Sensitivity of Prior Specification within Testlet Model

by

Jongmin Ra

(Under the direction of Allan S. Cohen and Seock-Ho Kim)

Abstract

Many IRT models have been developed to maintain the quality of items and estimate an individual's underlying latent ability, θ, more accurately. The conventional one-, two-, and three-parameter normal or logistic models assume local independence after controlling for an individual's ability, θ. Although this conventional assumption is straightforward, some studies have shown that it may not be accurate under some conditions as shown in testlets (Bradlow et al., 1999). Testlets composed of a set of items sharing common stimuli have been widely used in educational and psychological tests. With the demand for more accurate estimation of items and an individual's θ, the need for new estimation procedures has become obvious.

The purpose of this study is to examine the sensitivity of different prior distributions within the 3PLT model. First, the efficacy of the 3PLT model in the WinBUGS 1.4 program (Spiegelhalter, Thomas, Best, & Lunn, 2003) was compared to the 3PLT model in the SCORIGHT 3.0 (Wang, Bradlow, & Wainer, 2004) and the Gibbs (Du, 1998) programs, neither of which can manipulate pre-specified prior distributions. Later, the impacts of different prior distributions in the 3PLT model will be discussed.

Index words: Item response theory, Markov Chain Monte Carlo estimation, Prior specification, Testlet Model

Sensitivity of Prior Specification within Testlet Model

by

Jongmin Ra

B.A., Han Nam University, Daejeon, Korea, 2000

M.A., Southern Illinois University, Edwardsville, 2003

M.S., The University of Georgia, Athens, 2011

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

Sensitivity of Prior Specification within Testlet Model

by

Jongmin Ra

Approved:

Major Professors:	Allan S. Cohen Seock-Ho Kim
Committee:	Gary J. Lautenschlager Jonathan Templin

Electronic Version Approved:

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

Dedication

Mysterious Joviality

Mirage whirling within hearts
Inserts puffs of loving touch to souls
Now, in the hours of grey the moment

Jadegreen eyes whispering lovable smiles
Usher us into mosaic roads
Numbed hearts stuck in the dark
Glide into the rainbow

Seize the moment
Of missing, loving touch
Never forget the moments
Gracious tears flowed into sun,

and to whom I loved.

Acknowledgments

I would like to sincerely thank my advisors, Professor Seock-Ho, Kim and Allan S. Cohen for their guidance, understanding, patience, and supports for my doctoral study. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own research interests, and at the same time all kinds of supports. I am deeply grateful to them for the long discussions that helped me sort out technical details of my works.

Table of Contents

Page

Appendix

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 STATEMENT OF PROBLEM

There have been enormous statistical advances made in the analysis of standardized educational and psychological tests. Parallel with this, the practical advantages of the Bayesian approach were recognized in item response theory (IRT) and have been adopted to provide more detailed information about item parameters and an individual's underlying latent ability.

Recently, testlets comprising a set of items from a common stimulus (Rosenbaum, 1988; Wainer & Kiely, 1987; Wainer & Lewis, 1990) have emerged in educational tests as a remedy for multiple-choice items which are often criticized for decontextualization (Li, 2004). Once Bradlow, Wainer, and Wang (1999) suggested a two-parameter normal testlet model so as to include the testlet effect in the model, subsequent studies (Wainer, Bradlow, & Du, 2000; Wang, Bradlow, & Wainer, 2002) showed that testlet models effectively account for local dependence existing among items sharing the same stimulus and also yield accurate model parameter recovery.

However, issues of prior specification on testlet models have been neglected and need to be investigated, especially, under the three-parameter logistic testlet (3PLT) model. It is well known that prior distribution affects the rate of convergence when sample sizes are small. Furthermore, it is not appropriate to assume all item parameters follow the normal distribution (Wang & Wilson, 2005a, 2005b; Irvine & Kyllonen, 2002; Li, 2004).

Wang and Wilson's (2005a, 2005b) study showed that it is not always realistic and appropriate to assume item parameters are normally distributed under the one-parameter logistic testlet model. Other studies (e.g., Irvine & Kyllonen, 2002; Li, 2004) also revealed that the distribution of item parameters may not follow the normal distribution, even though items are sometimes randomly selected from an item bank. Although there is a great deal of research examining the substantive and statistical characteristics of prior distributions, there have been little research investigating the sensitivity of prior specification within the testlet models.

1.2 THE PURPOSE OF THE STUDY

Many IRT models have been developed to maintain the quality of items and estimate an individual's underlying latent ability, θ, more accurately. The conventional one-, two-, and three-parameter normal or logistic models assume local independence after controlling for an individual's ability, θ. Although this conventional assumption is straightforward, some studies have shown that it may not be accurate under some conditions as shown in testlets (Bradlow et al., 1999). Testlets composed of a set of items sharing common stimuli have been widely used in educational and psychological tests. With the demand for more accurate estimation of items and an individual's θ, the need for new estimation procedures has become obvious.

Recent developments in Markov chain Monte Carlo (MCMC) analyses facilitated the implementation of Bayesian analysis of complex data sets with testlets. Despite a large volume of research on estimation techniques, the effects of the characteristics of the data sets, and violations of model assumptions within testlet models, few studies are available on the sensitivity of prior distributions within testlet models.

The purpose of this study is to examine the sensitivity of different prior distributions within the 3PLT model. First, the efficacy of the 3PLT model in the WinBUGS 1.4 program (Spiegelhalter, Thomas, Best, & Lunn, 2003) was compared to the 3PLT model in the SCORIGHT 3.0 (Wang, Bradlow, & Wainer, 2004) and the Gibbs (Du, 1998) programs, neither of which can manipulate pre-specified prior distributions. Later, the impacts of different prior distributions in the 3PLT model will be discussed.

1.3 SIGNIFICANCE OF THE STUDY

Teslets have commonly been used in psychological and standardized educational tests such as the Graduate Record Examination (GRE) and the Test of English as a Foreign Language (TOEFL). Implementing testlets in conventional IRT models involve a variety of challenging measurement problems such as local dependence within testlets on ability and item parameter estimation, and test reliability. As the Bayesian method with MCMC for complicated IRT models is becoming increasingly common, (e.g., Albert, 1992; Béguin & Glas, 2001; Bradlow, et al. 1999; Patz & Junker, 1999a, 1999b), relevant research is needed to ensure that the most efficient, accurate, and flexible prior distributions are incorporated into testlet models.

Overview of later chapters

This study is organized as follows. Chapter 2 provides some theoretical background for this study. Previous studies related to testlet models and issues of local dependence and prior specification are reviewed. Chapter 3 outlines the specifications of components in testlet models used for this study, data generation methods, implementing prior distributions, research design and the evaluation criteria. Chapter 4 discusses the estimation of parameters with the WinBUGS 1.4 program, shows some simulation results, applies the proposed models and methods to real test data, and summarizes the results. Chapter 5 contains discussions of the results from the simulation study and the real data analysis and discusses limitations and possible future work.

Chapter 2

THEORETICAL BACKGROUND

This chapter serves to provide a general background and theoretical framework for this study. There are three sections in this chapter. Section I describes violations of local independence. Testlet models and some topics related to prior distributions are described in sections II and III.

2.1 LOCAL INDEPENDENCE

IRT models have been widely used in standardized educational tests to measure an individual's θ and psychometric properties of items (Loevinger, 1947; Lord & Novick, 1968). IRT models commonly assume local independence, in which an individuals' response to items are independent and based only on an individual's θ (Lord, 1980). When a set of items on a test are locally independent for given individuals, the probabilities of a response pattern on those items are equal to the product of probability associated with the individual's response to the individual items (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Roger, 1991).

Recently, testlets composed of a set of items sharing common stimuli have been widely used in standardized educational tests. (Li, 2004; Rosenbaum, 1988; Wainer & Kiely, 1987; Wainer & Lewis, 1990; Wainer & Wang, 2000). Testlets (e.g, reading passages, essays, mathematical reasoning, algebra tests, and analytical reasoning) have advantages of reducing impact of item ordering, of reducing time and cost, and securing test content and balancing content (Ariel, Veldkamp, & Breithaupt, 2006; Wainer, Kaplan, & Lewis, 1992; Wainer, Lewis, Kaplan, & Braswell, 1991).

As a consequence, the testlet approach is considered as a realistic method of measuring an individual's θ (Wainer, Sireci, & Thissen, 1991). However, testlet-based tests are likely to violate the local independence assumption (Wainer & Thissen, 1996). The assumption of the local independence in IRT has emerged as a crucial problem in testlet-based tests since items within testlets rely on a common stimulus (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989). The presence of local dependence (LD) is an indication that the items on tests do not measure individual's θ accurately (Ackerman, 1992).

Also, ignoring LD among items within testlets results in inflated estimates of score reliability and test information (Sireci et al., 1991; Thissen et al, 1989; Wainer & Lukhele, 1997; Wainer & Thissen, 1996; Wang & Wilson, 2005a, 2005b; Yen, 1993). Thissen et al. (1989) showed that lower validity correlation coefficient obtained when traditional IRT procedures applied for a testlet-based test.

Study (Wainer & Thissen, 1996) showed that the possible effects of ignoring the presence of LD on measurement is that test information function was overestimated and standard error of measurement (SEM) was underestimated. A study by Wainer and Wang (2000) also showed that standard IRT models assuming the local independence assumption result in an overstatement of precision of the θ estimates as well as a bias in item difficulty and discrimination parameter estimates when the assumption of the local independence was violated. More specifically, overestimated guessing parameters occurred for both reading comprehension and listening comprehension items when testlet-associated local item dependence was ignored for reading comprehension. Also, underestimated item discrimination parameters occurred for listening comprehension items while overestimated item discrimination parameters occurred. Furthermore, studies (Wainer & Lukhele, 1997; Sireci et al., 1991) showed that reliability was overestimated when LD was ignored within testlets.

2.2 TESTLET MODEL

A collection of items sharing common stimuli in which possibilities of correction within an individual's responses exist is called a testlet or an item bundle (Wainer & Kiely, 1987; Wainer & Lewis, 1990; Rosenbaum, 1988). Testlets are suggested as the unit of construction, and are commonly used for computerized adaptive tests (Wainer et al., 1992). Furthermore, testlets are regarded as more realistic and even better for measuring contextualized problem-solving skills that are difficult to develop in a single item (Bao, 2007; Wainer, Lewis, & Braswell, 1991).

Tests containing testlets can minimize content exposure (Ariel et al., 2006), reduce time and cost (Bradlow et al., 1999; Wainer & Wang, 2000), and increase construct validity (Zenisky, Hambleton, & Sireci, 2002). Despite the advantages of testlets, LD within the same testlets is likely to be introduced when testlets are included in tests (Rosebaum, 1988; Sireci et al., 1991; Thissen et al., 1989). Thus, covariances among the items in testlets often are not solely explained by the traits of interest. When ignored, this additional within-testlet covariation results in overestimates of the true reliability.

In order to avoid LD problems, researchers have paid close attention to LD within testlets (Bradlow et al., 1999; Du, 1998; Lee, Kolen, Frisbie, & Ankernmann, 2001; Thissen et al., 1989). One approach for dealing with LD is to calculate a single score over all items in testlets and then fit polytomous models [e.g., Samejima's (1969) graded response model, Bock's (1972) nominal response model, and Muraki's (1992) generalized partial credit model] to testlets (Lee & Frisbie, 1999; Lee et al., 2001; Sireci et al., 1991; Thissen et al., 1989; Wainer & Thissen, 1996). When a set of items within testlets is treated as units of analysis, the score for each testlet would be computed as the sum of the correct answer to items nested in that testlet.

Studies (Sireci et al., 1991; Thissen et al., 1989; Zenisky et al., 2002) suggested that the problem of LD can be effectively avoided if a set of items within testlets is treated as the units of analysis, assuming local independence across testlets, but not within testlets.

Studies (Thissen et al., 1989; Wainer & Wang , 2000; Wang & Wilson, 2005a, 2005b; Yen, 1993) showed that fitting polytomous IRT models to testlets provides limited information because information about item-level discrimination and response pattern is lost.

A second limitation to this approach is related to item selection in computer adaptive testing (Wainer & Wang, 2000). If only the summed scores for testlets are used for parameter estimation, then an individual's responses to items within a testlet would not be able to provide any information about the levels of an individual's θ until he or she has responded to all of the items in that testlet. This implies that an individual's responses to the initial items in a testlet could not be used in selecting the subsequent items in that testlet. This could present practical difficulties in the development of computer adaptive testing.

The other approach is to consider LD in testlets as an additional random effect in the model. Bradlow et al. (1999) first suggested the two-parameter normal ogive model for a mixture of binary independent and testlet items and demonstrated its accuracy and effectiveness via 2 × 3 factorial simulation study. The number of examinees ($N = 1000$), test length ($n = 60$), and percentage of items nested within testlet (50%) were held constant across study conditions. They considered the testlet effect as a random effect in addition to the latent ability of interest to be measured by the test. The variance of testlet effects were assumed to be constant across different testlets. A random testlet effect can be explained as the interaction between individuals and testlets. Once an individual's θ and random testlet effect are controlled, an individual's responses are independent.

In addition, Wang and Wilson (2005a, 2005b) also proposed the Rasch testlet model for dichotomous responses. It is a variation of the two-parameter testlet model if the item discrimination power is kept to be constant for all items and all items are assumed to be scored dichotomously. Later, this model was extended to the more general testlet models (Du, 1998; Wainer, Bradlow, & Du, 2000; Wang, Bradlow, & Wainer, 2004). Du (1998) and Wainer et al. (2000) extended their previous model (Bradlow et al., 1999) by including the guessing parameter, c_i, to the two-parameter probit model and by allowing random variation across different testlets.

The three-parameter logistic testlet (3PLT: Du, 1998; Wainer et al., 2000) was further extended to include dichotomously and polytomously scored items (Wang et al., 2004). The study (Wang et al., 2004) was composed of a simulation study and two applications using operational data from the Test of Spoken English and the North Carolina Test of Computer Skills. The simulation component of the study examined the success of the model in recovering the true parameters. Three factors were manipulated: Number of categories for each item (2, 5, 10), testlet length (3, 6, 9), and testlet variance ($\sigma^2_\gamma = 0, 0.5, 1$). Response data for 1000 simulees were simulated for a 30-item test across five replications for each condition. Of the 30 items, 12 were independent dichotomous items, and 18 were testlet items. The 3PLT model (Du, 1998; Wainer et al., 2000) is as follows :

$$P(y_{ij} = 1|\theta_j, a_i, b_i, c_i, \gamma_{jt(i)}) = c_i + (1 - c_i)\left[\frac{\exp(a_i(\theta_j - b_i - \gamma_{jt(i)}))}{1 + \exp(a_i(\theta_j - b_i - \gamma_{jt(i)}))}\right]$$

In this model, $P(y_{ij} = 1)$ is the probability that an individual j, answers item i correctly; θ_j is the ability of an individual j; b_i is the difficulty parameter of item i, a_i denotes the discrimination parameter of item i, c_i denotes the guessing parameter of item i; and c_i in testlet model is reparameterized as $\frac{\exp(q_i)}{1+\exp(q_i)}$, which becomes $q_i = \log(\frac{c_i}{1-c_i})$. $\gamma_{jt(i)}$ is a random effect which represents the interaction of individual j with a testlet t_i.

$\gamma_{jt(i)}$ is constant within a testlet for individual j , but the value of $\gamma_{jt(i)}$ differs for each individual. The variances of γ are allowed to vary across testlets and indicate the amount of LD in each testlet. If the variance of $\gamma_{jt(i)}$ is zero, items within the testlet can be considered conditionally independent. The larger the variance, $\gamma_{jt(i)}$, the greater the proportion of total variance in the test score that is attributable to the testlet.

Advantages of testlet models are that they are flexible because an individual's response patterns can be considered while keeping the same traditional scoring rubric systems and the same concept of item parameters (Wang & Wilson, 2005a, 2005b). Thus, information contained in the response patterns for individual items within testlets is not lost as it is with polytmous models. Testlet models are embedded in a Bayesian hierarchical framework and inferences use MCMC techniques.

2.3 PRIOR

2.3.1 PRIORS IN ONE-, TWO-, AND THREE-PARAMETER IRT MODELS

It is well known that incorporating prior distributions into the Bayesian framework yields more precise item parameters by preventing parameters from drafting out of reasonable ranges (Baker & Kim, 2004; Lord, 1980). The Bayesian methods with the MCMC algorithm make it possible to build more complex IRT models because estimation of models is comparatively easier with MCMC than with either joint maximum likelihood estimation and marginal maximum likelihood estimation (De Ayala, 2009).

The Bayesian method with MCMC has been increasingly used for complicated IRT models (Albert, 1992; Bradlow et al., 1999; Patz & Junker, 1999a, 1999b; Wainer et al., 2000). Albert (1992) used a full Bayesian method based on Gibbs sampling to estimate the two-parameter normal ogive IRT model, and later Patz and Junker (1999a, 1999b) discussed Metropolis-Hastings sampling algorithms to estimate two-, and three-parameter logistic models and mixed models.

Studies (Bazán, Branoco, & Bolfarine, 2006; Swaminathan & Gifford, 1982, 1985, 1986; Swaminathan, Hambletion, Sireci, Xing, & Rivazi, 2003; Mislevy, 1986) showed that estimation of item parameters can be accurately made, and estimation can be carried out with smaller sample sizes by incorporating prior distributions.

All the prior distributions regarding parameters of interest need to be incorporated into the model parameters, but prior distributions in many cases are either vague or non-existent in the Bayesian approach (De Finetti, 1974; Gao & Chen, 2005; Wainer, Bradlow, & Wang, 2007). This makes it very difficult to specify a unique prior distribution. Thus, the specification of prior distributions in the Bayesian approach has emerged as an important issue (Bazán et al, 2006).

Studies (Albert & Ghosh, 2000; Swaminathan et al., 2003) have explained how to use informative prior distributions in IRT models. If appropriate a priori information about parameters was available, tight prior distributions have substantial effects on estimates (Harwell & Baker, 1991). Tight prior distributions implied small variance and led parameters to shrinkage toward the mean of the prior (Baker & Kim, 2004).

Different prior distributions for item and individual parameters have been used in the same IRT models. In previous studies, there seems to be consensus with respect to the prior distribution for θ. It is commonly assumed that ability follows the standard normal distribution, $\theta \sim N(0, 1)$. Fixing the location and scale parameters of ability distribution ensures identifiability of the curve parameters.

Regarding item parameters, different prior distributions have been investigated for item parameters (Rupp, Dey, & Zumbo, 2004). Studies (Patz & Junker, 1999a) revealed that it is difficult to assign dependent priors for those parameters, even if a multivariate normal prior distribution is specified. Thus, independent prior distributions for the parameters of item discrimination a, item difficulty b, and pseudo-guessing parameter, c, are preferred (Bazán et al, 2006).

