as ethnicity, gender, disability status, and native language (Bridgeman, Cline, & Hessinger, 2003; Lawrence, 1993; Neustel, 1998; Sireci, 2005). Most of these studies have examined the effect of speededness by comparing the means of the examinee's performance as a function of these or similar manifest characteristics. The development of the mixture IRT models has enabled researchers to examine latent groups of examinees by dividing them into speeded and non-speeded latent classes rather than basing an assumption of speededness simply on number of omissions or number of errors on items at the end of the test. In this regard, Bolt, Cohen, and Wollack (2002) and Cohen et al. (2002) used a two-class mixture IRT model to investigate the association between speededness and examinees' background characteristics. This approach was implemented in two steps. First, a specially constrained two-class mixture Rasch model was used to identify speeded and non-speeded examinees. Second, the association between latent class membership and manifest examinee characteristics such as gender, ethnicity, and age, was examined by correlation or regression analysis.

Although the two-step approach has been shown to be useful, it is also possible that some errors are inadvertently allowed into the estimation. Such errors can potentially attenuate the relationships between latent group membership and examinee characteristics as the

two-step procedure doesn't consider the impact of these covariates on parameter estimates. Adams, Wilson, and Wu (1997), for example, showed that attenuation occurred in a two-step regression analysis due to measurement errors. Employing collateral information, however, led to smaller mean squared errors of the ability estimates than with the two-step procedure. That is, inclusion of covariates in the regression analysis may help to reduce these errors and, thereby also may help to reduce the standard errors of the parameter estimates.

In the current study, the two steps will be incorporated into a single step, and manifest examinee characteristics will be used to help place examinees into either speeded or non-speeded classes. This will be done by expressing the IRT model as a multilevel model. Fox and Glas (2001) and Adams, Wilson, and Wu (1997) showed that IRT models can be expressed as multilevel models and can be used to impose a regression model with examinee covariates on examinees' abilities. One advantage of expressing IRT models as multilevel models is to take into account the nested structure of the data. Another advantage is to examine the relationship between individual-level variables and aggregation-level variables. In addition, using latent ability scores as dependent variables instead of observed scores also provides the possibility of separating the impact of examinee's ability and item difficulty and modeling response variation and measurement errors (Fox & Glas, 2001). The model in the current study incorporates the IRT model and the two-class latent class model (i.e., speeded and non-speeded latent classes) in the framework of the generalized linear mixed model. Person and item covariates will be used directly in the model to explain test speeded effects. This one-step procedure will estimate all the model parameters simultaneously.

1.3 Significance of the Study

Von Davier and Yamamoto (2007) note that using background data that are related to latent ability should result in smaller conditional variance for latent ability. For the case of mixture IRT models, there are two latent variables, one is continuous (i.e., ability) and the other is categorical (i.e., latent class). As Von Davier and Yamamoto (2007) note, "Given one or

more conditioning background variables, the multinomial distribution of the class variable will be more concentrated around certain latent classes as compared to the overall distribution, assuming that class membership and background variables are not independently distributed" (p. 110). Therefore, adding covariates related to latent class should improve the estimation accuracy of the class membership. Smit et al. (1999, 2000) estimated latent class membership using a dichotomous mixture IRT model which incorporated background variables as covariates into the mixture IRT model. These were used to help in predicting latent class membership. Studies such as these indicate that the incorporation of background variables into the model can improve the accuracy of classification of examinees into latent classes.

The proposed model in the current study will use background data from the examinees and items and will incorporate these variables into the model for helping determine latent class membership. In this way, it will be possible to investigate the association between test speededness and these kinds of background variables simultaneously along with estimation of mixture IRT model parameters. This one-step procedure should provide more accurate estimation than the two-step procedure. Van Nijlen and Janssen (2009) note that, if person covariates fully explain the latent classes, an interaction effect between the person covariate and items will make the use of mixture IRT models unnecessary. Thus a simpler model could be used.

Chapter 2

Literature Review

The effects of time limits on examinee performance, referred to as speededness effects (Evans & Reilly, 1972), typically have negative impacts on test performance. One result of setting time limits on a test is that some examinees will not have enough time to finish the test. Such differential effects of response latencies, in general, are not usually considered part of the construct of interest for most tests (Lord & Novick, 1968).

2.1 Conventional Methods of Evaluating Test Speededness

Approaches to assessing test speededness have generally been of two types: non-IRT-based methods and IRT-based methods. The conventional methods of evaluating test speededness are usually not based on IRT models. The methods reviewed in this proposal are only those related to paper-and-pencil tests, usually consisting of multiple-choice items. The conventional methods can be further divided into single and double- or multiple-administration approaches. Gulliksen (1950) and Stafford (1971) proposed methods to assess speededness in a single administration design. Both of their methods are functions of the number of not-reached items. Gulliksen evaluated speededness using two ratios, one of which is the ratio of the standard deviation of the number of items answered incorrectly s_w and the standard deviation of the number of items not answered correctly s_x. Items not answered correctly include not-reached, omitted and incorrect answers. The second ratio involves the standard deviation of the number of items not reached s_{nr} and the standard deviation of the number of items answered incorrectly. According to Gulliksen (1950), a test is primarily speeded,

when the ratio $\frac{s_w}{s_x}$ becomes very small (0.1 or less), and a test is primarily a power test, when $\frac{s_{nr}}{s_x}$ is very small (i.e., 0.1 or less).

Stafford (1971) proposed a speededness quotient as the ratio between the number of not-reached items and the sum of the number of items not reached, omitted and answered incorrectly. This quotient can be expressed as $\frac{NR_i}{W_i+O_i+NR_i}$, where NR_i is the number of not-reached items, W_i is the number of incorrect answers, and O_i is the number of omitted items.

Swineford (1974) suggested a rule of thumb for determining whether or not a test is speeded: A test is unspeeded if at least 80% of the examinees reach the last item and virtually all examinees reach at least 75% of the items. Swineford notes that this standard is arbitrary. Secolsky (1989) found this definition provided reasonable distinctions of speeded and non-speeded tests on TOEFL data. On average, about 99.7% of the examinees reached at least 75% of the items and 80% or more of the examinees reached all the items. Secolsky notes that the extent of speededness may be underestimated since these criteria didn't include examinees who randomly guessed or responded with the same responses to the last 25% of the test. Secolsky recommended that, if these criteria were to be accepted, more direct investigations of speededness such as survey or observational study should be conducted to determine the extent of speededness. Rindler (1979) notes that this standard provides a dichotomous measure of power, but it provides relatively little information about the violation of the standard. Further, it offers little help in measuring the degree of speededness.

For multiple test administrations, Cronbach and Warrington (1951) proposed a speededness index, tau, which compares the performance of the same examinee under timed and untimed test administrations of parallel tests: $tau = \frac{r_{A_sB_p}r_{A_pB_s}}{r_{A_pB_s}r_{A_sB_p}}$, where r is the correlation corrected for attenuation between the speeded (s) and power (p) test administrations.

The Cronbach and Warrington approach, though conceptually interesting, is typically administratively impractical as it requires multiple test administrations. In part, this is because not all tests have a parallel form. However, even if a parallel form were available,

it may not be possible to have enough time to administer both forms to the same or even similar examinees. Moreover, examinees usually are not motivated to take a second test, if they have already taken the first.

A major problem with the single test administration methods is they are only appropriate for rights-only scored test. Secolsky (1989) notes, "they are not sensitive enough to the possibility that some portion of the examinee group did not have enough time to truly attempt the items near the end of the test. In reality, some non-ignorable portion of the examinee sample may have responded with random or patterned responses to the items at the end of the test or test section as the time limit approached" (p. 2). A patterned response would be one with the same pattern of responses such as over the last few items on the test. In addition, these methods underestimate test speededness by ignoring the random guessing that may occur near the end of the test (Secolsky, 1989).

2.2 IRT-Based Methods of Evaluating Test Speededness

2.2.1 Bejar's Speededness Indices

In view of the shortcomings of the conventional methods noted above, a number of efforts have focused on assessing speededness within an IRT framework. Bejar (1985) developed two IRT-based indices for detecting the speededness of right-scored tests. Bejar assumed that the most difficult items under speeded conditions would be more likely subject to random guessing. Therefore, an examinee's performance on these items would not be solely determined by that examinee's latent ability. Instead, performance would also be a function of a speededness factor, thereby introducing construct-irrelevant variance into the test score. Bejar's method assumes examinees do not answer the items in sequence, but cycle through the test and leave the most difficult items last. Bejar described an item-level index and an examinee-level index. The item-level index, distributed as a chi-square, is one in which examinees are classified into 15 equally-spaced ability level intervals. The item-level index,

Q, is given as

$$Q = \sum_{j=1}^{15} \frac{[N_j(O_{ij} - E_{ij})]^2}{E_{ij}(1 - E_{ij})}, \tag{2.1}$$

where N_j is the number of examinees in cell j, Q_{ij} is the observed proportion of examinees in cell j who answered item i correctly, and E_{ij} is the predicted proportion of examinees in cell j who answered item i correctly, based on the estimated item parameters from the IRT model used. E_{ij} is calculated as:

$$E_{ij} = \frac{1}{N_j} \sum_{k \in j}^{N_j} P_i(\theta_k), \tag{2.2}$$

where θ_k is the ability estimate obtained from the IRT model. Bejar's Q index assumes that examinees leave the most difficult items for last and the most difficult items are vulnerable to test speededness. It compares the predicted number of correct responses and the number of actual correct responses for each item.

The calculation of the examinee-level index is based on the same assumption as the item-level index. That is, examinees are assumed to not have answered the items sequentially, but to have left the most difficult items for last. To calculate the examinee-level index, all the items are rearranged according to their difficulties with the most difficult items placed at the end of the test. Then the test is divided into two parts, an easy part which contains the first 75% of the items and a hard part which contains 25% of the items. This division is arbitrary, but is adaptable to ETS' rule of thumb for speededness, which states that all the examinees reached 75% of the items and 80% of the examinees reached all the items (Swineford, 1974). The examinee-level index compares the predicted and observed performance only on the hard items since it assumes that only the hard items are vulnerable to test speededness. The computation of the predicted number of correct answers is based on the IRT model used, and is computed as

$$E(\theta_e) = \sum P_i(\theta_e), \tag{2.3}$$

where P_i is the probability of getting an item correct based on the IRT model, and θ_e is the estimated ability calculated from the easy items. The sum is taken over the hard items. The

observed number of correct responses is

$$O = \sum_i u_i, \tag{2.4}$$

and u_i takes on a value of 1 if item i is answered correctly and a value of 0 if not. One problem with Bejar's method is that it is circular (Bejar, 1985), because the IRT model parameters that were used to calculate the expected performance are themselves contaminated by test speededness. Secolsky (1989) noted that Bejar's index may incorporate sources of error that are not solely attributable to test speededness.

2.2.2 Mixture Item Response Model

Recent work in item response modeling has included the development of mixed item response theory models (MixIRTM) for use in detection of speededness (Bolt, Cohen, & Wollack, 2002; Yamamoto, 1989; Yamamoto & Everson, 1997). This work is based on mixture IRT models (e.g., Mislevy & Verhelst, 1990; Rost, 1990). Below, we introduce the MixIRTM followed by descriptions of how it has been used for detection of speededness.

The use of IRT models requires the strong assumption that item difficulties are constant for all persons in the population. In fact, this assumption may be violated by some examinees in the population (Rost, 1990). In some cases, for example, the population may consist of subpopulations which differ only qualitatively, such as in the particular strategies used for responding. The mixing proportions of the subpopulations are usually not known beforehand. The combination of item response model and latent class analysis is a useful statistical tool for taking into accounts both qualitative differences among examinees. An important assumption of this model is that the IRT model holds within each latent class. The mixture IRT models make it possible to divide examinees into latent classes which differ most with respect to item parameters, thus maximizing the between-group differences. Unlike latent class analysis, the mixture IRT models "accounts for ability differences within the mixture components, whereas latent class analysis assumes no differences in conditional response probabilities within each class" (Von Davier & Yamamoto, 2007, p. 115).

Rost (1990) proposed a mixture Rasch model (MRM) which assumes a Rasch model holds within each latent class, but ability levels of examinees may vary within classes and item difficulties may differ among classes. In the MRM, each examinee is characterized by a latent class parameter, g, and an ability parameter, θ_{jg} , for examinee j in group g. The conditional probability of a correct response in the MRM is given as

$$P(x_i = 1|\theta_j, \beta_i, g) = \frac{1}{1 + \exp[-(\theta_j - \beta_{ig})]}, \tag{2.5}$$

where g is the latent group an examinee belongs to; β_{ig} is the item difficulty for item i in latent class g. To remove the indeterminacy of the scale, the sum of the item difficulties within each class can be constrained to be zero, $\sum_i \beta_{ig}$=0. An alternative is that the expected value of the ability estimate is constrained to be zero, $E(\theta)$=0. The unconditional probability that an examinee j answers item i correctly is :

$$P(x_{ij} = 1|\theta_j, \beta_{ig}) = \sum_{g=1}^{G} \frac{\exp(\theta_j - \beta_{ig})}{1 + \exp(\theta_j - \beta_{ig})}, \tag{2.6}$$

with the underlying assumptions that $\sum_g \pi_g$=1 and $0 < \pi_g < 1$. The dichotomous mixture Rasch model has also been extended to include polytomous (Rost, 1991). Under this model, the probability of observing a response vector of $(x_1, x_2..., x_I\ )$ with $x_i \in \{0, 1, ..., m_i\}$ by an examinee is given as:

$$P(X|\theta, g) = \prod_{i=1}^{I} \frac{\exp(x_i\theta - \beta_{ix_ig})}{1 + \sum_{y=1}^{x_i} \exp(y\theta - \beta_{iyg})}, \tag{2.7}$$

where $\beta_{iyg} = \sum_{y=1}^{x} \alpha_{iyg}$ are the class-dependent cumulative item parameters. This model is equivalent to the partial-credit model by Masters (1982).

2.2.3 Mixture IRT Models for Test Speededness

Bolt et al. (2002) proposed a two-class mixture Rasch model with ordinal constraints to model speededness effects. In the Bolt et al. approach, a constrained version of a mixture Rasch model (MRM) developed by Rost (1990) is used to identify two groups of examinees, a

non-speeded group and a speeded group. The mixture Rasch model describes the probability of an examinee getting an item correct as

$$P(x_i = 1|\theta, \beta_i, g) = \frac{1}{1 + \exp[-(\theta - \beta_{ig})]}, \tag{2.8}$$

where g is the latent group an examinee is classified as belonging to and is the difficulty for item i in latent class g. Bolt et al. assumed that items located near the end of the test were most likely to be affected by speededness and items at earlier locations were less likely to be so affected. In the Bolt et al. model, therefore, difficulties of items at the beginning of the test are constrained to be equal. Items at the end of the test, on the other hand, are assumed to be affected by speededness and so are constrained to be harder for the speeded examinees than for the non-speeded examinees. The MRM used by Bolt et al. differs from the model by Rost by the use of these constraints. In the Bolt et al. formulation, items in the middle of the test were not included in the model, although this condition can be relaxed. Wollack, Cohen, and Wells (2003) demonstrated the usefulness of the Bolt et al. model, by showing that scale stability for a college-level English placement test could be maintained over an 11-year period, if item parameters were estimated from responses of the non-speeded group only. An important limitation of the two-class mixture Rasch model with ordinal constraint is it is based on the assumption that the test speededness mainly occurs near the end of the test. As is noted below, the choice of speededness point is arbitrary.

Yamamoto (1987, 1989) proposed a Hybrid model to incorporate examinees' strategy switching. The Hybrid model is a combination of a standard IRT model and a latent class model (Lazarsfeld & Henry, 1968). Examinee responses are modeled with an IRT model up to the strategy switching point, and then with a latent class model thereafter. The qualitative aspects of examinee's performance are captured by the latent class model. Yamamoto (1990, 1995) extended this model to assume that speededness is reflected in multiple latent classes, each differing in the number of consecutive items at the end of the test that are answered randomly. The extended Hybrid model is still based on the assumption that a speeded examinee is assumed to switch response strategies from the use of the latent ability modeled

by the IRT model to a random guessing strategy modeled by the latent class model. The probability of getting an item correct for an examinee is

$$p(x_i = 1|\theta, \beta_i, k) = (1 + \exp(\theta - \beta_i))^{m_{ik}} c_i^{m_{ik}+1}, \tag{2.9}$$

where k is the last speeded item; m_{ik}= -1 if $i \leq k$ and $m_{ik} = 0$, if $i > k$; x_i is the examinee's response to item i, β_i is the item difficulty parameter; θ is the examinee's ability; c_i is the expected proportion correct under a patterned (i.e., same responses to all of the last few items) or a random response strategy. The likelihood of response vector x_v, given θ_v, is

$$P(x_v|\theta, \beta_i, k_v) = \prod_{i=1}^{k_v} P(\theta_v, \beta_i)^{x_{iv}} Q(\theta_v, \beta_i)^{1-x_{iv}} \prod_{i=k_v+1}^{I} c_i^{x_{iv}} (1 - c_i)^{1-x_{iv}}. \tag{2.10}$$

The extended Hybrid model provides a method for reducing the effect of test speededness on the estimates of item and ability parameters. However, the Hybrid model shares the same shortcoming with the two-class mixture Rasch model with ordinal constraint in that it assumes test speededness occurs only at the end of the test. The Hybrid model is not capable of capturing a switch in response strategy, in other words, if test speededness occurs at earlier locations of the test, since the Hybrid model can only allow one strategy switching point.

2.2.4 Gradual Process Change Model

Both the Hybrid model and the Bolt et al. model consider test speededness effects as beginning not later than a specific point on the test for all speeded examinees. This is a useful assumption, but it is not the only way to view the beginning of speededness effects. It is also possible that each examinee may become speeded at a different point on the test (Wollack et al., 2003). This latter assumption is accommodated in the gradual process change model (GPCM; Goegebeur et al., 2008). In the GPCM, the speededness point is modeled as an examinee-specific effect. The model is composed of two parts, one of which is the usual IRT model and the other, the speededness part. The IRT model holds for that part of the test

that is not speeded for examinee j. When the test becomes speeded for examinee j, the probability of success decreases through to the end of the test. The speededness parameters in the model are examinee specific, so the gradual change model includes both the speededness point and the speededness rate of each individual examinee. The probability of a correct response under the gradual change model can be written as

$$P_{ij} = c_i + (1 - c_i) * P_{ij|\theta_j} * min\{1, [1 - (\frac{i}{I} - \eta_j)]^{\lambda_j}\}, \tag{2.11}$$

where c_i is the guessing parameter, I is test length ($i = 1, , I$), η_j is the point on the test expressed as a fraction of the number of items at which the effects of speededness begin for examinee j, and λ_j is the rate at which speededness influences the response of examinee j. If $i \leq \eta_j$, there is no speededness effect, and either λ_j equals zero or η_j equals one. The equation of the gradual change model reduces to the regular IRT model with a guessing parameter under a non-speeded condition. When λ_j is not zero, then η_j indicates the point on the test at which the IRT model no longer completely accounts for the probability of a correct response. The gradual process change model is more flexible than the Hybrid model and the two-class mixture Rasch model with ordinal constraint in that it estimates a speededness rate and a speededness point for each examinee. So, once an examinee becomes speeded, the model will account for the gradual change of strategy switching from a response process described by $P_{ij}|\theta$ to one that includes a gradual switching to a random response. The GPCM also allows examinees to become speeded at different locations, i, instead of at a single arbitrary point like the Bolt et al.two-class mixture Rasch model. The gradual process change model as in Equation (2.11) can also be extended to a mixture gradual process change model (MixGPCM) by inclusion of a latent class model into the model. The assumption with such a model is that different latent classes of speeded or non-speeded examinees may exist, defined by differential use of response strategies. Wang and Cohen (2008) described a mixture Rasch version of this model with ordinal constraints. The probability of a correct response

under the MixGPCM is

$$P_{ij} = \frac{\exp(\theta_j - \beta_{ig})}{1 + \exp(\theta_j - \beta_{ig})} * min\{1, [1 - (\frac{i}{I} - \eta_j)]^{\lambda_j}\}, \tag{2.12}$$

To identify a speeded and non-speeded group, certain constraints were applied on λ_j. That is, when λ_j is zero, a class of non-speeded examinees was identified by the model. When λ_j is greater than zero, a class of speeded examinees was identified.

2.2.5 Multilevel Item Response Model

The IRT models describe the relationship between the examinee's latent ability and responses based on the characteristics of the items. Sometimes examinee characteristics may affect their performance on the test. As the usual IRT models do not incorporate these person characteristics into the model, it is common to use a two-step analysis. In the first step, latent ability is estimated using the IRT model. In the second step, the estimated ability obtained from the first step is used as the dependent variable and the examinee's characteristic variables are used as predictors.

There are some potential problems that can arise using this two-step procedure. One problem is that in regression analysis, the dependent variable is assumed to be measured without error, but the ability estimate under IRT models is estimated with heterogeneous errors. These errors, in turn, will result in non-random errors intruding into the regression model (Kamata, 1998). Another problem with the two-step approach is that the marginal maximum likelihood estimation of ability is not consistent, thereby posing another potential threat to the precision of the regression model. A one-step procedure, however, might be able to improve the precision of estimation.

In addition to providing a one-step procedure, the model we describe in this section further takes into account the fact that much educational data are nested. For example, students are usually nested within classrooms, teachers, schools, districts, neighborhoods, etc. Both the classical statistical analysis and the regular IRT models assume that observations are independent. The reality is examinees from the same nested unit may be more similar

in ability than examinees from other nested units. Neglect of the nested structure of the data will result in biased estimation of the individual effect (Raudenbush & Bryk, 2002). Using multilevel models, the effect of the individual unit (e.g., student) may be divided into a within-group effect and between-group effect.

To overcome this biasing effect, a multilevel item response model has been proposed that incorporates an IRT model into a multilevel structure. As is shown below, this type of model can help take care of the biasing effect on estimation of ability that arises when the multilevel structure is ignored. Further, the threat posed by inconsistent estimates of ability can be removed. The result is a one-step procedure that can help to reduce the standard errors of the ability estimates by simultaneously estimating ability and the effects of examinee covariates.

Previous studies of multilevel IRT models showed that the multilevel IRT model could discriminate better at the individual level and also could explain more variance at the higher levels than the standard multilevel model (Fox & Glas, 2001). The multilevel IRT model also is capable of including covariates at each level. By adding examinee covariates into the IRT models, the effects of examinee characteristic variables at different levels on examinee's ability can be taken into account, thus producing improved estimate of ability.

Kamata (1998, 2001) reformulated the Rasch model as a two-level generalized linear model. The person parameters in Kamata's model were treated as random effects and the items were treated as fixed. In Kamata's model, the first level is the item-level model or the structural model, and the second level is the person-level model. Consistent with usual practice with IRT models, a Bernoulli distribution is assumed for the item response data. Under the Bernoulli distribution, the expected value and the variance of the observed response is

$$E(y_{ij}|p_{ij}) = p_{ij} \tag{2.13}$$

and

$$var(y_{ij}|p_{ij}) = p_{ij}q_{ij} = p_{ij}(1 - p_{ij}), \tag{2.14}$$

where p_{ij} is the probability of examinee j getting item i correct. A logit link is used as the link function. The logit for getting an item correct is

$$\begin{aligned}
\eta_{ij} &= \log \frac{p_{ij}}{1-p_{ij}} \\
&= \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + ... + \beta_{(k-1)j}X_{(k-1)ij} \\
&= \beta_{0j} + \sum_{q=1}^{k-1} \beta_{qj}X_{qij} \\
&= \beta_{0j} + \beta_{qj}, \qquad (2.15)
\end{aligned}$$

where β_{qj} is the coefficient and β_{0j} is the intercept; X_{qji} is a dummy variable for person j on item i, with a value of 1, when $q = i$ and a value of 0, when $q \neq i$. X_{qij} is dropped $X_{qij} = 1$, because $q = i$, and when $q \neq i$, all other $X_{qij} = 0$. So the above equation is basically

$$\begin{aligned}
\eta_{ij} &= \log(\frac{p_{ij}}{1-p_{ij}}) \\
&= \beta_{0j} + \beta_{qj}X_{qij} \\
&= \beta_{0j} + \beta_{qj} * 1 \\
&= \beta_{0j} + \beta_{qj}. \qquad (2.16)
\end{aligned}$$

One of the dummy variables is dropped for person j so that the intercept is the expected item effect for the dropped item and the other coefficients are obtained as the difference between the effect of item i and the effect of the reference item (i.e., the dropped item).

The second level of this model is the person-level. At this level, the intercept β_{0j} is assumed to be a random effect across persons and the item effects are assumed to be fixed across persons. These specifications can be reflected in the following equation:

$$\beta_{0j} = \gamma_{00} + \mu_{0j} \qquad (2.17)$$

$$\beta_{qj} = \gamma_{q0}, \qquad (2.18)$$

where μ_{0j} is the random person effect and is normally distributed with a mean of zero and unit variance. The item parameters are fixed across persons but vary across items, as there

is no random effect term added to the item coefficients excluding the intercept. γ_{00} is the difficulty of the reference item and γ_{00} is the effect of item q.

The combined model can be written as

$$\eta_{ij} = \gamma_{00} + \mu_{0j} + \gamma_{q0}. \tag{2.19}$$

The multilevel Rasch model can be shown to be equivalent to the usual Rasch model: The probability of getting an item i correct by person j under the multilevel Rasch model is

$$p_{ij} = \frac{1}{1 + \exp\{-[\mu_{0j} - (-\gamma_{q0} - \gamma_{00})]}. \tag{2.20}$$

This is equivalent to the Rasch model which is expressed as

$$p_{ij} = \frac{1}{1 + \exp(-\theta_j - \delta_i)}, \tag{2.21}$$

where $\delta_i = \gamma_{q0} - \gamma_{00}$. The above multilevel Rasch model can be extended to a latent variable regression model with person-level predictors, where the level 1 model is the same as above:

$$\begin{aligned} \eta_{ij} &= \log(\frac{p_{ij}}{1 - p_{ij}}) \\ &= \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + ... + \beta_{kj}X_{kij} \\ &= \beta_{0j} + \sum_{q=1}^{k} \beta_{qj}X_{qij} = \beta_{0j} + \beta_{qj}, \end{aligned} \tag{2.22}$$

In the second level model, person-level predictors can be added into the model to account for the effects of predictors on the latent variable:

$$\beta_{0j} = \gamma_{00} + \gamma_{10}W_{1j} + ... + \gamma_{q0}W_{qj} + \mu_{0j}, \tag{2.23}$$

where W_{qj} is the person-level predictor and γ_{q0} is the coefficient of the predictor.

Kamata's formulation of the multilevel Rasch model treats persons as random but items as fixed effects. Noortgate, De Boeck, and Meulders (2003) developed a cross-classified multilevel logistic IRT model in which both items and persons are treated as random. That is, the model assumes both items and examinees are random samples from their respective

populations. The responses are regarded as nested within pairs of persons and items. For educational measurement applications, it is common to have only one observation in each cell of the person by item matrix. In the Noortgate et al. model, the logit for each cell has one fixed component and two random components. One of the components is related to the item effect and the other to the person effect. The level 1 model of this cross-classified IRT model is similar to Kamata's level 1 model and is expressed by the following equation:

$$\eta_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_{ij}, \tag{2.24}$$

where β_{0j} is ability for person j and β_{ij} is item easiness. (In this model, β_{ij} is the negative of the usual IRT item difficulty.)

The level-2 model then takes into account the random effects introduced by the variation among persons and the variation among items. At level 2, covariates related to persons and items also can be added to help explain random variance among persons and items. The level 2 model can be expressed as

$$\beta_{0j} = u_{1j}, \tag{2.25}$$

$$\beta ij = \gamma_0 + u_{2i}, \tag{2.26}$$

where u_{1j} is the random effect associated with the person j. Since the mean of ability is constrained to be zero to solve the identification problem, the intercept of the ability estimate is set to zero. Since both u_{1j} and u_{2i} have a mean of 0, the intercept γ_0 can be interpreted as the mean logit. It can also be interpreted as the mean of the item parameters across the whole test. u_{2i} is the random effect associated with items. The combined cross-classified IRT model can then be expressed as

$$\eta_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \gamma_0 + u_{1j} + u_{2i}. \tag{2.27}$$

To improve estimation, the variances of the item residuals are constrained to be equal and the covariances are constrained to be zero. So, the covariance matrix of the residuals is a diagonal matrix with equal values on the diagonal and zeros off the diagonal.

Although the description of the cross-classified IRT model by Noortgate et al. (2003) was based on the regular Rasch model, the model can also be extended to other IRT models such as the 2-parameter logistic (2PL) or the 3-parameter logistic (3PL) model. For the 2PL cross-classified IRT model, the rationale is the same as the Rasch model. The only difference is the 2PL model has one more item parameter which can be interpreted as item discrimination (Noortgate et al., 2003). Birnbaum (1968) proposed a 2PL IRT model which can be expressed as:

$$\eta_{ij} = \alpha_i \theta_j - \beta_i, \tag{2.28}$$

where α_i can be regarded as the item discrimination parameter, θ_j is still ability for person j and β_i is the item difficulty parameter for item i.

Based on this model, the 2PL cross-classified IRT model is

$$\eta_{ij} = \gamma_{00} + u_{1j} u_{3i} + u_{2i}, \tag{2.29}$$

where u_{1j} is the random effect associated with the person's ability; u_{2i} is the random effect associated with the item easiness; and u_{3i} is the random effect associated with the item discrimination. The interaction term $u_{1j}u_{3i}$ indicates the interaction effect of ability and item discrimination on the probability of getting an item correct.

Covariates for items and persons can be added into the cross-classified IRT model to help explain the probability of responses. The following is an example of a cross-classified IRT model with person and item covariates:

$$\eta_{ij} = \beta_0 + \sum_{a=1}^{A} \beta_a M_{ai} + \sum_{b=1}^{B} \beta_b N_{bj} + \sum_{w=1}^{W} \beta_w W_{wij} + u_{1j} + u_{2i}, \tag{2.30}$$

where M_{ai} is the item covariate; N_{bj} is the person covariate; and W_{wij} is the person by item interaction covariate. The βs are the coefficients related to each covariate, u_{1j} and u_{2i} are the random effects associated with the person and item, respectively. To estimate the model, a normal distribution with a mean of zero and a variance of τ is assumed for both item and person random effects. The cross-classified IRT models can also be extended to more

levels such as would be the case if persons are grouped or items are grouped based on some additional characteristics.