Either informative and noninformative prior distributions on item discrimination parameter, a has been used. The reason to use informative prior distributions on a, is that a correct answer in testing always implies a higher ability. Thus, a is constrained to be grater than 0. Informative prior distributions on a was also implemented because the existence of the joint posterior distribution is not guaranteed when an improper prior is used (Bazán et al, 2006; Ghosh, Ghosh, Chen, & Agresti, 2000). Several studies have been done using informative prior distributions. Johnson and Albert (1999) specified the normal distribution for μ_a and σ_a^2, $N(\mu_a, \sigma_a^2)$, with or without hyper-parameters. Studies (Kim, Cohen, Baker, Subkoviak, &

Leonard, 1999; Patz & Junker, 1999a, 1999b; Sahu, 2002) have used the log-normal distribution for μ_a and σ_a^2, $LN(\mu_a, \sigma_a^2)$, with or without hyper-parameter distributions. In addition, other studies (Spiegelhalter, Thomas, Best, & Gilks, 1996; Sahu, 2002) used the half-normal distribution for μ_a and σ_a^2 with a constraint $a > 0$, $N(\mu_a, \sigma_a^2)I(0,)$.

Additional studies (Swaminathan & Gifford, 1985, 1986) have used a χ_v with v degree of freedom for the item discrimination parameter, a. Bafumi, Gelman, Park, and Kaplan (2005) has used a gamma and an inverted gamma distribution, with parameter m and n, $IG(m,n,)$ for a. Other studies (Albert, 1992; Fox & Glass, 2001, 2003) besides the ones mentioned above have been done using improper noninformative prior distributions for the parameters a and b.

Regarding b, it is common to assign the normal distribution for $b \sim N(\mu_b, \sigma_b^2)$. Moreover, when little prior information is available about b, relatively large values are assigned to σ_b^2. Studies (Patz & Junker, 1999a, 1999b; Swaminathan & Gifford, 1982, 1985, 1986) have used uniform distribution for μ_b and σ_b^2, $N(\mu_b, \sigma_b^2)$, in which μ_b follows uniform distribution and σ_b^2 follows inverse chi-square distribution.

Regarding c, studies (e.g., Patz & Junker, 1999a, 1999b; Swaminathan & Gifford, 1986) have specified the $Beta(s_i, t_i)$ distribution mentioned by Novick and Jackson (1974), where $s_i = m * M$ and $t_i = m(1 - M) - 2$ (m = the number of observations the prior information is worth and M = mean value). Another study (Mislevy, 1986) employed a normal distribution on transformed c, that is $q = \log(\frac{c}{1-c})$.

2.3.2 PRIORS IN TESTLET MODELS

Bradlow et al. (1999) specified noninformative prior distributions for the unknown means and variances. Thus, the distribution of the parameters of interest can be determined by the data. Therefore, with the noninformative prior distributions, the MCMC via Gibbs sampling is drawn from the posterior distribution to make inference about parameters of interest. Within testlet models, it is common to use a normal distribution, $N(0, 1)$, for parameters of

θ (Bradlow et al., 1999; Du, 1998; Li, Bolt, & Fu, 2006; Wainer et al., 2000; Wang et al., 2004).

In addition, the mean of the testlet parameters for a particular testlet across all individuals is usually set to 0 in order to identify the scale of the parameters. Thus, a normal distribution for $\gamma_{jt(i)}$ is commonly used, $\gamma_{jt(i)} \sim N(0, \sigma_\gamma^2)$. Bradlow et al. (1999) specified the normal distribution for μ_a and σ_a^2, $N(\mu_a, \sigma_a^2)$ with hyper-parameters for a. Li (2004) used the half normal priors, $N(\mu_a, \sigma_a^2)I(0,)$ for a. Regarding b, several studies (Bradlow et al., 1999; Du, 1998; Li et al., 2006; Wang et al., 2004) implemented a normal prior $N(\mu_b, \sigma_b^2)$ for b.

In terms of hyper-prior distributions, all noninformative hyper-priors are $\mu_a \sim N(0, \sigma_{\mu_a}^2)$, $\mu_b \sim N(0, \sigma_{\mu_b}^2)$, and $\mu_q \sim N(0, \sigma_{\mu_q}^2)$ for prior means, and $\sigma_a^2 \sim \chi_{g_a}^{-2}$, $\sigma_b^2 \sim \chi_{g_b}^{-2}$, and $\sigma_q^2 \sim \chi_{g_q}^{-2}$ for prior variances, where $\chi_{g_a}^{-2}$, $\chi_{g_b}^{-2}$, and $\chi_{g_q}^{-2}$ are inverse chi-square random variables with g_a, g_b, and g_q degrees of freedom which are defined as 0.5 to reflect a small amount of information. Either $\sigma_{\mu_a}^2$, $\sigma_{\mu_b}^2$, and $\sigma_{\mu_q}^2 = 100^2$ (Li et al., 2006; Wang et al., 2002) or $\sigma_{\mu_a}^2$, $\sigma_{\mu_b}^2$, and $\sigma_{\mu_q}^2 = 0$ (Du, 1998; Bradlow et al., 1999; Wang et al., 2002, 2004) was used to indicate a lack of information.

As shown in previous studies, choosing a prior distribution of the parameters of a model is a tedious task (Carlin & Louis 2000) because there is possibility of not reflecting uncertainty about the parameters of interest. Second, the posterior distribution is available but not derivable in closed form in the non-conjugate priors, in general. Last, it is difficult to describe uncertainty about the parameter of interest in the form of a particular distribution. In particular, uncertainty about the parameters of the prior distribution requires more informative model such as in empirical Bayesian methods. Thus, it is rare for anyone to make any claims that a particular prior can logically be defended as truly noninformative. Instead, the focus is on investigating various prior distributions and comparing them to see if any have advantages in some practical senses (Kass & Wasserman, 1996).

2.3.3 CONVERGENCE AND BURN-IN PERIOD

It is necessary to confirm whether convergence is reached because a non-converged MCMC algorithm may lead to incorrect information about estimates (Shinharay, 2004). If the chain does not converge, the simulated draws from this chain would not represent the posterior distribution of parameters of interest. Thus, the inference about parameters based on the distribution of these draws would be invalid. Therefore, it is very important to assess convergence of Markov chains before any Bayesian inferences are made.

A number of convergence diagnostics have been developed (Cowles & Carlin, 1996; Brooks & Robert, 1998). The most popular diagnostics are time-series plots, autocorrelation plots, and the Gelman-Rubin statistic, R. A time-series plot, also called a “history plot”, is a scatter plot showing the generated values of a parameter at each iteration number in a chain of sample values. Clear trends in the plot indicate that successive simulated values of parameters are highly correlated and a chain has not converged. Time-series plots provide a simple way to check the stability of simulated parameter values.

An autocorrelation plot is a plot of the correlation between sequential draws of a parameter in Markov chain. It is a commonly-used tool for checking randomness in a data set. This randomness is ascertained by computing autocorrelations for data values at varying time lags. Autocorrelation plots are not strictly a convergence diagnostic tool, but they help indirectly to assess convergence. A MCMC algorithm generating highly correlated parameter values will need a large number of iterations to converge to the appropriate posterior distribution. In other words, such autocorrelation can cause inefficient MCMC simulation. Solution to high autocorrelation is to “thin” the chains by keeping every k_{th} simulation draw from each sequence and discarding the rest.

The R statistic (Gelman & Rubin, 1992) suggests monitoring convergence based on multiple chains with different starting points. Once convergence is reached, between-chain variance and within-chain variance for each parameter should be almost equivalent because variation within the chain and variation between the chains should coincide. There, R near

1 for all parameters of interest means the MCMC algorithm has converged. However, one drawback it has is that its value depends on the choice of starting value. However, it is not straightforward in determining the convergence of algorithm with a single definitive convergence diagnostic tool. Therefore, using multiple tools is recommended in order to increase the chance of correctly assessing convergence (Sinharay, 2004).

In the context of testlet situation, previous researchers employed different ways of checking convergence of algorithm and the appropriate length of the burn-in period. However, it is possible to categorize various convergence algorithm into two possible classes. One is to import other computer software [e.g, Bayesian Output Analysis (BOA:Smith, 2001) program, Convergence Diagnosis and Output Analysis Software for Gibbs sampling output (CODA:Best, Cowles, & Vines, 1995), and SCORIGHT]. For instance, Sinharay(2004) implemented BOA or CODA program on the output of SCORIGHT (Wang et al., 2004). He also used the Gelman-Rubin convergence statistic (Gelman & Rubin, 1992) to determine the number of burn-in period. The other approach is to rely on the outputs of the Win-Bugs program. Bao's study (2007) is one of many examples. Bao (2007) mainly used the information available in the WinBUGS program. Those information are history plots (trace plots) showing random sampling within the same part of the same space for all chains, Brooks-Gelman-Rubin (BGR) showing the convergence of both the pooled and within interval widths to stability, and auto-correlation function showing where the autocorrelation has decreased to zero. Density plots are examined to investigate whether enough iterations have been completed. If enough iterations are run, the error due to the nature of MCMC being an empirical approximation to the posterior is less than 5% of the estimated posterior standard deviation (Spiegelhalter, Thomas, Best., & Lunn, 2003).

As expected, researche in testlet models selected different number of iterations and burn-in period (Gelman, Carlin, Stern, & Rubin, 2003; Raftery & Lewis, 1996; Sinharay, 2004; Sinharay, Johnson, & Stern, 2006; Sinharay & Stern, 2002). Information for previous research in the testlet model is provided in Table 2.1. Gelman et al. (2003) suggested discarding the first half of the iterations to be conservative. Raftery and Lewis (1996) suggested there were fewer than 500 burn-in periods for convergence diagnostic, and the recommended lengths of chain were typically less than 15,000 iterations. Sinharay, Johnson, and Stern (2006) suggested five chain of 6,000 iterations after discarding 2,000 iterations as burn-in periods and drew every 20th for one-, two-, and three-parameter logistic models. Sinharay (2004) also recommended several chains of 50,000 iterations with 1,000 burn-in periods or one longer chain having 120,000 iterations with 20,000 burn-in periods for convergence for the testlet model.

Table 2.1: Summary of Previous Studies

Studies	N	ITEMS	ITERATION	REPLICATION	MODEL
Bao	5,000	50 items	4,000	10	2PLT
(2007)		30, (10 items for 2 testlets)	(1,500, brun-in)		
Baldwin	2,000	50 items	30,000	50	3PLT
(2008)		(No specific information)	10,000 (burn-in)		
		60 items			
Bradlow et al.	1,000	30, (5 items for 6 testlets)	10,000	No	2PNO
(1999)		30, (6 items for 5 testlets)	(5,000, burn-in)		
		30, (10 items for 3 testlets)			
Du	1,000	70 items	12,000	No	3PLT
(1998)	N	30, (10 items for 4 testlets)	(7,000, burn-in)		
	2,000	24 items			
Li et al.		(5 items for 1 testlet)	15,000	No	2PNO
(2006)		(6 items for 2 testlets)	(1,000, burn-in)		
		(7 items for 1 testlet)			
	1612	60 items	50,000		
Sinharay		35, (5 items for 3 testlets)	(10,000, burn-in)	No	3PLT
(2004)		(4 items for 1 testlet)	120,000		
		(6 items for 1 testlet)	(20,000, burn-in)		
Wang & Wilson	2,000	20 items	15,000	100	1PLT
(2005)		(5 items for 4 testlets)	(1,000, burn-in)		
		30 items			
Wang et al.	1,000	12, (3 items for 6 testlets)	3,000	5	3PLT
(2002)		12, (6 items for 3 testlets)	(2,000, burn-in)		
		12, (9 items for 2 testlets)			

Note:1PLT (one-parameter logistic testlet model); 2PLT (two-parameter logistic testlet model); 3PLT (three-parameter logistic testlet model); 2PNO (two-parameter normal testlet model).

Chapter 3

METHODS

3.1 Computer Programs

The WinBUGS 1.4 program is the main computer program in this study. Note that both the SCORIGHT 3.0 (Wang et al., 2004) and the Gibbs (Du, 1998) computer programs were used for comparison purposes in the analysis of real data. Both computer programs [Gibbs (Du, 1998) and SCORIGHT (Wang et al., 2004)] use MCMC to fit the 3PLT model and allow users options for choosing the number of chains and iterations of MCMC. Differences existing between the two programs, however, concern availability of possible models and options for choosing the number of thins, in which only every n*th* iterations are used to decrease autocorrelation. The SCORIGHT 3.0 program allows users to implement the two-parameter logistic testlet model, whereas the Gibbs does not. Also, options for choosing the number of thins are available in the SCORIGHT 3.0 program but not in the Gibbs program. In addition, both computer programs do not provide any diagnostic method for MCMC convergence and any options for changing prespecified prior values.

However, the WinBUGS 1.4 computer program (Spiegelhalter et al., 2003) is more flexible than the SCORIGHT 3.0 and the Gibbs programs. The MCMC employing Gibbs sampling in the WinBUGS 1.4 program was implemented to estimate the 3PLT model parameters. Under MCMC, model parameters are estimated by repeatedly sampling each parameter from its posterior distribution, conditional on the data and the most recent estimates of all other parameters. After an initial burn-in period, it is possible to create a Markov chain in such a way that the sampled values are drawn from the parameter's full conditional distribution. The value of each parameter is estimated as the mean of the Markov chain.

Sampling from posterior distribution requires the specification of prior distribution for all MCMC parameters. Both real and simulated data sets are analyzed in the study. A detailed description is presented later.

3.2 RESEARCH DESIGN

Since test design affects quality of estimates about items, ability, and final inferences (Bradlow et al., 1999), several simulation conditions will be considered. The WinBUGS 1.4 program (Spiegelhalter et al., 2003) was implemented to fit the 3PLT model with the same prior distributions as in the SCORIGHT 3.0 program (Wang et al., 2004) and the Gibbs program (Du, 1998) for comparison purposes with real data. Later, different prior distributions on items parameters will be implemented in the WinBUGS program (Spiegelhalter et al., 2003).

3.2.1 DATA GENERATION

The simulation study will be performed to evaluate the sensitivity of prior distributions in 3PLT model by using the WinBUGS program. Item responses for the hypothetical individuals will be obtained based upon a testlet response theory model (Bradlow et al., 1999; Wainer et al., 2000; Wang et al., 2002). Item responses will be randomly generated by imitating a testlet-based test. Factors that are varied across the simulation are prior and hyper-prior distributions of item parameters and the testlet variances.

For parameters of items, the discrimination parameter, a, will be generated using the half normal distribution, $a \sim (\mu_a, \sigma_a^2)I(0,)$; the difficulty parameter, b will be generated using the normal distribution, $b \sim N(\mu_b, \sigma_b^2)$; and the transformed guessing parameter, $q = \log(\frac{c}{1-c})$ will be generated by using the normal distribution, $q \sim N(\mu_q, \sigma_q^2)$. Parameters were obtained from results of the Florida Comprehensive Assessment Test (FCAT). An example input file for the WinBUGS program to generate data is presented in Appendix B.

Ability and random effects for the testlet effect were assumed to be independent of each other (Bradlow et al., 1999; Li et al., 2006; Wainer et al., 2000; Wang et al., 2002). Since it is assumed that the distribution of the ability is known up to a scale parameter, the generated ability parameters follow a normal distribution, $N(0,1)$.

Two sample sizes will be employed; both 1000 and 2000 individuals will be simulated. In addition, the variance of the testlet effect parameters over individuals will be used to quantify the magnitude of the interaction. Since the researchers are generally concerned with the means of the testlet parameters, which are customarily set to 0 to make the scale of the model identifiable in the estimation process, the degree of the testlet effect will be determined by the variances of the testlet parameter values, $(0, \sigma_\gamma^2)$. Thus, testlet parameters will be generated using a normal distribution, $N(0, \sigma_\gamma^2)$. The magnitudes of the testlet effects are determined by the ratio of the random-effect variance of testlets to the random-effect variance of ability (Li et al., 2006; Wainer et al., 2007). In this study, three conditions of different degrees of testlet effects will be simulated: no testlet effect , σ_γ^2=0; moderate testlet effect, σ_γ^2=0.5, and strong testlet effect, σ_γ^2=1. These conditions of testlet effects were similar to those simulation conditions specified in various studies (Bao, 2007; Bradlow et al., 1999; Wang et al., 2002; Wang & Wilson, 2005a, 2005b), in which σ_γ^2 was specified as 0.0, 0.5, and 1.0.

When the parameter values are in place, the probability of getting each item correct will be calculated using the 3PLT model. These parameters of generated latent ability, testlet effect, and defined item parameters will be used to compute the corresponding probability. Two conditions will be studied:

1. the three-parameter logistic (3PL) model assuming local dependence,

2. the 3PLT model assuming local dependence and testlet effect function homogeneous across

Condition 1 assumes that all the items in a test are independent of one another. Condition 2 presumes that the testlet parameter applies constantly to all items in testlets: a constant testlet variance will be generated across all testlets. Regarding conditions 2, variances of testlet effects related to this model will be no testlet effect(0.0), medium (0.5), and large(1.0).

A total of 60 dichotomous items will be generated in a test. Common test structure is set by fixing the test composed of first 5 independent dichotomous items among 60 items. A different number of testlets will have different number of items: 3, 6, and 10 items for each testlet, respectively. The reason each testlet has different numbers of items is that the number of items in a testlet affects the degree of variance of testlets. Also, in general, larger the number of items there are clustered in a testlet, the more likely the testlet effect can be shown (Bradlow et al., 1999). Two chains of iterations and 20 replications will be conducted.

3.2.2 PRIOR AND HYPER-PRIOR DISTRIBUTION

Prior specification is an important step in Bayesian analysis because statistical analysis in the Bayesian approach needs to include prior distributions in the model specification. Also, it is reasonable to use vague information about hyper-parameters in the absence of a strong theory regarding the prior distribution of items and individuals.

Ability parameters, θ will be estimated with $N(0,1)$. Item parameters will be estimated as follows: $N(\mu_a, \sigma_a^2)I(0,)$ for a, $N(\mu_b, \sigma_b^2)$ for b, and $N(\mu_q, \sigma_q^2)$, in which $q = \log(\frac{c}{1-c})$. Values of $\sigma^2_{\mu_a}$, $\sigma^2_{\mu_b}$, and $\sigma^2_{\mu_q}$ are specified as 0.01, 0.001 which indicates different amount of information.

$\sigma_a^2 \sim \chi^{-2}_{g_a}$, $\sigma_b^2 \sim \chi^{-2}_{g_b}$, and $\sigma_q^2 \sim \chi^{-2}_{g_q}$ for prior variance, where $\chi^{-2}_{g_a}$, $\chi^{-2}_{g_b}$ and $\chi^{-2}_{g_q}$ are inverse chi-square random variables with g_a, g_b, and g_q degrees of freedom, which are defined as 0.4, 0.5, and 2.0 to reflect different amount of information. Parameter of testlet effect will be estimated with $(0, \sigma_\gamma^2)$, in which σ_γ^2 follows $\chi^{-2}_{g_\gamma}$ and will be defined as 0.4, 0.5, and 1.0. Furthermore, Figure 3.1 shows inverse chi-squared distribution with the different degrees of freedom.

A normal distribution imposed on prior distributions in the 3PLT model has two parameters, the mean, μ, and the variance, σ^2. The normal distribution is as follows:

$$P(x_1, x_2, \cdots, x_p|\mu, \sigma^2) \propto \frac{1}{\sigma^n} \exp(-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2).$$

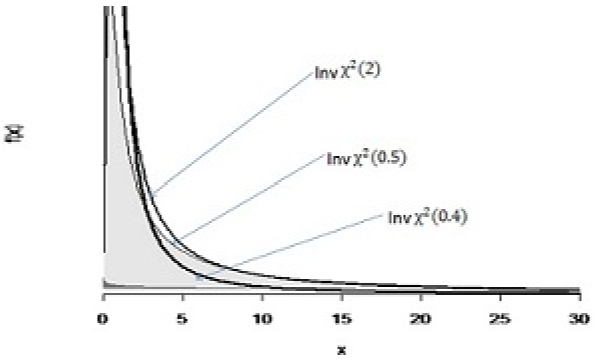

Figure 3.1: Inverse Chi-squared distribution with different degrees of freedom

Conjugate prior distribution having the same functional form as the likelihood function leads to posterior distribution belonging to the same distribution as prior distribution. Conjugate prior distributions for these parameters, μ and σ^2 are needed. Assuming μ is fixed, then the conjugate prior for σ^2 is an inverse gamma distribution that is a general case of the inverse chi-squared distribution (Spiegelhalter et al., 2003, p. 58) which is as follows:

$$f(\sigma^2|\alpha,\beta) \sim IG(\alpha,\beta)$$

Then

$$P(\sigma^2|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}\exp(-\frac{\beta}{x}).$$

The posterior distribution obtained when likelihood and prior distribution combined is as follows:

$$P(x|\mu,\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}\exp(-\frac{\beta}{\sigma^2})(\frac{1}{\sqrt{2\pi\sigma^2}})^p\exp(-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2).$$

The inverse variance term, $\frac{1}{\sigma^2}$, is usually called the precision and is denoted by τ (Spiegelhalter et al., 2003, p. 58). Thus, when σ^2 is reparameterized in terms of precision, τ, the conjugate prior becomes a gamma distribution as follows:

$$f(\tau|\alpha,\beta) \sim G(\alpha,\beta), P(\tau|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\tau)^{\alpha-1}\exp(-\tau\beta).$$

Then, the posterior distribution is as follows:

$$P(x|\mu, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \exp(-\tau\beta)(\frac{\tau}{2\pi})^{p/2} \exp(-\frac{\tau}{2} \sum (x_i - \mu)^2).$$

It is also possible to compute the probability of getting new data given old data by marginalizing out parameters:

$$P(\theta|x, \mu, \alpha, \beta) = \int P(\theta|x, \mu, \alpha, \beta) P(\tau|x, \alpha, \beta) d\tau = \int P(\theta|x, \mu, \tau) P(\tau|x, \alpha, \beta) d\tau$$

$$= \int \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \exp(-\tau\beta)(\frac{\tau}{2\pi})^{p/2} \exp(-\frac{\tau}{2} \sum (x_i - \mu)^2) d\tau$$

Then

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(2\pi)^{p/2}} \int \tau^{(\alpha + \frac{p}{2}) - 1} \exp^{-\tau(\beta + \frac{1}{2} \sum (x_i - \mu)^2)} d\tau.$$

By normalizing constant, this integral becomes:

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(2\pi)^{p/2}} \frac{\Gamma(\alpha + \frac{p}{2})}{(\beta + \frac{1}{2} \sum (x_i - \mu)^2)^{\alpha \frac{p}{2}}}$$

$$= \frac{\Gamma(\alpha + \frac{p}{2})}{\Gamma(\alpha)} \frac{1}{(2\pi\beta)^{\frac{p}{2}}} \frac{1}{(1 + \frac{1}{2\beta} \sum (x_i - \mu)^2)^{\alpha + \frac{p}{2}}}.$$

This integral make a normal distribution having a heavier tailed distribution, which becomes a student's t-distribution. In this model, μ is a location parameter, β is a dispersion parameter, and α is a shape parameter, or degrees of freedom. The multivariate Student t distribution can be reparameterized if k is α, and λ is $\frac{\alpha}{\beta}$:

$$P(x|\mu\lambda, k) = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} (\frac{\lambda}{k\pi})^{\frac{p}{2}} \frac{1}{(1 + \frac{\lambda}{k}(x_i - \mu)^2)^{\frac{k+1}{2}}}.$$

This multivariate Student's t distribution becomes a multivariate Cauchy distribution if the degrees of freedom, α, is 1, and becomes a multivariate normal distribution if the degrees of freedom, α, goes $\propto$. Also, it is commonly known that degrees of freedom need to be larger than 2 to ensure the existence of the variance matrix. However, there is no mathematical reason why the degrees of freedom should be an integer (Heikkinen, & Kanto, 2002), even though Student (1908) considered the distribution only with integer degrees of freedom.

First, the same prior and hyper-prior distribution were used for the comparison purposes. Then, different prior specifications in WinBUGS will be employed so as to compare the sensitivity of prior distributions.

3.2.3 CONVERGENCE AND ITERATIONS

With item parameter estimation via MCMC methods, convergence of the parameter estimation needs to be examined. If the parameter estimates do not converge, incorrect inference about parameters of interest will result. Thus, it is necessary to determine the number of iterations to discard, during which the parameter estimation stabilizes or converges. It is important to decide how many MC iterations are necessary to obtain robust posterior estimation with appropriate burn-in periods. Some of the initial observations should be discarded to avoid the impact of starting states on estimating parameters of interest because unstable iterations affect MC errors (Bazán et al., 2006; Gelman et al., 2003).

Various tools are commonly used such as simple graphical methods, methods using ratio of dispersions, methods based on spectral analysis, method based on the theory of Markov chains available in CODA (Best et al., 1995) and BOA (Smith, 2001). Furthermore, The convergence diagnostic provided in the WinBUGS 1.4 program, including the Gelman-Rubin convergence statistic, R, (Gelman & Rubin, 1992; Brook & Gelman, 1998) and sample history, were computed from multiple chains to determine the number of burn-in periods.

It is also important to recognize that the error in posterior estimation can be caused by not only the standard deviation, but also the sampling error, referred to as MC error (Spiegelhalter et al., 2003). Spiegelhalter et al. (2003) also suggested that the simulation should be run until the MC error for each parameter of interest is less than about 5% of the sample standard deviation. The smaller the MC error, the larger the MCMC iterations.

Post and confshrink statistics available in the SCORIGHT program were also employed to assess convergence when the real data was analyzed: confshrink estimates potential scale reduction with an estimates and an approximate 97.5% upper bound (Wang et al., 2004).

3.3 SIMULATION STUDY

Simulation study was summarized in this section. In order to examine impacts of prior distributions in the 3PLT model, item responses were randomly generated by mimicking a testlet-based test. It was assumed that there were 60 items in a test. Three simulation factors were considered in the simulation study; magnitude of random effect due to testlets, magnitude of prior distribution, magnitude of hyper-prior distribution, different number of items in a testlet, and different number of sample sizes.

The variance of testlet effect, σ^2_γ, was varied in order to simulate varying degrees of dependence. The no testlet effect condition ($\sigma^2_\gamma = 0$) was also included as a baseline for comparisons. Three levels of testlet effect were 0.0, 0.5, and 1.0. The magnitudes of prior distributions were 0.01 to 0.001. The same magnitude of prior distribution was assigned to all item parameters. Degrees of freedoms were 0.4 to 2.0. The number of different items in a testlet was 3 to 10. The number of samples sizes was 1000 and 2000. Thus, the total number of simulation conditions resulted in $3 \times 2 \times 3 \times 3 \times 2 = 108$ conditions (see Table 3.1). Prior to running analysis, the estimates are rescaled on to the same metric by fixing 5 common items among 60 items. An example WinBUGS estimation file is presented in Appendix C.

Table 3.1: Design of Simulation Study

MODEL	N	ITEM	σ^2_γ	$\sigma^2_{\mu.}$ for μ_a, μ_b, μ_q	d.f of $\sigma^2_{g_a}, \sigma^2_{g_b}, \sigma^2_{g_q}$
	$N = 1000$	3 (10 testlets)	σ^2_γ=0.0		
3PLT		6 (5 testlets)	σ^2_γ=0.5	$\sigma^2_{\mu.} = 0.01, 0.001$	d.f = 0.4, 0.5, 2.0
	$N = 2000$	10 (3 testlets)	σ^2_γ=1.0		
Combination	2	3	3	2	3

3.4 MODEL EVALUATION

In each simulation condition, the simulation studies were replicated 25 times which were between low and large number of replication based on previous studies (see Table 2.1). The success of the model was evaluated with two criteria, the root mean squared error (RMSE) of the estimates from the true values and correlation between the true and the estimated parameters which used in the other study (Wang et at., 2002). The RMSE was the discrepancy between the estimated values and the true values. The RMSE was defined as

$$\text{RMSE } (\hat{T}_i) = \sqrt{\tfrac{1}{r}\textstyle\sum_{r=1}^{r}(\hat{T}_i - T)^2}$$

where T is a true parameter and $\hat{T}_i$ is the estimated value of the parameter from the ith replication, and the the simulation is replicated r times, which is 25 in this simulation study. $\bar{\hat{T}}_i$ is the mean of the estimated parameters. The RMSE $(\hat{T}_i)$ can be further dissected into two parts: the standard error of estimated parameters and the bias of the estimated parameters:

$$\begin{aligned}\text{RMSE } (\hat{T}_i) &= \text{Bias}(\hat{T}_i) + \text{SE } (\hat{T}_i)\\ &= \tfrac{1}{r}\textstyle\sum_{i=1}^{r}(\hat{T}_i - T) + \sqrt{\tfrac{1}{r}\textstyle\sum_{r=1}^{r}(\hat{T}_i - \bar{\hat{T}}_i)^2}\ .\end{aligned}$$

Chapter 4

RESULTS

This chapter presents the results from simulation studies and real application study. First, the design of the simulation study is described. Simulation study aims to explore impacts of prior distributions on the parameter estimates. In order to investigate the impacts of different means of prior distributions, different degrees of freedom, different number of items nested in testlets in estimating parameters, 108 conditions were considered (see Table 3.1). The second section presents results obtained from the real data.

4.1 Convergence

Convergence of the parameter posterior distribution to a stationary distribution is crucial to MCMC estimation. Using WinBUGS, two chains of length of 50000 were run and approximately 6 hours to complete with the sample size ($N = 1000$) and 14 hours with the sample size ($N = 2000$). The first 10000 iterations in each chain were discarded (burn-in iterations). All the sampling histories, BGR diagrams, and autocorrelation plots suggested the Markov chains converge to stationary posterior distributions.

Convergence was examined through visual inspection of several convergence diagnostic plots available in WinBUGS. The first plot is a "sampling history plot" for each parameter. Figure 4.1 illustrates the histories of the item discrimination, item difficulty, item guessing and testlet parameters of item 6. The sampling histories showed that each chain displayed convergence to a stationary distribution. Similar results were observed for the other items and testlets.

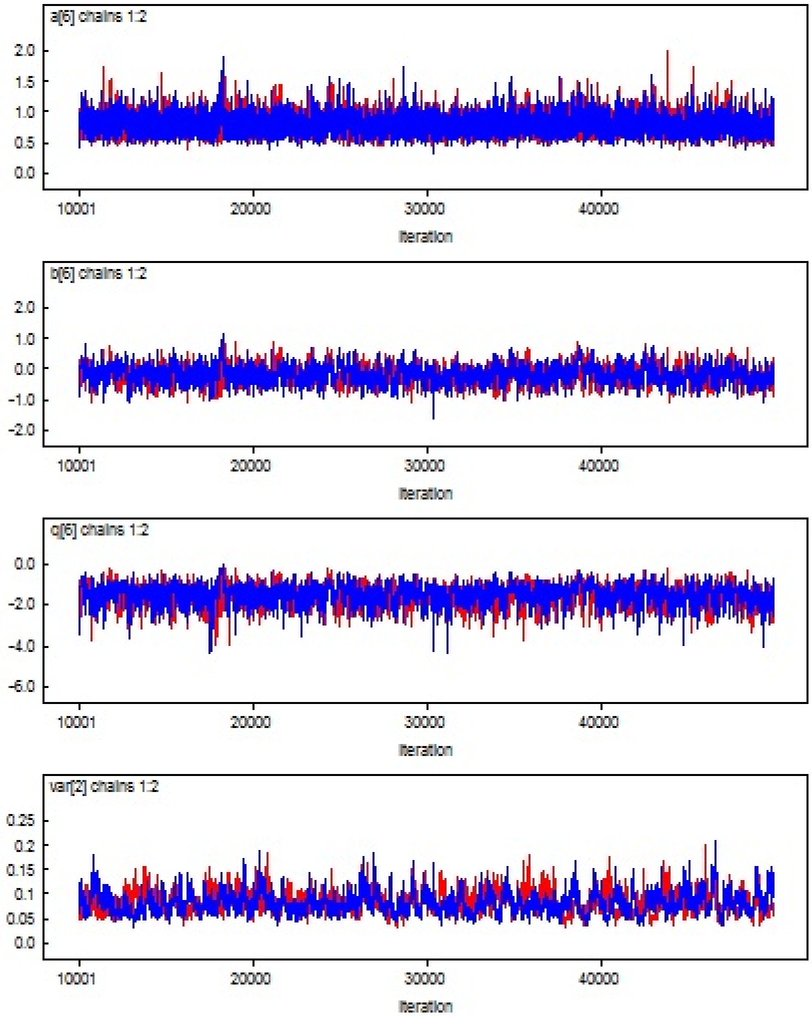

Figure 4.1: Sampling History Plots of Item Parameters Associated with Item 6

In WinBUGS, “BGR diagram” is often used to show the Gelman-Rubin convergence statistic for multiple chains. It includes three lines in different colors. The green (G) and blue (B) lines reflect the pooled and within-chain posterior variances, respectively. The ratio of these two variances, that is, the Gelman-Rubin statistic, is represented by the red (R) line. Figure included the “BGR diagram” for the item discrimination, item difficulty, item guessing, and testlet parameters of item 6. As seen, the red line (Gelman-Rubin statistic) converged to 1, indicating equality between the pooled and within-chain variances. Thus, these plots demonstrated the convergence of the two chains with 50000 iterations was attained for all the parameters of item 6. Similar results were obtained for the other parameters. Figure 4.2 included the BGR diagrams Similar results were observed for the other items.

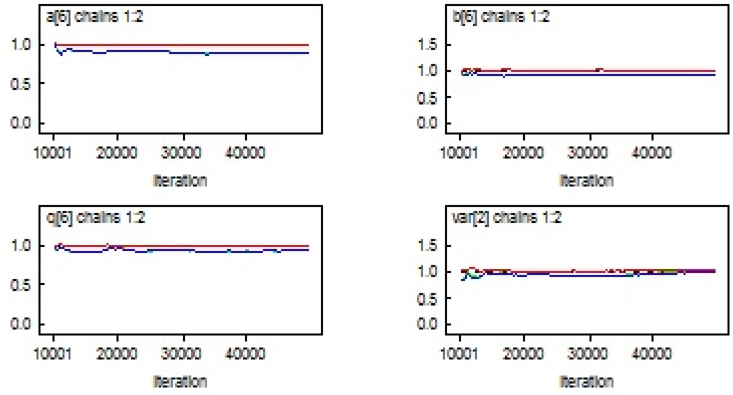

Figure 4.2: BGR Diagrams for the Parameters of Item 6

Autocorrelation plots are also helpful in evaluating convergence. High correlations between adjacent states imply a slow rate of convergence, thus requiring more iterations to achieve stationary posterior distributions for the model parameters. Figure 4.3 provided the autocorrelation plots for the parameters of item 6. As can been seen, the correlations among the successive draws were reduced to 0, indicating the length of 50000 iterations was sufficient to ensure convergence. Similar autocorrelation plots were found for other item parameters.

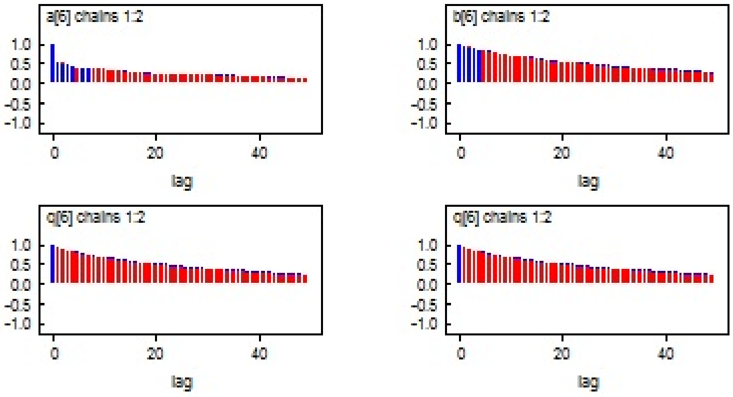

Figure 4.3: Autocorrelation Plots for the Parameters of Item 6

Based on the preliminary analysis with real data and simulation study, it had been decided that 40000 samples should be drawn from each posterior distribution after 10000 samples were discarded as burn-in periods.

4.2 Parameter Recovery

Once the estimations were done, the results of the WinBUGS runs illustrated the simulation conditions under which those models could recover the parameters used to generated the data, give the model that generated the data matched the model used. The quality of model parameter recovery using MCMC estimation is an important factor in determining whether the 3PLT model could be implemented successfully. As a result, parameter recovery was examined. The recovery of the true parameter values are estimated using the root mean square error (RMSE).

Variances of Testlet Parameter. Table 4.1 represented the magnitudes of average estimated testlet parameters across 25 replications. When variances of the testlet parameters were 0, all conditions tended to overestimate impacts of testlet effects. For instance, testlet parameters ranged 0.138 to 0.214 under the $\sigma^2_{\mu \cdot} = 0.01$ and 0.134 to 0.254 when $\sigma^2_{\mu \cdot} = 0.001$. In addition, as the number of items nested in testlets increased, values of testlet parameters also increased under the $\sigma^2_{\mu \cdot} = 0.01$ while values of testlet parameters decreased when items nested in testlets increased under $\sigma^2_{\mu \cdot} = 0.001$. However, For the moderate and large testlet effect cases, the patterns were not as evident as in the no testlet effect with the respect to the number of items nested in testlets. However, when variances of the testlet parameters were 0.5 and 1.0, all conditions tended to underestimate impacts of testlet effects. These patterns were clearly showed in Figure 4.4 and Figure 4.5. However, values of testlet parameters were close to the true values when the number of examinees increased.