2.2.6 A Mixture Cross-Classified Item Response Model

The multilevel IRT models introduced in the previous section evaluate the examinee's ability as a random effect, but sometimes, the population of examinees consists of subpopulations which do not share the same ability distribution. These subpopulations are not known a priori, but need to be detected. The regular multilevel IRT models are not capable of modeling this heterogeneity in the examinee population. One model which has the potential to model this heterogeneity is a multilevel mixture IRT model (MMixIRTM). The MMixIRTM is a combination of the regular multilevel IRT model and the mixture IRT model. The model is useful in this regard as it allows different specifications of ability distributions for the different latent subpopulations. The MMixIRTM models the heterogeneity in the population by identifying latent classes of examinees which are homogeneous within each class. The multilevel IRT model holds within each latent class, but model parameters may differ across classes. Cho (2007) incorporated the mixture model into the multilevel IRT model and extended the model to include both student-level and school-level mixtures. Latent classes of students and schools who performed differentially on particular items are detected. Mixtures at student level are classified based on students' response patterns and school-level latent classes are identified according to the proportions of student-level latent classes (Cho, 2007).

The multilevel mixture IRT model described here is motivated by Kamata's hierarchical generalized linear model and by Noortgate et al.'s cross-classification multilevel logistic models. Unlike the regular multilevel IRT models, the multilevel mixture IRT model in this study treats both item and person as random such that covariates related to persons and items can be added into the model to help explain the effects of test speededness. In the case of mixture IRT models, the inclusion of covariates can be used to help classify the examinees into different latent groups (Cho, Cohen, & Kim, 2006; Smit et al., 1999, 2000).

Level-One Model. Under the generalized linear model framework, Rijmen, Tuerlinckx, De Boeck, and Kuppens (2003) showed that the Rasch model can be expressed as

$$\eta_{ij} = \theta_j + \beta_i, \tag{2.31}$$

where θ_j is the person's ability and β_i is the item easiness (that is, -β_i is the usual item difficulty). At the first level, the item responses are regarded as nested under each item $\times$ person pair. In this way, the level-1 model captures the variation among the responses, and is formulated as

$$\log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \eta_{ij} = \theta_j + \sum_{i=1}^{I} \beta_{ig} X_{ij} = \theta_j + \beta_{ig}, \tag{2.32}$$

with $Y_{ij} \sim \text{Bernoulli}(P_{ij})$. β_{ig} is the item easiness parameter for item i in latent class g and θ_j is the ability estimate of examinee j; X_{ij} is a dummy variable which is coded as 0 if person j responds to item i and a value of 0 if not. As is shown in the right hand side of the above equation, the indicator X_{ij} can be dropped and only β_{ig} for item i in class g is needed.

Level-Two Model. At this level, random effects associated with persons and items can be specified in the model. The level-2 model is given as

$$\theta_j = \gamma_{0j} + u_{1j} \tag{2.33}$$

$$\gamma_{0j} \sim N(\mu_g, 1) \tag{2.34}$$

$$u_{1j} \sim N(0, \sigma_1^2) \tag{2.35}$$

with $\mu_1 = 0$.

$$\beta_{ig} = \gamma_{ig} + u_{2i} \tag{2.36}$$

$$\gamma_{ig} \sim (\overline{\beta_g}, 1) \tag{2.37}$$

$$u_{2i} \sim N(0, \sigma_2^2). \tag{2.38}$$

That is, γ_{0j} is the fixed person effect, and is normally distributed with a mean of μ_g and variance of one; μ_g is the mean of the ability estimates; and, u_{1j} is the random person effect, which is also normally distributed with a mean of zero and variance of σ_1^2. The mean ability

of Class 1 is fixed to zero for identification. γ_{ig} is the fixed item effect for item i in latent class g and is assumed to have a normal distribution with a mean of $\overline{\beta_g}$, where $\overline{\beta_g}$ is the mean of the item easiness estimates, and unit variance. u_{2i} is the random item effect which is normally distributed with a mean of zero and variance of σ_2^2.

The fixed and random effects can also be equivalently specified as the following:

$$\theta_j \sim N(\mu_g, \sigma_1^2) \tag{2.39}$$

$$\beta_{ig} \sim N(\overline{\beta_g}, \sigma_2^2) \tag{2.40}$$

The level-2 model described above is shown without either person or item covariates. This is referred to as an unconditional model. This unconditional model can also be extended into a latent variable regression model by incorporating person and item covariates to explain differences in person ability and item easiness. Person-by-item covariates can also be included at this level and can be used, for example, to help detect problematic items, such as those exhibiting DIF. The level-2 model with person and item predictors is given as

The level-2 model can be extended into a latent variable regression model by incorporating person and item predictors. The level-2 model with person and item predictors is given as

$$\theta_j = \gamma_{0j} + u_{1j} \tag{2.41}$$

$$\gamma_{0j} = b_{0j} + \sum_{n=1}^{N} b_n N_j \tag{2.42}$$

$$b_{0j} \sim N(0, 1) \tag{2.43}$$

$$u_{1j} \sim N(0, \sigma_1^2) \tag{2.44}$$

$$\beta_{ig} = \gamma_{ig} + u_{2i} \tag{2.45}$$

$$\gamma_{ig} = \lambda_{ig} + \sum_{m=1}^{M} \lambda_{mg} M_i \tag{2.46}$$

$$\lambda_{ig} \sim N(0, 1) \tag{2.47}$$

$$u_{2i} \sim N(0, \sigma_2^2). \tag{2.48}$$

where γ_{0j} is the fixed person effect, and is usually assumed to be normally distributed with a mean of μ_{jg} and variance of one; b_{0j} is the intercept when the person covariates have no impact on examinees' ability estimates.

b_n is the coefficient of the person covariates on the ability estimate for examinee j. u_{1j} is the random person effect which is assumed to be normally distributed with a mean of zero and variance of σ_1^2. γ_{ig} is the fixed item effect which also has a normal distribution with a mean of $\bar{\beta}_{ig}$ and variance of 1. λ_{ig} are the item parameter estimates in latent class g, when the item covariates have no influence on the item easiness estimates and λ_{mg} is the coefficient of the item covariates on the estimation of the item easiness in latent class g. u_{2i} is the random item effect which is normally distributed with a mean of zero and variance of σ_2^2. The random person and item effects are assumed to be equal across the latent groups. The item variances within each latent class are constrained to be equal and the covariances are constrained to be zero within each latent class. Therefore, the variance and covariance matrix for the item parameters within each latent class are diagonal matrices with the same values on the diagonal and zeros off the diagonal. The level-2 model with covariates is equivalent to the following specification:

$$\theta_j \sim N(\mu_{jg}, \sigma_1^2) \tag{2.49}$$

$$\mu_{jg} = b_{0g} + \sum_{n=1}^{N} b_n N_j \tag{2.50}$$

and

$$\beta_{ig} \sim N(\overline{\beta_{ig}}, \sigma_2^2) \tag{2.51}$$

and

$$\overline{\beta_{ig}} = \lambda_{0g} + \sum_{m=1}^{M} \lambda_{mg} M_i. \tag{2.52}$$

Combined Model. In this section, we describe the combined unconditional cross-classified multilevel mixture IRT model (cc-MMixIRTM). The full unconditional model (i.e., the model without person or item covariates) is given as

$$\eta_{ij} = \gamma_{0j} + \gamma_{ig} + u_{1j} + u_{2i}, \tag{2.53}$$

with

$$\gamma_{0j} \sim N(\mu_g, 1) \tag{2.54}$$

$$u_{1j} \sim N(0, \sigma_1^2) \tag{2.55}$$

$$\gamma_{0ig} \sim N(\overline{\beta_g}, 1) \tag{2.56}$$

$$u_{2i} \sim N(0, \sigma_2^2). \tag{2.57}$$

Similarly, the combined conditional cc-MMixIRTM with item and person covariates can be expressed by the following equation:

$$\eta_{ij} = b_{0j} + \sum_{n=1}^{N} b_n N_j + \lambda_{ig} + \sum_{m=1}^{M} \lambda_{mg} M_i + u_{1j} + u_{2i}, \tag{2.58}$$

where b_{0g} is the mean of ability estimate within latent g, when the person covariate has no effect on ability estimate. b_{0g} is assumed to follow a normal distribution with a mean of 0 and variance of 1. b_{pg} is the regression coefficient associated with the person covariate within latent class g. λ_{0g} is the mean of item easiness estimates within latent class g. λ_{mg} is the regression coefficient associated with the item covariate within latent class g. The combined conditional model is equivalent to the following:

$$\eta_{ij} = \theta_j + \beta_{ig} \tag{2.59}$$

$$\theta_j \sim N(\mu_{jg}, \sigma_1^2) \tag{2.60}$$

$$\mu_{jg} = b_{0g} + \sum_{n=1}^{N} b_n N_j \tag{2.61}$$

$$\beta_{ig} \sim N(\bar{\beta_{ig}}, \sigma_2^2) \tag{2.62}$$

$$\bar{\beta_{ig}} = \lambda_{0g} + \sum_{m=1}^{M} \lambda_{mg} M_i. \tag{2.63}$$

The mean of the ability for latent class 1 is still constrained to zero ($b_{0g} = 0$).

The conditional cross-classified IRT model given above presents an additional advantage from treating items as random instead of fixed. When items are treated as fixed and covariates related to items are included in the conditional model, there is no error term related to the

item effect included in the model. The absence of errors related to item parameters indicates these item covariates explain all the variance in the item parameters. This is usually an unrealistic assumption (see, e.g., De Boeck, 2008). Treating items as random, however, allows an error term associated with item effects to be included in the model. The inclusion of item covariates will explain some part of the variance in the item parameters but not all of it.

2.2.7 A Mixture Multilevel IRT Model for Test Speededness

In this section, we describe an application of the multilevel cross-classified mixture IRT (cc-MMixIRT) model to detection of speededness in a paper-and-pencil test.

Speededness effects can arise, when tests are timed and examinees feel there is insufficient time to answer all items on the test. Several attempts have been made to detect speededness effects. In this application, we apply the cc-MMixIRT model to the test speededness problem. To do this, we use assumptions about speededness similar to those by Bolt et al. (2002): We assume items at earlier locations of the test are not affected by test speededness, but items near the end of the test are more affected by test speededness. The Bolt et al. model is actually a two-class mixture IRT model in which one class is defined by model constraints to be composed of non-speeded examinees and the second class is defined to be composed of speeded examinees (i.e., examinees whose responses reflect a speededness effect).

In the mixture IRT model, item parameters for the non-speeded items are fixed to be equal in both the speeded and non-speeded groups. For items in the speeded locations of the test, that is, for items near the end of the test, the item easiness parameters are assumed to be larger for the non-speeded group than for the speeded group. The item variances are constrained to be equal across the whole test and there is assumed to be no covariation among items under the local independence assumption of IRT. In the mixture model, the item variances within each latent class are the same but may differ across classes. The unconditional multilevel mixture model for test speededness can be expressed by the following

equations:

$$\eta_{ij} = \theta_j + \beta_{ijg} \tag{2.64}$$

$$\theta_j \sim N(\mu_g, \sigma_1^2) \tag{2.65}$$

$$\beta_{ijg} \sim N(\bar{\beta}_g, \sigma_2^2) \tag{2.66}$$

To reflect the assumption about test speededness, constraints are imposed on item parameters; that is, $\bar{\beta}_1 = \bar{\beta}_2$ for items at earlier locations of the test where items are assumed unaffected by test speededness and $\bar{\beta}_1 > \bar{\beta}_2$ for items near the end of the test where items are assumed to most likely be affected by test speededness. For the conditional model, the same constraints are imposed on item parameters. That is, $\bar{\beta}_1 = \bar{\beta}_2$ for non-speeded items at early locations of the test and $\bar{\beta}_1 > \bar{\beta}_2$ for speeded items at the end of the test.

Smit, Kelderman, and Van der Flier (1999, 2000) show that the incorporation of collateral variables in the mixed Rasch model and the 2 parameter IRT model improved the classification accuracy of subjects into latent classes and reduced the standard errors of parameter estimates. When the sample size increases or when the association between collateral variables and the latent class becomes stronger, the accuracy of the class membership assignment approaches 1. Lubke and Muthén (2007) investigated the effect of covariates on latent class membership by regressing latent class variable on covariates. Results indicated correct assignment of latent classes increased with the increase of covariate effects. Even a small covariate effect reduced the classification error. It is also possible to improve the recovery of factor means by including covariate.

To differentiate the speeded and non-speeded group in this study, person covariates were included to help predict latent classes. In the example presented here, an unconditional cross-classification model with no covariates was run to classify examinees into speeded and non-speeded groups. Then cross-tabulations were examined between latent groups and examinee background variables. Those variables which had a significant association with the latent group membership were subsequently included in a latent variable regression model to help predict the latent classes. Probabilities of latent classes are modeled as a multinomial

logit regression due to the inclusion of covariates (Cho, Cohen, & Kim, 2006). The likelihood of a response vector X_j based on the mixture cross-classified IRT model is expressed as:

$$P(X_j) = \sum_{g=1}^{G} \pi_g \prod_{i=1}^{k} p^{x_{ij}} (1-p)^{1-x_{ij}} \tag{2.67}$$

and the probabilities of mixtures with covariates can be expressed as:

$$\pi_{jg}|X_j = \frac{\exp(\lambda_{0g} + \sum_{c=1}^{C} \lambda_{cg} X_{jc})}{\sum_{g=1}^{G} \exp(\lambda_{0g} + \sum_{c=1}^{C} \lambda_{cg} X_{jc})} \tag{2.68}$$

Where π_{jg} is the probability for examinee j to be in latent class g; λ_{0g} is the class-specific intercept when the covariate X_{jc} has no effect on the probability of group membership; λ_{cg} is the class-specific effects of covariate X_{jc} on the probability of group membership. λ_{0g} and λ_{c1} are constrained to be 0 for identification (Cho, Cohen, & Kim, 2006).

Chapter 3

Methods and Research Design

3.1 Estimation for the Multilevel Mixture IRT Model Under a Bayesian Framework

In this chapter, we present the methods used for estimating the parameters of the new model, followed by a real data example showing how the model can be used, and then a design for the simulation study to evaluate the performance of the new model. The cross-classified multilevel mixture IRT (cc-MMixIRT) models in the present study were estimated using a Markov chain Monte Carlo (MCMC) algorithm as implemented in the computer program WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003).

Ricci and Ye (2009) compared penalized quasi-likelihood (PQL) estimation, and used MCMC algorithms for estimating the model parameters of the cross-classified IRT model. Results indicated that using an MCMC algorithm provided better parameter estimates than using PQL estimation in terms of lower estimation bias and Type I error rates of both item and person covariates. The MCMC algorithm also has been found useful for estimating complex IRT models (Baker, 1998; Kim & Bolt, 2007; Patz & Junker, 1999) such as mixture IRT models (Bolt et al., 2001).

To implement MCMC estimation, Markov chains, which are sequences of random samples for each of the parameters being estimated in the model, were constructed using Gibbs sampling. In MCMC estimation, random samples are repeatedly drawn from the full posterior distributions of the model parameters. After sufficient iterations (burn-in) have been run so that the chains can be assumed to have converged to a stationary distribution, then the remaining iterations are used to approximate the expectations of the model parameters.

With MCMC estimation, we are interested in obtaining the joint posterior distribution of all the model parameters and the observed responses. The joint posterior distribution is proportional to the joint prior distribution and the probability distribution of the observed responses. In the hierarchical part of the multilevel model we assume the prior distributions for the model parameters are not known and have their own prior distributions.

As in equation 3.2, the probability of getting an item correct depends only on the model parameters; the hyper-parameters affect the probability only through the model parameters. As an example, for the unconditional multilevel mixture IRT model, the prior distribution can be expressed as

$$P(\mu_g, \bar{\beta}_g, \sigma_1^2, \sigma_2^2, \theta_j, g, \beta_{ig}) = P(\mu_g, \bar{\beta}_g, \sigma_1^2, \sigma_2^2) \times P(\theta_j, g, \beta_{ig} | \mu_g, \bar{\beta}_g, \sigma_1^2, \sigma_2^2), \tag{3.1}$$

where μ_g, $\bar{\beta}_g$,σ_1^2, σ_2^2 are the hyper-priors, that is, the prior distributions of the upper levels of the hierarchical priors; and θ_j, g, and β_{ig} are the model parameters. The joint posterior distribution for the unconditional model is

$$\begin{aligned} & P(\mu_g, \bar{\beta}_g, \sigma_1^2, \sigma_2^2, \theta_j, g, \beta_{ig} | Y_{ij}) \\ & \quad \propto P(\mu_g, \bar{\beta}_g, \sigma_1^2, \sigma_2^2, \theta_j, g, \beta_{ig}) \times P(Y_{ij} | \mu_g, \bar{\beta}_g, \sigma_1^2, \sigma_2^2, \theta_j, g, \beta_{ig}) \\ & \quad = P(\mu_g, \bar{\beta}_g, \sigma_1^2, \sigma_2^2) \times P(\theta_j, g, \bar{\beta}_g, \lambda_{0c}, \lambda_{cg} | \mu_g, \bar{\beta}_g, \sigma_1^2, \sigma_2^2) \times P(Y_{ij} | \theta_j, g, \bar{\beta}_g, \lambda_{0c}, \lambda_{cg}), \end{aligned} \tag{3.2}$$

and the posterior distribution for the conditional model with gender covariate is

$$\begin{aligned} & P(\mu_g, \bar{\beta}_g, \sigma_1^2, \sigma_2^2, \lambda_{0c}, \lambda_{cg}, \theta_j, g, \beta_{ig} | Y_{ij}) \\ & \quad \propto P(\mu_g, \bar{\beta}_g, \sigma_1^2, \sigma_2^2, \lambda_{0c}, \lambda_{cg}, \theta_j, g, \beta_{ig}) \times P(Y_{ij} | \mu_g, \bar{\beta}_g, \sigma_1^2, \sigma_2^2, \lambda_{0c}, \lambda_{cg}, \theta_j, g, \beta_{ig}) \\ & \quad = P(\mu_g, \bar{\beta}_g, \sigma_1^2, \sigma_2^2) \times P(\theta_j, g, \bar{\beta}_g, \lambda_{0c}, \lambda_{cg}) | \mu_g, \bar{\beta}_g, \sigma_1^2, \sigma_2^2) \times P(Y_{ij} | \theta_j, g, \bar{\beta}_g, \lambda_{0c}, \lambda_{cg}), \end{aligned} \tag{3.3}$$

where $P(Y_{ij}|\theta_j, g, \beta_{ig})$ is the probability of answering item i correctly by examinee j under the mixture Rasch model.

3.1.1 Priors

As in all Bayesian analysis, the specification of prior distributions should be consistent with knowledge about the problem. When there is no prior knowledge about the distributions of

parameters, then non-informative priors should be used. The parameter estimates obtained using non-informative priors are basically the same as those obtained by the maximum likelihood estimation method. The use of priors may reduce the variability measures and pull the parameter estimates toward the prior means, particularly with small sample sizes (Kim, 2007). It is often practical to start with a simple and relatively non-informative prior distribution on the hyper-priors and then try to use more informative prior distributions if that seems appropriate. For the unconditional model with no covariate,

$$\eta_{ij} = \theta_j + \beta_{ijg} \tag{3.4}$$

with

$$\theta_j \sim N(\mu_g, \sigma_1^2) \tag{3.5}$$

and

$$\beta_{ijg} \sim N(\bar{\beta}_g, \sigma_2^2), \tag{3.6}$$

the following priors were used in this study:

$$\theta_j \sim N(\mu_g, \sigma_1^2) \tag{3.7}$$

$$\mu_1 = 0 \tag{3.8}$$

$$\mu_2 \sim N(0, 1) \tag{3.9}$$

$$\sigma_1^2 \sim uniform(0, 2) \tag{3.10}$$

$$\beta_{ig} \sim (\bar{\beta}_g, \sigma_2^2) \tag{3.11}$$

$$\bar{\beta}_g \sim N(0, 1) \tag{3.12}$$

$$\sigma_2^2 \sim uniform(0, 2) \tag{3.13}$$

The constraint on μ_1 was used for identification. For the conditional model with item and person covariates, the priors were the same as those for the unconditional model.

3.1.2 Checking Convergence

The convergence of an MCMC algorithm refers to situations in which the Markov chain reaches its stationary distribution. The stationary distribution is the posterior distribution following burn-in. Burn-in is that part of the MCMC chain that is discarded as the sampled parameters are not assumed to be from the correct or target distribution. That is, it is assumed to be the correct target distribution that is obtained when the MCMC chain has converged. This stationary distribution is then used to generate subsequent values in the MCMC chain. Once convergence has been realized, subsequent sampled values are used to estimate the posterior distribution for each parameter being estimated.

An important characteristic of a stationary distribution is that the sampled values drawn from the stationary distributions (i.e., after burn-in) are independent of the initial values. Generally, it is not clear about how many iterations in the MCMC chain need to be run to reach a stationary distribution. Further, the length of the burn-in may differ for different parameters given the data. Several tools exist for assessing convergence in the computer software WinBUGS (Spiegelhalter et al., 2003). The trace plots of the parameter estimates can be examined across iterations to see if the patterns of estimates have stabilized. When examining trace plots, for converged chains, there should be no horizontal bands with upward or downward trends. WinBUGS also provides autocorrelations between estimates of a parameter in a chain. Low or high autocorrelations imply fast or slow mixing within a chain indicating whether a chain is converging quickly or slowly. The lower the autocorrelation, the closer the chain is to converging. If the autocorrelation approaches zero, the simulation is considered to have reached a stationary state and the model converged.

Additional tools in the form of statistical tests for checking convergence are available in the computer program Bayesian Output Analysis Program (BOA; Smith, 2005). The following tools from BOA also were employed to check the convergence of the model. One of these, the Geweke convergence diagnostic is appropriate for the analysis of individual chains. Geweke (1992) proposed a method to compare the means of the sampled values from

two non-overlapping parts (usually the first 0.1 and last 0.5 portions) of the chain to check the independence between the two means. The Z-statistic calculated based on the difference between the two means should be not statistically significant, if the model converges (Smith, 2005). In this study, this statistic was evaluated at $\alpha = .05$.

Another diagnostic statistic used from BOA is the Heidelberg and Welch diagnostic. This diagnostic consists of two parts. The first part calculates the Cramer-von-Mises statistic using the whole chain to test whether the chain has reached its stationary state, indicating the model has converged. If there is evidence of nonstationarity, then the first 10% of the iterations are discarded and the test is repeated. The test continues to be repeated until the chain converges or at least 50% of the iterations are discarded. The Heidelberg and Welch diagnostic reports the number of iterations discarded and the number of iterations needed to keep to reach the stationary distribution. The second part of the diagnostic is a halfwidth test in which the portions of the chain which passed the stationary test are used to calculate half the width of the (1-α)% credibility interval around the posterior mean. If the halfwidth test fails, a longer run is required to accurately estimate the posterior mean of the parameter.

3.2 A Simulation Study

The purpose of this simulation study was to determine how conditional and unconditional forms of the mixture cross-classified model performed for detection of test speededness. Recall that the unconditional form was the cc-MMixIRT model without covariates and the conditional form was the model with covariates.

3.2.1 Simulation Conditions

Two factors were simulated in the simulation study. One factor to be manipulated was sample size. The usual result with IRT models is that item parameter estimates and scores tend to be estimated more accurately as sample size increases. In addition, for Bayesian estimation, when the sample size is small, the parameter estimates tend to be pulled more strongly

toward the means of the priors. When the sample size is large, however, the likelihood based on the data will usually dominate the parameter estimates. Li, Cohen, Kim, and Cho (2009) showed that when a test had 30 items, a sample size of 600 recovered the model parameters well for the mixture Rasch model. Therefore, the sample sizes of 1,000, 2,000 and 3,000 were simulated in this study. A pilot run using sample sizes of 1,000 and 3,000 were run on the unconditional model. The results for the two sample sizes were compared with the results for a sample size of 2,000. The standard errors decreased with the increase in the sample size, but the differences in the standard errors between the sample of 2,000 and the sample of 3,000 were very small (two points in the second decimal place). Therefore, a sample of 3,000 was viewed as sufficient for estimation of recovery performance in a large sample.

Another factor manipulated was the proportion of speeded and non-speeded examinees. Three proportions of test speededness were simulated: little speededness, moderate speededness and high speededness. To simulate these proportions, 10%, 20% and 30% of the samples included speeded examinees. The classification accuracy of the mixture cross-classified Rasch model was checked by comparing these proportions with the proportions recovered from the estimated obtained from the generated data.

There were a total of 9 conditions for the unconditional model: 3 sample size $\times$ 3 proportions of speededness and a total of 9 conditions for the conditional model: 3 sample size $\times$ 3 proportions of speededness. Thus, there were a total of 18 conditions for both models with 20 replications for each condition, resulting in a total of 360 simulated data sets analyzed in the simulation study.

3.2.2 Data Simulation Procedures

In the simulation study, the response data of the unconditional model were generated based on equations (2.64, 2.65, and 2.66). For the conditional model, gender was simulated as a covariate on the latent group membership; its impact was modeled as in equation (2.68).

During the generation of the response data for the conditional model, the gender effects were fixed.

Item parameters estimated from a real data set by the unconditional model and the conditional model were used as generating values for the unconditional and conditional models, respectively. There were 28 items in the real data analysis. The first 20 items were assumed not to be affected by test speededness. Therefore, the item parameters for the first 20 items were fixed and were assumed to be equal across latent classes. The last 8 items were assumed to be most affected by test speededness and their item parameters were constrained to be unequal across latent classes. Therefore, the item parameters for the first 20 items were not estimated during the estimation of the model. Only the parameters of the last 8 items were estimated.

A one-group cross-classified Rasch model was applied to a real data set and the item parameter estimates from Item 1 to Item 20 from the real data were fixed in both the unconditional and conditional model. The item easiness parameters from Item 21 to Item 28 were generated by fixing the variances of the random item and person effects obtained from a real data. After generating the item easiness parameters for the last 8 items, the item easiness parameters for all the items were fixed for each of the conditions during the generation of the response data. The generating variances of the random item and person effects are presented in Table 3.1. The generating items parameters for the unconditional model and the conditional model are presented in Tables 3.2 and 3.3, respectively. Abilities for each of the latent classes under both models were generated from a normal distribution with a mean of 0 and variance of 1.

The proportions of speeded and non-speeded examinees were manipulated by specifying the latent group membership of each examinee during data generation procedure. The proportions of speeded and non-speeded examinees simulated are listed in table 3.4. Response data were generated based on the unconditional model and the conditional model.

Table 3.1: Generating Variances of the Random Item and Person Effects for the Unconditional and Conditional Models.

	σ_i^2	σ_p^2
Unconditional Model	1.116	1.149
Conditional Model	1.146	1.193

Table 3.2: Item Easiness Generating Parameters for the Unconditional Model.

	Item Parameters	
Items	Speeded Group	Non-Speeded Group
1	0.889	0.889
2	1.802	1.802
3	1.665	1.665
4	0.755	0.755
5	1.693	1.693
6	1.979	1.979
7	1.146	1.146
8	-0.205	-0.205
9	1.463	1.463
10	0.702	0.702
11	1.566	1.566
12	0.912	0.912
13	-0.176	-0.176
14	-0.399	-0.399
15	0.432	0.432
16	0.835	0.835
17	-0.641	-0.641
18	1.622	1.622
19	0.760	0.760
20	-1.012	-1.012
21	-0.694	0.512
22	-2.168	1.073
23	-2.875	-1.156
24	-1.508	-0.226
25	-1.070	-0.832
26	-3.720	-0.636
27	-2.861	-1.370
28	-2.021	-0.086

Table 3.3: Item Easiness Generating Parameters for the Conditional Model.

	Item Parameters	
Items	Speeded Group	Non-Speeded Group
1	0.889	0.889
2	1.802	1.802
3	1.665	1.665
4	0.755	0.755
5	1.693	1.693
6	1.979	1.979
7	1.146	1.146
8	-0.205	-0.205
9	1.463	1.463
10	0.702	0.702
11	1.566	1.566
12	0.912	0.912
13	-0.176	-0.176
14	-0.399	-0.399
15	0.432	0.432
16	0.835	0.835
17	-0.641	-0.641
18	1.622	1.622
19	0.760	0.760
20	-1.012	-1.012
21	-0.319	0.592
22	-1.744	1.134
23	-2.428	-1.022
24	-1.106	-0.123
25	-3.246	-0.709
26	-2.414	-0.520
27	-1.603	-1.229
28	-2.871	0.012

Table 3.4: Generating Proportions of Speeded and Non-Speeded Examinees.

Speeded	Non-Speeded
10	90
20	80
30	70

3.2.3 Recovery Analysis

A recovery analysis was done to examine the performance of the mixture cross-classified Rasch model for test speededness. The objective of this analysis was to determine the extent to which the generating parameters could be recovered from the simulated data sets generated by the unconditional and conditional models under the different simulation conditions. The recovery analysis compared the generating values with the estimates for each of the following: item easiness parameters, variances of the random person and item effects, and proportions of speeded and non-speeded examinees. For the recovery of the item easiness parameters, the root mean square error (RMSE), and bias were computed across replications and items. The bias of item easiness parameter is

$$bias_{\beta} = \frac{\sum_{i=1}^{I}\sum_{r=1}^{R}(\hat{\beta_{ir}} - \beta_{ig})}{RI}, \tag{3.14}$$

and the RMSE is expressed as:

$$RMSE_{\beta} = \sqrt{\frac{\sum_{i=1}^{I}\sum_{r=1}^{R}(\hat{\beta_{ir}} - \beta_{ig})^2}{RI}}, \tag{3.15}$$

where β_{ir} is the estimated item easiness parameter for item i in replication r and β_{ig} is the generating item easiness parameter for item i. R is the number of replications and I is the number of items. The bias and RMSE of the variances of the random item and person effects were calculated in a similar way:

$$bias_{\sigma_i^2} = \frac{\sum_{r=1}^{R}(\hat{\sigma_{ir}^2} - \sigma_{ig}^2)}{R} \tag{3.16}$$

$$bias_{\sigma_p^2} = \frac{\sum_{r=1}^{R}(\hat{\sigma_{pr}^2} - \sigma_{pg}^2)}{R}, \tag{3.17}$$

where

- σ_i^2 is the variance of the random effect for item i;
- $\hat{\sigma_{ir}^2}$ is the estimated variance of the random item effect for replication r;
- σ_{ig}^2 is the generating variance of the random item effect;

- σ_p^2 is the variance of the random effect for person p;
- $\hat{\sigma_{pr}^2}$ is the estimated variance of the random person effect for replication r;
- σ_{pg}^2 is the generating variance of the random person effect;
- R is the number of replications.

The RMSEs for the variances of the random item and person effects are specified as:

$$RMSE_{\sigma_i^2} = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\sigma_{ir}^2} - \sigma_{ig}^2)^2}{R}} \tag{3.18}$$

$$RMSE_{\sigma_p^2} = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\sigma_{pr}^2} - \sigma_{pg}^2)^2}{R}} \tag{3.19}$$

The recovery of the proportions of speeded and non-speeded examinees was examined by calculating the proportions of examinees who were correctly classified into the speeded and non-speeded groups.