In addition, Appendix A.1 provided average RMSE for variances of testlet parameters and the smaller RMSE values indicates better estimation performance. Testlets consisted of 10 items had the smallest RMSE values, ranging from 0.099 to 0.026 regardless of different

Table 4.1: Magnitude of Variances of Testlet Effect

		N = 1000			N = 2000		
$\sigma^2_{\mu.}$ = 0.01							
d.f.	Item	0.0	0.5	1.0	0.0	0.5	1.0
0.4	3	0.138	0.363	0.652	0.103	0.376	0.751
	6	0.146	0.356	0.701	0.143	0.396	0.765
	10	0.191	0.364	0.727	0.168	0.403	0.775
0.5	3	0.166	0.369	0.718	0.150	0.393	0.771
	6	0.167	0.372	0.748	0.159	0.398	0.768
	10	0.191	0.403	0.776	0.170	0.407	0.783
2.0	3	0.185	0.370	0.701	0.149	0.395	0.747
	6	0.192	0.390	0.748	0.157	0.402	0.790
	10	0.214	0.407	0.758	0.159	0.413	0.790
$\sigma^2_{\mu.}$ = 0.001							
0.4	3	0.155	0.366	0.688	0.144	0.386	0.734
	6	0.141	0.376	0.765	0.100	0.395	0.765
	10	0.134	0.374	0.731	0.099	0.404	0.775
0.5	3	0.254	0.373	0.751	0.150	0.399	0.751
	6	0.241	0.373	0.770	0.148	0.401	0.770
	10	0.212	0.387	0.782	0.145	0.406	0.803
2.0	3	0.194	0.385	0.743	0.157	0.395	0.751
	6	0.157	0.390	0.765	0.154	0.404	0.791
	10	0.154	0.418	0.795	0.151	0.405	0.848

number of items within testlets across different hyper-prior distributions when the number of same sizes is 1,000. Overall, testlets consisted of 10 items still had smaller RMSE values compare to testlets having different number of items when variances of the testlet parameters were 0. When the sample sizes increased to 2,000, RMSE values dramatically decreased. Figure 4.6 and Figure 4.7 showed these trends.

The parameter recovery of testlet parameter of γ seemed to be not as good as those of item and person parameters. That might be due to the facts that each testlet provided relatively little information to estimate its person-testlet interaction parameter, γ, since items nested within testlet had the same testlet parameter. The testlet structure, 3 testlets of size 10

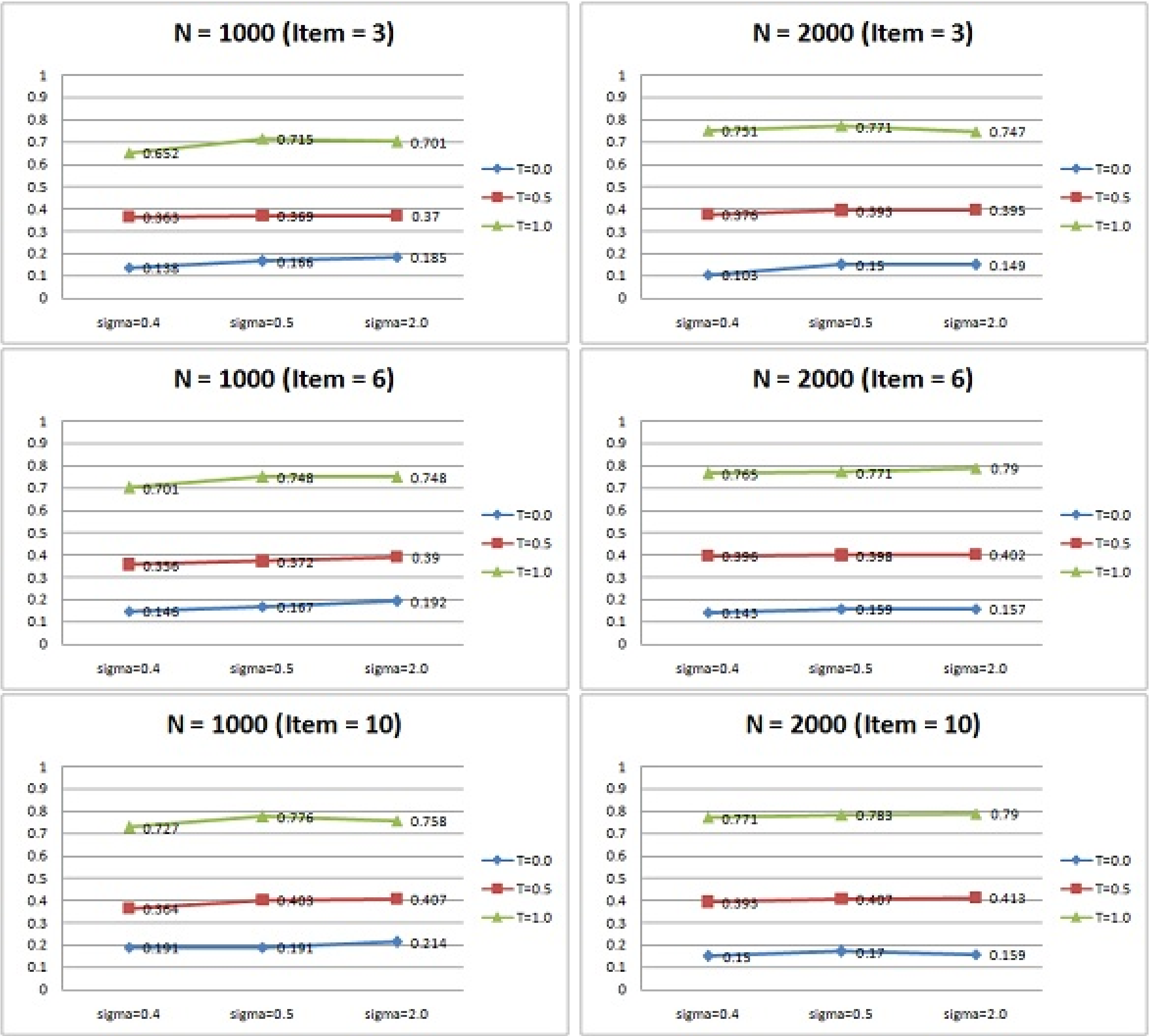

Figure 4.4: Average Variances of Testlet with $\sigma^2_{\mu.} = 0.01$

versus 5 testlets of size of 6 did not have a consistent directional impact on the summary measures chosen when parameters were estimated with $\sigma^2_{\mu.} = 0.01$.

Person Parameter. Correlations between average estimated θ estimates and true θ values are presented in Table 4.2. Higher correlation indicates better estimation performance for the model. All the conditions produce very similar correlations under each of the three testlet effect conditions. For example, when the variance of the testlet parameters are 0, the mean

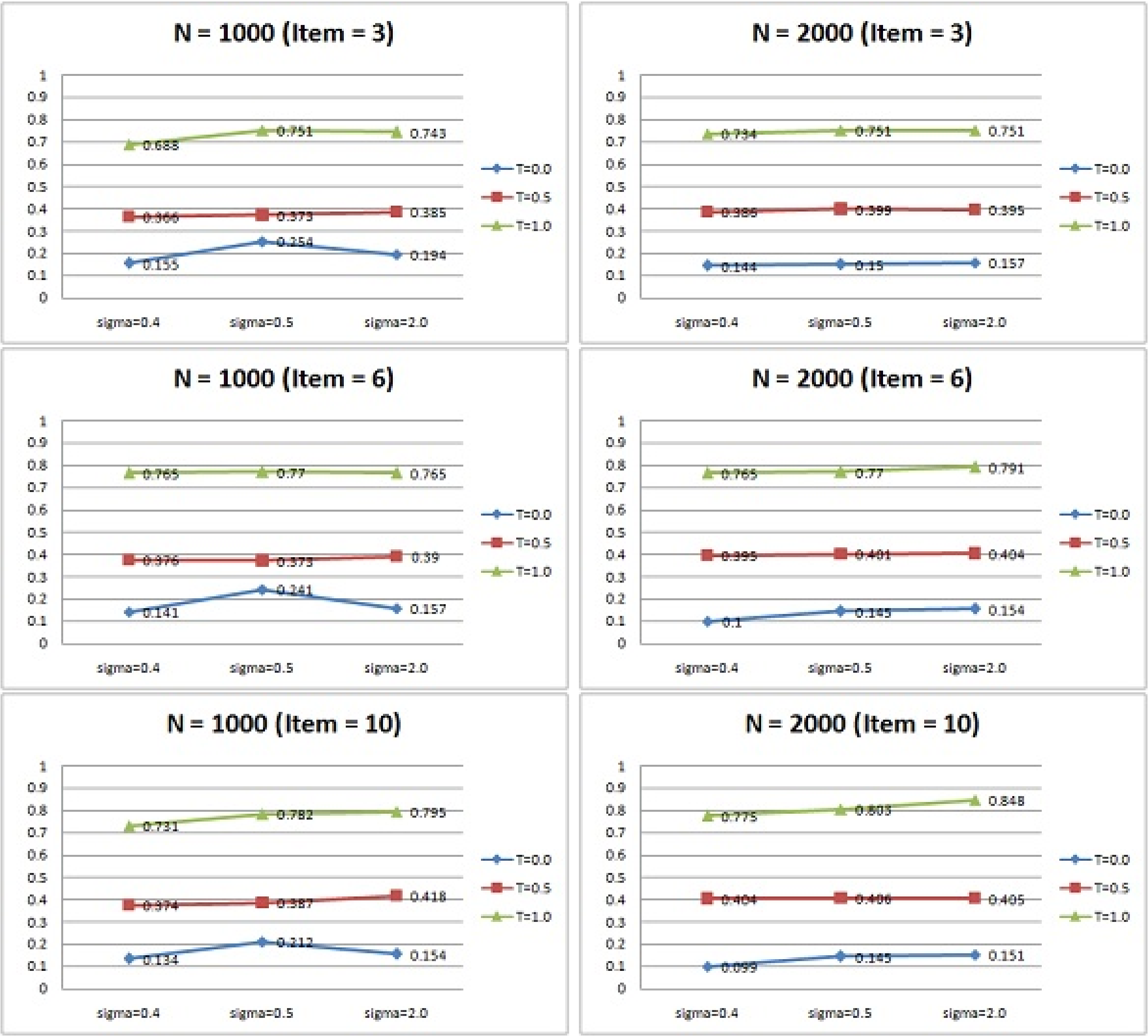

Figure 4.5: Average Variances of Testlet with $\sigma^2_{\mu\cdot} = 0.001$

corelation of the estimated person parameter and the true person parameter $r(\hat{\theta}, \theta) = .915$. According to the information given in the above table, the correlations of true and estimated person parameters were around .86 ∼ .93. When comparing $r(\hat{\theta}, \theta)$ of $\sigma^2_{\mu\cdot} = 0.01$ to that of $\sigma^2_{\mu\cdot} = 0.001$, the better $r(\hat{\theta}, \theta)$ were obtained in the context of $\mu = 0.001$.

Item Parameter. Performance of different prior distributions are also evaluated by examining how well it recovers the true item parameters. For each sample, correlation between

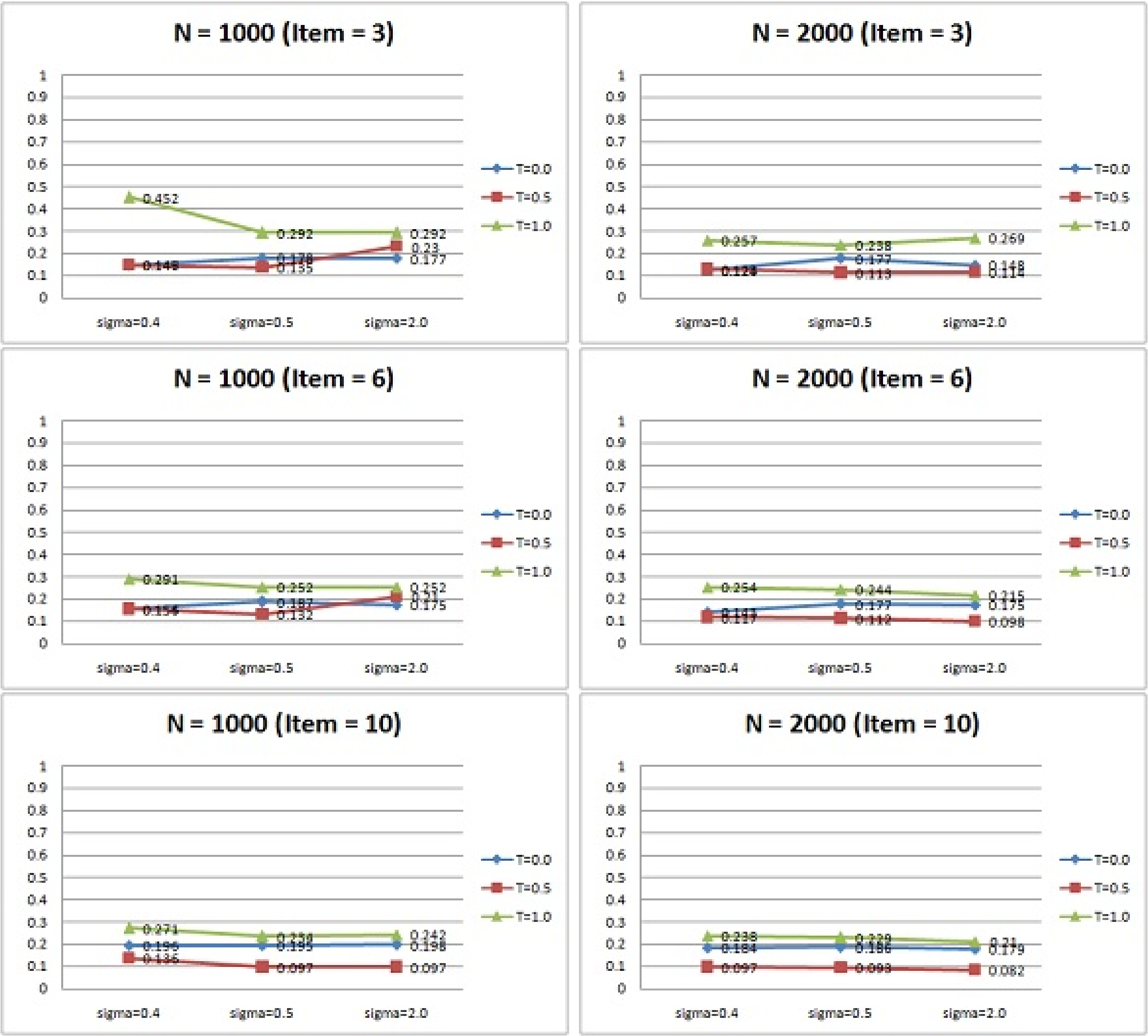

Figure 4.6: RMSE of Testlet Parameters with $\sigma^2_{\mu.} = 0.01$

the estimated item parameters and the true item parameters were computed. Table 4.3, 4.4, and 4.5 presented the summary statistics of the correlations for item parameters.

Item Discrimination Parameter. With respect to $\sigma^2_{\mu.} = 0.01$, Table 4.3 showed that the mean correlation for the item discrimination parameters was much higher when there was a large testlet effect with $N = 1000$, ranging from .962 to .973. Overall mean correlations for the item discrimination parameters were smaller when the item discrimination parameters

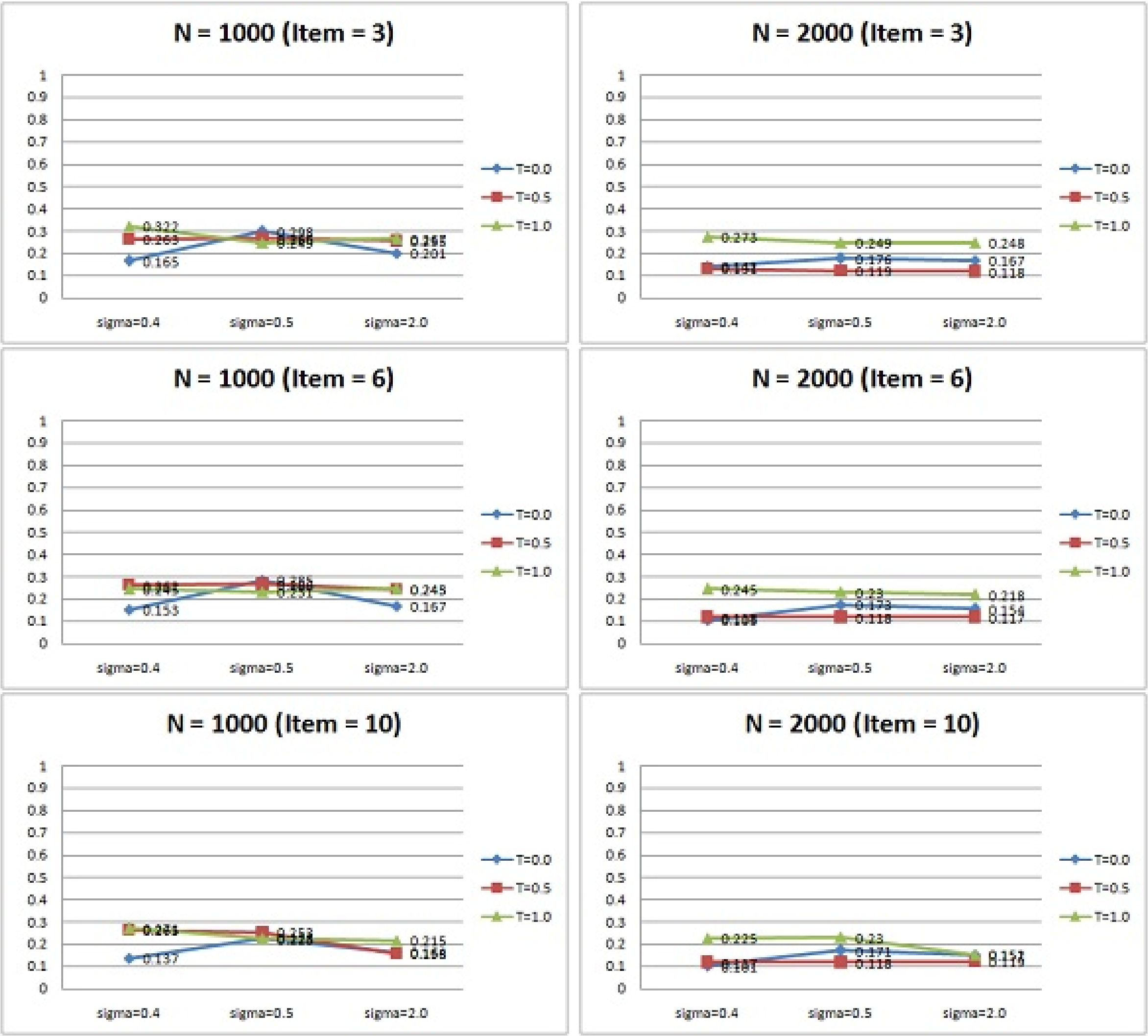

Figure 4.7: RMSE of Testlet Parameters with $\sigma^2_{\mu.} = 0.001$

were estimated under the mild variance of the testlet effects. For instance, when the variance of the testlet parameter were 0.5, mean $r(\hat{a}, a)$ was .944, ranging from .807 to .997. When there was no testlet effect, mean $r(\hat{a}, a)$ was .965, ranging .960 to .971.

With respect to degrees of freedom, higher mean $r(\hat{a}, a)$ was obtained with large numbers of items nested in testlets. The large $r(\hat{a}, a)$ was obtained when the number of items within testlets was not considered; the average $r(\hat{a}, a)$ was .945 for $d.f. = 0.4$, .957 for $d.f. = 0.5$

Table 4.2: Correlation of True and Average Estimated Person Parameters

		$N = 1000$			$N = 2000$		
$\sigma^2_{\mu.} = 0.01$							
$d.f.$	Items	0	0.5	1.0	0	0.5	1.0
0.4	3	.929	.888	.888	.921	.925	.922
	6	.929	.887	.883	.921	.925	.922
	10	.929	.866	.863	.921	.926	.922
0.5	3	.929	.888	.873	.921	.925	.922
	6	.929	.887	.888	.921	.925	.922
	10	.866	.866	.871	.921	.925	.922
2.0	3	.866	.888	.866	.921	.925	.922
	6	.929	.866	.862	.921	.925	.922
	10	.929	.865	.888	.921	.926	.922
$\sigma^2_{\mu.} = 0.001$							
0.4	3	.929	.930	.929	.922	.935	.933
	6	.928	.931	.929	.922	.934	.933
	10	.929	.930	.929	.922	.935	.934
0.5	3	.929	.931	.930	.922	.934	.934
	6	.929	.931	.930	.922	.935	.933
	10	.929	.931	.930	.922	.935	.933
2.0	3	.929	.930	.931	.922	.935	.933
	6	.929	.931	.930	.922	.936	.934
	10	.929	.931	.930	.922	.936	.933

and .975 for $d.f. = 2.0$. However, when the number of sample sizes were increased to 2000, the mean $r(\hat{a}, a)$ across all conditions were increased. When there was no testlet effect, mean $r(\hat{a}, a)$ was .977, ranging from .971 to .979.

Besides correlation for the item discrimination parameters, Figure 4.8 and 4.9 showed patterns of RMSE of item discrimination parameters. Also, Appendix A.2 presented summary statistics of average RMSE for item discrimination parameter estimates. When the degrees of prior distribution was specified to $\sigma^2_{\mu.} = 0.01$, the lower RMSE of item discrimination parameters was obtained when the item discrimination parameter was estimated with $d.f. = 0.4$ (0.153); 0.161 for $d.f. = 0.5$ and 0.163 for $d.f. = 2.0$. When the sample sizes were

Table 4.3: Correlation of True and Average Estimated Item Discrimination Parameters

		$N = 1000$			$N = 2000$		
$\sigma^2_{\mu.} = 0.01$							
$d.f.$	Items	0	0.5	1.0	0	0.5	1.0
0.4	3	.960	.807	.963	.978	.978	.968
	6	.964	.907	.962	.978	.976	.962
	10	.971	.997	.970	.978	.972	.961
0.5	3	.962	.907	.962	.979	.972	.967
	6	.964	.907	.968	.978	.974	.961
	10	.971	.997	.971	.971	.968	.960
2.0	3	.962	.979	.963	.979	.971	.964
	6	.964	.994	.973	.978	.973	.961
	10	.971	.997	.972	.978	.975	.960
$\sigma^2_{\mu.} = 0.001$							
0.4	3	.960	.961	.962	.978	.961	.962
	6	.963	.963	.962	.978	.962	.963
	10	.970	.969	.963	.977	.961	.963
0.5	3	.962	.963	.964	.978	.969	.964
	6	.965	.965	.931	.978	.974	.962
	10	.972	.969	.962	.977	.972	.961
2.0	3	.965	.961	.962	.978	.963	.964
	6	.965	.964	.963	.978	.966	.964
	10	.970	.965	.963	.977	.973	.965

increased to 2000, values of RMSE dramatically decreased. The smallest RMSE values were obtained when item discrimination parameters were estimated with $d.f. = 0.5$ (0.052); 0.053 for $d.f. = 0.4$ and 0.053 for $d.f. = 2.0$.

Item Difficulty Parameter.Table 4.4 also showed that the correlations between true and average estimated item difficulty parameters, with the mean correlations ranging from .774 to .929. Appendix A.3 presents summary statistics of the average RMSE. When item difficulty parameters were estimated with $\mu = 0.01$, the higher mean correlation for the item difficulty parameters was obtained under the condition when the variance of the testlet parameters were 0 with $N = 1000$, ranging from .911 to .926.

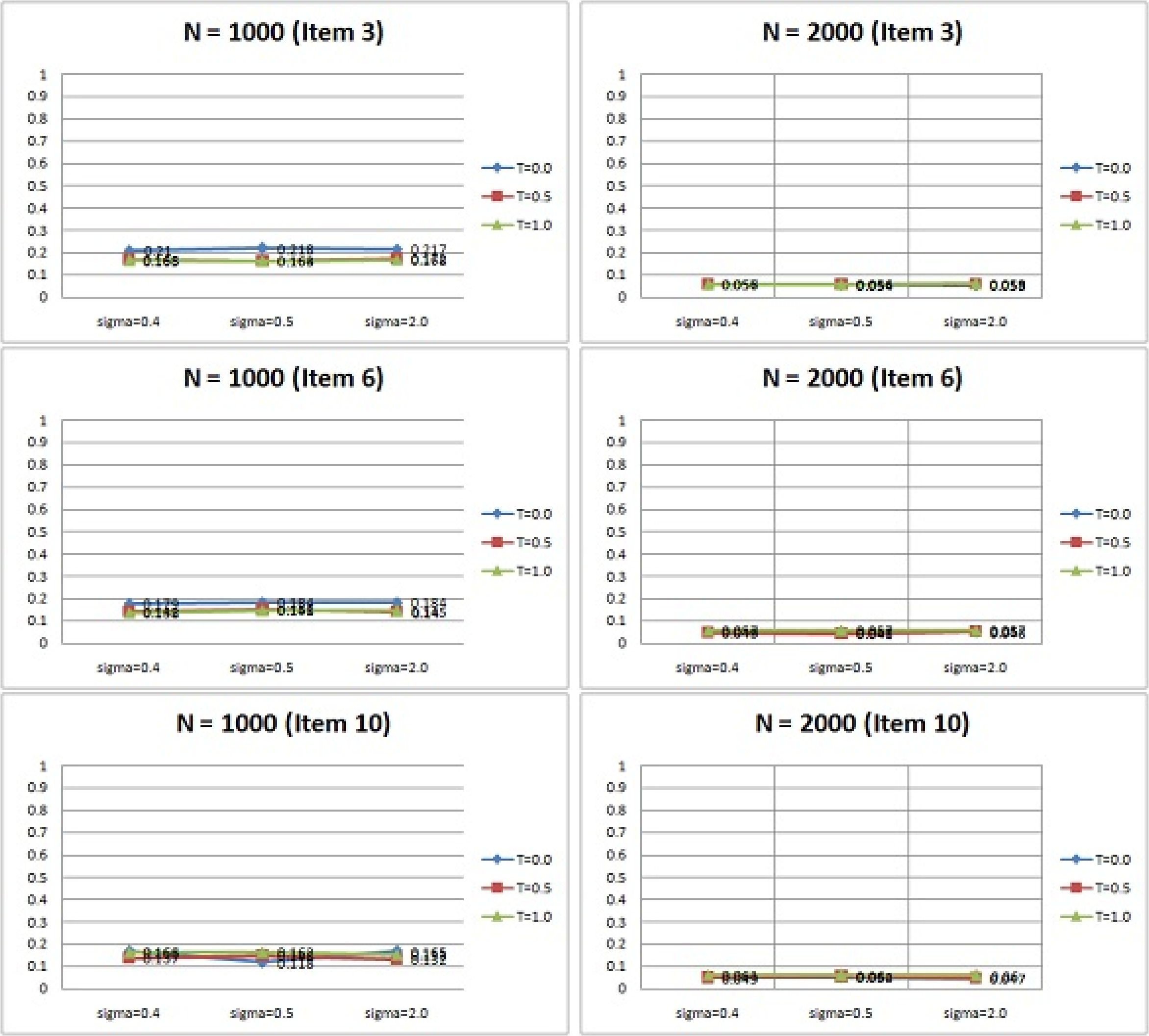

Figure 4.8: RMSE of Item Discrimination Parameters with $\sigma^2_{\mu\cdot} = 0.01$

Overall mean correlations for the item difficulty parameters were smaller when the item difficulty parameters were estimated under the large variance of the testlet effects. For instance, when the variance of the testlet parameter were 0.5, mean $r(\hat{b}, b)$ was .860, ranging from .807 to .915. When the variance of the testlet parameters was 1.0, mean $r(\hat{b}, b)$ was .884, ranging .774 to .911.

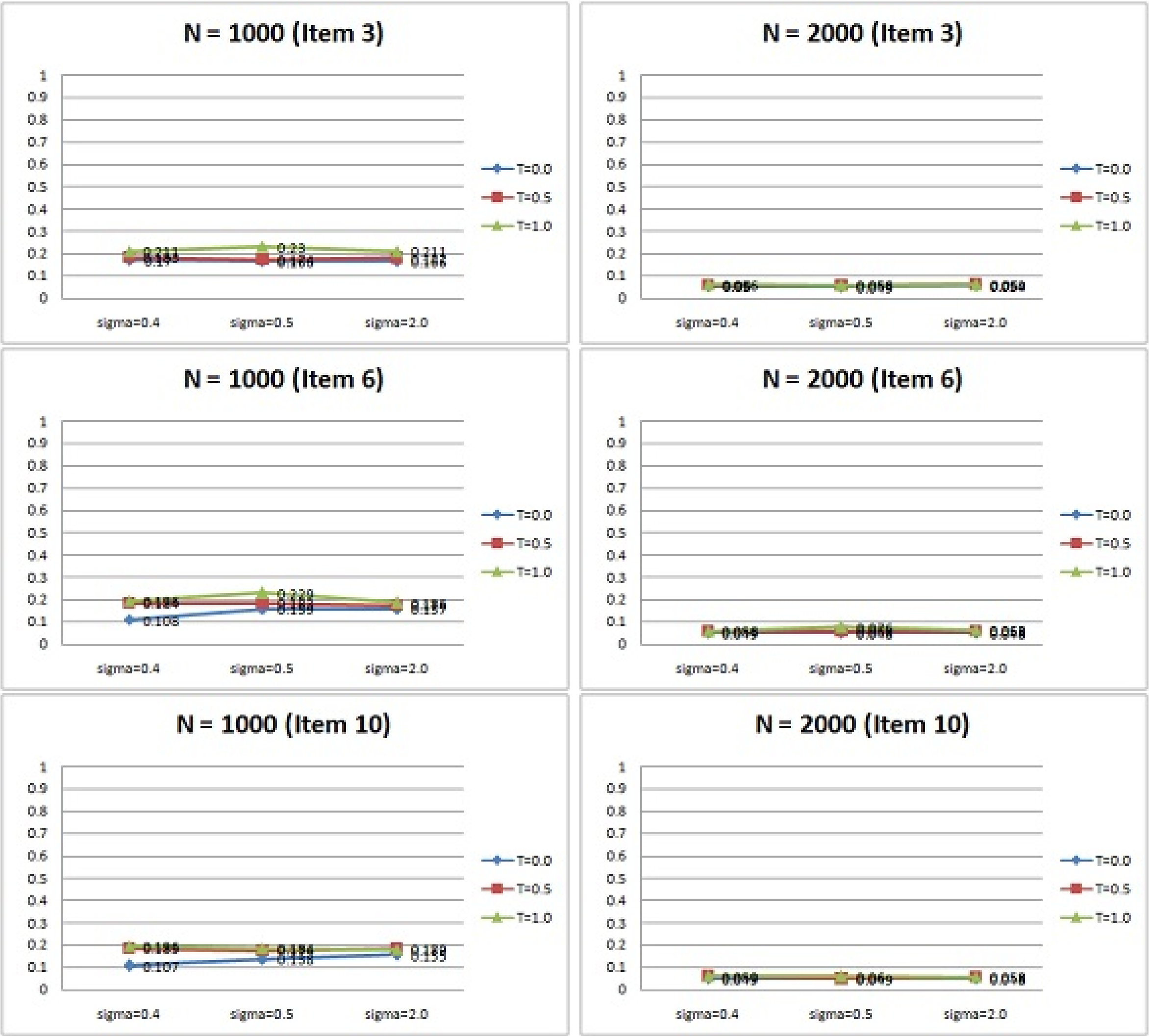

Figure 4.9: RMSE of Item Discrimination Parameters with $\sigma^2_{\mu.} = 0.001$

With respect to degrees of freedom, mean $r(\hat{a}, a) = .975$ obtained from $d.f. = 2.0$ under the large testlet effect condition, which was higher than other values obtained from $d.f. = 0.5$ ($r(\hat{b}, b) = .957$) and $d.f. = 0.0$ ($r(\hat{b}, b) = .945$). However, when the number of sample sizes increased to 2000, the mean $r(\hat{b}, b)$ across all conditions increased as well. When there was no testlet effect, mean $r(\hat{b}, b)$ was .926, ranging from .921 to .929. In general, higher correlation was obtained from $d.f. = 2.0$, $r(\hat{a}, a) = .927$.

Table 4.4: Correlation of True and Average Estimated Item Difficulty Parameters

		N = 1000			N = 2000		
$\sigma^2_{\mu.}$ = 0.01							
d.f.	Items	0	0.5	1.0	0	0.5	1.0
0.4	3	.926	.807	.893	.921	.917	.918
	6	.925	.823	.774	.924	.923	.924
	10	.924	.915	.897	.925	.924	.925
0.5	3	.924	.828	.892	.927	.915	.916
	6	.922	.829	.891	.927	.915	.916
	10	.911	.827	.894	.929	.921	.923
2.0	3	.916	.902	.901	.929	.915	.915
	6	.922	.904	.911	.927	.917	.916
	10	.922	.911	.903	.925	.926	.917
$\sigma^2_{\mu.}$ = 0.001							
0.4	3	.916	.913	.915	.921	.918	.917
	6	.916	.923	.917	.924	.925	.925
	10	.918	.924	.917	.925	.935	.925
0.5	3	.924	.914	.921	.927	.923	.926
	6	.924	.912	.924	.927	.921	.926
	10	.925	.914	.924	.929	.935	.925
2.0	3	.926	.916	.921	.929	.928	.925
	6	.925	.917	.921	.927	.932	.926
	10	.925	.917	.924	.925	.936	.927

As the testlet effect increased, the $r(\hat{b}, b)$ revealed similar values. Figure 4.10 showed overall patterns across all conditions.

Figure 4.11 showed overall patterns of item difficulty parameters. When $\sigma^2_{\mu.}$ was increased to 0.001, the higher mean value of correlation was obtained when the item difficulty parameter was estimated with *d.f.* = 2.0 ($r(\hat{b}, b)$ = .922); ($r(\hat{b}, b)$ = .918 for *d.f.* = 0.4 and $r(\hat{b}, b)$ = .920 for *d.f.* = 0.5.) When the sample sizes were increased to 2000, the same results were obtained; $r(\hat{b}, b)$ = .924 for *d.f.* = 0.4, $r(\hat{b}, b)$ = .926 for *d.f.* = 0.5, and $r(\hat{b}, b)$ = .928 for *d.f.* = 1.0.

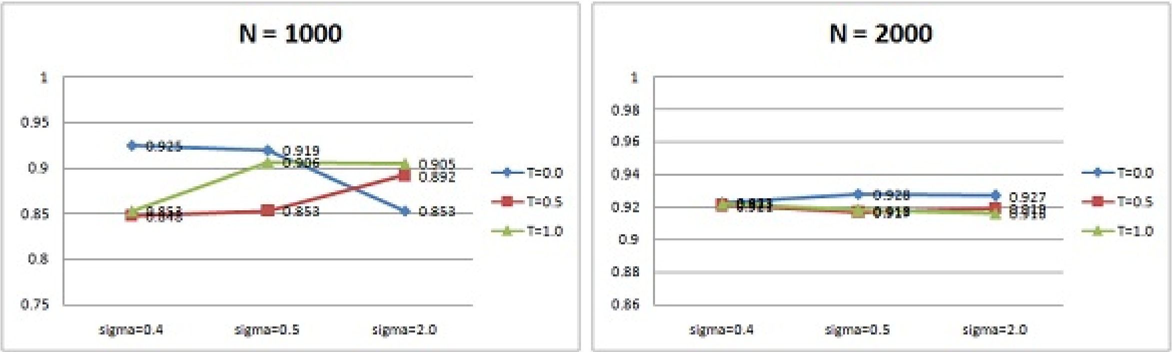

Figure 4.10: Average Correlation of Item Difficulty with $\sigma^2_{\mu.} = 0.01$

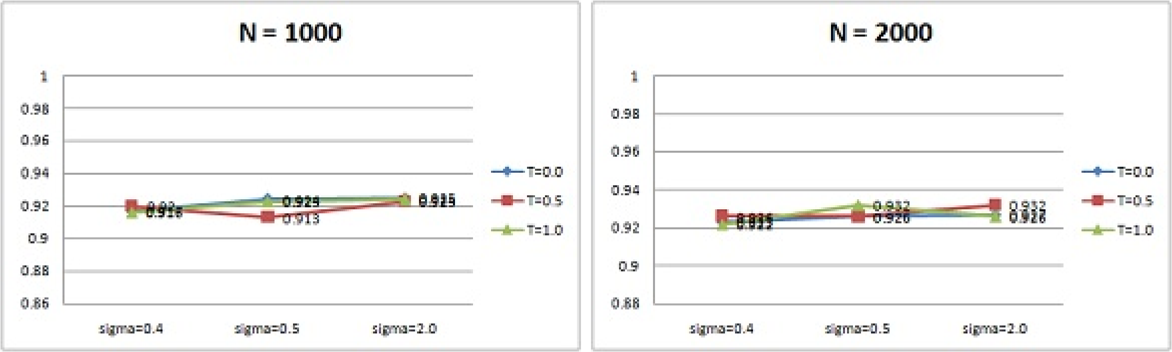

Figure 4.11: Average Correlation of Item Difficulty with $\sigma^2_{\mu.} = 0.001$

Besides correlation for the item discrimination parameters, Figure 4.12 and 4.13 showed patterns of RMSE of item difficulty parameters. In addition, Appendix A.3 presented summary statistics of the average RMSE for item difficulty parameter estimates. When $\sigma^2_{\mu.}$ was increased to 0.01, the lower RMSE of item difficulty parameters was obtained when the item difficulty parameter was estimated with $d.f. = 2.0$; .434 for $d.f. = 0.0$, .425 for $d.f. = 0.5$, and .412 for $d.f. = 2.0$ (Figure A.3). When the sample sizes were increased to 2000, values of RMSE were dramatically decreased. The smallest RMSE values were obtained when item difficulty parameters were estimated with $d.f. = 0.5$; .343 for $d.f. = 0.4$, .342 for $d.f. = 0.5$, and .353 for $d.f. = 2.0$.

When the prior distribution was increased to $\sigma^2_{\mu.} = 0.001$ (Figure **??**), lower RMSE value was obtained when the item difficulty parameters were estimated with $d.f. = 0.5$; .367 for $d.f. = 0.4$, .355 for $d.f. = 0.5$ and .364 for $d.f. = 2.0$. However, when the sample sizes were increased to 2000, the lower RMSE value was estimated with $d.f. = 2.0$; .330 for $d.f. = 0.4$, .331 for $d.f. = 0.5$, and .314 for $d.f. = 2.0$.