Before calculating the bias and RMSE for the item easiness parameters, the estimated item easiness parameters were first transformed onto the same metric with the generating item parameters. This was done by subtracting the difference in the means of ability parameters as expressed by the following equation:

$$\beta_T^* = \beta_T - (\mu_T - \mu_B), \tag{3.20}$$

where T is the target scale or the estimated scale in the simulation study; B is the base scale or the scale of the generating data, β_T is the estimated item easiness parameter, μ_T is the estimated mean of the ability parameter from the generated data sets, and μ_B is the generating mean of the ability parameter.

Chapter 4

Results

4.1 Simulation Results

The simulation results are presented in this section separately for the unconditional model and the conditional model.

4.1.1 Simulation Results of the Unconditional Model

There were 9 conditions for the unconditional model: 3 sample sizes $\times$ 3 proportions. The recovery of item easiness parameters and the variances of random item and person effects under each simulation condition were evaluated by bias and RMSE statistics for each of the estimated parameters. The recovery analysis across 20 replications for sample size and proportion of speededness conditions is summarized in Table 4.1 to Table 4.5.

The recovery of the item easiness parameters for the speeded group and the non-speeded group was evaluated separately since the speeded group had a much smaller number of examinees than the non-speeded group. Because of the difference in sample size, the expected result was that larger biases and larger RMSEs might be expected for the items in the speeded group than in the non-speeded group. The proportions of speeded and non-speeded examinees were compared with the generating proportions and the differences between them were computed.

Recovery of Item Easiness Parameters for the Unconditional Model. The bias and RMSE statistics for item easiness parameters for the speeded and non-speeded groups for the unconditional model are presented in Table 4.1. The estimated item easiness parameters were equated by adjusting the differences between the estimated ability mean and

the generating ability mean. The bias and RMSE of the item easiness parameters for the speeded group were, in general, much larger than those of the non-speeded group. The absolute values of the bias of the item easiness parameters for the speeded group ranged from 0.022 to 0.180 and the absolute values of the bias of the item easiness parameters for the non-speeded group ranged from 0.001 to 0.102. The RMSE of the item easiness parameters for the speeded group ranged from 0.128 to 0.630 and the RMSE of the item easiness parameters for the non-speeded group ranged from 0.098 to 0.586. These results seem reasonable since the speeded group has a much smaller sample size than the non-speeded group. As expected, the bias of item parameters decreased with an increase in sample size. For example, for the 1,000 examinee condition, when the proportion of speededness increased from 10% to 20%, the bias in the item easiness parameters decreased from 0.18 to 0.064 for the speeded group and from 0.102 to 0.021 for the non-speeded group.

Some patterns can be identified within the speeded and non-speeded groups, respectively. For example, within the speeded group, the bias and RMSE decreased with the increase of sample sizes or with the increase of proportions of speeded examinees. The bias and RMSE were lower with increases in proportions of speededness, because more examinees were involved in the estimation of item parameters, when the proportions increased. For the non-speeded group, the bias and RMSE were very similar across the simulation conditions. There was a small amount of reduction in the bias and RMSE of item easiness parameters with the increase in sample size or in the proportion of speeded examinees. The smallest sample size for the non-speeded group across all the simulation conditions was 700 examinees. This sample size was sufficiently large to be able to obtain accurate recovery of the generating parameters.

Recovery of Random Item Effects. In Table 4.2, the bias and RMSE for the variances of the random item effects were compared under each simulation condition. The biases and RMSEs for the variances of the random item effects are much smaller than those of the fixed effects of the item parameters. A general trend is observed, that is, the bias and RMSE of the

random item effects are reduced with the increase of either the sample size or the proportion of speeded examinees. This pattern is more obvious with the increase of sample size than with the increase of proportion of speeded examinees. The reduction in bias and RMSE was small when the proportions increased from 20% to 30%.

Recovery of Variances of Person Effects for Unconditional Model. Table 4.3 presents the bias and RMSE of the variances of the random person effects. The bias and RMSE of the variances of the random person effects are smaller than those of item easiness parameters and the variances of the random item effects, a result which is due to a much larger sample size of examinees than the number of items. The bias of the variances of the random person effects ranges from 0.023 to 0.002 and the RMSE ranges from 0.075 to 0.036. This indicates a very good recovery for the variances of random person effects. The patterns in the bias and RMSE of the random person effect were slightly different from those in the bias and RMSE of the item effects. The bias and RMSE of the random person effects decrease with the increase of sample sizes. However, the bias and RMSE of the random person effects are similar among the three proportions with each sample size. That is, the bias and RMSE of the random person effects are similar when the proportions of speeded examinees increased from 10% to 30% and the total sample size remain the same.

Recovery of Latent Classes for Unconditional Model. Table 4.4 presents the differences in the simulated proportions of speeded examines and the estimated proportion of speeded examinees by the unconditional model. The differences between the simulated and recovered proportions of speeded examinees were the largest (4.0%) for the 1,000 examinees conditions and 10% speededness and the differences were the smallest (0.1%)for the 3,000 examinees conditions and 30% speededness. The recovery of latent classes was affected by the proportions of speededness. With the increase of proportions of speededness, the recovery of latent classes was improved.

Summary of Recovery Analysis for Unconditional Model. In summary, the recovery of model parameters associated with person effects was better than the recovery of

parameters associated with item effects. One possible explanation may be due to the larger number of examinees and smaller number of items. In the simulation study, 8 items were assumed to be affected by speededness. So only these 8 items were used to estimate the parameters associated with item effects.

4.1.2 Simulation Results of the Conditional Model

For the conditional model, gender effect on the latent group membership was included in the model. Gender was indicated in the data by a dummy code with 0 = male and 1 = female, and its impact on the latent group membership was modeled as a multinomial logit regression as in equation (2.68). As in the unconditional model, 9 conditions were simulated for the conditional model: 3 sample sizes $\times$ 3 proportions of speeded examinees. The recovery of item easiness parameters and the variances of the random item and person effects for each simulation condition across 20 replications were evaluated by the bias and RMSE of these parameters. The recovery of proportions of speeded and non-speeded examinees was examined by calculating the differences in the generating proportions of the speeded examinees and the recovered proportions of speeded examinees in the generated data sets across 20 replications of each simulation condition. The recovery results for the conditional model were summarized in Table 4.5 to Table 4.8.

Recovery of Item Easiness for the Conditional Model. The bias and RMSE indices for the item easiness parameters for the speeded and non-speeded group by the conditional model are shown in Table 4.5. The item easiness parameters were equated by adjusting the differences in the estimated ability mean and the generating ability mean. The bias statistics of the item easiness parameters for the speeded and non-speeded groups were larger for the conditional model than for the unconditional model. However, the RMSE indices for the item easiness parameters for both the speeded group were smaller than that of the unconditional model and the RMSE indices for the item easiness parameters for the non-speeded group were similar to that of the unconditional model.

Table 4.1: Bias and RMSE of Item Easiness Parameters by Latent Groups: Unconditional Model.

	Proportion											
	10				20				30			
	Speeded		Non-Speeded		Speeded		Non-Speeded		Speeded		Non-Speeded	
N	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
1000	0.180	0.630	0.102	0.198	0.064	0.470	0.021	0.209	-0.074	0.677	0.048	0.586
2000	-0.154	0.479	-0.033	0.183	-0.081	0.330	-0.014	0.155	-0.052	0.252	-0.012	0.130
3000	0.072	0.375	0.026	0.129	0.022	0.275	0.017	0.128	-0.031	0.213	-0.001	0.098

Table 4.2: Bias and RMSE of the Variances of Random Item Effect: Unconditional Model.

	Proportion					
	10		20		30	
N	Bias	RMSE	Bias	RMSE	Bias	RMSE
1000	0.067	0.215	-0.014	0.116	-0.040	0.119
2000	0.049	0.184	0.019	0.122	0.036	0.086
3000	0.022	0.149	0.019	0.102	0.009	0.123

Table 4.3: Bias and RMSE of the Variances of Random Person Effect: Unconditional Model.

	Proportion					
	10		20		30	
N	Bias	RMSE	Bias	RMSE	Bias	RMSE
1000	0.017	0.075	0.023	0.067	0.021	0.063
2000	0.005	0.042	-0.002	0.049	-0.012	0.037
3000	-0.003	0.044	-0.008	0.030	0.003	0.036

Table 4.4: Differences Between the Generating and Recovered Proportions of Speeded Examinees: Unconditional Model.

	Proportion		
N	10	20	30
1000	4.0	2.0	0.1
2000	0.2	1.0	1.0
3000	1.0	1.0	0.1

For the speeded group, the bias of item easiness parameters decreased with the increase of the proportions of speeded examinees for sample sizes of 1,000 examinees and 2,000. For example, for the sample size of 1,000, the bias decreased from 0.315 to 0.042 for the speeded group, when the proportion increased from 10% to 30%. For the sample size of 3,000 examinees, the bias statistics were similar across the three speededness proportions. Bias statistics across the three speededness proportions for the sample of 3,000 examinees were all around 0.050.

The RMSEs for item easiness parameters in the speeded group decreased with the increase in sample size or with the increase in proportion of speeded examinees. For the 10% speededness condition, the RMSEs of the item easiness parameters of the speeded group decreased from 0.655 to 0.325, when the sample size increased from 1,000 to 3,000. The RMSEs decreased from 0.655 to 0.241 when the proportion of speededness increased from 10% to 30% for the 1,000 examinee sample size.

For the non-speeded group, the bias and RMSE for the item easiness parameters decreased a little bit with the increase of sample sizes, but they are similar across the three speededness proportions for the same sample size. For example, the bias index for the item easiness parameters for the sample size of 1000 examinees and 1 10% speededness proportion was 0.168 and it decreased to 0.007 when the sample size inceased to 3000 with the same speededness proportioin. The bias statistics decreased from 0.168 to 0.044 for the non-speeded group across the nine simulation conditions. The bias has a larger reduction when the sample size increased from 2,000 to 3,000 examinees than from 1000 to 2000. For example, the bias statistics of the non-speeded group for the 2000 examinee condition were all larger than 0.110. When the sample size increased to 3000, all bias statistics decreased to less than 0.100.

The RMSE of the item easiness parameters for the non-speeded group decreased with an increase in sample sizes and remained similar across the three proportions of speeded examinees. The RMSEs of the non-speeded group decreased from 0.277 to 0.124, when the sample size increased from 1,000 to 3,000 for the speededness proportion of 10%.

Recovery of Variances of Random Item Effects for Conditional Model. Table 4.6 presents the bias and RMSE statistics for the variances of random item effects. Bias and RMSE statistics for the random item effects were smaller when the sample size increased from 1,000 to 3,000. In addition, bias decreased with an increase in the proportions of speeded examinees. The bias and RMSE indices for the random item effects by the conditional model were similar to those for the unconditional model for most simulation conditions.

Recovery of Random Person Effects for Conditional Model. The bias and RMSE statistics for the variances of the random person effects are presented in Table 4.7. As was the case with the unconditional model, bias and RMSE for the random person effects were much smaller than those of the item easiness parameters or the random item effects. The bias and RMSE of the random effects for the conditional model also were smaller those for the unconditional model. This finding was consistent with the previous research which found that inclusion of a covariate on latent group membership improved the accuracy of parameter estimates (Von Davier & Yamamoto, 2007). The RMSEs of the random person effects decreased with the increase in sample size, but did not change much as the proportion of speeded examinees increased. The reduction in bias and RMSE was very small, when sample size increased from 2,000 to 3,000, suggesting that a sample size of 2,000 probably would be large enough to get stable estimates of the random person effects.

Recovery of Latent Classes for Conditional Model. Presented in Table 4.8 are the differences between the simulated the recovered proportions of speeded examinees. After the inclusion of a covariate on latent group membership, the recovery of the classifications of examinees into latent classes improved. This result was consistent with previous studies which found the incorporation of background variables associated with latent class membership into the model can improve the accuracy of classification of examinees into latent classes (Smit et al., 1999, 2000). The differences in the simulated and recovered proportions of speeded examinees were relatively small, but did decrease somewhat with an increase in sample size. The differences between the generating proportions and the recovered proportions for sample

sizes of 2,000 and 3,000 examinees were quite small, suggesting that a sample size of 2,000 should be sufficient for stable estimates of the proportions of speeded examinees for the conditional model.

4.2 Example: Analysis of Speededness on a College-Level Mathematics Placement Test

Although the association between gender and latent class membership has been studied in previous speededness research, however, the association between gender and latent class membership is not a consistent one in the speededness literature. Bolt et al. (2002) found no significant association between gender and latent class membership on a mathematics course placement test at a Midwestern university. Cohen et al., (2002) examined the association between speeded and non-speeded groups and student's background and academic achievement and found no association between gender and latent class membership for students with weaker mathematics backgrounds. For students with a stronger mathematics background, however, a significant association was found between gender and latent class membership. Males were found to have a higher mean math score than females. Both of these studies implemented the analysis in two steps. As suggested earlier, this two-step approach may attenuate the association between gender and latent class membership, since estimation errors in the first step may not be accounted for in the second step. In this study, we examined the effect of gender and its association with latent classes in a single step on a new set of data. For purposes of this study, the gender effect was assumed not to differ across latent classes.

Speededness Models. Several models used for detection of speededness effects in paper-and-pencil tests were fit to the data in this example. These include the following models: the cross-classified multilevel mixture Rasch model as in equations (2.53 and 2.58), the Hybrid Rasch model as in equation (2.9), the two-class mixture Rasch model as in equation (2.8) and the mixture gradual process model with a Rasch measurement model as in equation (2.12).

Table 4.5: Bias and RMSE of Item Easiness Parameters by Latent Groups: Conditional Model.

	Proportion											
	10				20				30			
	Speeded		Non-Speeded		Speeded		Non-Speeded		Speeded		Non-Speeded	
N	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
1000	0.315	0.655	0.168	0.277	0.130	0.336	0.057	0.152	0.042	0.241	0.113	0.160
2000	0.249	0.402	0.110	0.182	0.192	0.278	0.110	0.142	0.146	0.219	0.113	0.153
3000	0.054	0.325	0.007	0.124	0.063	0.183	0.044	0.090	0.051	0.139	0.066	0.092

Table 4.6: Bias and RMSE of the Variances of Random Item Effects: Conditional Model.

	Proportion					
	10		20		30	
N	Bias	RMSE	Bias	RMSE	Bias	RMSE
1000	0.155	0.267	0.060	0.156	-0.044	0.116
2000	0.093	0.132	0.037	0.109	0.034	0.055
3000	0.028	0.143	-0.009	0.097	-0.008	0.079

Table 4.7: Bias and RMSE of the Variances of Random Person Effects: Conditional Model.

	Proportion					
	10		20		30	
N	Bias	RMSE	Bias	RMSE	Bias	RMSE
1000	0.002	0.059	-0.028	0.073	0.009	0.063
2000	-0.001	0.040	-0.001	0.032	-0.001	0.052
3000	-0.001	0.034	0.012	0.030	-0.009	0.037

Table 4.8: Differences Between the Generating and Recovered Proportions of Speeded Examinees: Conditional Model.

	Proportion		
N	10	20	30
1000	1.8	1.7	1.6
2000	0.7	0.4	0.2
3000	0.1	0.1	0.1

The item parameter estimates from these models were also compared with those obtained by the Rasch model.

For the cross-classified multilevel mixture Rasch model, two separate models were estimated: An unconditional mixture multilevel Rasch model and a conditional mixture Rasch model. The unconditional model was estimated first followed by estimation of the conditional model. Gender was used as the covariate on the mixing proportions of the conditional model to help classify examinees into speeded and non-speeded groups.

To reflect the assumption of the Hybrid model and the MRM with an ordinal constraint, the item difficulty parameters for the first twenty items were constrained to be equal. The item difficulties of the last eight items were constrained to be larger for the speeded group than those for the non-speeded group. For the mixture gradual process change model, the speededness rate parameters λ_j were constrained to be zero for the non-speeded group and estimated for the speeded group.

4.2.1 Methods

Data. Data used in this study were from a mathematics placement test administered to entering college students at a large Midwestern university system. All the items were multiple-choice with five alternatives. There were three forms of the placement test, Form A, Form B and Form C. Form A and Form C had 35 items, and Form B had 40 items. Students were counseled to take Forms A and B if they had fewer than 2 $\frac{1}{2}$ years of high school mathematics or if they had not taken any trigonometry. Students who had at least 2 $\frac{1}{2}$ years of high school mathematics and who had taken trigonometry were counseled to take Forms B and C. Only items on Forms A and B were used in this study. Three tryout items in Form A and 4 items in Form B were embedded at different locations in the test. These items were excluded from the analysis, leaving 32 operational items on Form A and 36 operational items on Form B that were analyzed in this study.

The first 20 operational items on Form A and the last 8 on Form B were included in the analysis. Items in the middle of the test were not used for the analysis. Although these items may be affected by test speededness, the effects of speededness were assumed to not be enough to be modeled. The first 20 items were assumed to be not affected by test speededness and the last 8 items were assumed to be most affected by test speededness. This is the same logic as used by Bolt et al. (2002).

Only the sample of 14,878 examinees who took Form A and Form B were included in the analysis. Those who had missing values on gender also were excluded from the analysis resulting in a sample of 13,336. A random sample of 2,017 examinees (approximately 15 percent) was selected from the sample of 13,336 examinees. There were 809 males and 1,207 females in the sample.

Modeling Speededness. To model test speededness effects, we assumed the first 20 items were not affected by test speededness and the last 8 items were most affected by test speededness. An examination of the simple statistics shows that the proportions of correct responses to the first 19 items were in general higher than the last 10 items and the proportions of incorrect responses to the non-speeded items were lower than the speeded items. This was the case except for Items 8, 13, 14, and 17. The high proportions of incorrect responses to these items may not be due to test speededness but due to other characteristics of these items. Item 20 was assumed not to be affected by test speededness, although, it had a relatively high proportion of incorrect responses. It is possible that this may indicate the beginning of speededness effects. The impact of test speededness on Item 20, however, may not be as strong as the impact on the last 8 items, since there was not a large number of omitted items over these last 8 items. The last 8 items did have a much larger proportion of omitted responses, however, than did the non-speeded items. Mroch and Bolt (2006) have noted that omissions for items near the end of the test are a possible indicator of test speededness. Female examinees tended to have slightly lower proportions of correct responses and higher proportions of omitted responses than male examinees over the last 8 items. This

may be an indication that female students might be more affected by test speededness on this test than male students.

Estimation of Model Parameters. Parameters of the conditional and unconditional models were estimated on the real data sample using MCMC estimation as implemented in the software WinBUGS (Spiegelhalter et al., 2003). For MCMC estimation, the means of the posterior distributions for the sampled parameter values following burn-in were used as the parameter estimates. Examinees were assigned to one of the latent classes at each iteration. The posterior probability of latent class membership was used as the estimate of latent class membership.

The unconditional model was run for 10,000 iterations. The Heidelberg and Welch diagnostic from BOA (Smith, 2005) suggested the model converged after 3,000 iterations. The same MCMC algorithm was run on the conditional, Rasch, MRM, Hybrid and MixGPCM for 10,000 iterations. The Heidelberg and Welch diagnostic suggested the MRM converged after the first iteration, the Hybrid model and the MixGPCM converged after 1,000 iterations. All models converged at less than 5,000 iterations. Therefore, a conservative burn-in of 5,000 iterations was used. These iterations were discarded and the remaining 5,000 iterations were used to estimate the model parameters. The amount of time needed to estimate the unconditional model for 10000 iterations is 1.55 hours on an HP BL460c 2.00 GHz server blade with a Quad-Core Intel Xeon processor and 3.25GB RAM running a Windows 2003 server operating system.and the amount of time needed to estimate the conditional model is 1.39 hours. The amount of time needed to estimate the MRM is about 1.51 hours and 1.59 hours for the Hybrid model and 8.50 hours for the MixGCM.

4.2.2 Results of Real Data Analysis

Descriptive Statistics for Test Items. The descriptive statistics for these 28 items for male and female students are reported in Table 4.9. The mean raw score for male students was 0.83 units higher than for female students. An independent samples t-test between male

Table 4.9: Descriptive Statistics For Male and Female Students.

	N	*M*	*SD*
Male	809	16.66	5.06
Female	1208	15.83	5.12

and female students produced a t statistic of 3.60 with a p-value of .00. This indicates that male and female students differed significantly in mean total raw score over these 28 items.

In addition to the above descriptive statistics, the proportions of correct, incorrect and omitted responses to the 28 items were also examined. The proportions of correct, incorrect and omitted responses to the 28 items by all the examinees and male and female examinees are reported in Table 4.10. For the total sample, the proportion of correct responses decreased beginning with Item 21, where the ordinal constraints were imposed. The proportions of incorrect responses for the total sample increased from about Item 20 and the proportion of omissions increased steadily from Item 21. Similar patterns of correct, incorrect and omitted responses can be observed for male and female examinees. Following Item 20, where test speededness was assumed to begin (and so the ordinal constraints were imposed), the proportions of correct responses were higher for male examinees than female examinees.

Credibility Intervals on Posterior Estimates. One advantage of Bayesian estimation is that it will produce not only a point estimate of the parameters, but also an interval around this estimate called a credibility interval. The credibility interval is the interval which is assumed to contain the parameter of interest. A significance test can be performed on the model parameters based on this interval. For posterior estimates, if the credibility interval includes 0, then the parameter estimate is considered not to be significantly different from 0. Otherwise, it is considered to be significantly different from 0. Results for the unconditional model are summarized in Table 4.11.

Table 4.10: Proportions of Correct, Incorrect and Omissions for the Total Sample, Male Examinees and Female Examinees.

	Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Male	Incorrect	31.6	15.9	19.7	30.4	17.7	11.1	29.8	54.1	17.8	36.0	16.7	30.3	51.7	60.1	40.8	33.6	64.0	16.2	29.5	65.6	29.3	54.4	68.2	72.3	37.5	46.1	46.4	69.1
	Omission	.0	.0	.0	.0	.2	.0	.0	.0	.0	.6	.1	.1	.1	.5	1.0	.2	.4	.1	.2	.1	3.3	3.0	3.7	4.3	4.9	4.6	4.6	5.6
	Correct	67.9	81.6	81.0	62.7	80.5	82.0	74.3	45.4	74.9	65.1	76.8	67.7	44.9	43.0	59.5	67.6	37.3	77.8	62.5	26.5	62.4	35.6	23.3	21.9	53.5	42.5	39.5	24.8
Female	Incorrect	32.1	18.4	19.0	36.8	19.2	18.0	25.7	54.6	25.1	34.6	22.9	32.0	54.8	56.5	39.7	32.0	62.4	22.2	37.3	73.4	34.7	61.4	72.4	74.1	41.1	52.9	55.4	69.0
	Omission	.0	.0	.0	.4	.2	.0	.0	.0	.0	.2	.2	.3	.3	.6	.8	.4	.2	.0	.2	.1	2.9	3.0	4.2	4.0	5.4	4.6	5.1	6.2
Total	Correct	68.1	82.6	80.8	65.5	81.2	84.8	72.7	45.6	77.8	64.5	79.4	68.5	46.2	41.5	59.0	67.0	36.6	80.2	65.6	29.6	64.4	38.4	25.0	22.5	55.1	45.2	43.3	25.0
Sample	Incorrect	31.9	17.4	19.2	34.3	18.6	15.2	27.3	54.4	22.2	35.2	20.4	31.3	53.5	57.9	40.1	32.6	63.1	19.8	34.2	70.3	32.5	58.6	70.7	73.4	39.7	50.2	51.8	69.0
	Omission	.0	.0	.0	.2	.2	.0	.0	.0	.0	.4	.2	.2	.2	.5	.9	.3	.3	.0	.2	.1	3.1	3.0	4.0	4.1	5.2	4.6	4.9	5.9
	Correct	68.4	84.1	80.3	69.6	82.1	88.9	70.2	45.9	82.2	63.4	83.2	69.6	48.2	39.4	58.2	66.1	35.6	83.7	70.2	34.2	67.4	42.6	28.1	23.4	57.6	49.3	49.1	25.3

Table 4.11: Item and Person Effects: Results for the Unconditional Model.

	Parameter	Group	Estimates	*SD*	2.5%	97.5%
Fixed	Item effects($\overline{\beta_g}$)	Speeded(β_1)	-1.78	0.43	-2.62	-0.96
Effects		Non-speeded(β_2)	-0.30	0.36	-1.04	0.44
	Person effects(μ_g)	Speeded(μ_1)	0.32	0.11	0.12	0.55
		Non-speeded(μ_2)	0.00			
Variance of	Item effects(σ_2^2)		1.05	0.47	0.47	2.23
Random effects	Person effects(σ_1^2)		0.87	0.04	0.80	0.95

The item parameter estimates in Table 4.11 are for item and person effects under the unconditional model. Item difficulty estimates can be obtained by multiplying the item easiness effects by -1. As can be seen, the mean item easiness for the speeded group ($M = -1.78$) was lower than for the non-speeded group ($M = -0.30$). (These values for item difficulty would be 1.78 and 0.30 for the speeded and nonspeeded groups, respectively.) This ordering of means for item easiness was due to the way the ordinal constraints on the item parameters were implemented to model the speededness assumption in the last 8 items. Examinees in the speeded group also had a higher mean ability than those in the non-speeded group. The variance of the random person effects ($\sigma_1^2 = 0.87$) was smaller than the variance of the random item effect ($\sigma_2^2 = 1.05$). The smaller variance for the random person effects was probably due to the larger sample size (N=2,017) and to the small number of items (n=28).

The results for the conditional model are presented in Table 4.12. Recall that the probability of latent class membership was modeled as a multinomial logistic regression using gender as a covariate. Gender was dummy coded with male examinees coded as 0 and female examinees coded as 1.

Table 4.12: Item and Person Effects: Results for the Conditional Model.

	Parameter	Group	Estimates	*SD*	2.5%	97.5%
	Intercept(γ_{0g})	Speeded(γ_{01})	-1.51	0.37	-2.55	-0.89
Gender		Non-speeded(γ_{02})	0.00			
Effects	Slope(γ_{1g})	Speeded(γ_{11})	0.39	0.19	0.03	0.79
		Non-speeded(γ_{12})	0.00			
Fixed	Item effects($\overline{\beta_g}$)	Speeded(β_1)	-1.47	0.37	-2.24	-0.76
Effects		Non-speeded(β_2)	-0.24	0.33	-0.89	0.43
	Person effects(μ_g)	Speeded(μ_1)	0.28	0.10	0.10	0.50
		Non-speeded(μ_2)	0.00			
Variance of	Item effects(σ_2^2)		0.99	0.44	0.46	2.16
Random effects	Person effects(σ_1^2)		0.87	0.04	0.81	0.95

As was observed for the unconditional model, the mean item easiness for the speeded group ($M = -1.47$) was lower than for the non-speeded group ($M = -0.24$). The average ability of the speeded group (0.28), however, was higher than that of the nonspeeded group.

Consistent with results from Smit, et al. (1999, 2000), the standard errors and the credibility intervals for the model parameters would be smaller for the conditional model than for the unconditional model. The results for the fixed item effects for the unconditional model were 0.43 (see Table 4.11) and for the conditional model, 0.37 (see Table 4.12). The standard error of the fixed item effects of the non-speeded group was reduced slightly from 0.36 in the unconditional model to 0.33 in the conditional model. The standard error of the random item effects was also slightly reduced from 0.47 in the unconditional model to 0.44 in the conditional model. Any reduction of standard errors of the fixed person effects or the random person effects between conditional and unconditional models was not obvious. The variance of the random item effect was smaller in the conditional model than the unconditional model indicating the inclusion of covariate helped explain part of the random variances in items.

The credibility interval shown in Table 4.12 from 2.5% to 95%, does not include zero for gender, indicating that gender had a significant impact on the classification of an examinee's latent group membership. The 95% credibility interval for the intercept and the slope of the gender effect also did not include zero, which likewise indicates that male and female examinees differed significantly in the probability of latent class membership. Since male examinees were coded as 0 and female examinees were coded as 1, the positive slope (0.39) indicates that female examinees had a higher probability of being classified into Class 2 which was the speeded group. Thus, female examinees tend to be more affected by test speededness than male examinees. The simple descriptive statistics of the proportions of incorrect and omitted responses by male and female examinees in Table 4.10 indicate a similar conclusion.

To examine the performance of the mixture cross-classified Rasch model for test speededness, the estimates of the parameters and the classifications by the unconditional and conditional models were compared with those of the Rasch model, two-class MRM, Hybrid model and the MixGPCM. The unconditional model, conditional model, MRM and Hybrid model all analyzed 28 items, the first 20 and the last 8 items. The MixGPCM included 50 of the items on the test, including the first 32 items on Form A and the last 18 items on Form B. This is because the purpose of the MixGPCM was to model the gradual change in speededness over the course of the test for both latent classes.

The proportions of speeded and non-speeded examinees classified by each of the models are reported in Table 4.13. It is possible to directly compare the mixture cross-classified IRT models with the other models from this table.

As shown in Table 4.13, after adding gender as a covariate, the proportions of speeded examinees increased from 11.25% for the unconditional model to 13.98%. The MRM identified the largest proportion of examinees as speeded and the MixGPCM identified the smallest. The proportion of speeded examinees classified by the unconditional mixture cross-classified Rasch model was smaller than the MRM (26.03%) and the Hybrid model (13.63%), but relatively close to the MixGPCM (10.1%). The proportion of speeded examinees classified by the

Table 4.13: Proportions of Speeded and Non-Speeded Examinees For all Five Models.

	Unconditional Model	Conditional Model	MRM Model	Hybrid Model	MGCM Model
Speeded	11.25	13.98	26.03	13.63	10.10
Non-speeded	88.75	86.02	73.97	86.37	89.90

Table 4.14: Cross-Tabulation of Group Membership by the Unconditional and Conditional Models.

		Unconditional Model		Total
		Speeded	Non-speeded	
Conditional Model	Speeded	225 (11.2%)	57 (2.8%)	282 (14.0%)
	Non-speeded	2 (.1%)	1733 (85.9%)	1735 (86.0%)
Total		227 (11.3%)	1790 (88.7%)	2017 (100%)

conditional model (13.98%) was similar to that for the Hybrid model (13.63%). After adding gender as covariate, approximately 2.8% more of the non-speeded examinees were identified as speeded by the conditional model than by the unconditional model. A cross-tabulation of the group memberships identified by the unconditional and conditional models indicating the changes in latent classes is shown in Table 4.14.

The chi-squares between latent classes for each model and for gender are reported in Table 4.15. Chi-squares were calculated in a two-step approach. The five models (i.e., the MRM, Hybrid model, MixGPCM, the unconditional cross-classified model and the conditional cross-classified model) were applied to classify examinees into speeded and non-speeded groups. Then Chi-squares computed between gender and latent class for each model failed to show a significant association with latent class for the MRM, the unconditional cross-

Table 4.15: Chi-squares Between Gender and Group Membership.

	Unconditional	Conditional	MRM	Hybrid	MGCM
χ_2	1.692	18.811	1.694	4.657	0.046
P-value	0.197	0.000	0.196	0.034	0.830

classified multilevel mixture Rasch model and the MixGPCM. Gender did have a significant association, however, for latent classes from the Hybrid model ($p <= .05$). The association between gender and latent classes for the conditional cross-classified multilevel mixture Rasch model, however, was also significant ($p < .001$). This finding was consistent with the previous research and suggested that the two-step approach, as was used with the unconditional model, MRM, and MixGPCM, permitted measurement errors in the first step for these models to intrude into the estimates at the second step, resulting in attenuated relationships between gender and latent classes.

The percentages of speeded and non-speeded examinees for males and females were also examined. Females had a slightly higher percentage of speeded examinees than males. This was the case under both the conditional model and the unconditional model. The percentages of speeded and non-speeded examinees for male and female examinees are summarized in Table 4.16. Since gender was significantly associated with latent classes, it was interesting to further examine the composition of latent classes by gender based on the MRM, Hybrid and MixGPCM.

Results presented in Table 4.16 indicate that a higher percentage of female examinees than males were classified as speeded by both the unconditional and the conditional models. Further, after including gender in the model (i.e., the conditional model), the change in the percentage of speeded examinees was larger for female examinees than for male examinees. The percentages of female examinees in the speeded class increased from 12.0% to 16.7% and the percentage of male examinees decreased from 10.1% to 9.9%. More female examinees

Table 4.16: Proportions of Speeded and Non-Speeded Examinees by Gender For Each Model.

	Unconditional		Conditional		MRM		Hybrid		MixGPCM	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
Speeded	12.0	10.1	16.7	9.9	27.1	24.5	15.0	11.6	10.2	9.9
Non-speeded	88.0	89.9	83.3	90.1	72.9	75.5	85.0	88.4	89.8	90.1

Table 4.17: Item Difficulty Estimates for the Non-Speeded Group By Each Model.

Item Difficulty	Conditional	Unconditional	MRM	Hybrid	Rasch	MixGPCM
Item 21	-0.914	-0.888	-0.930	-0.693	-0.699	-0.893
Item 22	0.414	0.450	0.373	0.063	0.564	0.440
Item 23	1.136	1.162	1.116	0.562	1.283	1.141
Item 24	1.126	1.163	0.985	0.681	1.456	1.412
Item 25	-0.672	-0.624	-0.817	-0.625	-0.241	-0.353
Item 26	0.032	0.049	0.0142	-0.260	0.232	0.077
Item 27	0.012	0.060	-0.057	-0.278	0.324	0.234
Item 28	0.971	0.996	0.850	0.501	1.296	1.160

also were classified as speeded than male examinees by the MRM and Hybrid model. The percentages of speeded male and female examinees identified by the MixGPCM were about the same.

Correlations between Item Difficulty Estimates from the Different Models. The item difficulty parameters from the unconditional model and the conditional model were compared with those from the Rasch model, MRM, Hybrid model and the MixGPCM. Recall that the item parameters obtained by the mixture cross-classified Rasch model for test speededness were item easiness parameters. To compare these estimates with the item difficulty estimates from other models, the item easiness parameters should be transformed to item difficulty parameters by multiplying -1. The mixture unconditional and conditional cross-classified IRT models, the MixGPCM and MRM all produced two classes of item parameters. The Hybrid model and Rasch model only produced a single set of item parameter estimates. The estimates from these latter two models were somewhat like those for the non-speeded groups from the other three models. The unequated item difficulty parameters for the non-speeded group are presented in Table 4.17 and the item difficulties of the speeded group are reported in Table 4.18.

Table 4.18: Item Difficulty Estimates for the Speeded Group By Each Model.

Item Difficulty	Conditional	Unconditional	MRM	MixGPCM
Item 21	0.128	0.235	-0.129	-1.676
Item 22	1.144	1.136	1.075	-0.441
Item 23	1.842	1.878	1.762	0.944
Item 24	3.048	3.302	2.845	-0.594
Item 25	1.384	1.639	1.049	-3.421
Item 26	1.039	1.169	0.796	-0.555
Item 27	1.568	1.669	1.276	-2.838
Item 28	2.889	3.181	2.544	0.418

Table 4.19: Correlations Between Item Difficulties for Last 8 Items In the Non-Speeded Group.

	Unconditional	Conditional	MRM	Hybrid	MixGPCM	Rasch
Unconditional	1.000					
Conditional	.999	1.000				
MRM	.997	.998	1.000			
Hybrid	.989	.989	.983	1.000		
MGCM	.989	.987	.974	.980	1.000	
Rasch	.993	.992	.981	.985	.998	1.000

The correlations between item difficulty parameters in the speeded and non-speeded groups from the different models were also examined. As correlations do not need item parameters to be put on the same metric, no equating was needed to compare the item parameter estimates from each model. The correlations of the item difficulty parameters in the non-speeded group by each model are presented in Table 4.19 and the correlations in the speeded-group by each model are presented in Table 4.20.

Table 4.19 shows that correlations of item difficulty estimates for the non-speeded groups identified by each of the models were very high, with a maximum correlation of .999 and a

Table 4.20: Correlations Between Item Difficulties for Last 8 Items In the Speeded Group.

	Unconditional	Conditional	MRM	MixGPCM
Unconditional	1.000			
Conditional	.996	1.000		
MRM	.982	.994	1.000	
MGCM	.341	.385	.447	1.000

minimum correlation of .974. Table 4.20 presents similar correlations for the speeded group. The correlations for the non-speeded group were higher than those for the speeded group. This is reasonable since the Hybrid model, MRM and MixGPCM model reduce to the regular Rasch model when speededness effects are absent.

As can be seen from Table 4.20, the correlations of the item difficulty parameters for the speeded group were relatively high between estimates from the unconditional model, conditional model and MRM with a minimum of .982. However, the correlations for the speeded group between the MixGPCM and the other three models were much lower, ranging from .341 to .447. A possible explanation for these lower correlations might be that the assumptions about test speededness were quite different for the MixGPCM that for the other three models. The unconditional model, conditional model and MRM assume items near the end of the test are harder for the speeded examinees than for the non-speeded examinees. Therefore, inequality constraints reflecting this assumption were imposed directly on the item parameter estimates for the last 8 items. Although the same assumption was implemented for the MixGPCM, the probability of getting an item correct for the speeded examinees under this model also was reduced by a speededness component, $min\{1, [1 - (\frac{i}{I} - \eta_j)]^{\lambda_j}\}$. This difference likely was the cause of the large differences in the correlations from the other three models with those from the MixGPCM.

Figure 4.1: Proportion of omissions by the unconditional model.

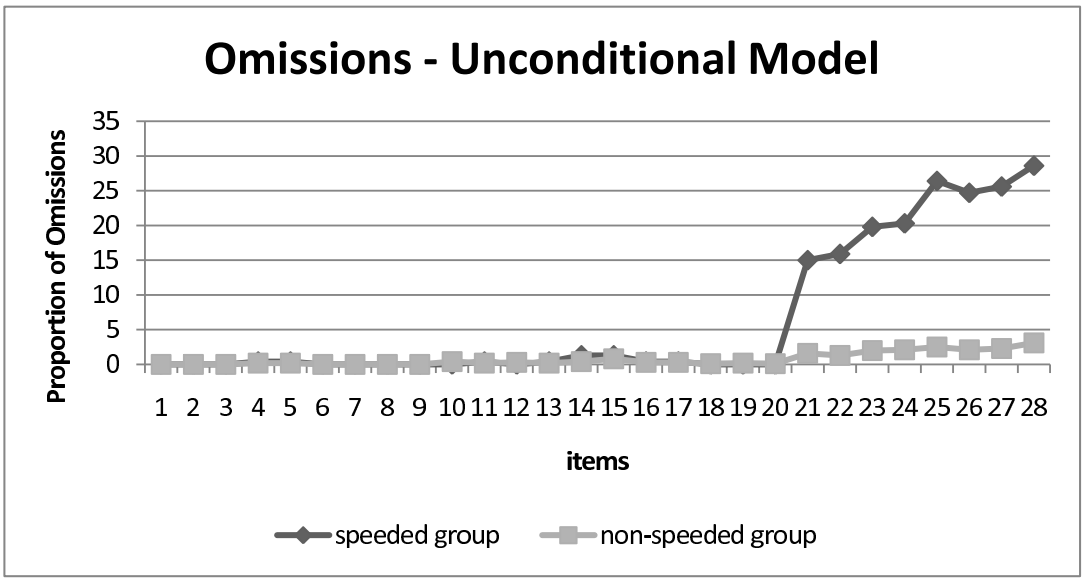


Figure 4.2: Proportion of omissions by the conditional model.

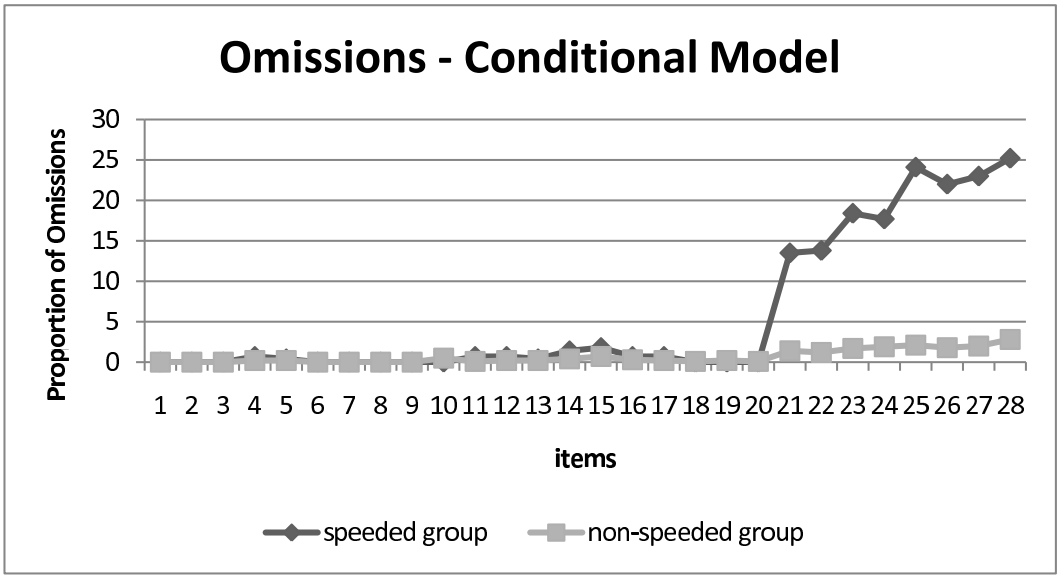


Figure 4.3: Proportion of omissions by the Hybrid model.

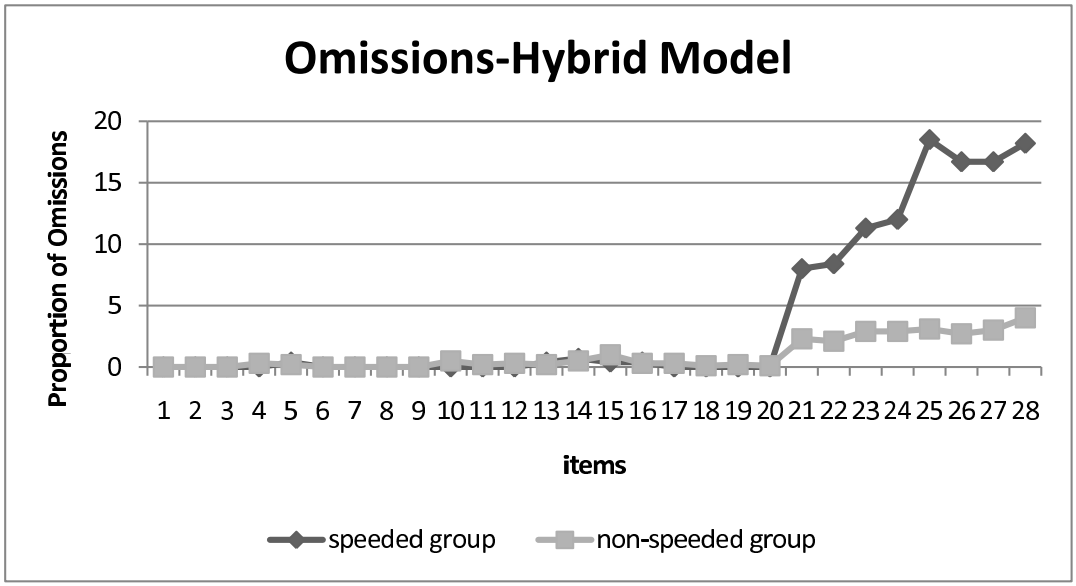


Figure 4.4: Proportion of omissions by the MRM model

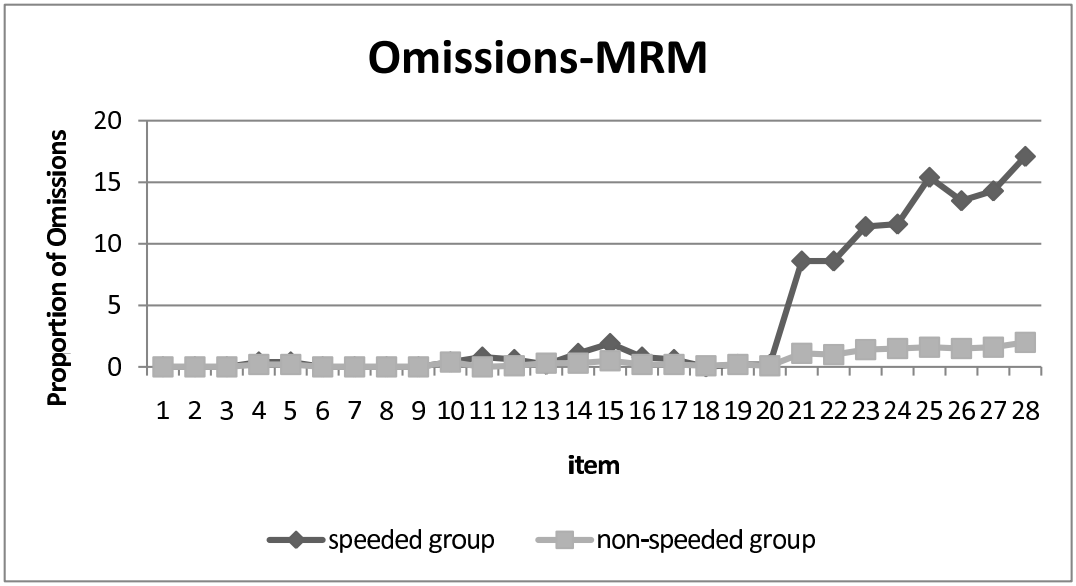
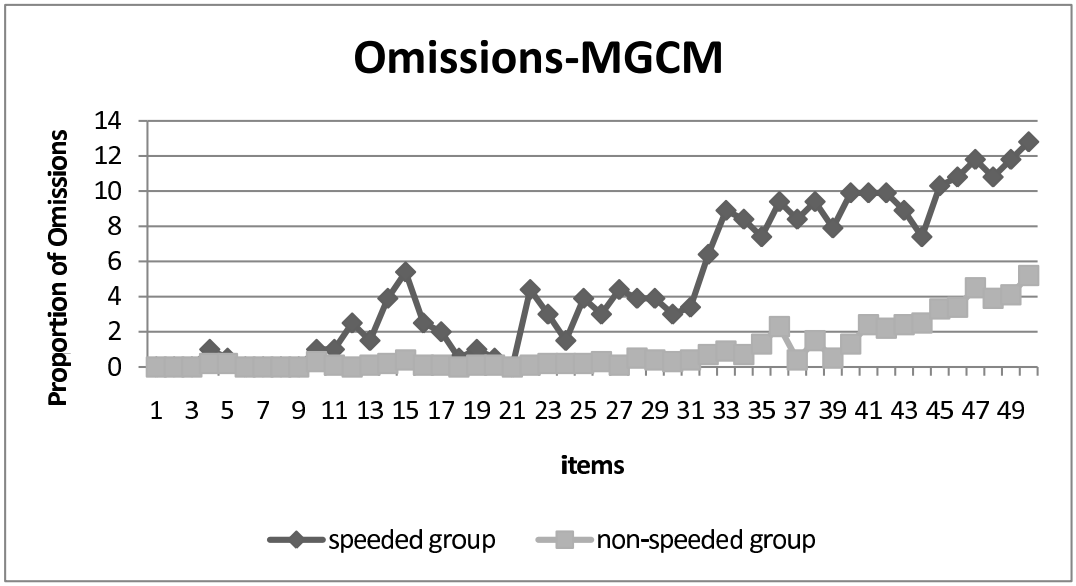
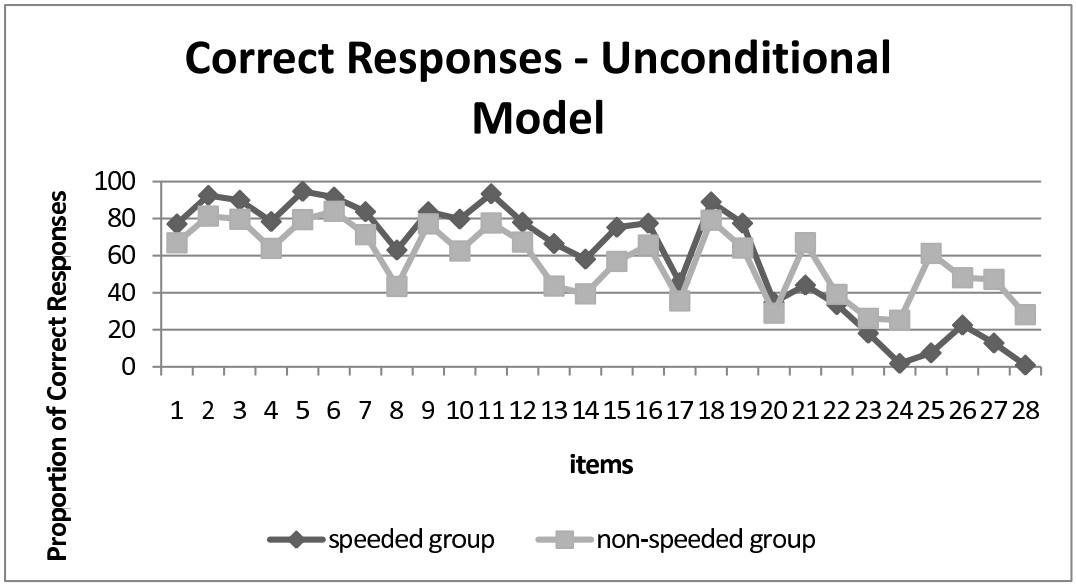


Figure 4.5: Proportion of omissions by the MixGPCM model.



Omitted Responses. The proportions of omitted responses to the 28 items by examinees in the speeded and non-speeded groups estimated for the unconditional, conditional, Hybrid and MRM are plotted in Figures 4.1 to 4.4, respectively. The proportions of omissions to the 50 items by the MixGPCM are presented in Figure 4.5. Both the speeded and nonspeeded groups had more omitted responses toward the end of the test for all models. The proportions of omitted responses to the first 20 items by the unconditional, conditional, Hybrid and MRM were similar between the speeded and non-speeded group. However, the proportion of omitted responses to the last 8 items was much higher for the speeded group than the non-speeded group. There was an abrupt increase in the omissions for the speeded group to items near the end of the test where test speededness effects were assumed to be present. This is not surprising as the last 8 items were separated by 40 items from the 20 items at the beginning of the test that were assumed to not have any speededness effects. For the MixGPCM, the omissions to the first few items were similar between the speeded

Figure 4.6: Proportion of correct responses by the unconditional model.



and non-speeded group. However, there was an increasing trend in the differences in the proportions of omissions between the speeded and non-speeded group toward the end of the test. This is reasonable since the MixGPCM models test speededness as a gradual change process rather than as an abrupt change as was modeled by the other four models.

Proportions of Correct and Incorrect Responses. The proportions of correct and incorrect responses by the speeded and non-speeded groups identified by each model are plotted in Figure 4.6 to 4.15.

Results plotted in Figures 4.6 to 4.8 show that for the first 20 items, which were assumed to be unaffected by test speededness, the proportions of correct responses were higher for the speeded group than for the nonspeeded group. This was true for the latent groups identified by the unconditional, conditional and MRM models. Figure 4.9 shows that the proportions of correct responses to the first 20 items were similar between the latent groups identified by the Hybrid. However, for the last 8 items, which were assumed to be most likely affected

Figure 4.7: Proportion of correct responses by the conditional model.

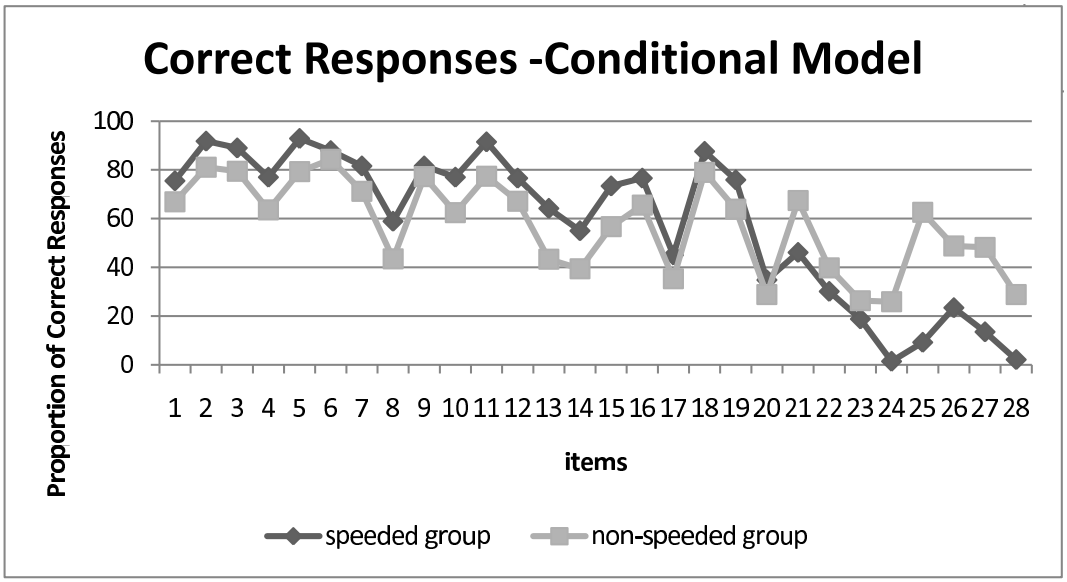


Figure 4.8: Proportion of correct esponses by the MRM model.

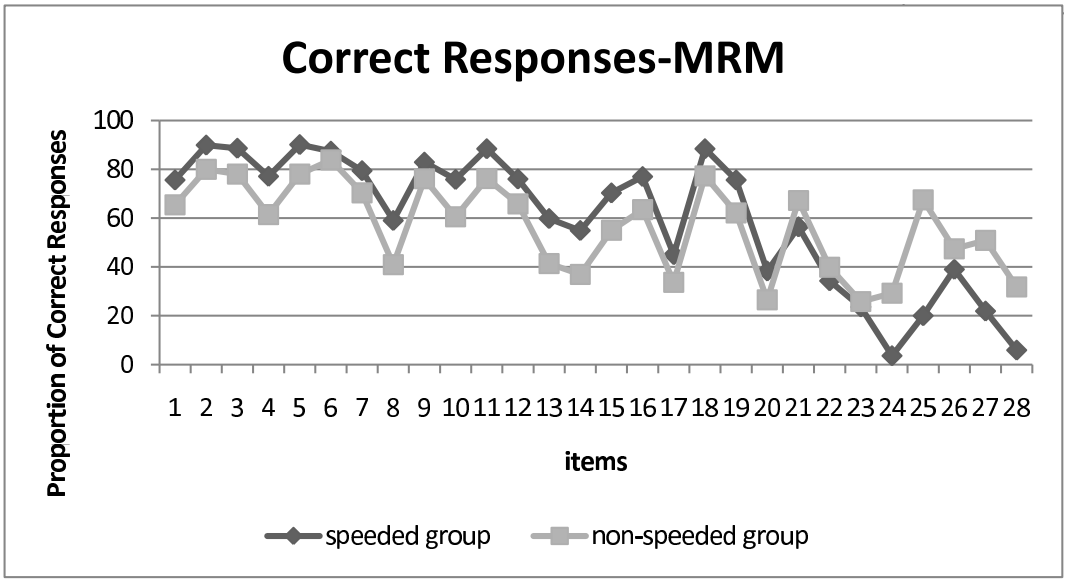


Figure 4.9: Proportion of correct responses by the Hybrid model.

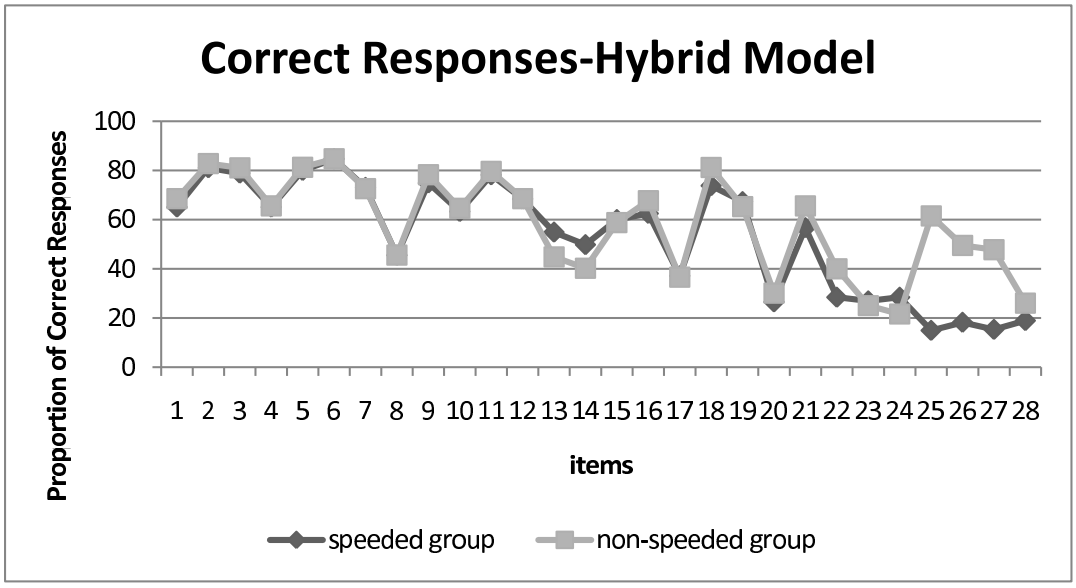
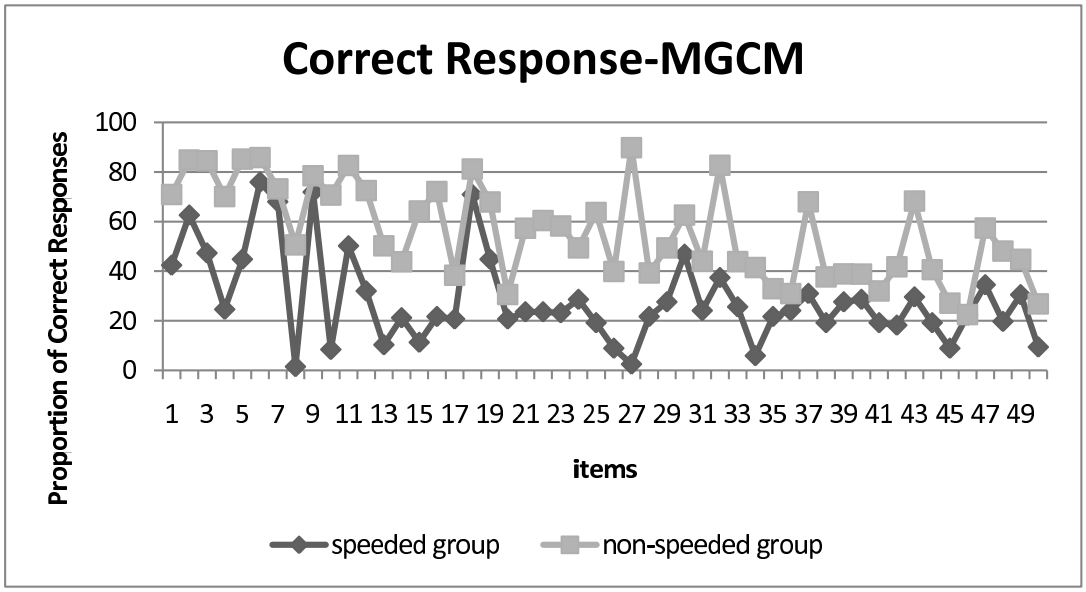


Figure 4.10: Proportion of correct responses by the MixGPCM model.



by test speededness, the nonspeeded group had consistently higher proportions of correct responses than the speeded group. This was true for the latent classes identified by all four models. The proportions of correct responses by the speeded and non-speeded groups identified by the MixGPCM were reported in Figure 4.10. A different pattern was found for this model in the proportions of correct responses. Unlike the patterns in the other four models, the nonspeeded group for the MixGPCM had a consistently higher proportion of correct responses from the beginning of the test to the end of the test.

As was suggested above for omissions, the differences in the patterns of correct and incorrect responses may be due to the way the speededness assumptions were implemented in the different models. The conditional and unconditional cross-classified mixture Rasch models, the Hybrid model and the MRM assume examinees become speeded at an arbitrary point and, from that point, at the same rate. The MixGPCM assumes examinees become speeded at different points with different rates. Therefore, even though examinees were classified into the speeded groups by the MixGPCM, they appeared to be affected by test speededness at different locations of the test. It is also likely that their speededness rates also might be different.

The results shown by the MixGPCM also suggest an additional conjecture. It appears that almost the same examinees were classed as speeded or nonspeeded by the MixGPCM as by the conditional, Hybrid, and MRM. This may indicate that examining the last 8 items on the test is sufficient for detection of speeded and nonspeeded groups.

The proportions of incorrect responses by the latent groups identified by each model are reported in Figure 4.11 to 4.15. Similar patterns to those for the correct responses can be found in the proportions of incorrect responses by the two cross-classified mixture Rasch models and the MRM. The proportions of incorrect responses to the first 20 items were higher for the non-speeded group than for the speeded group. For the items toward the end of the test, the proportions of incorrect responses by the speeded group were higher than for the non-speeded group. For the Hybrid model, the proportions of incorrect responses

Figure 4.11: Proportion of incorrect responses by the unconditional model.