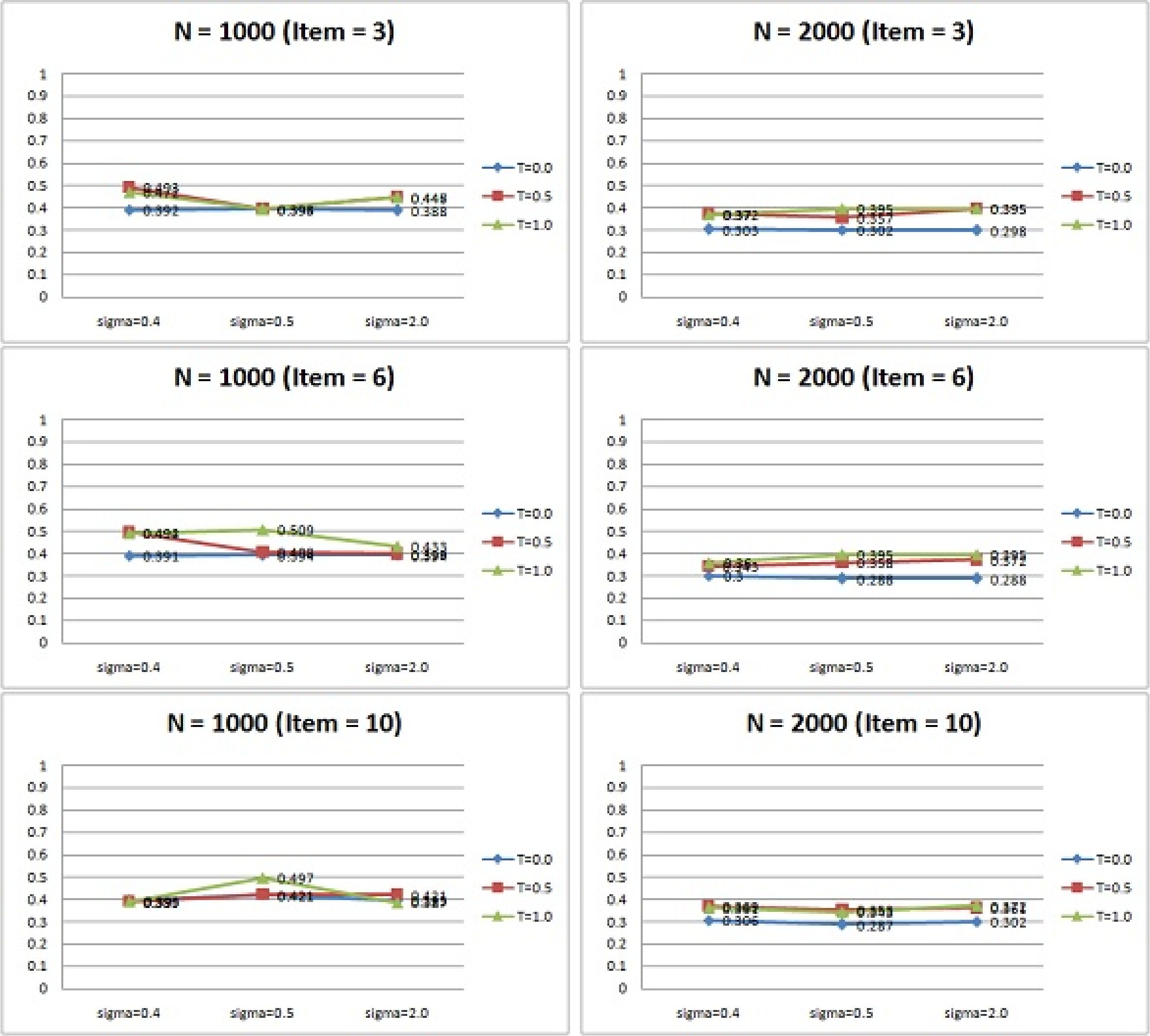

Figure 4.12: RMSE of Item Difficulty Parameters with $\sigma^2_{\mu.} = 0.01$

Item Guessing Parameter. Table 4.5 presented correlations between true and average estimated q parameters, ranging from .281 to .575. Table 4.5 demonstrated that the estimated

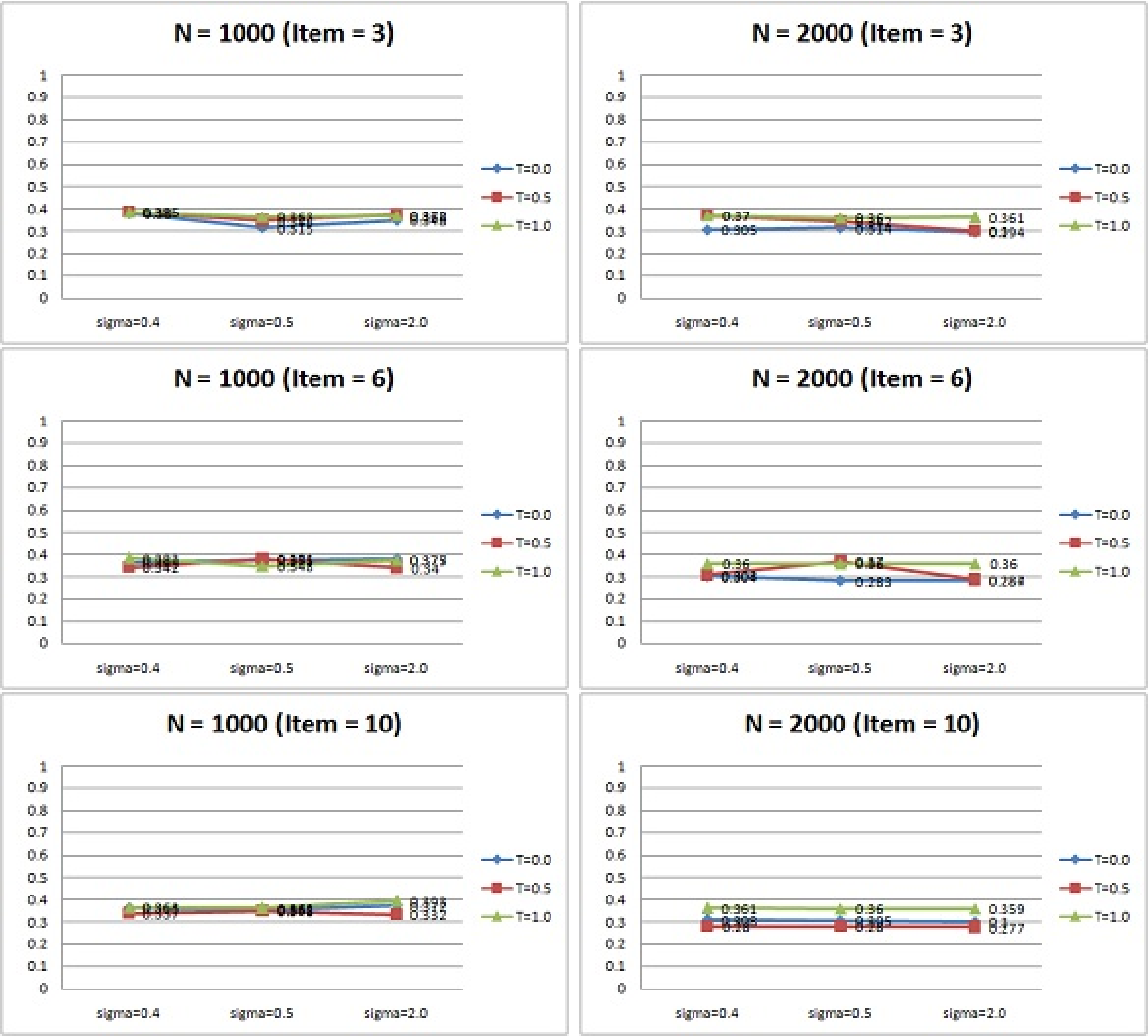

Figure 4.13: RMSE of Item Difficulty Parameters with $\sigma^2_{\mu \cdot} = 0.001$

guessing parameters were somewhat far away from the true parameter regardless of testlet effects. Also, Appendix A.4 shows the average RMSE of the guessing parameters.

When item guessing parameters were estimated with $\sigma^2_{\mu \cdot} = 0.01$ with $N = 1000$, the higher mean correlation for the item guessing parameters was obtained when there was no testlet effect, ranging from .330 to .458. Overall mean correlations for the item guessing parameters were smaller when the item guessing parameters were estimated under the large variance of the testlet effects. For instance, when the variance of the testlet parameter were 0.5, mean $r(\hat{q}, q)$ was .308, ranging from .252 to .351. When the variance of the testlet parameters was 1.0, mean $r(\hat{q}, q)$ was .348, ranging .281 to .428.

Table 4.5: Correlation of True and Average Estimated Item Guessing Parameters

		$N = 1000$			$N = 2000$		
$\sigma^2_{\mu.} = 0.01$							
d.f.	Items	0	0.5	1.0	0	0.5	1.0
0.4	3	.428	.320	.320	.412	.469	.529
	6	.429	.281	.281	.395	.416	.533
	10	.461	.337	.370	.369	.468	.528
0.5	3	.414	.351	.351	.461	.454	.438
	6	.418	.285	.370	.424	.457	.446
	10	.330	.327	.428	.423	.453	.437
2.0	3	.458	.252	.252	.451	.460	.530
	6	.419	.320	.343	.424	.453	.538
	10	.406	.302	.420	.394	.471	.503
$\sigma^2_{\mu.} = 0.001$							
0.4	3	.431	.420	.420	.402	.552	.529
	6	.417	.444	.481	.405	.565	.535
	10	.411	.470	.470	.412	.575	.576
0.5	3	.467	.428	.451	.462	.477	.534
	6	.436	.485	.470	.423	.560	.469
	10	.417	.431	.452	.427	.565	.511
2.0	3	.465	.452	.369	.415	.565	.531
	6	.449	.420	.419	.414	.570	.535
	10	.432	.440	.409	.450	.575	.541

With the respect to degrees of hyper-prior distributions, mean $r(\hat{q}, q)$ = .389 obtained from *d.f.* = 0.4 under the no testlet effect condition. When there was mild testlet effect, the estimated guessing parameter with *d.f.* = 0.5 had highest correlation; ($r(\hat{q}, q)$ = .313) for *d.f.* = 0.0, $r(\hat{q}, q)$ = .321 for *d.f.* = 0.5. and $r(\hat{q}, q)$ = .291 for *d.f.* = 2.0. However, when the number of sample sizes was increased to 2000, the mean $r(\hat{q}, q)$ across all conditions were increased. When there was no testlet effect, mean $r(\hat{q}, q)$ was .417, ranging from .369 to .461. Higher values $r(\hat{q}, q)$ was obtained from *d.f.* = 2.0; $r(\hat{q}, q)$ = .498 for *d.f.* = 2.0, and $r(\hat{q}, q)$ = .456 for *d.f.* = 0.5. Figure 4.14 and 4.15 showed the similar patterns. However, when compared values of $r(\hat{q}, q)$ obtained between $\sigma^2_{\mu.} = 0.01$ and $\sigma^2_{\mu.} = 0.001$, in general values of $r(\hat{q}, q)$ was higher when item guessing parameters were estimated with $\sigma^2_{\mu.} = 0.001$.

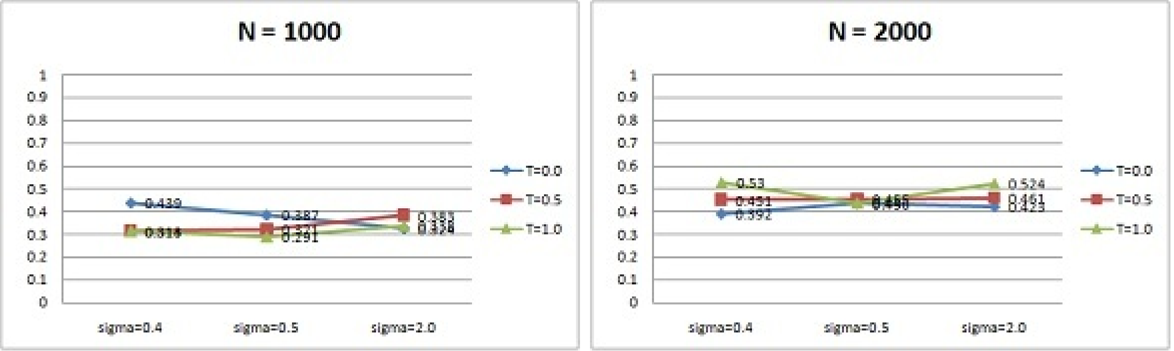

Figure 4.14: Average Correlation of Item Guessing with $\sigma^2_{\mu.} = 0.01$

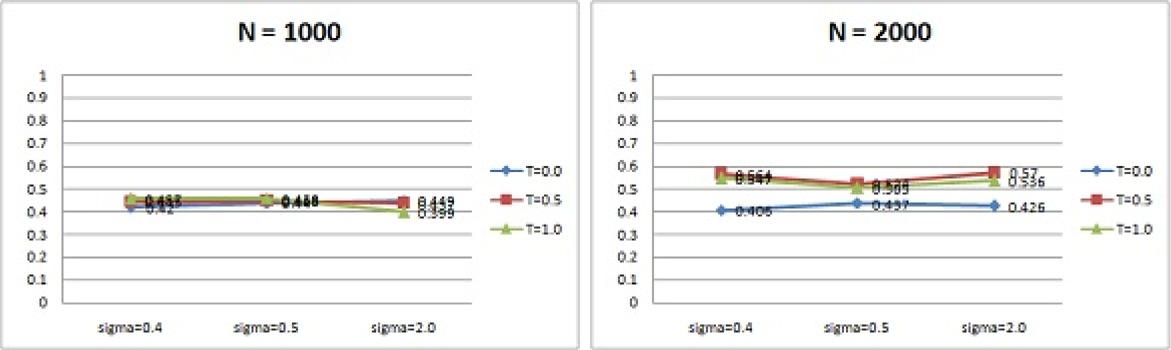

Figure 4.15: Average Correlation of Item Guessing with $\sigma^2_{\mu.} = 0.001$

Besides correlation for the item discrimination parameters, Figure 4.16 and 4.17 showed patterns of RMSE of item difficulty parameters. In addition, Appendix A.4 presented summary statistics of the average RMSE for item guessing parameter estimates. When $\sigma^2_{\mu.}$ was increased to 0.01, the lower RMSE value of item guessing parameters was obtained when the item guessing parameter was estimated with $d.f. = 0.4$; .397 for $d.f. = 0.4$, .429 for $d.f. = 0.5$, and .414 for $d.f. = 2.0$ (Figure A.3). When the sample sizes were increased to 2000, values of RMSE decreased. The smallest RMSE values were obtained when item guessing parameters were estimated with $d.f. = 0.4$; .288 for $d.f. = 0.4$, .315 for $d.f. = 0.5$, and .313 for $d.f. = 2.0$.

When the prior distribution was increased to $\sigma^2_{\mu.} = 0.001$ (Figure **??**), lower RMSE value was obtained when the item guessing parameters were estimated with $d.f. = 0.5$; .330 for $d.f. = 0.4$, .319 for $d.f. = 0.5$, and .348 for $d.f. = 2.0$. However, when the sample sizes were increased to 2000, the lower RMSE value was estimated with $d.f. = 2.0$; .293 for $d.f. = 0.4$, .292 for $d.f. = 0.5$, and .289 for $d.f. = 2.0$.

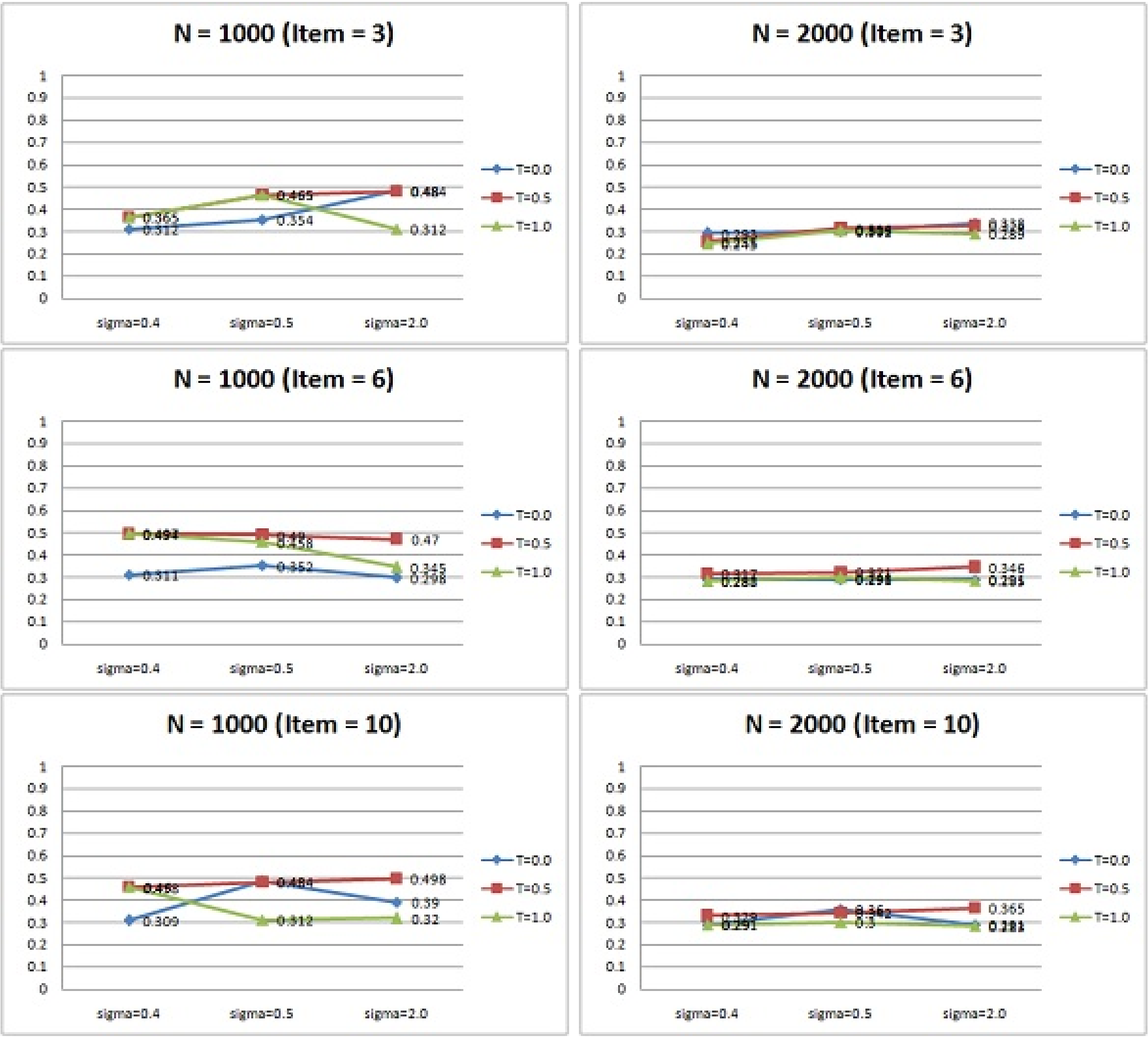

Figure 4.16: RMSE of Item Guessing Parameters with $\sigma^2_{\mu.} = 0.01$

Appendix A.5 described information about the RMSE values of items only nested testlets. Among 60 items, 5 items were fixed for the scaling purpose and 25 items were considered as independent items. Last 30 items were nested in different number of testlets. In general, testlets having a relatively large number of items reveal small degree of RMSE compared to