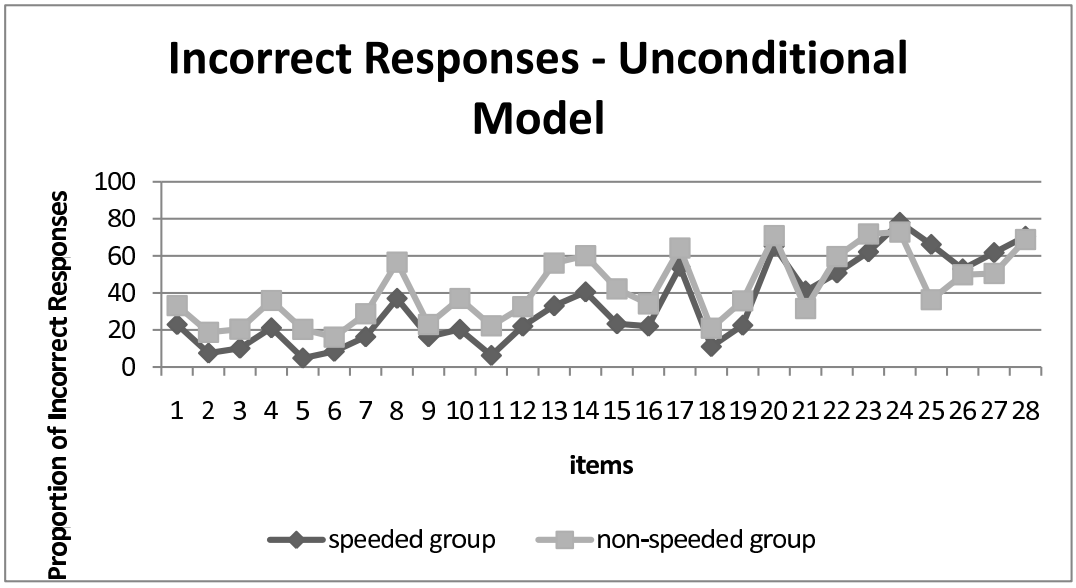


Figure 4.12: Proportion of incorrect responses by the conditional model.

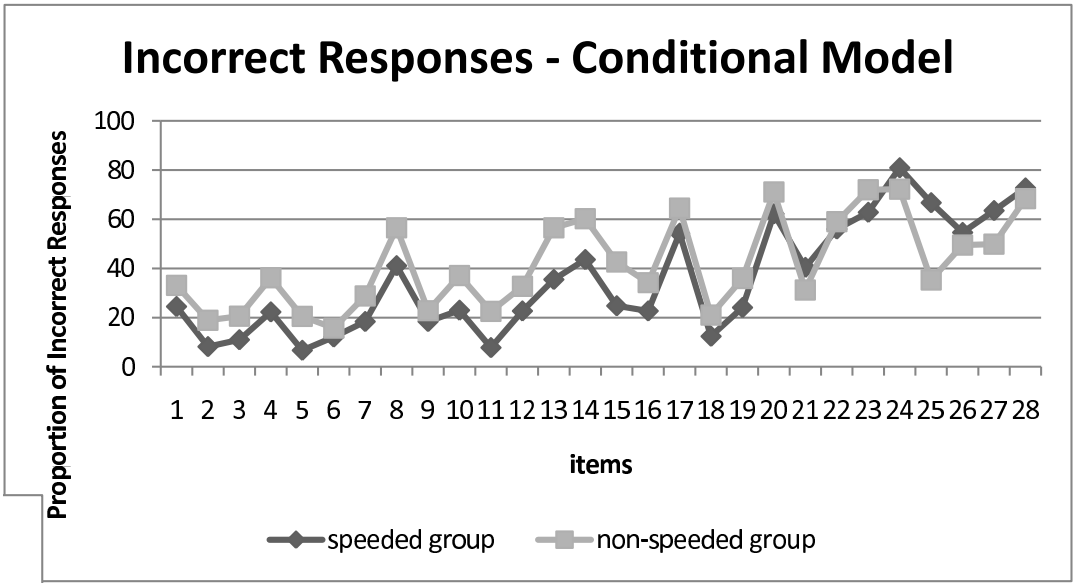


Figure 4.13: Proportion of incorrect responses by the Hybrid model.

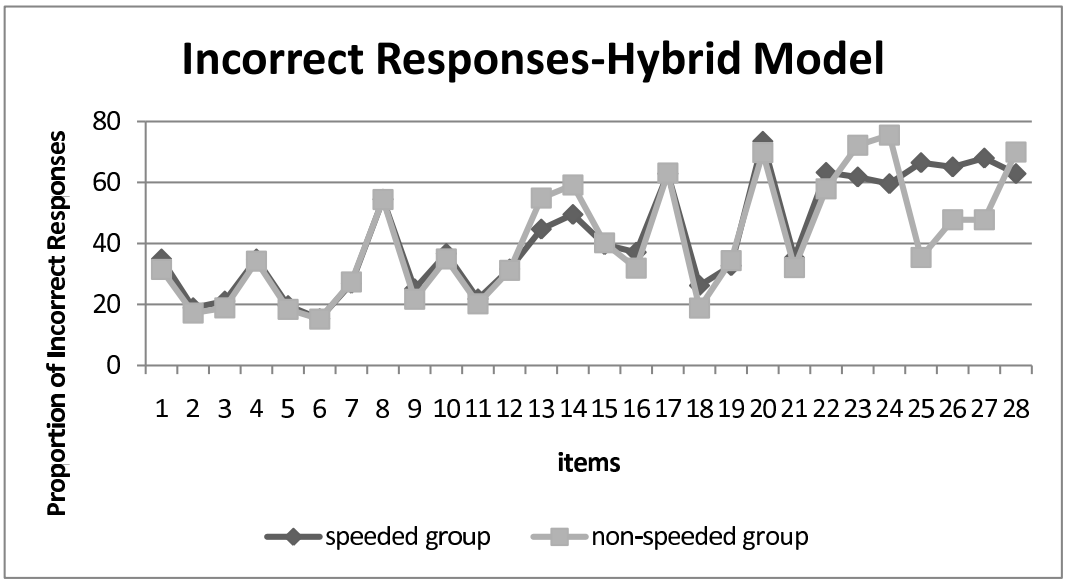


Figure 4.14: Proportion of incorrect responses by the MRM model.

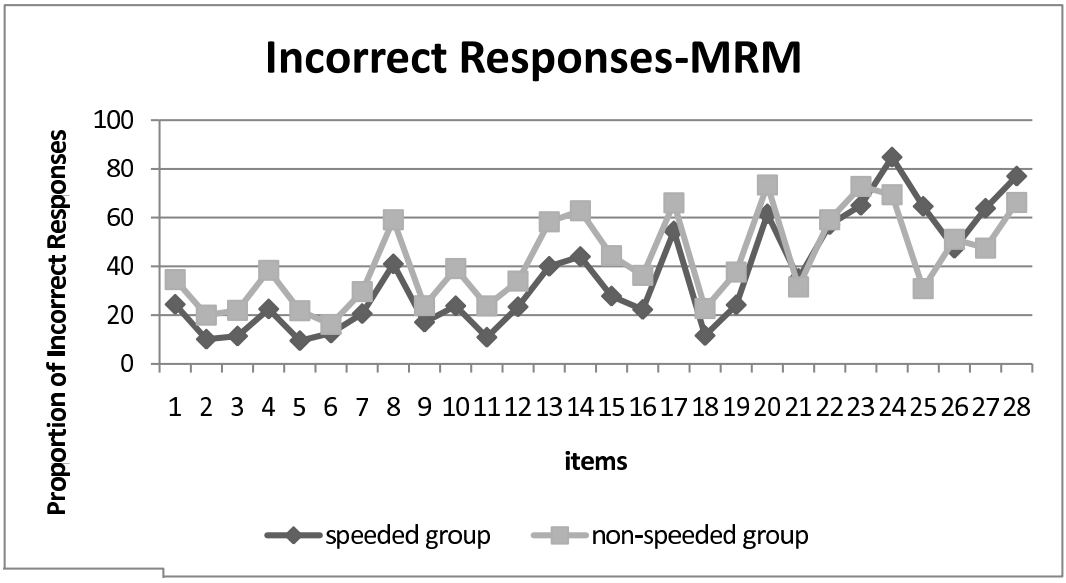
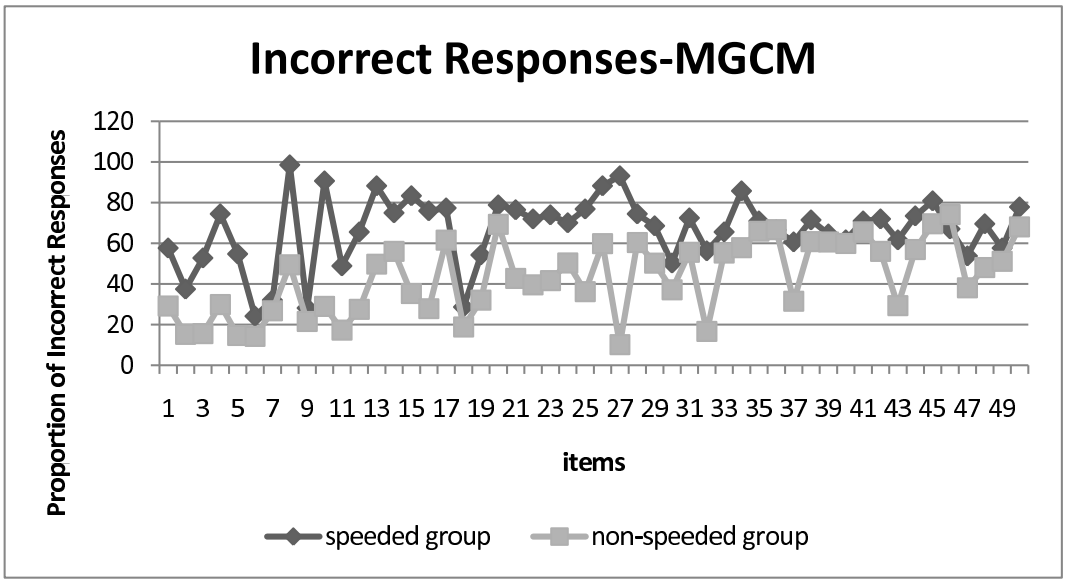


Figure 4.15: Proportion of incorrect responses by the MixGPCM model.



to items at earlier locations of the test were similar between the speeded and non-speeded groups. For items near the end of the test, similar pattern to those for the other models were found; that is, the proportions of incorrect responses were higher for the speeded group than for the non-speeded group. As for the proportions of correct responses, the MixGPCM showed a different pattern in the proportions of incorrect responses than was observed for the other models. The speeded group identified by the MixGPCM showed a consistently higher proportion of incorrect responses than the non-speeded group.

Raw Score for Speeded and Nonspeeded Groups. The raw scores for the first 20 items are reported for each latent class in Figures 4.16 t0 4.25. Figures of the raw scores calculated using the first 20 items show the raw scores of the speeded groups were higher than the nonspeeded groups by the unconditional and conditional cross-classified mixture Rasch models and the MRM. For the unconditional model, the speeded examinees had an average raw score of 15.30 for the first 20 items and the non-speeded examinees had an average raw

Figure 4.16: Raw scores of the first 20 Items by the non-speeded group-unconditional model.

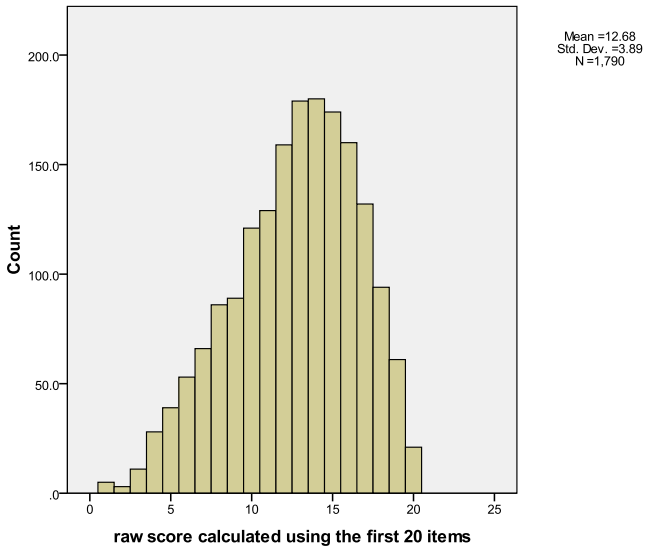


Figure 4.17: Raw scores of the first 20 items by the speeded group-unconditional model.

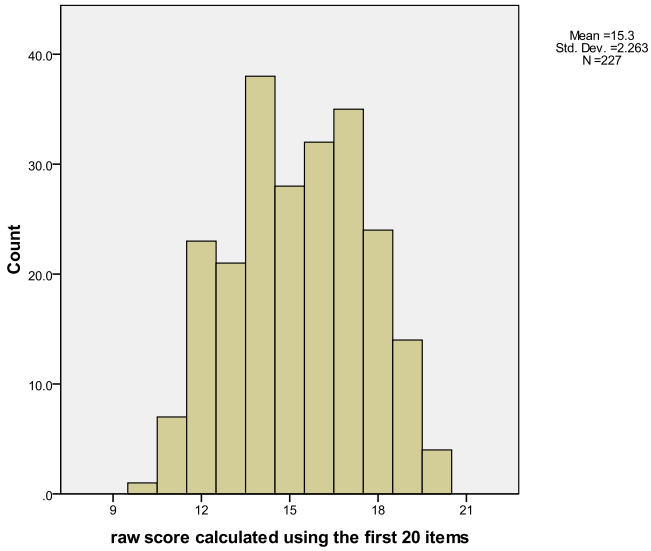


Figure 4.18: Raw scores of the first 20 items by the non-speeded group-conditional model.

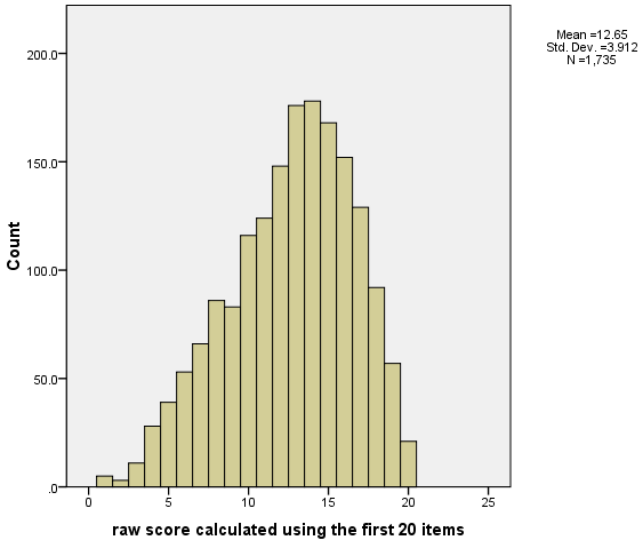


Figure 4.19: Raw scores of the first 20 items by the speeded group-conditional model.

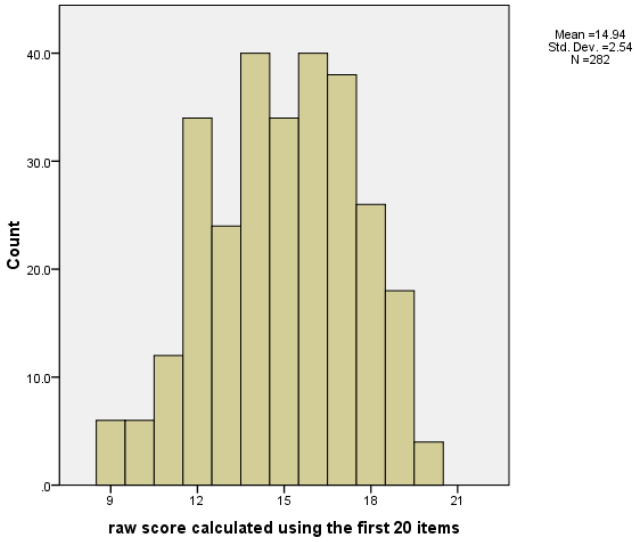


Figure 4.20: Raw scores of the first 20 items by the non-speeded group-Hybrid model.

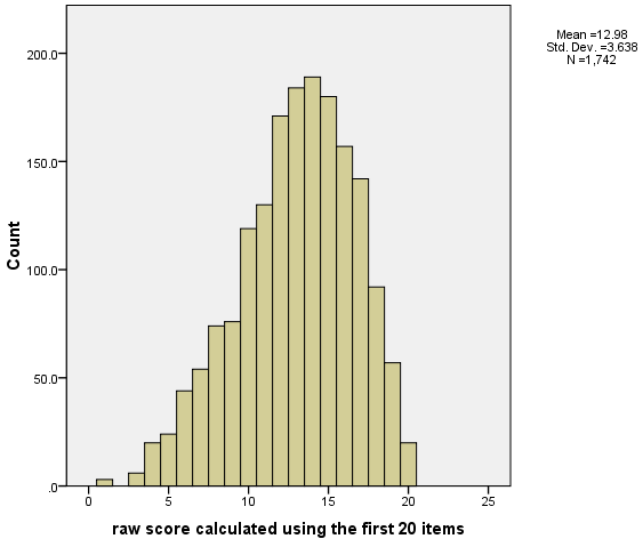


Figure 4.21: Raw scores of the first 20 items by the speeded group-Hybrid model.

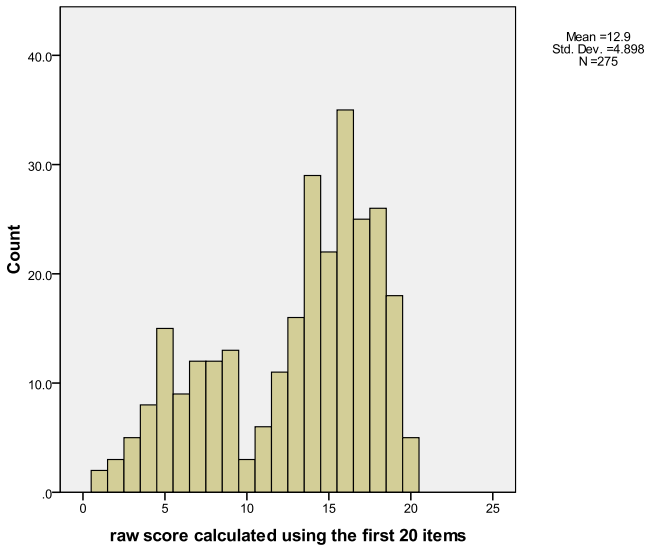


Figure 4.22: Raw scores of the first 20 items by the non-speeded group-MRM model.

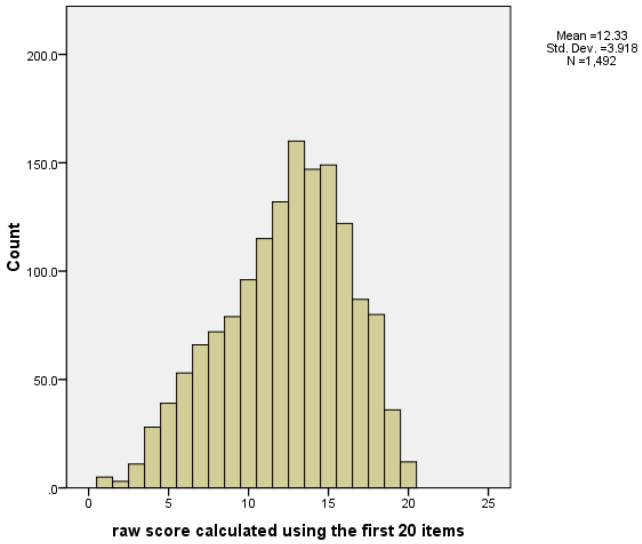


Figure 4.23: Raw scores of the first 20 items by the speeded group-MRM model.

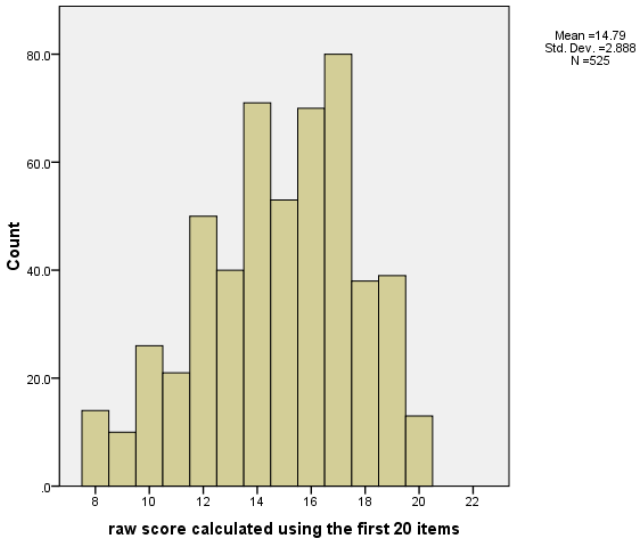


Figure 4.24: Raw scores of the first 20 items by the non-speeded group-MixGPCM model.

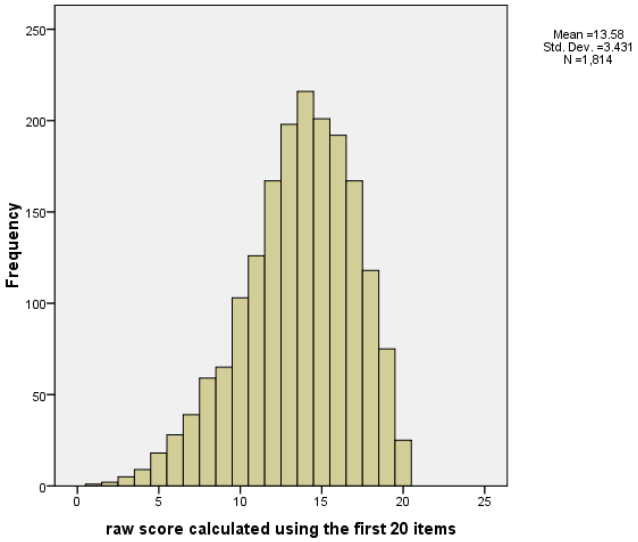
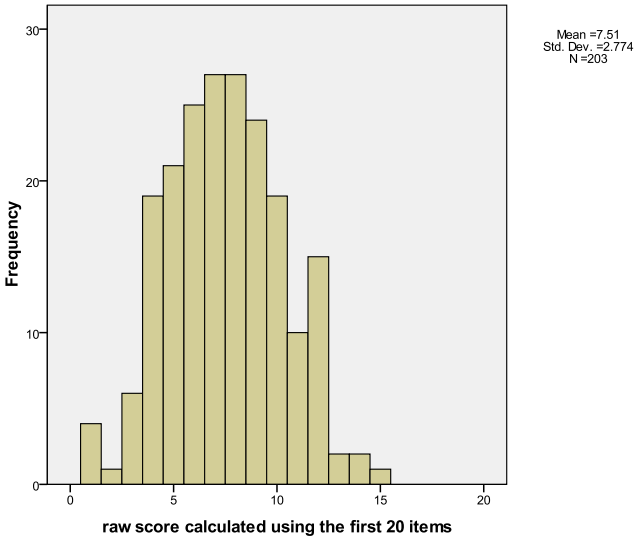


Figure 4.25: Raw scores of the first 20 items by the speeded group-MixGPCM model.



score of 12.68. The difference in the mean raw scores between the speeded examinees and the nonspeeded examinees was 2.62. For the conditional model, the speeded examinees had a mean raw score of 14.94 and the nonspeeded examinees had a mean raw score of 12.65. The difference in the mean raw scores was 2.29. The average raw scores of the speeded examinees identified by the MRM for the first 20 items was 14.79. The nonspeeded examinees had an average raw score of 12.33. The difference in the average raw score was 2.46. This pattern of differences indicated that, in the absence of test speededness, the speeded group tended to have a higher score than the nonspeeded group. However, the same order of differences in the mean raw scores was not observed in the results based on the Hybrid model. The average mean raw scores for the first 20 items was 12.90 for the speeded examinees identified by the Hybrid model and 12.98 for the nonspeeded examinees. The difference between these two means of .08 was not a meaningful difference. The raw scores for the first 20 items by the MixGPCM had a different pattern. The speeded examinees by the MixGPCM had an average raw score of 7.51 and the non-speeded examinees had an average raw score of 13.58. This difference was much larger than for the other models.

Histograms of the total raw scores calculated over the last 8 items, that were assumed to be most affected by test speededness, are presented in Figure 4.26 to 4.35.

The average raw scores over the last 8 items for speeded examinees and non-speeded examinees identified for the unconditional model were 1.41 and 3.42, respectively. The non-speeded examinees had an average raw score of 2.01 points higher than the speeded examinees for these items. For the conditional model, the speeded examinees had a mean raw score of 1.45 and the non-speeded examinees, a mean raw score of 3.48. The differences in the mean raw scores between the speeded and non-speeded examinees was 2.03. The mean raw scores of the speeded group and non-speeded group classified by the MRM were 2.05 and 3.60, respectively. The differences in these mean raw scores was 1.55 points. The same order of differences was observed for the Hybrid model. The speeded examinees identified by the Hybrid model had an average raw score of 2.08 and the non-speeded examinees have an

Figure 4.26: Raw scores of the last 8 items for the non-speeded group: unconditional model.

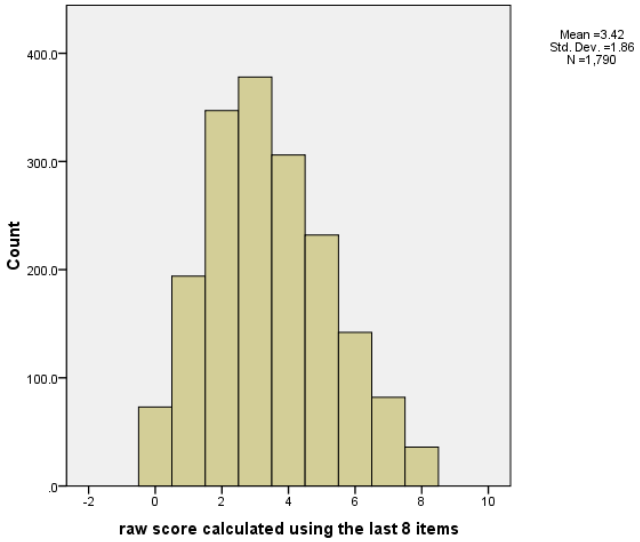


Figure 4.27: Raw scores of the last 8 items for the speeded group: unconditional model.

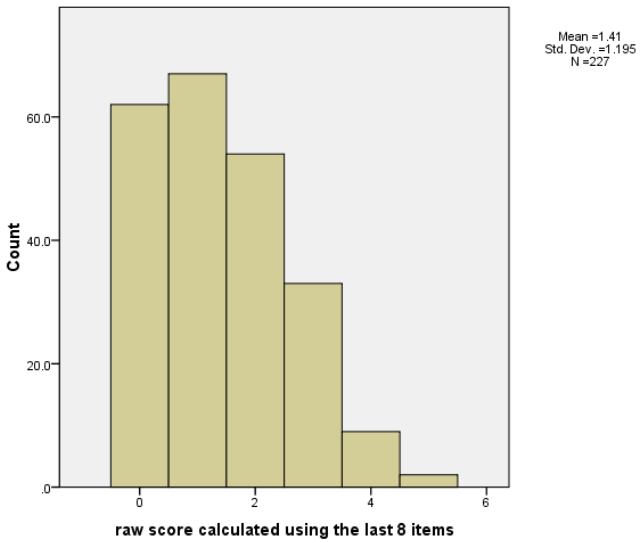


Figure 4.28: Raw scores of the last 8 items for the non-speeded group: conditional model.

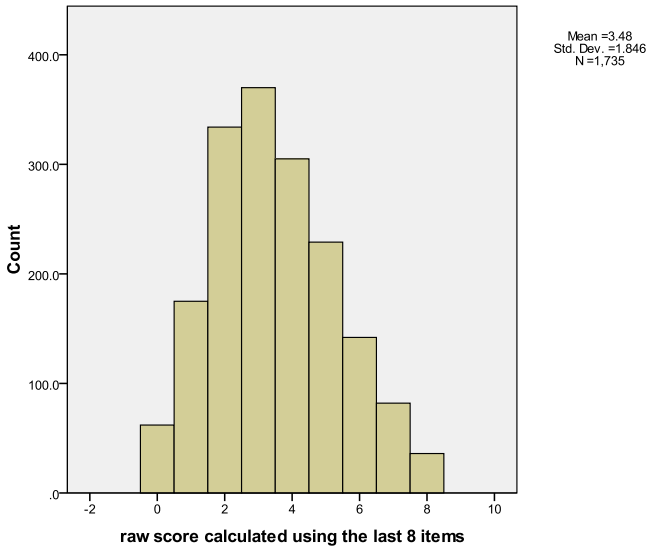


Figure 4.29: Raw scores of the last 8 items for the speeded group: conditional model.

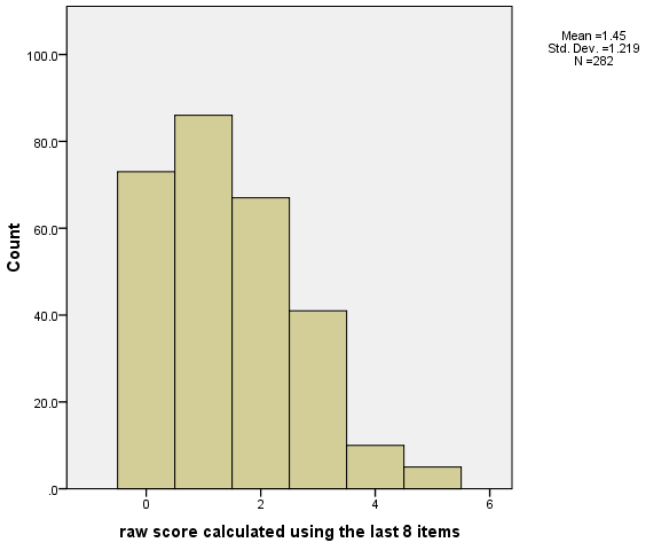


Figure 4.30: Raw scores of the last 8 items for the non-speeded group: Hybrid model.

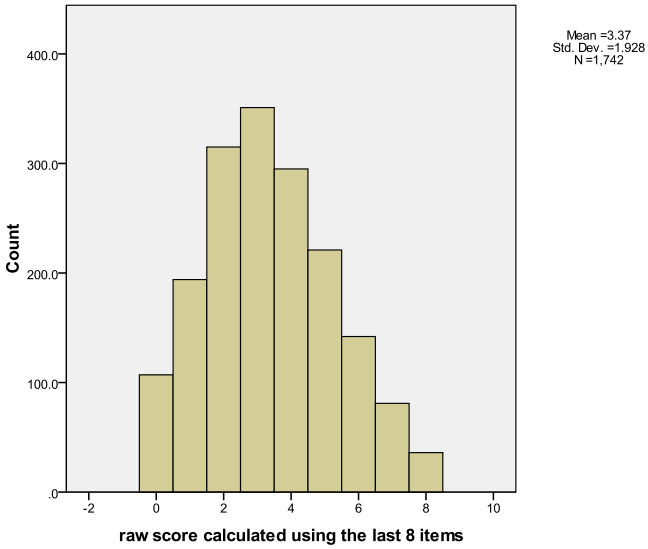


Figure 4.31: Raw scores of the last 8 items for the speeded group: Hybrid model.

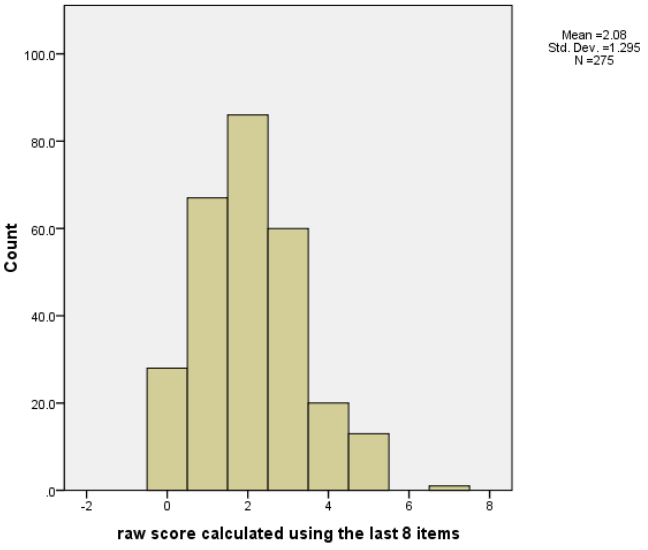


Figure 4.32: Raw scores of the last 8 items for the non-speeded group: MRM.

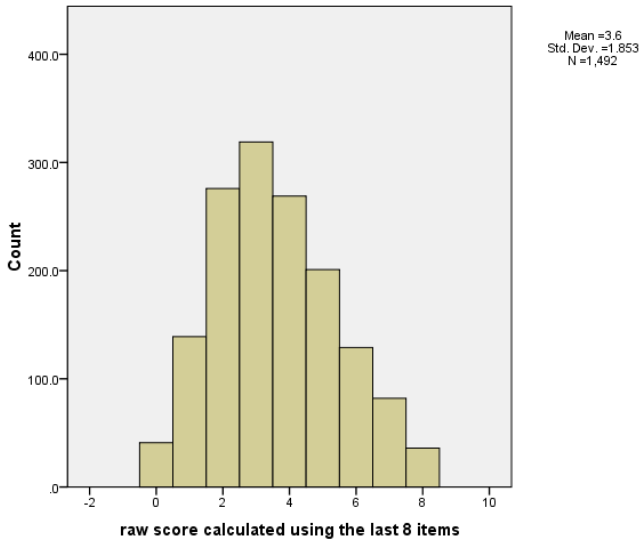


Figure 4.33: Raw scores of the last 8 items for the speeded group: MRM.

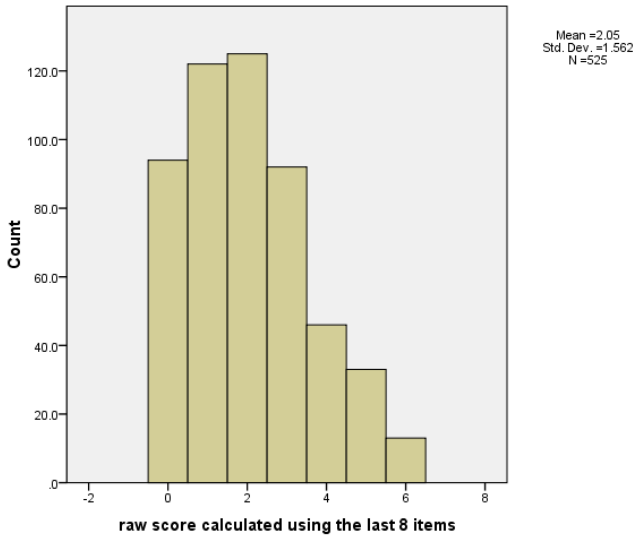


Figure 4.34: Raw scores of the last 8 items for the non-speeded group: MixGPCM.

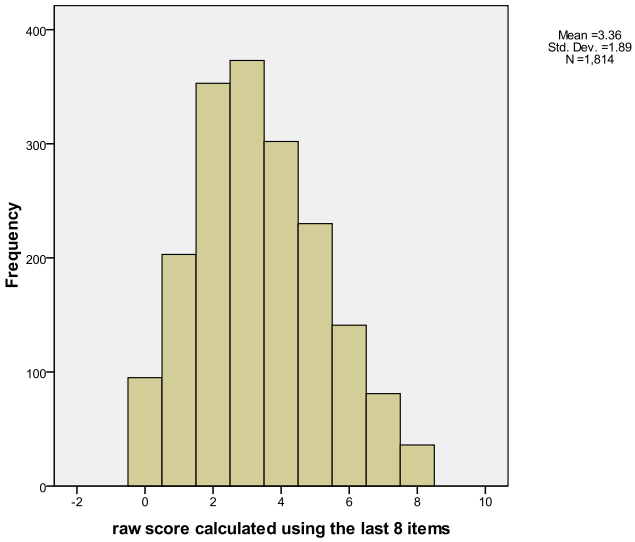
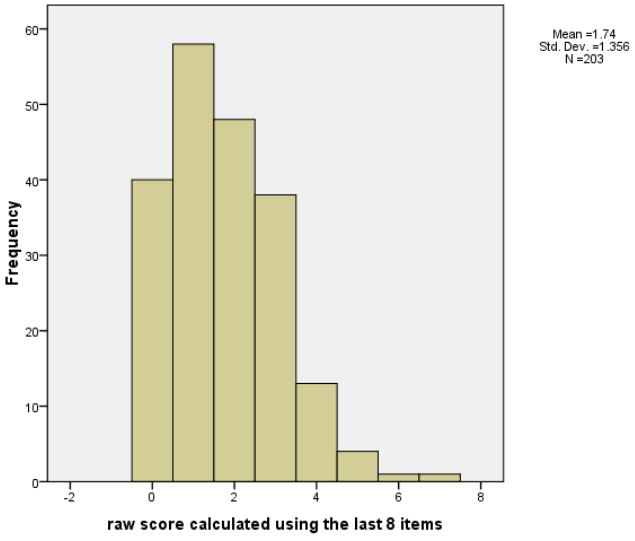


Figure 4.35: Raw scores of the last 8 items for the speeded group: MixGPCM.



average raw score of 3.37 with a difference of 1.29 points. Unlike the raw score patterns for the first 20 items, the pattern in the raw scores to the last 8 items by the MixGPCM was consistent with the patterns in the other models, that is, the average raw scores for the non-speeded group (3.36) was higher than the raw scores for the speeded group (1.74).

The differences observed for each model in raw scores between speeded and nonspeeded examinees was consistent with previous research (e.g., Cohen et al., 2002). This finding suggests that speeded examinees were of slightly higher ability. The differences in raw scores over the last 8 items was largely a function of the way that speededness was modeled. That is, the last 8 items were constrained to be harder for the speeded than for the non-speeded group, resulting in lower scores for the speeded examinees.

Chapter 5

Discussion

This dissertation described a cross-classified multilevel mixture IRT model for detection of test speededness effects. Unlike the usual IRT models, which treat person effects as random and item effects as fixed, the cross-classified IRT model treats both persons and items as random effects. This allows the inclusion of person or item covariates in the model. As was noted earlier, treating items as random effects is more realistic than treating them as fixed effects. This is because items are typically assumed to be sampled from a domain. Treating items as random, furthermore, allowed the inclusion of an error term in the latent variable regression. When items were treated as fixed, however, the item covariates are assumed to explain all the differences in item parameters, that is, without error. The assumption in this dissertation was that this may not be either realistic or appropriate.

Two forms of the mixture cross-classified IRT model were presented, one with no covariates (the unconditional model) and the other with covariate (the conditional model). The covariates applied to the conditional model were on latent group membership for the purpose of helping model latent group membership.

The proposed models were then estimated using an MCMC algorithm as implemented in the software WinBUGS (Spiegelhalter et al., 2003), and a simulation study was conducted to examine the recovery of the parameters of the model under various practical testing conditions. Two primary factors were manipulated in the simulation study, sample size and proportion of speeded examinees. A real data example was also presented to demonstrate the application of the mixture cross-classified IRT model for detection of test speededness.

5.1 Discussion of Simulation Study

For the unconditional model, the recovery of item easiness parameters improved with the increase in sample size or with an increase of proportion of speeded examinees. As sample size or the proportion of speeded examinees increased, the reduction in bias and RMSE of item easiness parameters was greater in the speeded group that in the non-speeded group. This is reasonable, since more examinees were involved in the estimation of item easiness parameters for the speeded group when the sample size or the proportion of speeded examinees increased. The corresponding reduction in the bias and RMSE for the sample of non-speeded examinees, however, was very small as these two factors increased. This was because the sample sizes of non-speeded examinees were large in all the simulation conditions. Therefore, the increases in sample sizes in the simulation study did not improve the estimates of the item parameters much.

The recovery of random item effects was better than the recovery of item easiness parameters. Similar patterns to those for the item easiness parameters were observed for random item effects. That is, the bias and RMSE of the random item effects decreased with the increase in either sample size or proportion of speeded examinees. Recovery of random person effects also was good across all simulation conditions. Although the bias and RMSE of the random person effects tended to be smaller with an increase in either sample size or the proportion of speeded examinees, the reductions were small. The random person effects were recovered well with a sample size of 1,000 examinees and 10% speeded examinees. The recovery of the classification of examinees into speeded and non-speeded groups improved with the increase in sample sizes. This was most obvious between the conditions of 1,000 examinees and 2,000 examinees. When the sample size increased from 2,000 to 3,000, the improvement in the recovery of classifications was small. The recovery of classification of examinees improved with an increase in proportions of speeded examinees but only for the conditions of 1,000 examinees. For the conditions of 2,000 and 3,000 examinees, improvement was not observed.

For the conditional model, gender was used as a covariate on the latent group membership. After the inclusion of gender as a covariate, the bias and RMSE of the item easiness parameters for the speeded group became smaller than for the unconditional model. Similar to the unconditional model, however, the bias and RMSE of the item easiness parameters for the speeded group decreased with an increase in either sample size or the proportion of speeded examinees. However, the bias and RMSE of the item easiness parameters for the non-speeded group was not smaller in the conditional model than in the unconditional model. Thus the inclusion of covariate on the latent class membership seemed to be of most help in improving recovery of item easiness parameters in the non-speeded group. A similar pattern to the unconditional model was also found in the non-speeded group. That is, the bias and RMSE of the item easiness parameters decreased with the increase of sample size and the proportion of speeded examinees. However, the improvement in the recovery of item easiness parameters for the non-speeded group was small, when the sample size increased from 1,000 to 2,000 or the proportions increased from 20% to 30%. The reduction in bias and RMSE of the item easiness parameters in the non-speeded group was more obvious, when the sample size increased from 2,000 to 3,000 indicating a large sample size was needed to improve recovery of item parameters for the non-speeded group.

5.2 Discussion of Results for the Real Data Example

Both an unconditional model and a conditional model were fit to the same set of college-level placement data. The unconditional model, that is, the cross-classified multilevel mixture Rasch model without item or person covariates, was compared with the conditional model, that included covariates for latent group membership. Gender was included as a covariate on the probabilities of latent class memberships in the conditional cross-classified model. This was done to help explain differences in person effects and to help classify examinees into latent classes.

Modeling Speededness. To model the effect of test speededness, the approach used in this dissertation imposed constraints on item parameters of both the unconditional model and the conditional model to reflect the assumption that items near the end of the test differed in easiness for speeded and non-speeded examinees. These same constraints were used for the MRM, Hybrid, and MixGPCM models as well. The constraints were implemented as inequality constraints on the item easiness parameters for the last 8 items on the test. The item easiness parameters were constrained to be lower than for the speeded group than for the non-speeded group.

Results for Conditional and Unconditional Models. The unconditional model classified about 11.3% of the examinees as speeded while the conditional model classified about 14.0% of the same examinees as speeded. After including gender as a covariate on group membership, more examinees who were classified as non-speeded by the unconditional model were subsequently classified as speeded by the conditional model, than the reverse. A significance test on gender indicated female examinees tended to be more affected by test speededness, as defined in the models studied here, than did male examinees. An examination of latent classes by gender showed there were more female examinees than male examinees in the speeded group classified by both the unconditional model and the conditional model. The differences in the proportions of speeded examinees between male and female examinees were increased after the inclusion of gender as a covariate.

The number of omitted responses to items at earlier nonspeeded locations of the test were similar for he speeded and nonspeeded examinees. However, speeded examinees had a consistently higher number of omissions near the end of the test than nonspeeded examinees for both models. For items at earlier locations of the test where test speededness was assumed to be absent, the number of correct responses were larger for the speeded group than for the non-speeded group and the number of incorrect responses were lower for the speeded group than for the non-speeded group. This situation was reversed for items near the end of the test where test speededness was assumed to be present. An examination of the raw scores

calculated using the first 20 items and the last 8 items respectively showed the same pattern. In general, the raw scores for the first 20 items were higher for the speeded group than the non-speeded group and the raw scores for the last 8 items were lower for the speeded group than for the non-speeded group.

Comparisons of Item Parameters with Rasch, MRM, Hybrid, and MixGPCM Models. The results of the unconditional and conditional cross-classified multilevel mixture Rasch models were compared with results from the MRM, Hybrid, MixGPCM and Rasch model. The MixGPCM identified the smallest proportion of speeded examinees (10.10%) among the five models, and the MRM classified the most examinees as speeded (26.03%). The proportion of speeded examinees identified by the conditional model (13.98%) was similar to that by the Hybrid model (13.63%), and the proportion of speeded examinees identified by the unconditional model (11.25%) was similar to that by the MixGPCM (10.10%).

The item parameters obtained with each model were also compared. For the item parameters in the speeded group, the unconditional model, the conditional model, the MixGPCM and the MRM produced two groups of item difficulties. The Hybrid and Rasch models produced a single set of item difficulties, and these were compared with the nonspeeded group. So, comparisons of item parameters for the nonspeeded group were made among all the five models. The item parameters for the speeded group, however, were only compared among the conditional model, the unconditional model, the MRM, and MixGPCM models.

Comparisons of item parameter estimates were made by calculating the correlations among the item parameters by each model. In the nonspeeded group, correlations of the item parameter estimates with those from the other five models were very high. The smallest was 0.974. The similarities of the item parameters for the two cross-classified models and the MRM might be due to the similarity in the way the speededness constraints were implemented. However, the item difficulties correlations in the speeded group were not consistently high among any of the models. The correlations among the two cross-classified multilevel mixture Rasch models and the MRM were high and similar to those in the non-speeded

group. The correlations between the MixGPCM and the other two models, however, were very low.

Comparisons of Omitted Responses. The responses to the speeded and non-speeded items by each latent group were also compared. The speeded group identified by the cross-classified multilevel mixture Rasch models, the MRM and the Hybrid model showed a consistently higher proportion of omitted responses to items near the end of the test than the nonspeeded group. In addition, The speeded groups had a lower proportion of correct responses and a higher proportion of incorrect responses to these same items. For items at earlier location of the test, the speeded groups had similar proportions of omissions, a higher proportion of correct responses and a lower proportion of incorrect responses than the nonspeeded groups. For the MixGPCM, responses showed a different pattern. The speeded group showed an increasing trend in the proportion of omitted responses over most of the 50 items. The differences in the proportions of omissions between the speeded and nonspeeded groups by the MixGPCM started at an earlier location of the test than the other models, as this model considered more item near the end of the test than just the last 8. For the MixGPCM, the nonspeeded group had a consistently higher proportion of correct responses and a lower proportion of incorrect responses than the speeded group throughout the whole test.

5.3 LIMITATIONS AND FUTURE STUDIES

Two factors were manipulated in the simulation study, sample size and proportion of speededness examinees. Other factors might be considered in future research. Results from the MixGPCM, for example, suggest that speededness effects might be evident earlier in the test. In this study, the item parameters of the first 20 items were fixed at values estimated using MULTILOG, and the speeded items, that is, the last 8 items, were estimated along with other model parameters. If the number of speeded items to be estimated changed, the accuracy of the item easiness parameters might also vary. It might be helpful to vary

the number of items to be estimated to look at its impact on the estimation of the model parameters.

Another limitation associated with the simulation study is the sample sizes. In this dissertation, the number of examinees simulated was large. It might be useful to simulate smaller sample sizes to determine recovery of model parameters.

Some other limitations of this dissertation include the assumptions about speededness effects. These were based on those used in Bolt et al. (2002), and as a result, they share limitations similar to those used with the MRM and Hybrid model. These models consider speededness from the same point in the test for all examinees. If test speededness occurred at earlier locations of the test, this model would be unable to capture it. The MixGPCM, however, does recognize that the choice of speededness point is arbitrary. Although it is a complex model, developing a cross-classified version of that model could provide a useful methodology.

Second, the mixture cross-classified IRT model assumes speeded examinees are affected by test speededness in the same way and so produces estimates for only one speeded group. In reality, examinees may differ in speededness patterns, and more than one speeded groups may exist. Speededness may be exhibited, for example, by differential difficulties in one speeded group, by different patterns of omissions in another speeded group, and possibly by some other type of response strategy in yet a third group. Extending the current cross-classified multilevel mixture IRT model to detect multiple speeded groups should be straightforward, but this conjecture will need to be studied.

Third, the current study only used the Rasch version of the cross-classified multilevel mixture IRT model. This model can be extended to the 2PL and 3PL models, but this would need to be studied. Using more highly parameterized IRT models than the Rasch model could result in examinees being placed into different latent classes (Alexeev, Templin, & Cohen, 2010). A 2PL or 3PL version of the cross-classified models, for example, might

produce different compositions of latent groups or might reveal other differences between speeded and non-speeded groups.

Fourth, in this study, only a single person covariate was included in the model. If more item or person covariates were to be included, the proportion and classification of latent classes might also change.

Finally, the cross-classified multilevel mixture IRT models also can be applied to data sets with testlets. Testlet effects could be modeled either at a higher level or as an item covariate. As for other extensions of this model, this kind of extension will also need to be studied.

Bibliography

[1] Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*, 47-76.

[2] Alexeev, N., Templin, J. & Cohen, A. S. (2010, April). *Detecting spurious latent classes with the mixture Rasch model.* Paper presented at the annual meeting of the National Council of Measurement in Education. Denver, CO.

[3] Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

[4] Attali, Y. (2005). Reliability of speeded number-right multiple-choice tests. *Applied Psychological measurement, 29*, 357-368.

[5] Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement, 22*, 153-169.

[6] Bejar, I. I. (1985). *Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language* (Research Report No. ETS-RR-85-11). Princeton, NJ: Educational Testing Service.

[7] Birnbaum, A. (1968). Some latent trait m dels and their use in inferring an examinee's ability. In F. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

[8] Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*, 381-409.

[9] 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.

[10] Bolt, D. M., Mroch, A. A., & Kim, J.-S. (2003, April,). *An empirical investigation of the Hybrid IRT model for improving item parameter estimation in speeded tests.* Paper presented at the annual convention of the American Educational Research Association, Chicago, IL.

[11] Bridgeman, B., Cline, F., & Hessinger, J. (2003). *Effect of extra time on GRE quantitative and verbal scores.* Princeton, NJ: Educational Testing Service.

[12] Briel, J. B., O'Neill, K. A., & Scheuneman, J. D. (1993). *GRE technical manual.* Princeton, NJ: Educational Testing Service.

[13] Briggs, D. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education, 21*, 89-118.

[14] Cho, S. (2007). *A multilevel mixture IRT model for DIF analysis.* Unpublished doctoral dissertation. University of Georgia: Athens, GA.

[15] Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2006, June). *An investigation of priors on the probabilities of mixtures in the mixture Rasch model.* Paper presented at the annual meeting of the Psychometric Society, Montreal, CN.

[16] 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.

[17] Cronbach, L. J., & Warrington, W.G. (1951). Time limit tests: Estimating their reliability and degree of speeding. *Psychometrika, 14*, 167-188.

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

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

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

[21] Fox, J. P., & Glas, C. (2001). Bayesian estimation of a multilevel IRT model usingGibbs sampling. *Psychometrika, 66*, 271-288.

[22] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* 4 (pp. 169-194). Oxford: Oxford University Press.

[23] 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.

[24] Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley and Sons.

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

[26] Kamata, A. (1998, April). *One-parameter hierarchical generalized linear logistic model: an application of HGLM to IRT*. Paper presented at the annual meeting of American Educational Research Association, San Diego, CA.

[27] Kim, S.-H. (2007). Some posterior standard deviations in item response theory. *Educational and Psychological Measurement, 67*, 258-279.

[28] Kim, J.-S., & Bolt, D. M. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice, 26*, 38-51.

[29] Lawrence, I. M. (1993). *The effect of test speededness on subgroup performance*. (Research report No. ETS-RR-93-49). Princeton, NJ: Educational Testing Service.

[30] Lazarsfeld, P. F., & Henry, N. W. (1968), *Latent structure analysis*. Boston: Houghton Mifflin.

[31] Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*, 353-373.

[32] Lord, F. M. (1956). A study of speed factors in tests and academic grades. *Psychometrika, 21* , 31-50.

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

[34] Lu, Y. & Sireci, S. G. (2007). Validity issues in test speededness. *Journal of Educational Measurement: Issues and Practice, 26*, 29-37.

[35] Lubke, G., Muthén, B. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling, 14*, 26-47.

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

[37] Mellon, S. J., Daggett, M., MacManus, V., & Moritsch, B. (1996). Development of GATB Forms E and F. In R. A. McCloy, T. L. Russell, & L. L. Wise (Eds.), *GATB improvement project final report*. Washington, DC: U.S. Department of Labor.

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

[39] Mroch, A. A., & Bolt, D. M. (2006, April). *An IRT-based response likelihood approach for addressing test speededness*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

[40] Neustel, S. (1998). *An investigation into the possible speededness of the MCAT* (Technical Report). Washington, DC: Association of American Medical Colleges.

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

[42] Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146-178.

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

[44] Ricci, P., & Ye, F. (April, 2009). *An investigation of cross-classification multilevel IRT models.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

[45] Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological methods, 8*, 185-205.

[46] Rindler, E. S. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement, 16*, 261-270.

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

[48] Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British journal of mathematical and statistical psychology, 44*, 75-92.

[49] Sager, C. E., Peterson, N. G., & Oppler, S. H. (1994). *An examination of the speededness of theGeneral Aptitude Test Battery power test*s. Washington, DC: American Institutes for Research.

[50] Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A mew method of measuring speededness. *Journal of Educational Measurement, 34*, 213-232.

[51] Secolsky, C. (1989). *Accounting for random responding at the end of the test in assessing speededness on the Test of English as a Foreign Language*. Princeton, NJ: Educational Testing Service.

[52] Sireci, S. G. (2005). Unlabeling the disabled: A psychometric perspective on flag-ging scores from accommodated test administrations. *Educational Researcher, 34*, 3-12.

[53] Smit, A., Kelderman, H. & Van Der Flier, H. (1999). Collateral information and mixture Rasch models. *Methods of Psychological Research Online, 4*, 19-32.

[54] Smit, A., Kelderman, H., & Van Der Flier, H. (2000). The mixed Birnbaum model: Estimation using collateral information. *Methods of Psychological Research Online, 5*, 31-43.

[55] Smith, B. (2005). *BOA: Bayesian output analysis program user manual* (Version 1.1). [Computer Software]. The University of Iowa, *http://www.public-health.uiowa.edu/boa*.

[56] Spiegelhalter, D.J., Thomas, A., Best, N.G., & Lunn,D. (2003). *WinBUGS 1.4 User manual* [Computer program]. Cambridge, UK: MRC Biostatistics Unit.

[57] Stafford, R. E. (1971). The speededness quotient: A new descriptive statistic for tests. *Journal of Educational Measurement, 8*, 275-278.

[58] Swineford, F. (1974). *Test analysis manual* (Statistical report SR-74-06). Princeton, New Jersey.

[59] 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.

[60] Van Nijlen, D., & Janssen, R. (2009, April). *Explaining guessing behavior by means of explanatory mixture models*. Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

[61] Von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In M. Von Davier & C. H. Carstensen (Eds), *Multivariate and mixture distribution Rasch models* (pp. 99-115). New York:Springer.

[62] Wang, A., & Cohen, A. S. (2008, July). *Evaluation of three test speededness models.* Paper presented at the 73rd annual meeting of the psychometric society, Durham, NH.

[63] Wilhelm, O., & Schulze, R. (2002). The relation of speeded and unspeeded reasoning with mental speed. *Intelligence*, 30, 537-554.

[64] 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.

[65] Yamamoto, K. (1987). *A model that combines IRT and latent class models*. Unpublished doctoral dissertation. University of Illinois, Champaign-Urbana.

[66] Yamamoto, K. (1989). *HYBRID model of IRT and latent class models*. (ETS Research Report RR-89-41), Princeton, NJ: Educational Testing Service.

[67] Yamamoto, K. (1990). *HYBIL: A computer program to estimate HYBRID model parameters*. Princeton, NJ: Educational Testing Service.

[68] Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (TOEFL Tech. Rep. No. TR-10). Princeton, NJ: Educational Testing Service.

[69] Yamamoto, K., & Everson, H. T. (1995). *Modeling the mixture of IRT and patterned responses using a modified hybrid model.* (Research Report Series RR-95-16), Princeton, NJ: Educational Testing Service.

[70] Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the hybrid model. In J. Rost and R. Langeheine. (Eds.), *Applications of latent trait and latent class models in the social sciences*, pp. 89-99. Waxmann: New York.

Appendix A

WinBUGS Code for the unconditional mixture cross-classified IRT model for test speededness

```
#N: number of examinees
    #T: number of items
    #gmem: latent group membership
    #sd.p: standard deviation of random person effect
    #sd.i: standard deviation of random item effect
    #tau.p: precision of random person effect
    #tau.i: precision of random item effect
    #mu: mean of ability
    #mbeta: fixed item effect or the mean of item parameter
    model
    for ( j in 1:N)
    for (k in 1:T)
    resp[j,k]~ dbern(p[j,k])
    logit(p[j,k])<- theta[j]+beta[gmem[j],k]
    gmem[j]~ dcat(pi[1:2])
    #level-2 model
    for ( j in 1:N)
    theta[j]~ dnorm(mu[gmem[j]], tau.p)
    #fix beta for items 1-20
    beta[1,1]<- 0.8893
```

beta[1,2]< − 1.802

beta[1,3]< − 1.665

beta[1,4]< − 0.7547

beta[1,5]< − 1.693

beta[1,6]< − 1.979

beta[1,7]< −1.146

beta[1,8]< − -0.2046

beta[1,9] < − 1.463

beta[1,10]< −0.7018

beta[1,11]< −1.566

beta[1,12]< −0.9116

beta[1,13]< − 0.1759

beta[1,14]< − -0.399

beta[1,15]< − 0.4316

beta[1,16]< − 0.8347

beta[1,17]< −-0.6414

beta[1,18]< −1.622

beta[1,19]< −0.7596

beta[1,20]< −-1.012

beta[2,1]< − 0.8893

beta[2,2]< − 1.802

beta[2,3]< − 1.665

beta[2,4]< − 0.7547

beta[2,5]< −1.693

beta[2,6]< −1.979

beta[2,7]< −1.146

beta[2,8]< −-0.2046

beta[2,9] <− 1.463

beta[2,10]<− 0.7018

beta[2,11]<− 1.566

beta[2,12]<− 0.9116

beta[2,13]<− -0.1759

beta[2,14]<− -0.399

beta[2,15]<− 0.4316

beta[2,16]<− 0.8347

beta[2,17]<− -0.6414

beta[2,18]<− 1.622

beta[2,19]<− 0.7596

beta[2,20]<− -1.012

for (k in 21:T)

beta[1,k]∼ dnorm(mbeta[1], tau.i)

beta[2,k]∼ dnorm(mbeta[2],tau.i)I(,beta[1,k])

mbeta[1]∼ dnorm(0,0.001)

mbeta[2]∼ dnorm(0,0.001)

#variance of person random effect using vague prior

sd.p ∼ dunif(0,2)

tau.p <− 1/(sd.p*sd.p)

var.p<− 1/tau.p

variance of random item effect tau.i.1 is the group 1 precision; tau.i.2 is group 2 precision

sd.i ∼ dunif(0,2)

tau.i<− 1/(sd.i*sd.i)

var.i<− 1/tau.i

mu[1]<− 0

```
mu[2]∼dnorm(0,0.001)
# Priors for mixture parameters (priors for person fixed effect)
pi[1:G]∼ ddirch(alpha[])
list(N=2017 , T=28 , G=2, alpha=c(0.5,0.5),
resp=structure(.Data=c(
1,1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,1,1,1,0,1,0,0,0,0,0,1,1,
1,0,1,1,1,1,0,1,0,0,1,1,1,1,1,1,1,1,1,0,0,0,0,1,0,0,1,0,
...
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1), .Dim=c(2017,28)))
```

Appendix B

WinBUGS Code for the conditional mixture cross-classified IRT model for test speededness

```
model
    for ( j in 1:N)
    for (k in 1:T)
    resp[j,k]~ dbern(p1[j,k])
    logit(p[j,k])<- theta[j]+beta[gmem[j],k]
    p1[j,k]<- max(0.00001,min(p[j,k],0.99999))
    #level 2 model
    for ( j in 1:N)
    theta[j] dnorm(mu[gmem[j]], tau.p)
    gmem[j]~ dcat(pi[j,1:G])
    # fix the first 20 items using item parameters from the unconditional model
    beta[1,1]<- 0.8893
    beta[1,2]<- 1.802
    beta[1,3]<- 1.665
    beta[1,4]<- 0.7547
    beta[1,5]<- 1.693
    beta[1,6]<- 1.979
    beta[1,7]<- 1.146
    beta[1,8]<- -0.2046
    beta[1,9]<- 1.463
```

beta[1,10]< − 0.7018

beta[1,11]< − 1.566

beta[1,12]< − 0.9116

beta[1,13]< − -0.1759

beta[1,14]< − -0.399

beta[1,15]< − 0.4316

beta[1,16]< − 0.8347

beta[1,17]< − -0.6414

beta[1,18]< − 1.622

beta[1,19]< −0.7596

beta[1,20]< −-1.012

beta[2,1]< − 0.8893

beta[2,2]< − 1.802

beta[2,3]< − 1.665

beta[2,4]< − 0.7547

beta[2,5]< − 1.693

beta[2,6]< − 1.979

beta[2,7]< − 1.146

beta[2,8] < − -0.2046

beta[2,9] < − 1.463

beta[2,10]< − 0.7018

beta[2,11]< − 1.566

beta[2,12]< − 0.9116

beta[2,13]< − -0.1759

beta[2,14]< − -0.399

beta[2,15]< − 0.4316

beta[2,16]< − 0.8347

```
beta[2,17]< − -0.6414
beta[2,18]< − 1.622
beta[2,19]< − 0.7596
beta[2,20]< −¡- -1.012
for (k in 21:T)
beta[1,k]∼ dnorm(mbeta[1], tau.i)
beta[2,k]∼ dnorm(mbeta[2],tau.i)I(,beta[1,k])
mbeta[1]∼ dnorm(0,1)
mbeta[2]∼ dnorm(0,1)
#variance of random person and item effect
sd.p∼dunif(0,2)
sd.i∼dunif(0,2)
tau.i< −1/(sd.i*sd.i)
tau.p< −1/(sd.p*sd.p)
var.i< −1/tau.i
var.p< −1/tau.p
mu[1]< −0
mu[2]∼ dnorm(0,1)
for(j in 1:N)
for (g in 1:G)
pi[j,g]< −alph[j,g]/sum(alph[j,1:G])
log(alph[j,g])< −gamma0[g]+gamma1[g]*gender[j]
for (j in 1:2000)
gender[j]∼ dbern(0.5)
#covariate identification
gamma0[1]< −0
gamma1[1]< −0
```

gamma0[2]∼ dnorm(0,1)

gamma1[2]∼ dnorm(0,1)

for (j in 1:N)

for(g in 1:G)

indprob[j,g]< −equals(gmem[j], g)

list(N=2017 , T=28 , G=2, gender=c(1,1,0,...0,0,1),

resp=structure(.Data=c(

1,1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,1,1,1,0,1,0,0,0,0,0,1,1,

1,0,1,1,1,1,0,1,0,0,1,1,1,1,1,1,1,1,1,0,0,0,0,1,0,0,1,0,

...