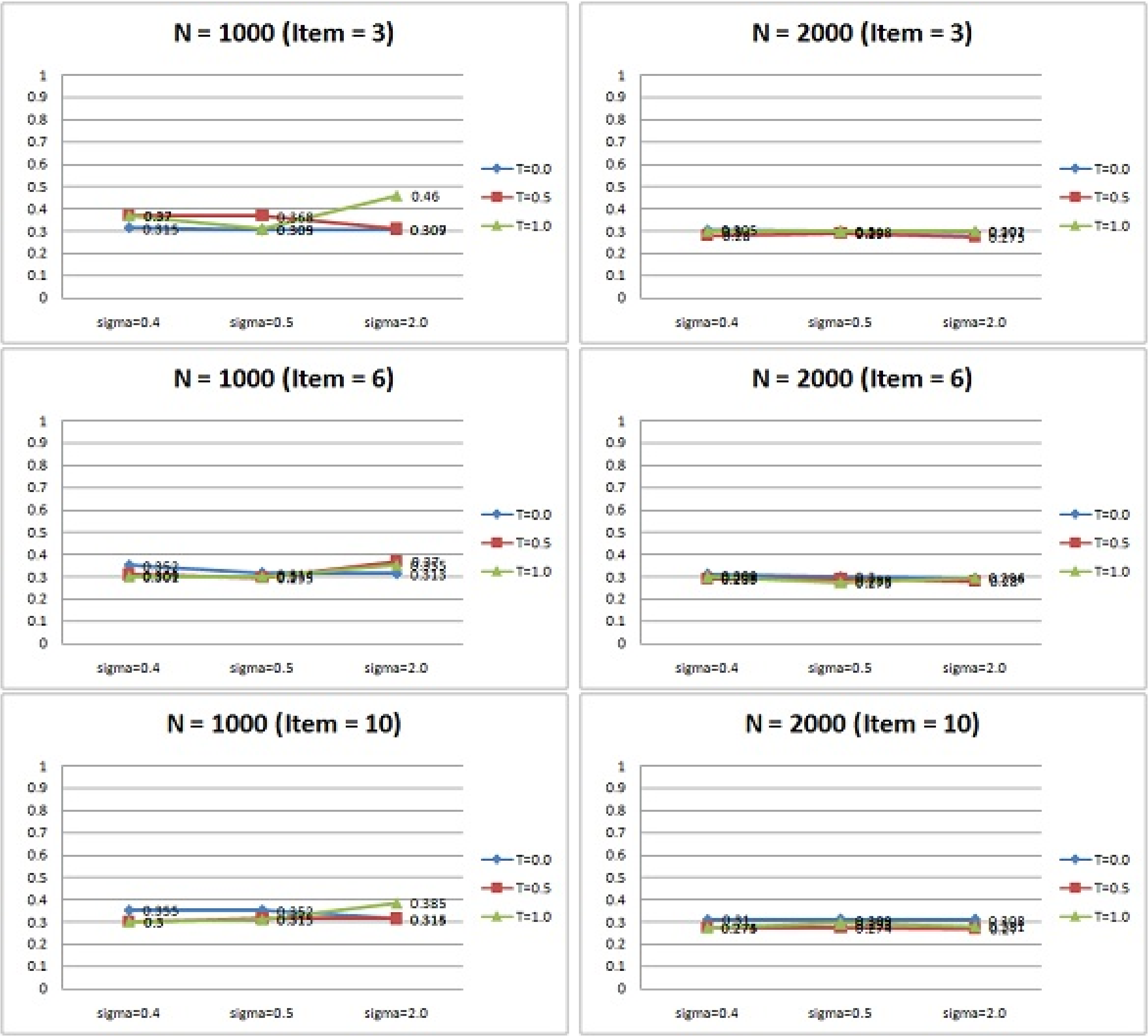

Figure 4.17: RMSE of Item Guessing Parameters with $\sigma^2_{\mu.} = 0.001$

other testlets having small number of items. Figure 4.18 showed that with the sample sizes 1000, lower RMSE of item discrimination occurred when the item discrimination parameter were estimated with $d.f. = 0.5$ under the condition of no testlet effect. When the testlet effects existed, lower RMSE values of item discrimination parameters happened when the item discrimination parameters were estimated with $d.f. = 2.0$. Similar results were also obtained when the item discrimination parameters were estimated with $\sigma^2_{\mu.} = 0.001$ (see

Figure 4.19). However, given the sample size ($N = 1000$), smaller RMSE values of item discrimination parameters occurred if those parameters were estimated with $\sigma^2_{\mu.} = 0.01$ under the condition of no testlet effect when compared RMSE values of item discrimination parameters between $\sigma^2_{\mu.} = 0.01$ and $\sigma^2_{\mu.} = 0.001$. However, if there were testlet effects, item discrimination parameters estimated with $\sigma^2_{\mu.} = 0.01$ had smaller RMSE values than those estimated with $\sigma^2_{\mu.} = 0.001$. When the sample sizes increased to 2000, no big difference between $\sigma^2_{\mu.} = 0.01$ and $\sigma^2_{\mu.} = 0.001$ occurred.

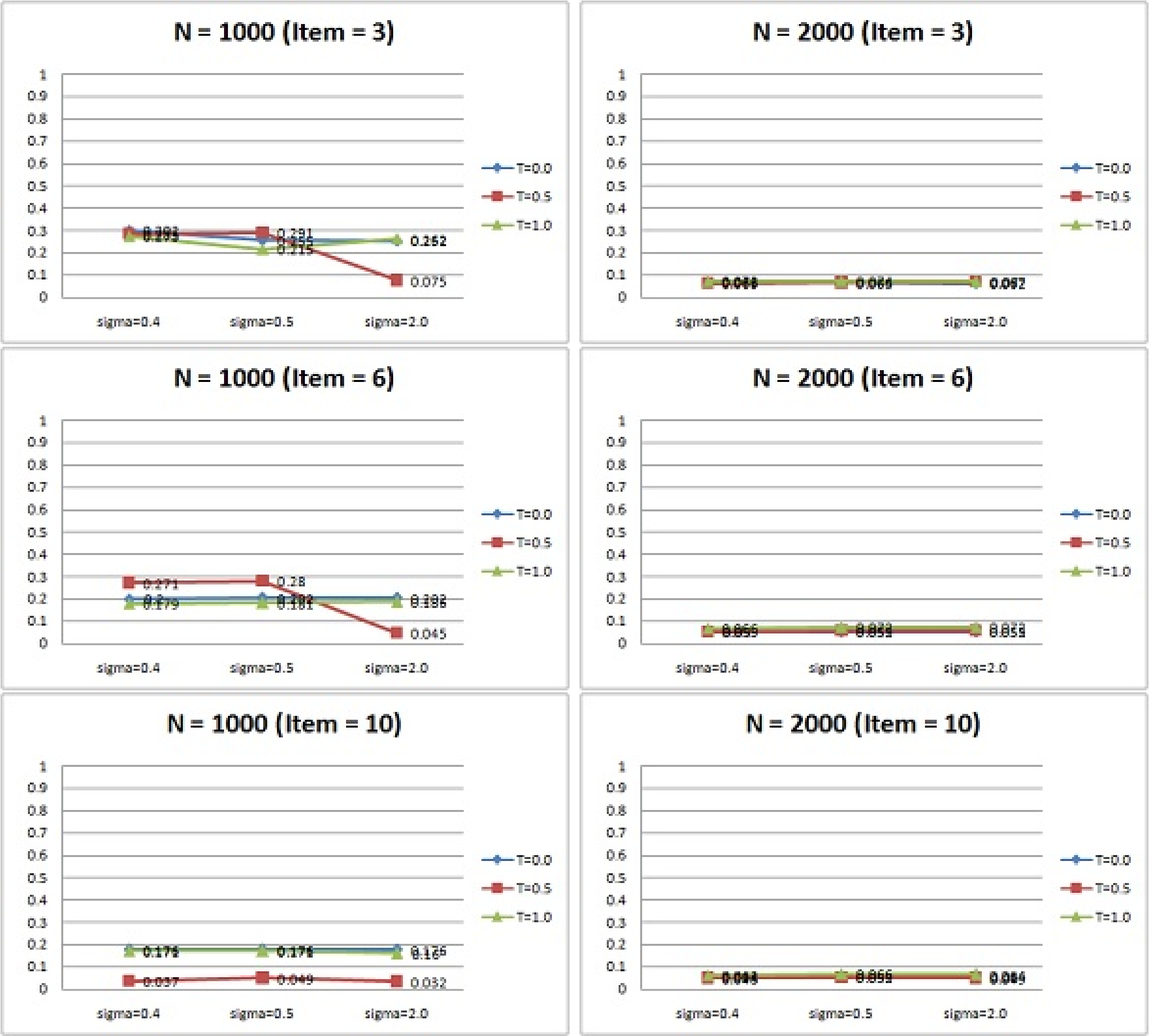

Figure 4.18: RMSE of Item Discrimination Parameters within Testlets with $\sigma^2_{\mu.} = 0.01$

Appendix A.5 and Figure 4.20 showed the patterns of RMSE of item difficulty parameters. In general, smaller RMSE values of item difficulty parameters were obtained when large

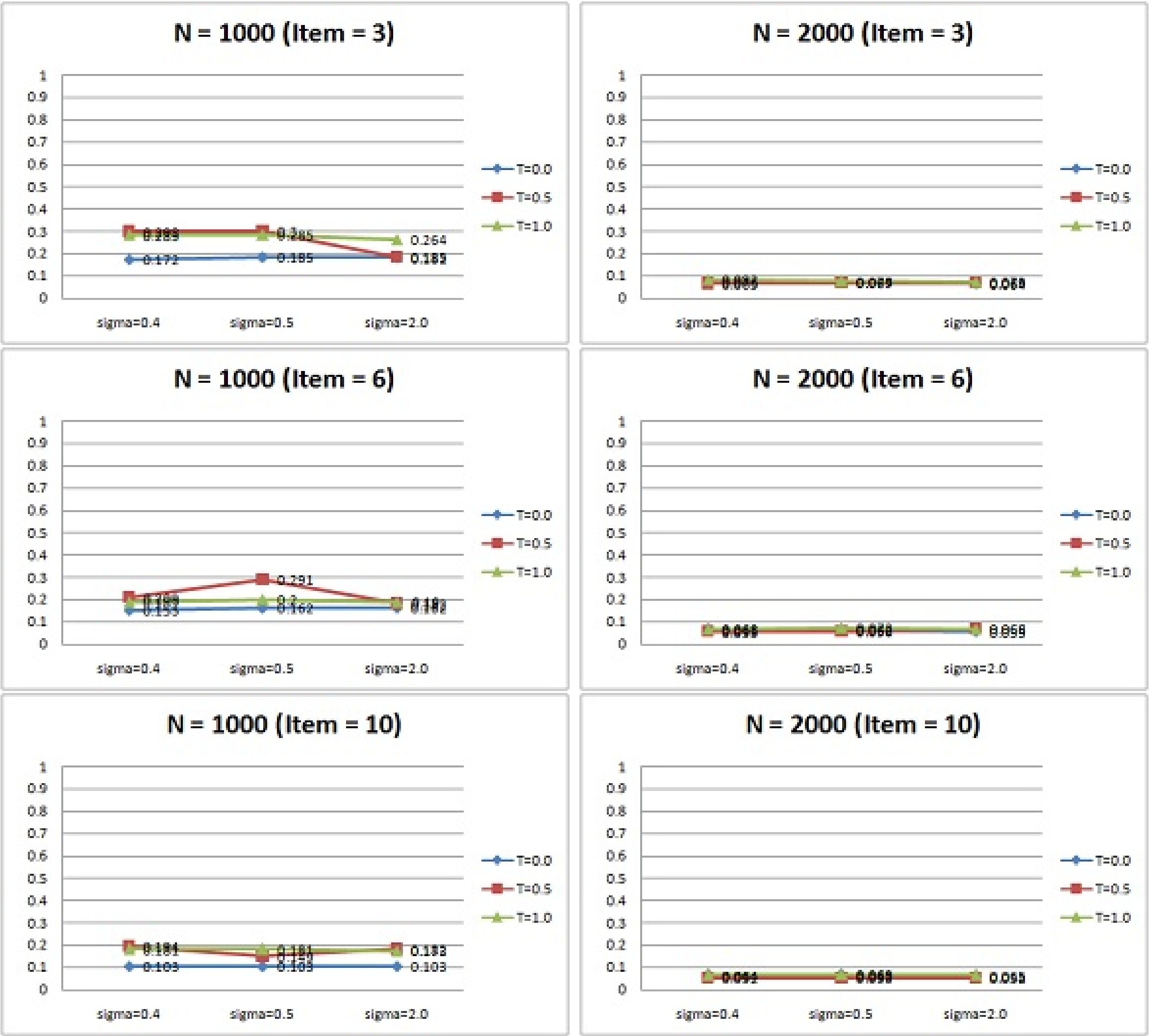

Figure 4.19: RMSE of Item Discrimination Parameters within Testlets with $\sigma^2_{\mu.} = 0.001$

number of items were nested in testlets. Lower RMSE values were obtained when the item difficulty parameters were estimated with $d.f. = 0.4$ across three different testlet effects. Given the sample sizes of 1000, smaller RMSE values of item difficulty parameters were obtained when the item difficulty parameters were estimated under no testlet effect. When there were testlet effects, lower RMSE values of item difficulty occurred when the item difficulty parameters were estimated with $d.f. = 2.0$. Similar results were also obtained when the item difficulty parameters were estimated with $\sigma^2_{\mu.} = 0.001$ (see Figure 4.21).

When values of RMSE were compared between $\sigma^2_{\mu.} = 0.01$ and $\sigma^2_{\mu.} = 0.001$, smaller RMSE values of item difficulty parameters occurred if those parameters were estimated with $\sigma^2_{\mu.} = 0.001$ across all conditions. However, when the sample sizes increased to 2000, relatively small RMSE of item difficulty parameters were obtained with $\sigma^2_{\mu.} = 0.01$.

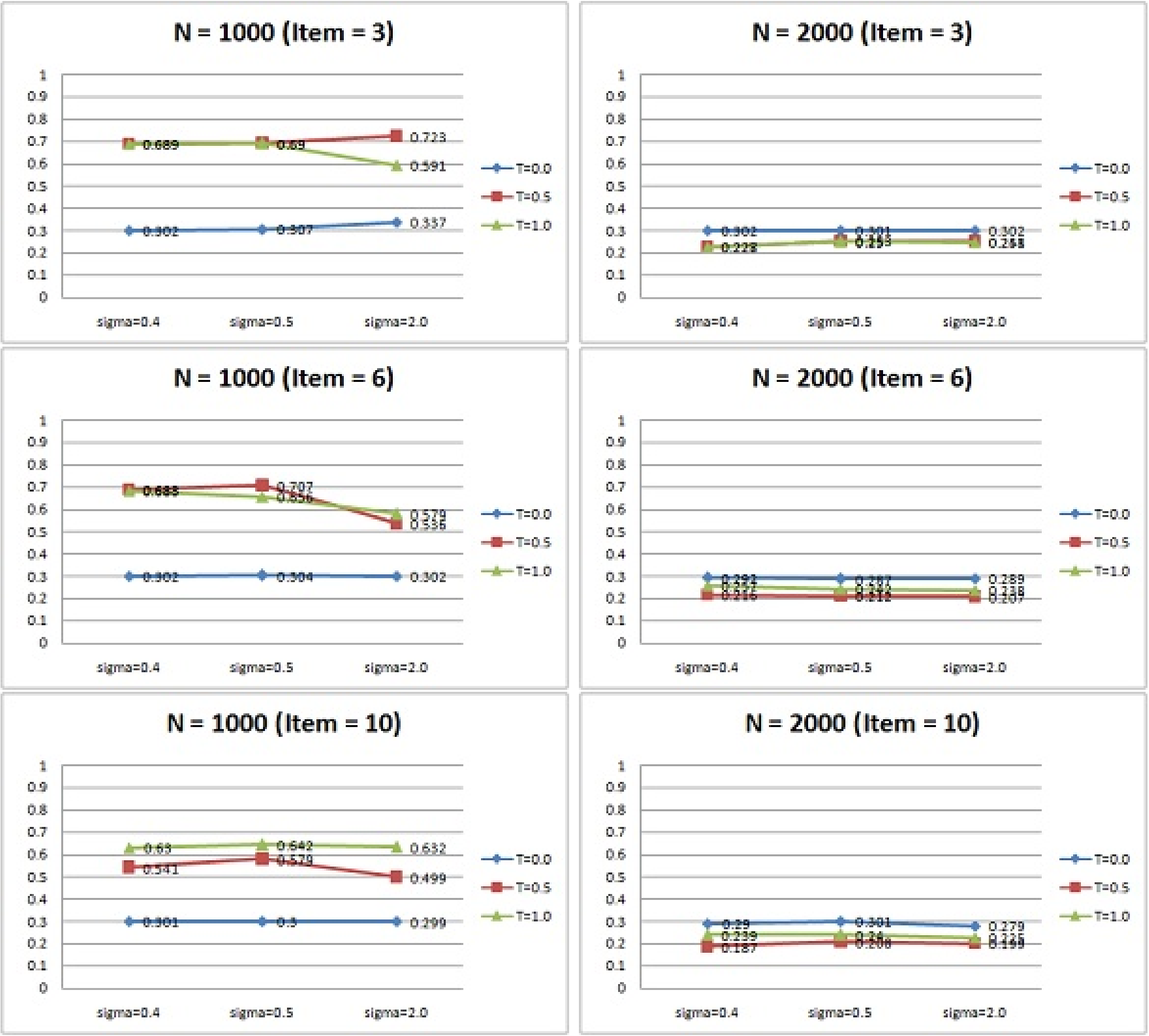

Figure 4.20: RMSE of Item Difficulty Parameters within Testlets with $\sigma^2_{\mu.} = 0.01$

With respect to item guessing parameters, Figure 4.22 showed the patterns of RMSE of item guessing parameters. The results showed that smaller RMSE values of item guessing