1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1), .Dim=c(2017,28)))

Appendix C

Convergence figures for one selected condition under the unconditional model

Figure C.1: The autocorrelation plot for pi for the condition of 1000 examinees with the proportion of 20% speededness

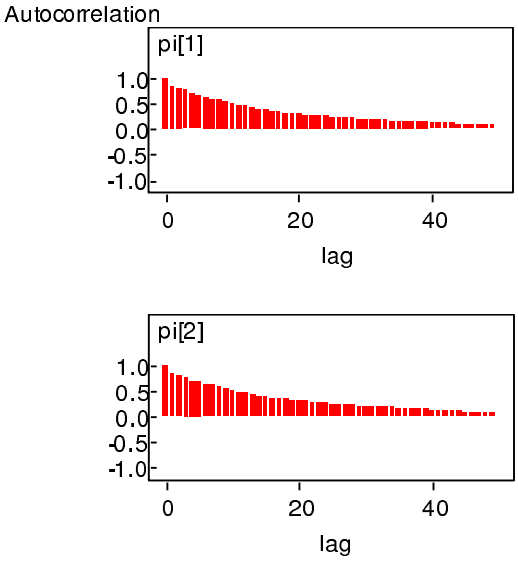


Figure C.2: The trace plot for pi for the condition of 1000 examinees with the proportion of 20% speededness

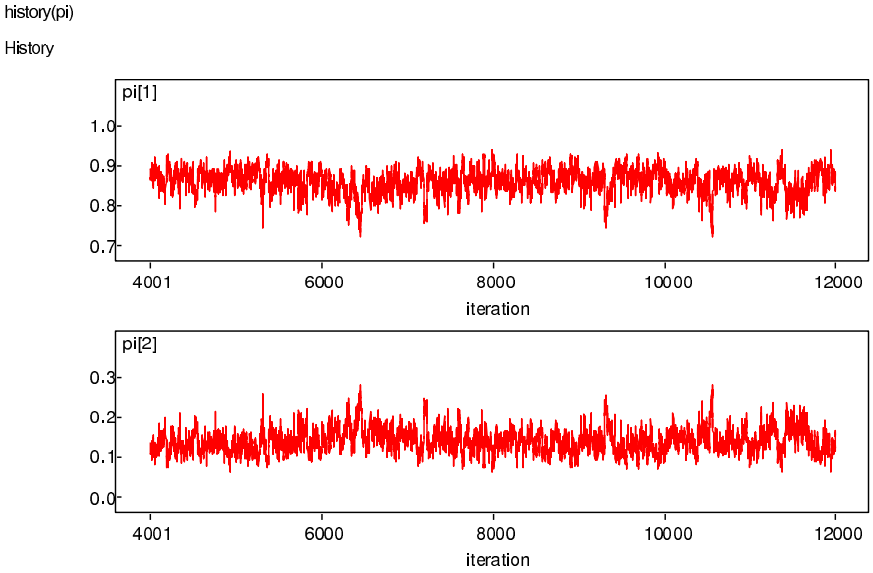
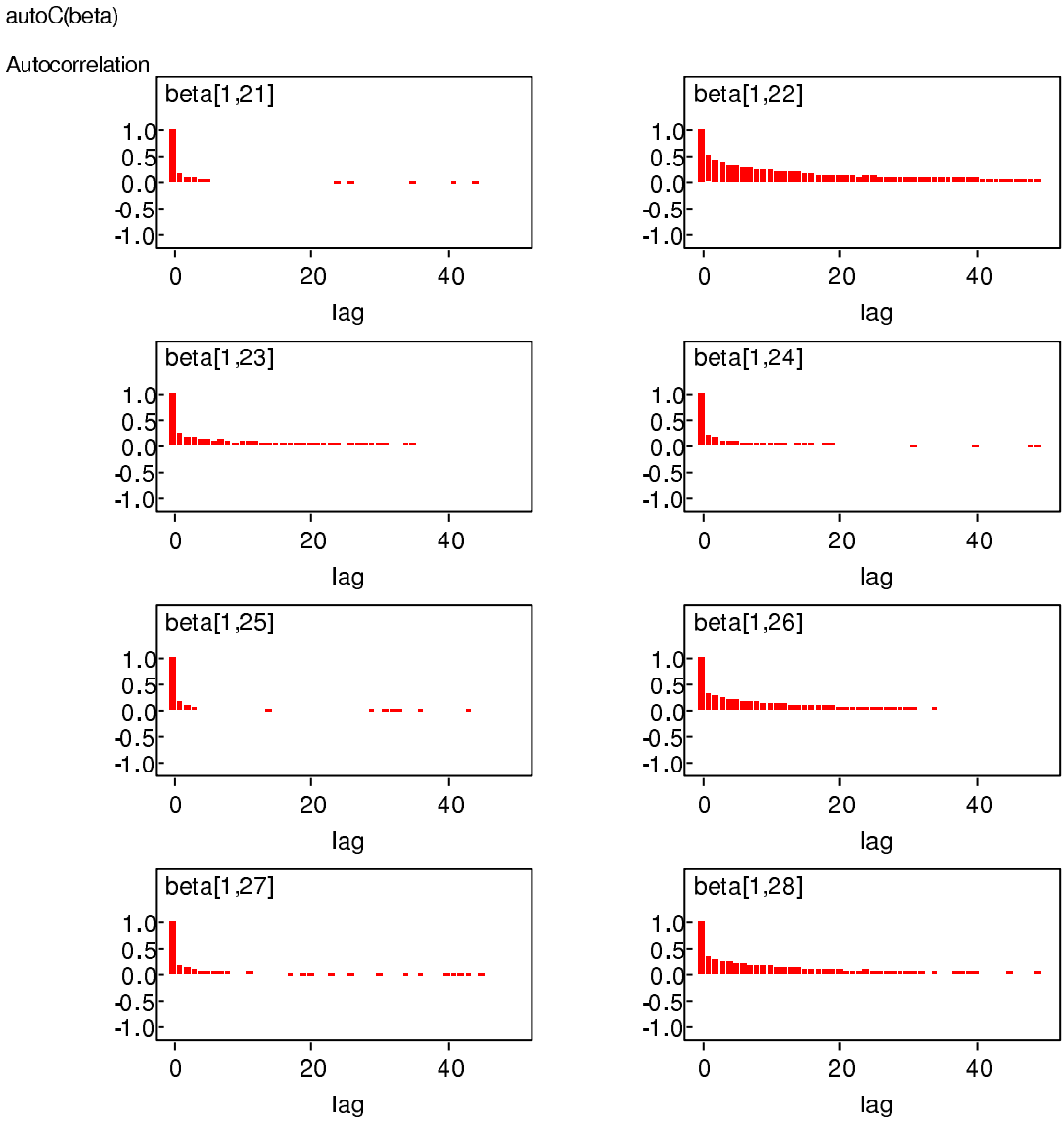


Figure C.3: The autocorrelation plot for beta for the condition of 1000 examinees with the proportion of 20% speededness



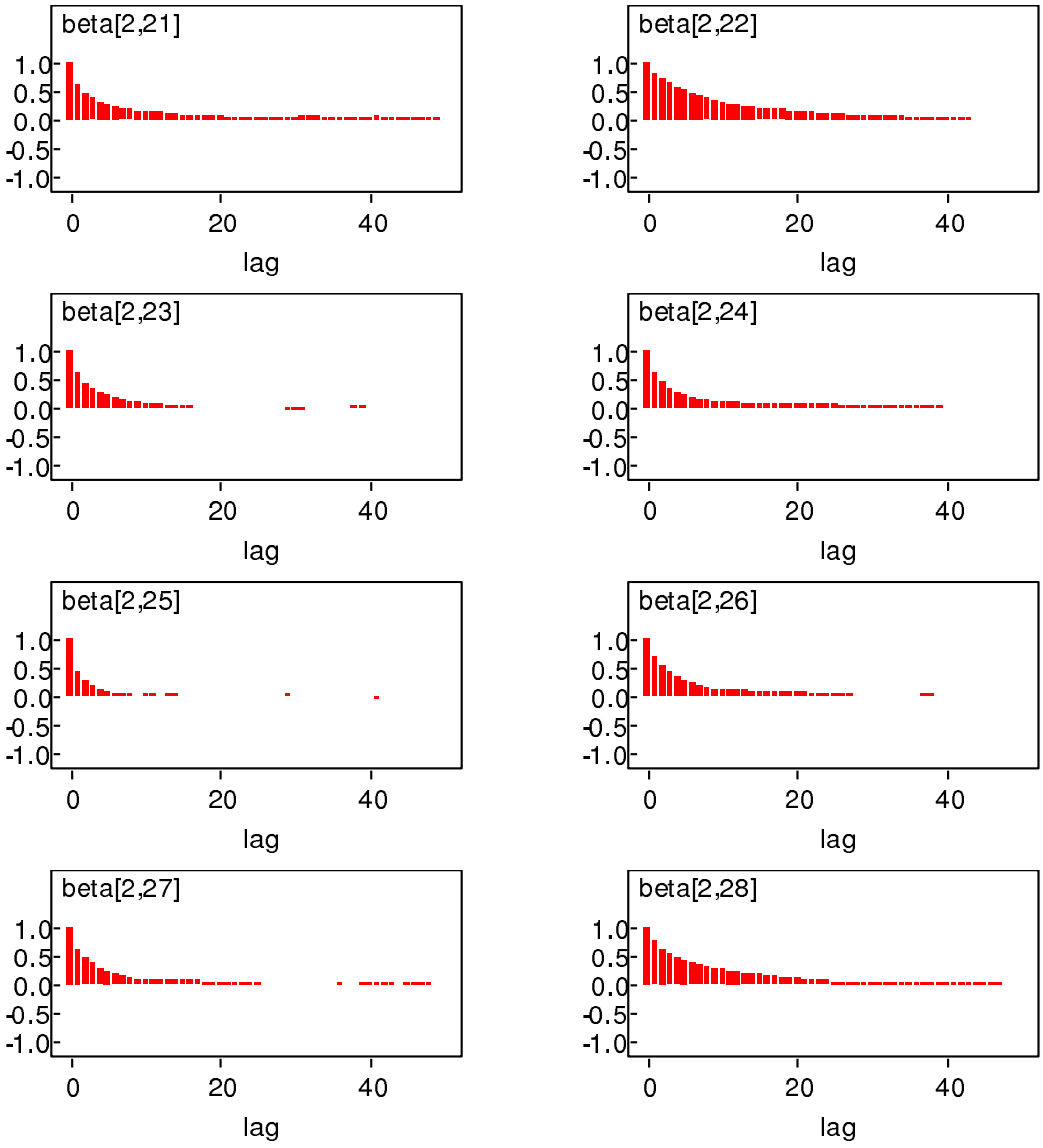
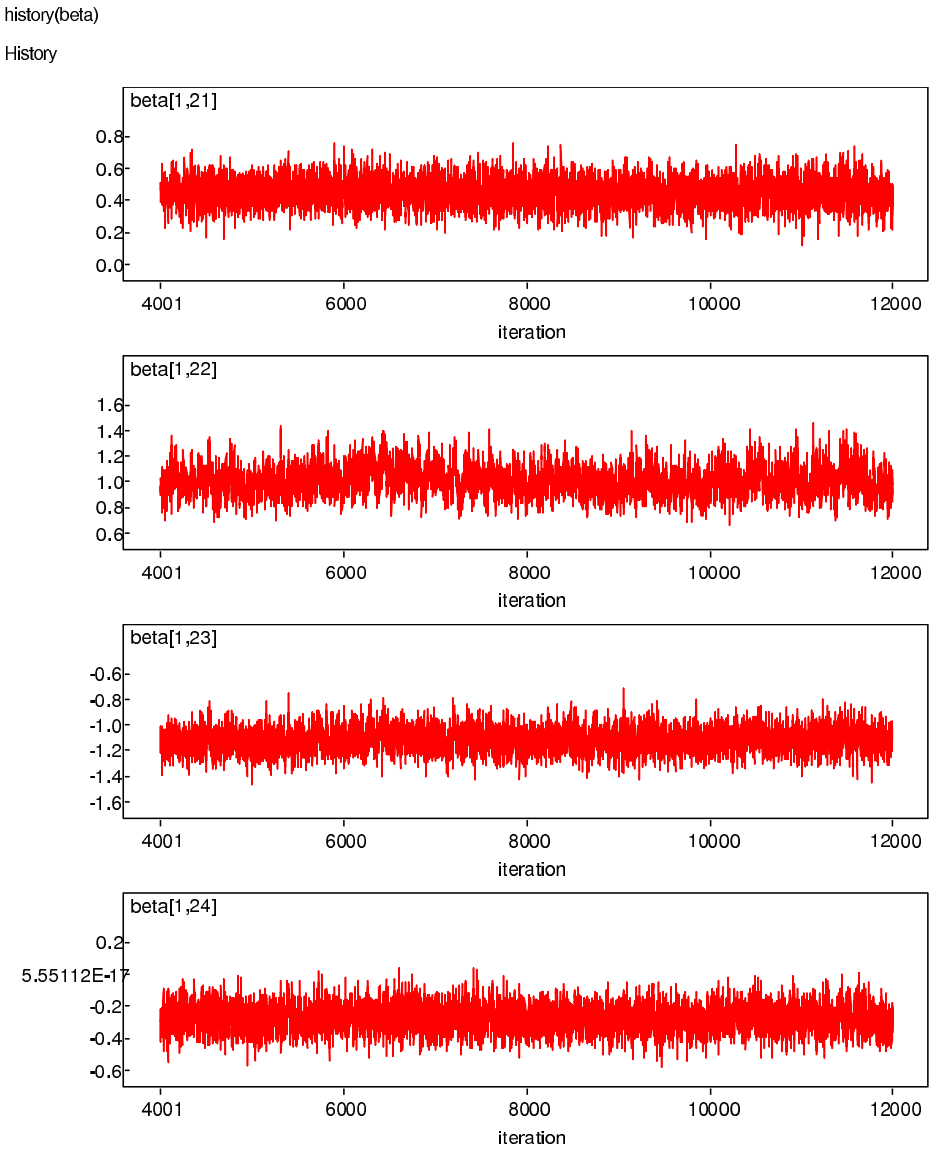
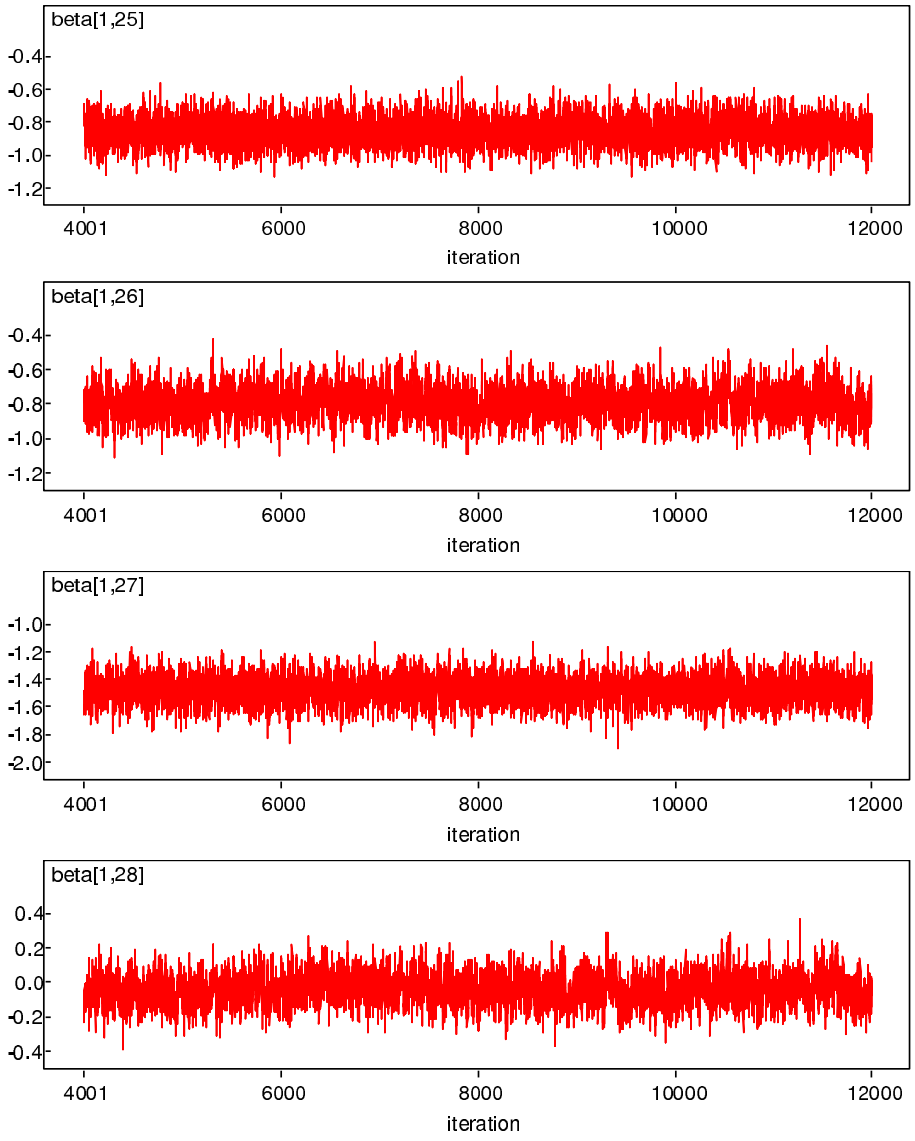
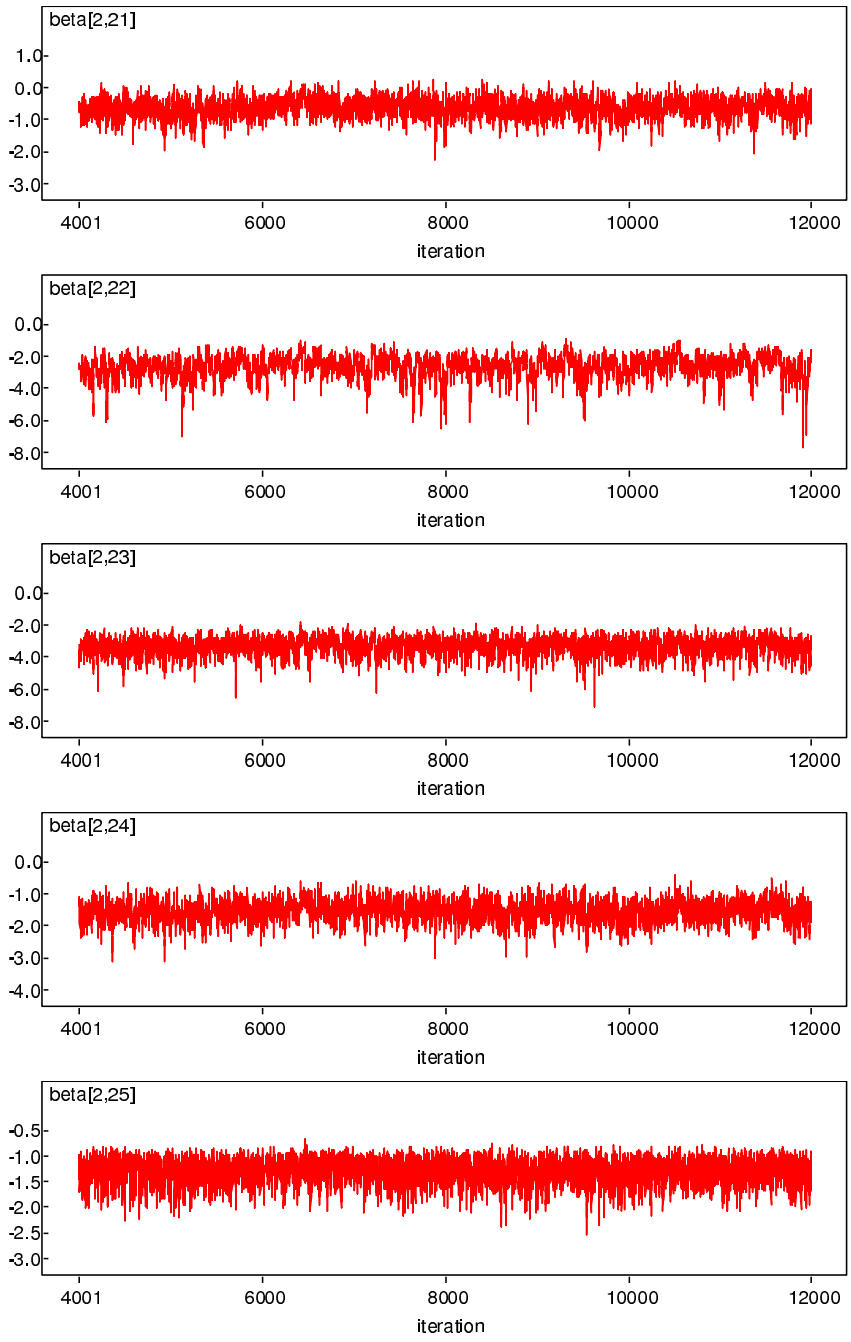

beta[2,21]
beta[2,22]
beta[2,23]
beta[2,24]
beta[2,25]
beta[2,26]
beta[2,27]
beta[2,28]
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

Figure C.4: The trace plot for beta for the condition of 1000 examinees with the proportion of 20% speededness




beta[1,25]
-0.4
-0.6
-0.8
-1.0
-1.2
4001
6000
8000
10000
12000
iteration
beta[1,26]
-0.4
-0.6
-0.8
-1.0
-1.2
4001
6000
8000
10000
12000
iteration
beta[1,27]
-1.0
-1.2
-1.4
-1.6
-1.8
-2.0
4001
6000
8000
10000
12000
iteration
beta[1,28]
0.4
0.2
0.0
-0.2
-0.4
4001
6000
8000
10000
12000
iteration


beta[2,21]
1.0
0.0
-1.0
-2.0
-3.0
4001
6000
8000
10000
12000
iteration
beta[2,22]
0.0
-2.0
-4.0
-6.0
-8.0
4001
6000
8000
10000
12000
iteration
beta[2,23]
0.0
-2.0
-4.0
-6.0
-8.0
4001
6000
8000
10000
12000
iteration
beta[2,24]
0.0
-1.0
-2.0
-3.0
-4.0
4001
6000
8000
10000
12000
iteration
beta[2,25]
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
4001
6000
8000
10000
12000
iteration

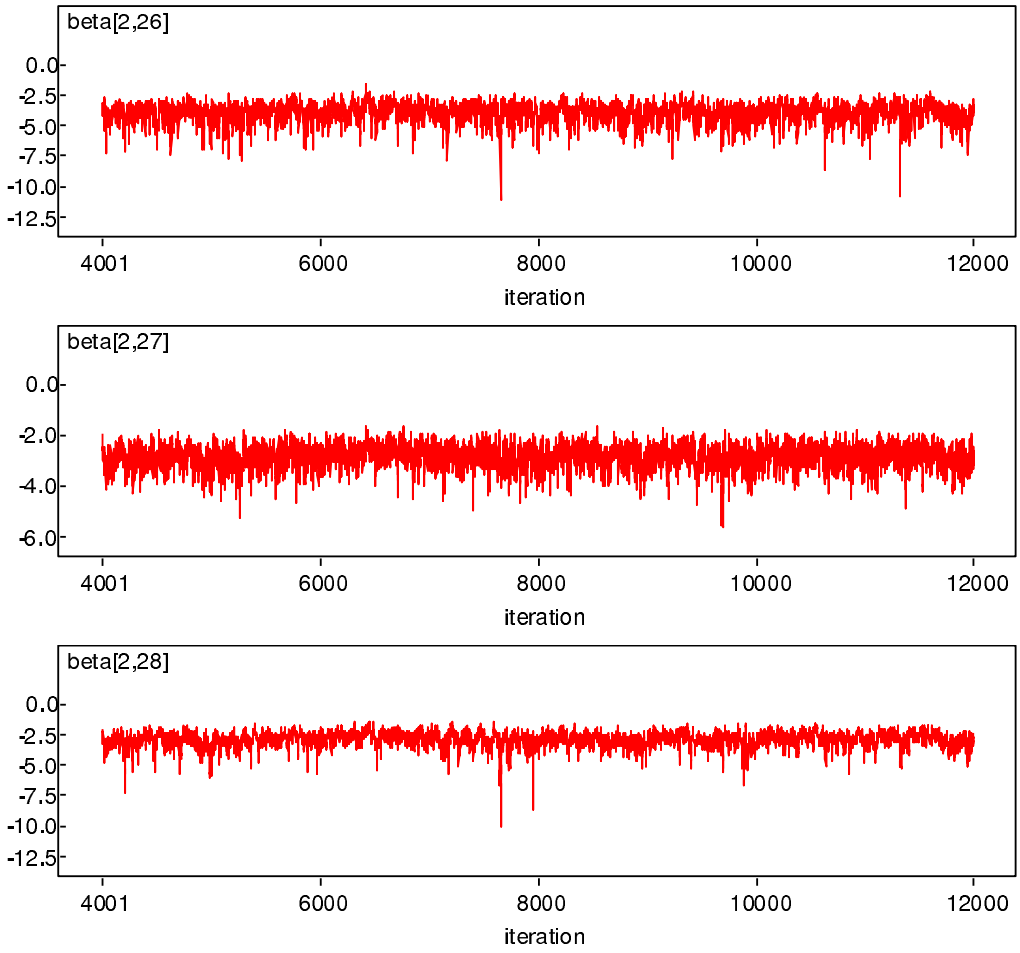


Figure C.5: The autocorrelation plot for mbeta for the condition of 1000 examinees with the proportion of 20% speededness

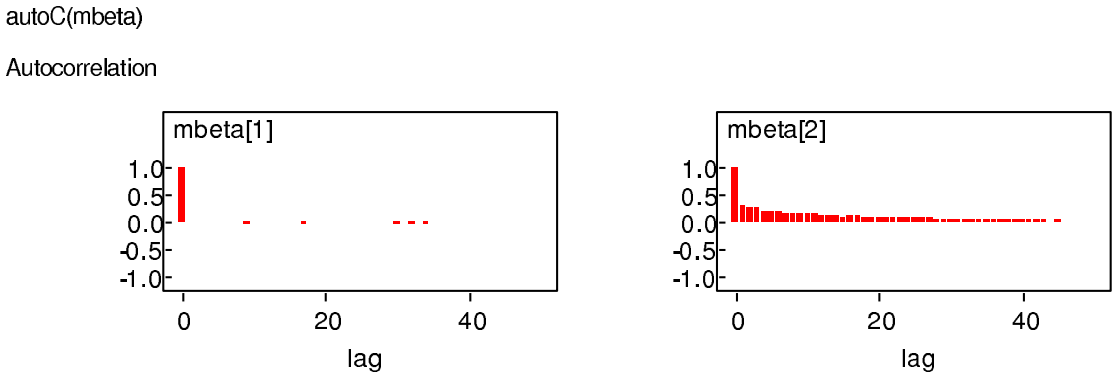


Figure C.6: The trace plot for mbeta for the condition of 1000 examinees with the proportion of 20% speededness

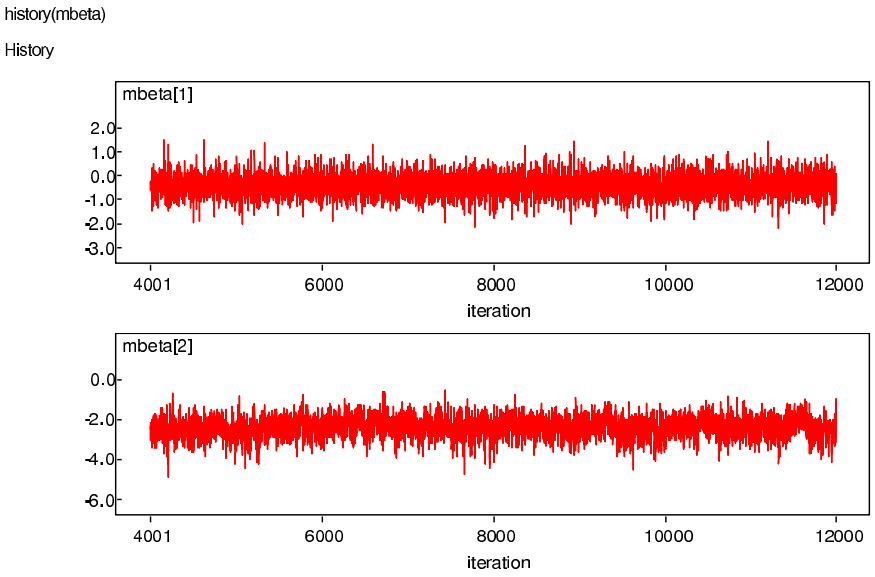


Figure C.7: The autocorrelation plot for the precision of random item and person effects for the condition of 1000 examinees with the proportion of 20% speededness

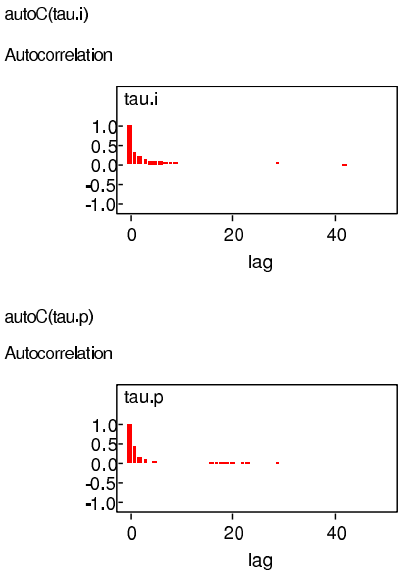


Figure C.8: The trace plot for the precision of random item and person effects for the condition of 1000 examinees with the proportion of 20% speededness

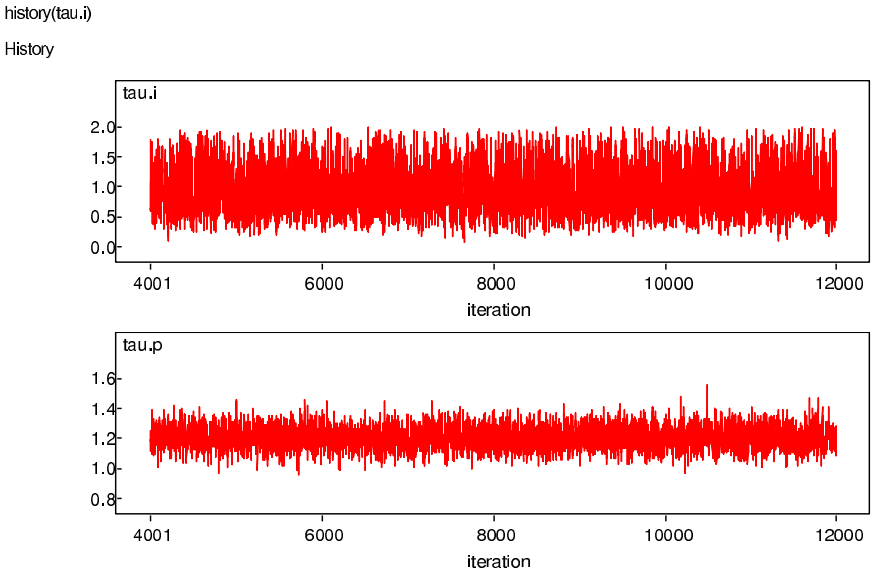


Figure C.9: The autocorrelation plot for mu for the condition of 1000 examinees with the proportion of 20% speededness

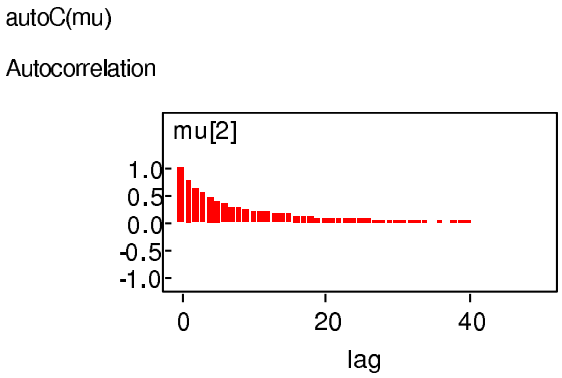
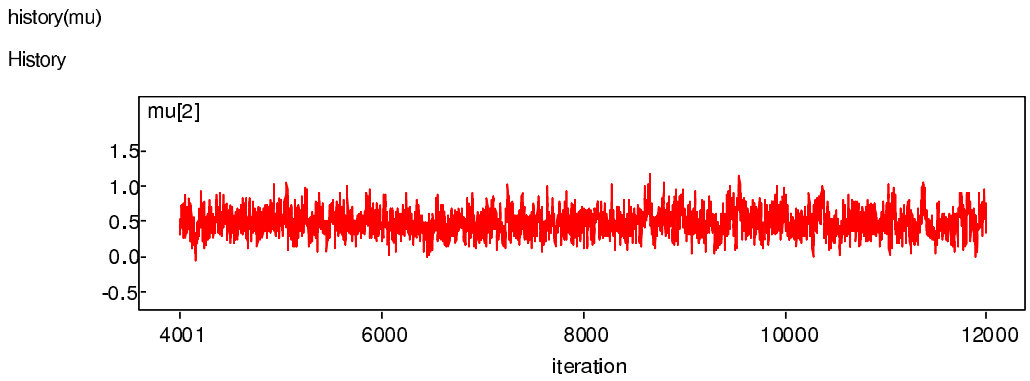


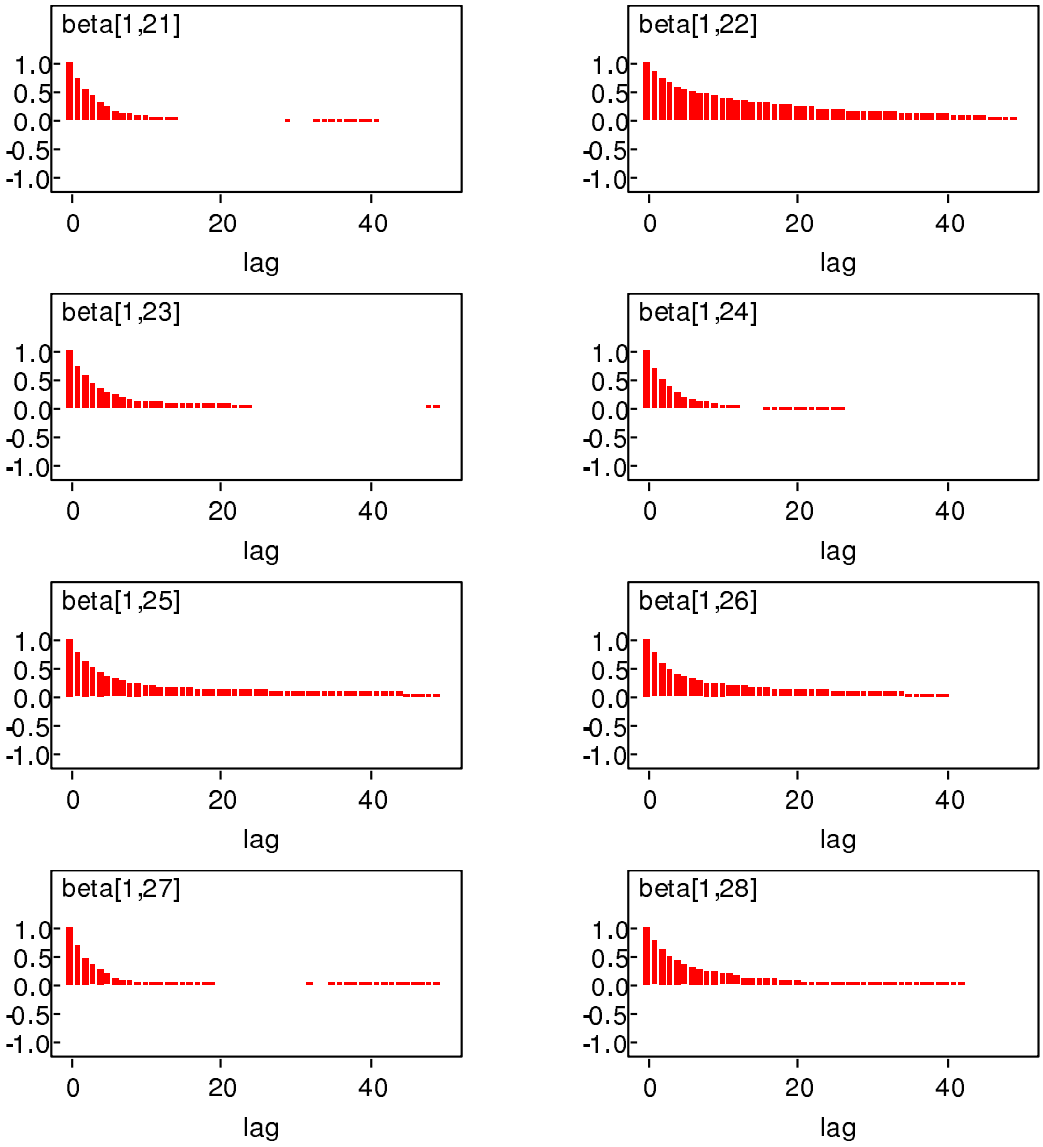
Figure C.10: The trace plot for mu for the condition of 1000 examinees with the proportion of 20% speededness



Appendix D

convergence figures for one selected condition under the conditional model

Figure D.1: The autocorrelation plot for betas for the condition of 1000 examinees with the proportion of 20% speededness



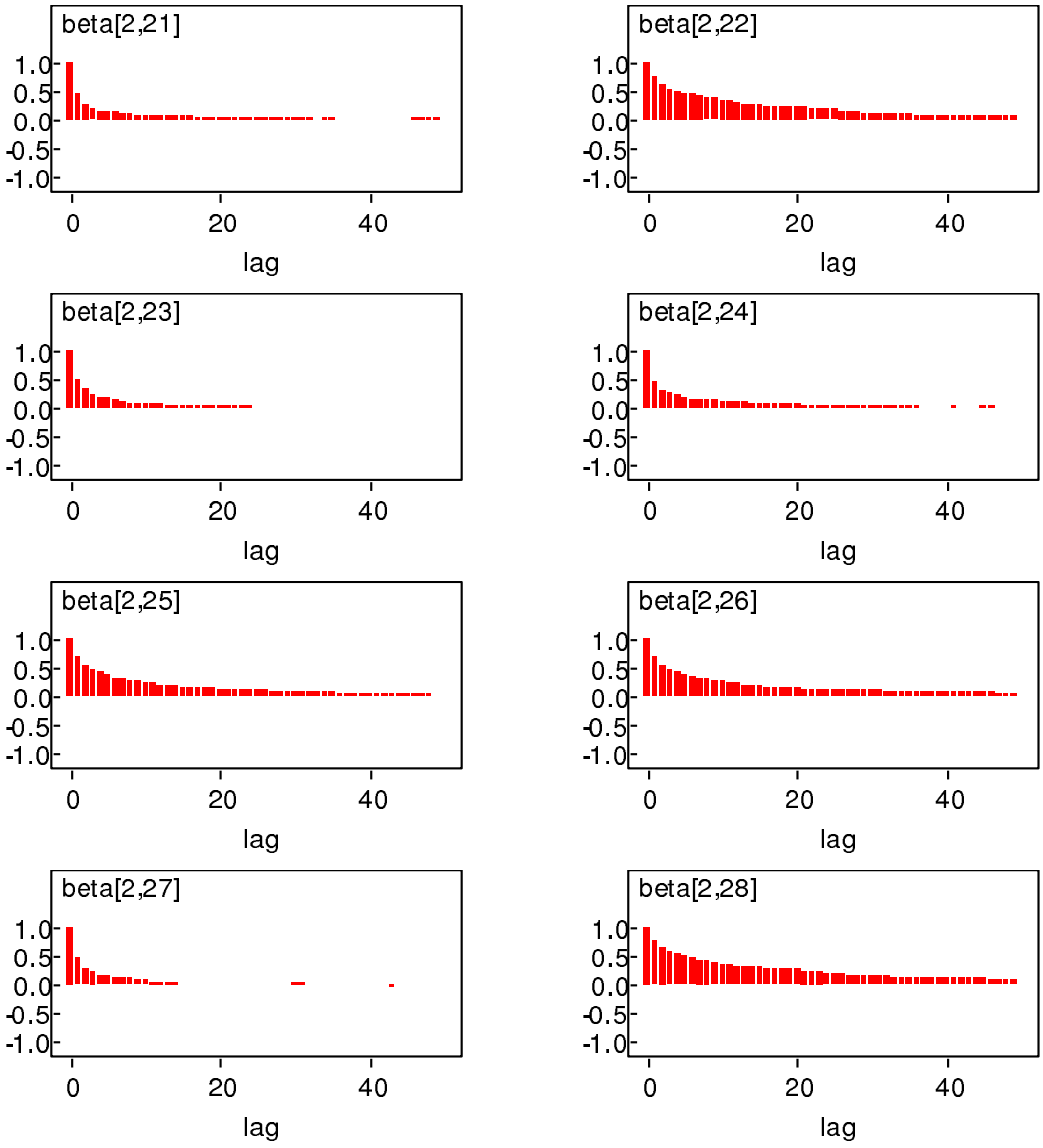
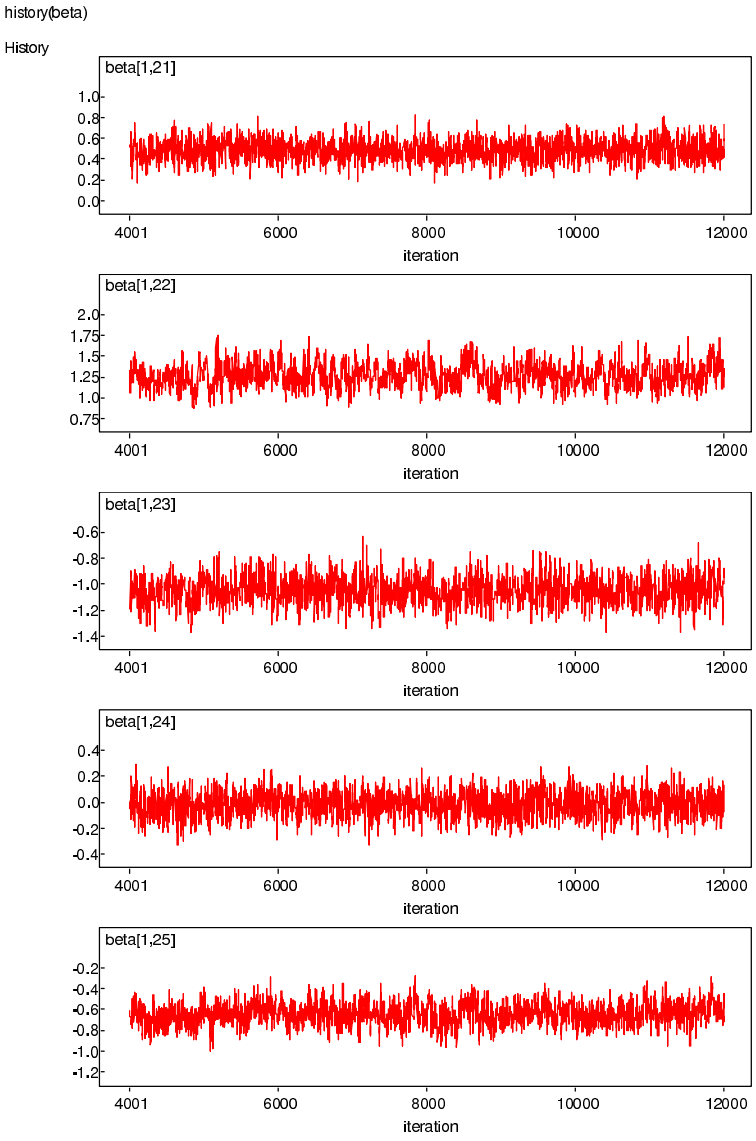
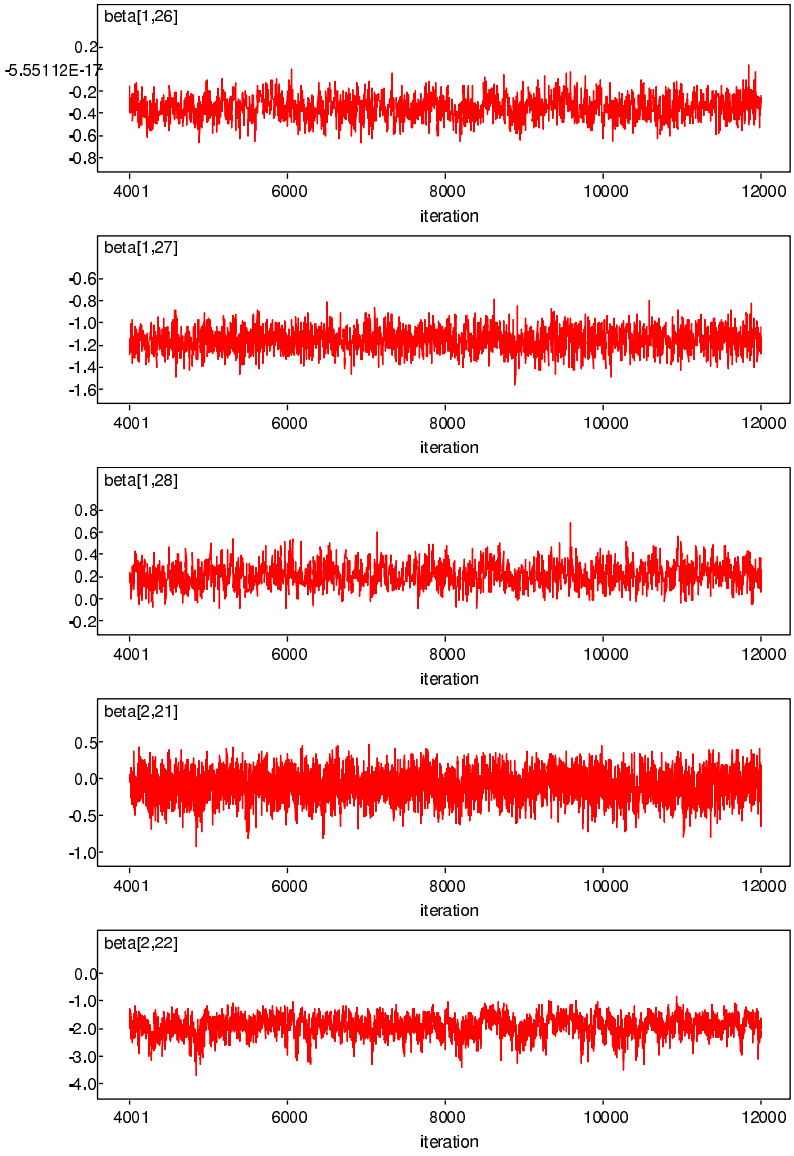

beta[2,21]
beta[2,22]
beta[2,23]
beta[2,24]
beta[2,25]
beta[2,26]
beta[2,27]
beta[2,28]
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

Figure D.2: The trace plot for betas for the condition of 1000 examinees with the proportion of 20% speededness




beta[1,26]
0.2
-5.55112E-17
-0.2
-0.4
-0.6
-0.8
4001
6000
8000
10000
12000
iteration
beta[1,27]
-0.6
-0.8
-1.0
-1.2
-1.4
-1.6
4001
6000
8000
10000
12000
iteration
beta[1,28]
0.8
0.6
0.4
0.2
0.0
-0.2
4001
6000
8000
10000
12000
iteration
beta[2,21]
0.5
0.0
-0.5
-1.0
4001
6000
8000
10000
12000
iteration
beta[2,22]
0.0
-1.0
-2.0
-3.0
-4.0
4001
6000
8000
10000
12000
iteration

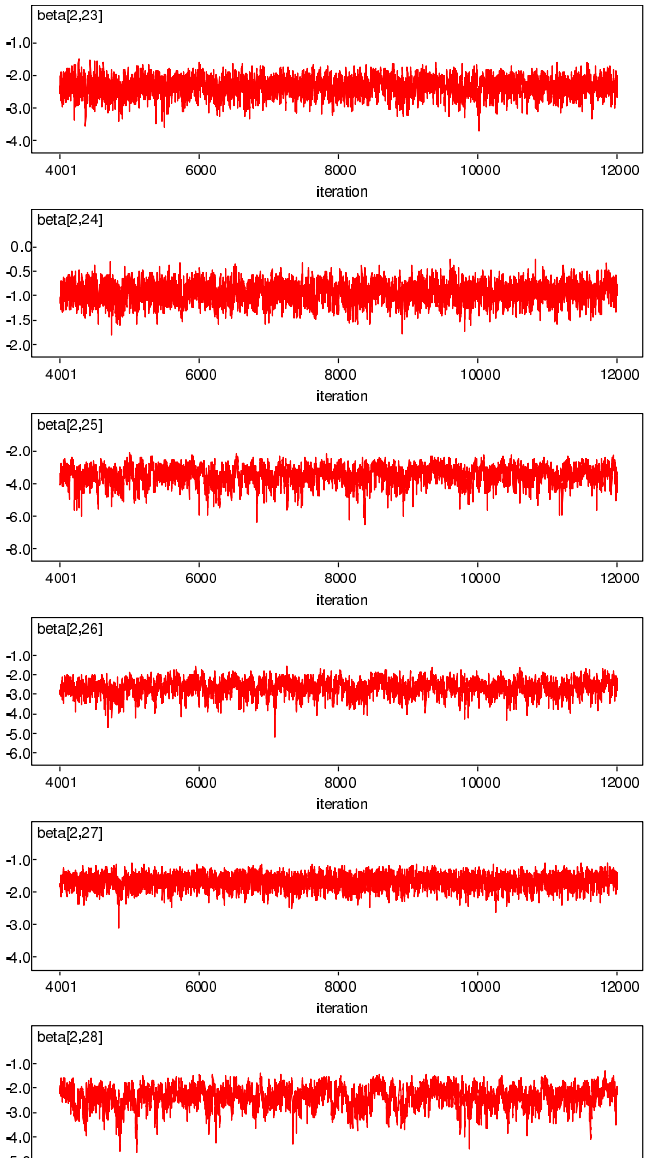

beta[2,23]
-1.0
-2.0
-3.0
-4.0
4001
6000
8000
10000
12000
iteration
beta[2,24]
0.0
-0.5
-1.0
-1.5
-2.0
4001
6000
8000
10000
12000
iteration
beta[2,25]
-2.0
-4.0
-6.0
-8.0
4001
6000
8000
10000
12000
iteration
beta[2,26]
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0
4001
6000
8000
10000
12000
iteration
beta[2,27]
-1.0
-2.0
-3.0
-4.0
4001
6000
8000
10000
12000
iteration
beta[2,28]
-1.0
-2.0
-3.0
-4.0

Figure D.3: The autocorrelation plot for mbeta for the condition of 1000 examinees with the proportion of 20% speededness

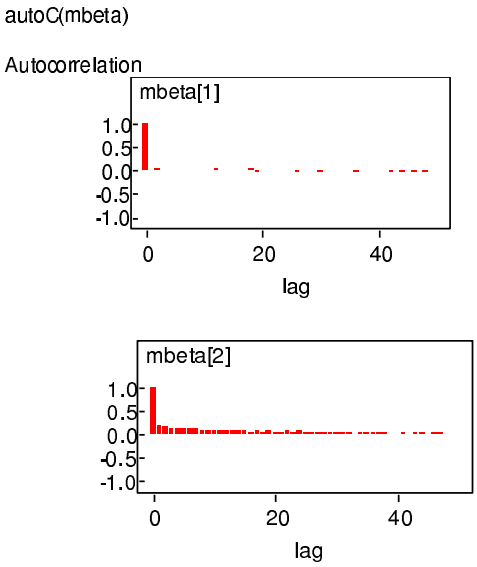


Figure D.4: The trace plot for mbeta for the condition of 1000 examinees with the proportion of 20% speededness

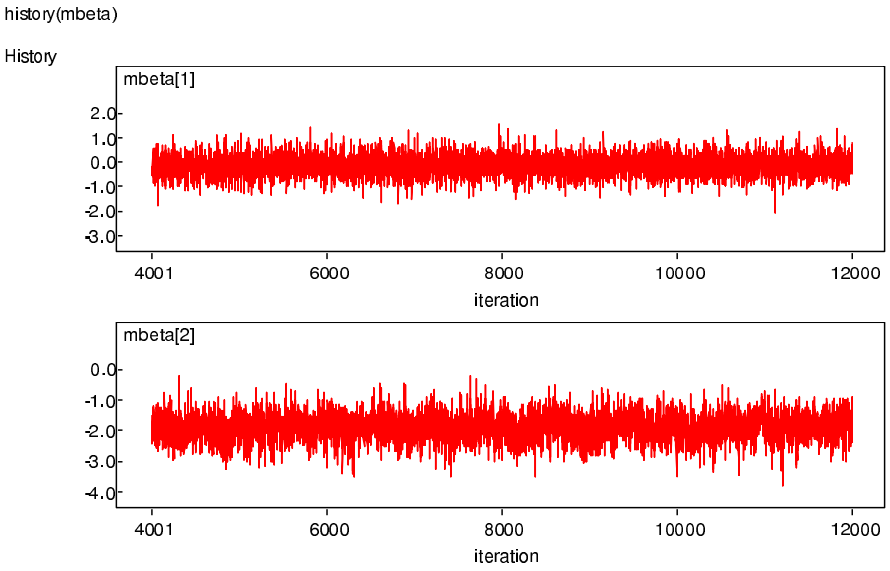


Figure D.5: The autocorrelation plot for tau.i and tau.p for the condition of 1000 examinees with the proportion of 20% speededness

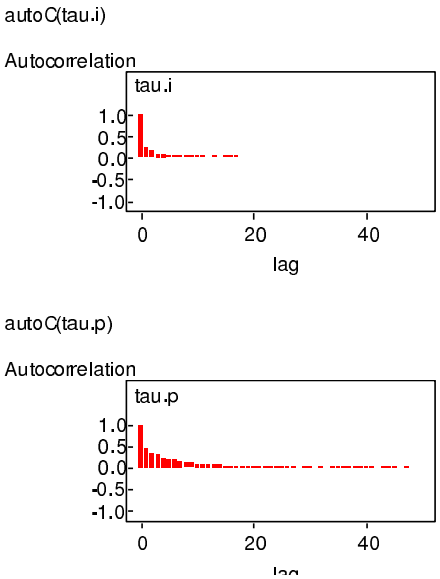


Figure D.6: The trace plot for tau.i and tau.p for the condition of 1000 examinees with the proportion of 20% speededness

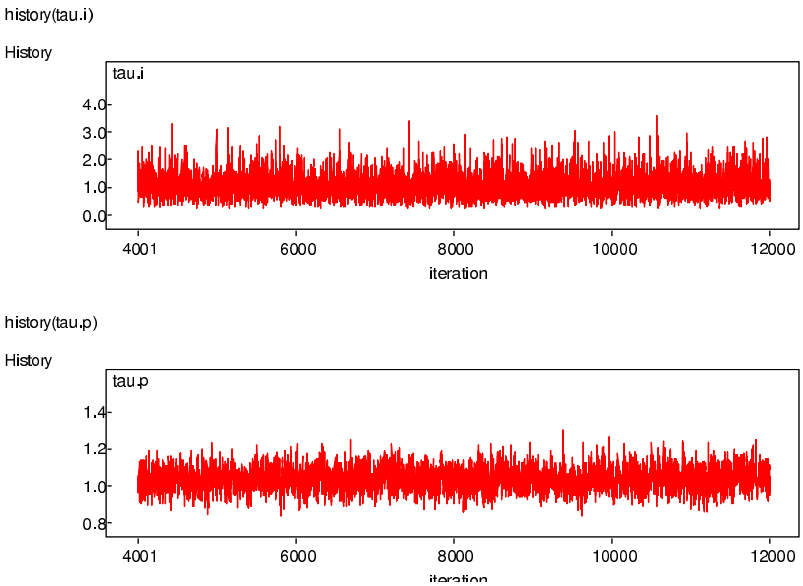


Figure D.7: The autocorrelation plot for mu for the condition of 1000 examinees with the proportion of 20% speededness

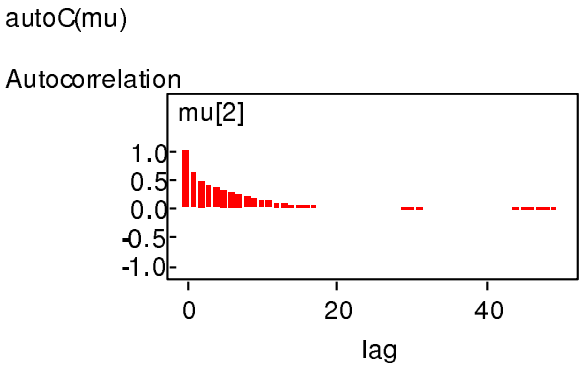


Figure D.8: The trace plot for mu for the condition of 1000 examinees with the proportion of 20% speededness

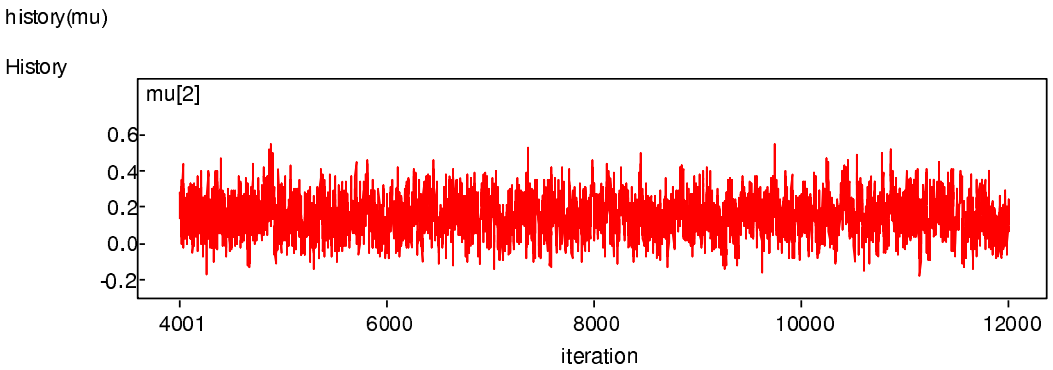


Figure D.9: The autocorrelation plot for gamma0 and gamma1 for the condition of 1000 examinees with the proportion of 20% speededness

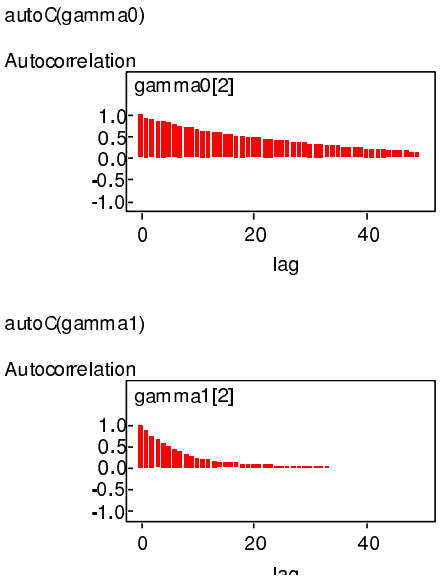


Figure D.10: The trace plot for gamma0 and gamma1 for the condition of 1000 examinees with the proportion of 20% speededness