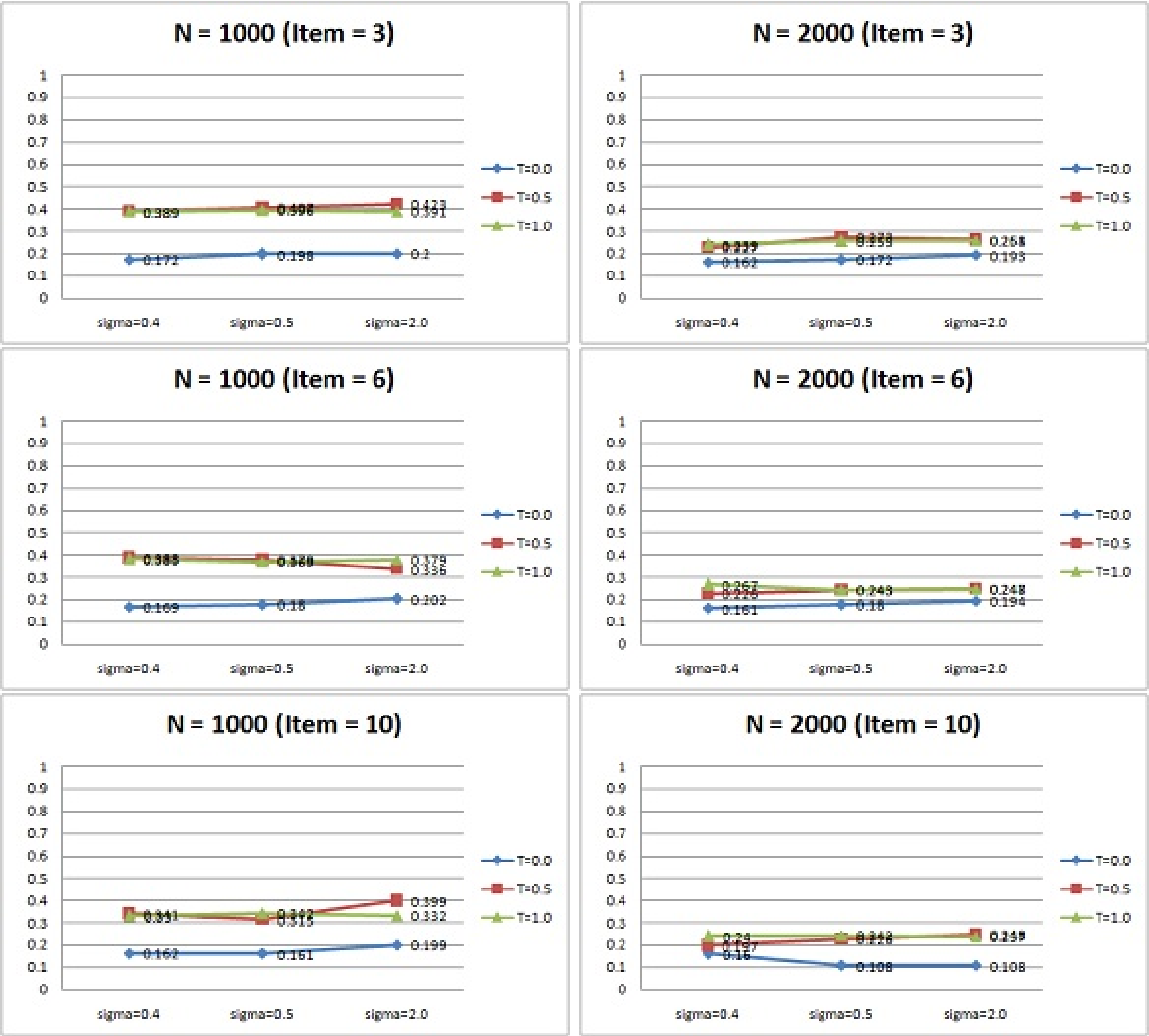

Figure 4.21: RMSE of Item Difficulty Parameters within Testlets with $\sigma^2_{\mu.} = 0.001$

parameters were obtained when large number of items were nested in testlets. Lower RMSE values were obtained when the item guessing parameters were estimated with $d.f. = 0.4$ across three different testlet effects in general. Given the sample sizes of 1000, smaller RMSE values of item guessing parameters were obtained when the item guessing parameters were estimated under the no testlet effect. When there were testlet effects, lower RMSE values of item guessing occurred when the item guessing parameters were estimated with $d.f. = 2.0$.

Similar results were also obtained when the item guessing parameters were estimated with $\sigma^2_{\mu.} = 0.001$ (see Figure 4.23).

When values of RMSE were compared between $\sigma^2_{\mu.} = 0.01$ and $\sigma^2_{\mu.} = 0.001$, relatively smaller RMSE values of item guessing parameters occurred if those parameters were estimated with $\sigma^2_{\mu.} = 0.01$. However, when the sample sizes increased to 2000, there were no big difference between $\sigma^2_{\mu.} = 0.01$ and $\sigma^2_{\mu.} = 0.001$.

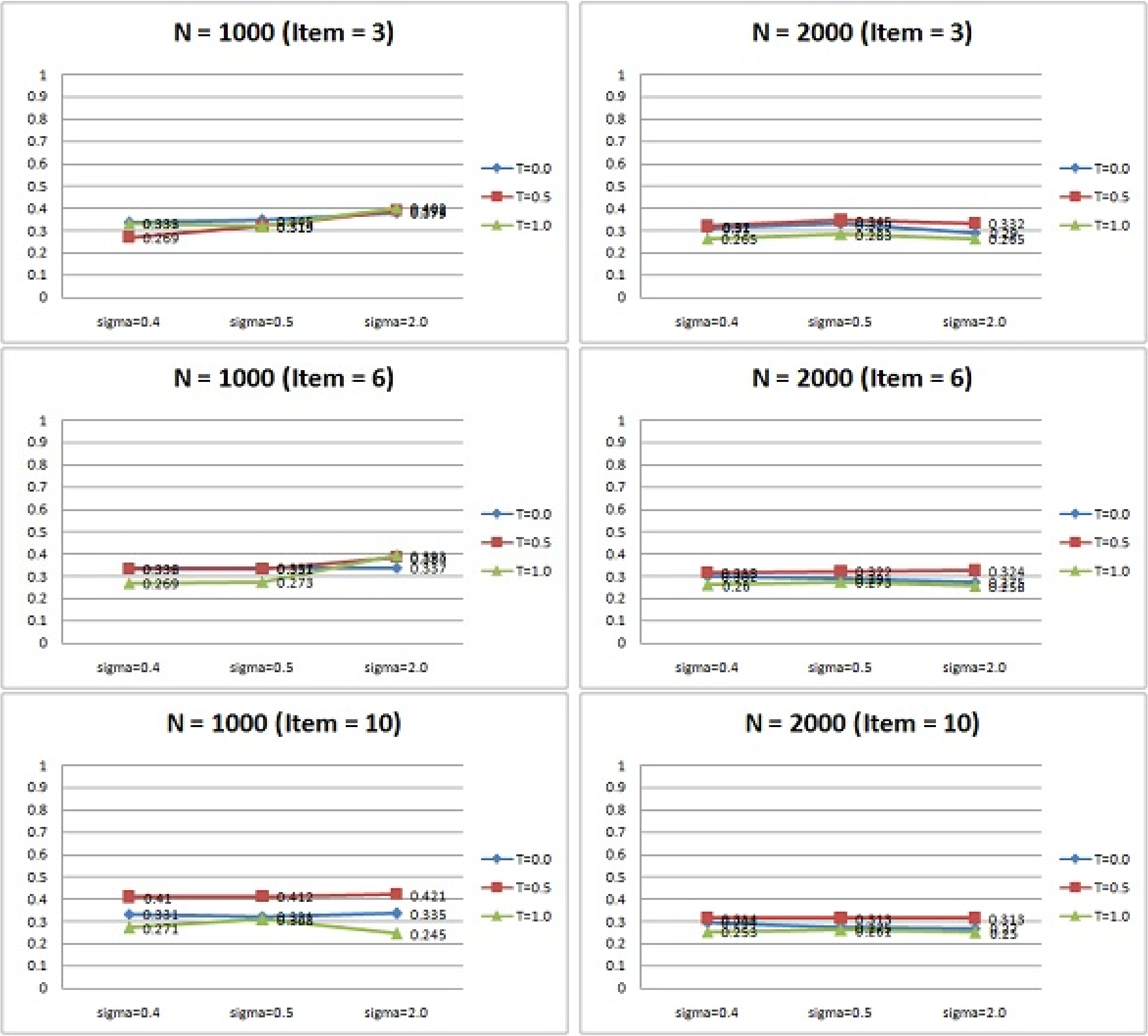

Figure 4.22: RMSE of Item Guessing Parameters within Testlets with $\sigma^2_{\mu.} = 0.01$

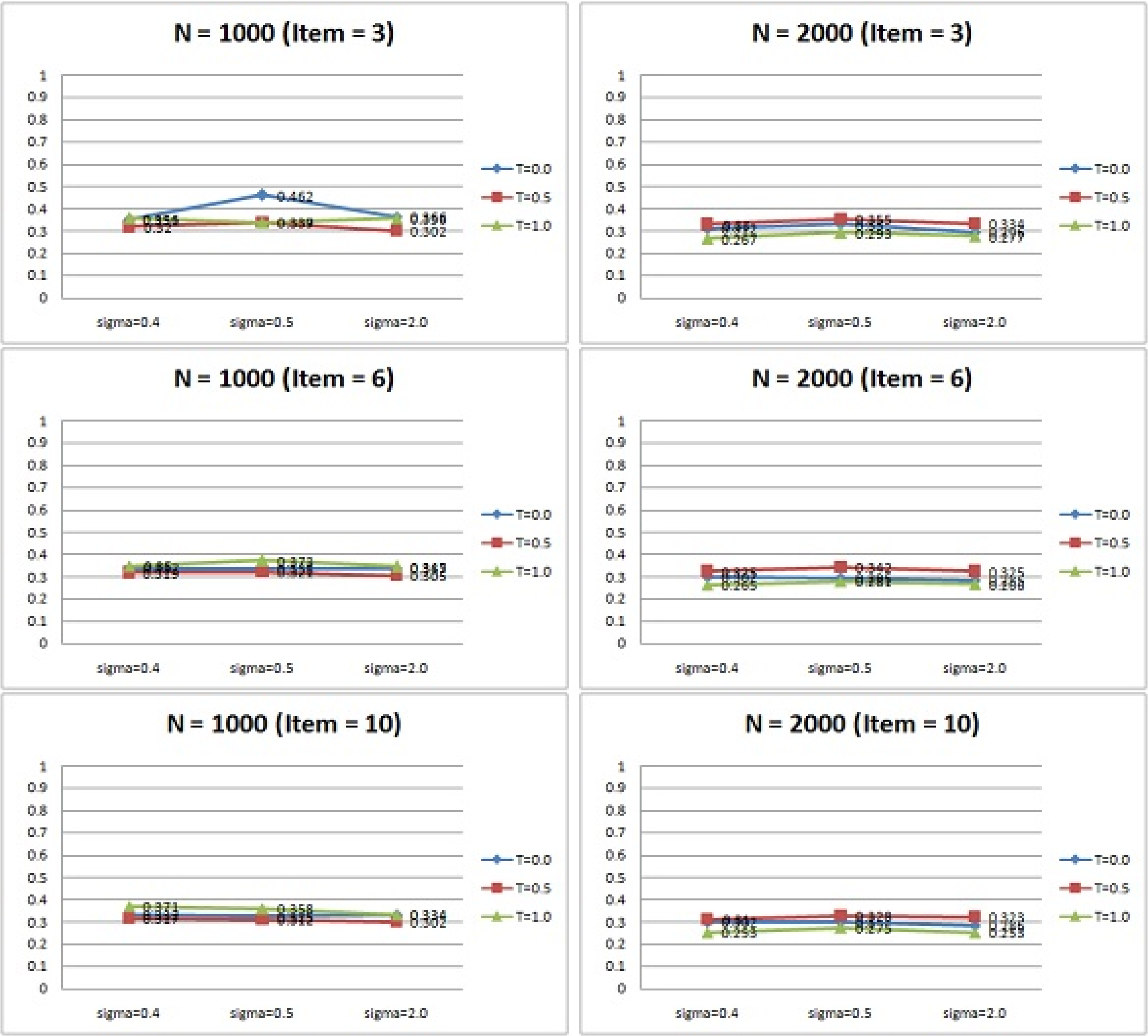

Figure 4.23: RMSE of Item Guessing Parameters within Testlets with $\sigma^2_{\mu.} = 0.001$

4.3 RESULTS OF REAL DATA ANALYSIS

The 2003 form of the Florida Comprehensive Assessment Test (FCAT) Reading Test for Grade 9 contains seven reading passages and 51 items. Each of these testlets consists of 6 to 9 items. The last 6 try-out items were discarded. The first six tests containing 45 items were used for this study. The six testelts were composed of six reading passages with 7, 9, 7, 8, 8, and 6 items, respectively. A sample of 1,000 examinees was randomly drawn from the total sample and used for this preliminary analysis.

4.3.1 CONVERGENCE

It is important to determine whether the Markov chain has reached its stationary distribution. If the chain does not converge, the simulated draws from this chain would not represent the posterior distribution of parameters of interest. The convergence diagnostics in preliminary results indicated that as many as 5,000 iterations are necessary to achieve convergence.

Gibbs and SCORIGHT. The Gibbs program does not provide any statistics to monitor convergence. However, SCORIGHT provides statistics as in post (i.e., posterior) and confshrink (i.e., confidence interval shrunk) for convergence when more than two chains are performed at the same time. Post statistic provides 2.5%, 50%, and 97.5% quantitles for the target distribution based on the Student-t distribution, whereas confshrink statistic, termed as the potential scale reduction, $\sqrt{\hat{R}}$, in Gelman and Rubin (1992), indicates how much estimated posterior intervals would shrink as the iterative simulations keep continuing and provides 97.5% quantiles of $\sqrt{\hat{R}}$ (Gelman & Rubin, 1992; Wang et al., 2004). $\sqrt{\hat{R}}$ is the square root of estimated variance divided by within chain variance, $\sqrt{\frac{V(\theta)}{W}}$, where V$(\theta) = (1 - \frac{1}{n})$ is estimated variance and W within chain variance (Gelman & Rubin, 1992; Wang et al., 2004). The value of confshrink should be around 1 which indicates reasonable convergence (Gelman & Rubin, 1992; Wang et al., 2004) because variation within the chain and variation between the chains should be equivalent. Otherwise, a longer iterations should be performed. The summary statistics of post and confshrink statistics with the real data are presented in Table 4.6.

WINBUGS. A number of convergence diagnostics such as plot history, autocorrelation plots, and the Gelman-Rubin statistic, R, from the WinBUGS program were also used to check convergence. It is easy to check the stability of simulated parameters by using the plot history, which shows the generated values of a parameter at each iteration in a chain

Table 4.6: Post and Confshrink Statistics from SCORIGHT

	5,000 iterations					15,000 iterations				
	Post			Confshrink		Post			Confshrink	
	2.5%	50%	97.5%	50%	97.5%	2.5%	50%	97.5%	50%	97.5%
a	−0.07	0.22	0.50	1.00	1.02	−0.08	0.22	0.52	1.00	1.00
b	−0.27	−0.13	0.02	1.03	1.11	0.27	−0.11	0.06	1.04	1.19
q	−1.86	−1.59	−1.32	1.57	3.00	−1.81	−1.56	−1.30	1.18	1.63
Testlet 1	0.10	0.21	0.31	1.27	1.99	0.11	0.21	0.31	1.01	1.02
Testlet 2	0.08	0.16	0.24	1.08	1.18	0.10	0.17	0.25	1.01	1.03
Testlet 3	0.15	0.33	0.52	1.23	1.78	0.21	0.34	0.48	1.00	1.02
Testlet 4	0.12	0.22	0.32	1.06	1.20	0.11	0.21	0.31	1.00	1.03
Testlet 5	0.14	0.22	0.32	1.01	1.04	0.11	0.21	0.30	1.01	1.02
Testlet 6	0.20	0.32	0.43	1.10	1.37	0.20	0.32	0.44	1.00	1.01

of sample values. The sample history of the first item estimated with a normal distribution is presented in Figure 4.24, in which two chains start from different values and then mix together. Similar results were observed for the other items. However, the sample history of the first item estimated with a log-normal distribution is shown in Figure 4.25, in which two chains start from the different values and then mix together quickly.

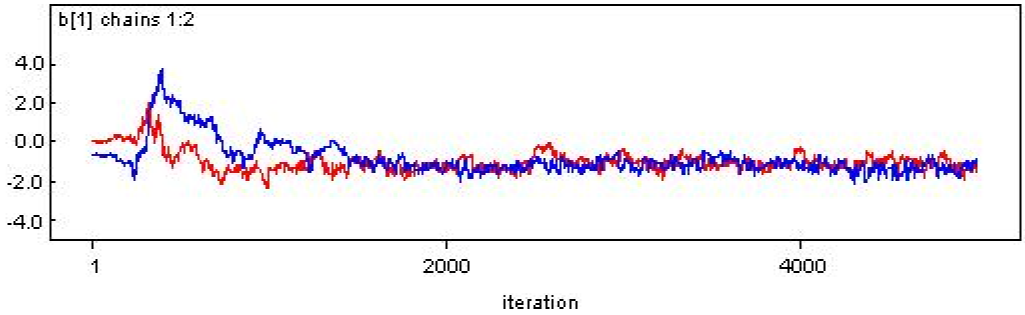

Figure 4.24: Sampling of History of the First Item with 5,000 Iterations under Normal Distribution

An autocorrelation plot shows correlation between each sequential draw of a parameter in a Markov chain. However, an autocorrelation plot does not evaluate convergence of MCMC

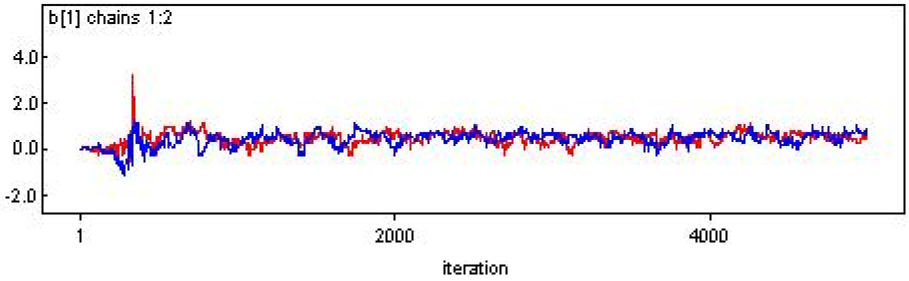

Figure 4.25: Sampling of History of the First Item with 5,000 Iterations under Lognormal Distribution

directly. Instead, an autocorrelation plot indirectly suggests appropriateness of MCMC convergence because autocorrelation causes inefficient MCMC. As shown in the autocorrelation plots in Figure 4.26, the autocorrelations for the first and the second chains decrease to nearly zero at 40. This indicates that the correlation between any two drawn values separated by independent. The second step parameter presents the worse case, in which the autocorrelation remains above 0.5 even at about lag 40. These high autocorrelations explain why the convergence is slow.

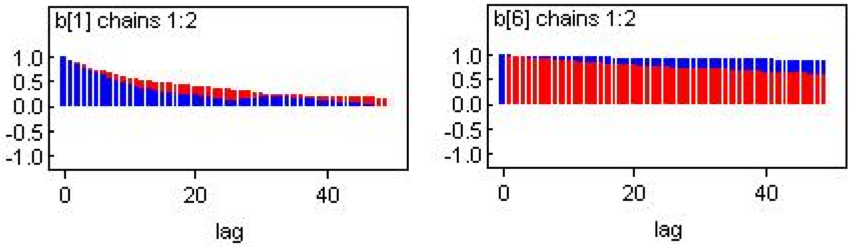

Figure 4.26: Autocorrelation of the First Item with 5,000 Iterations under Normal Distribution

The Gelman-Rubin convergence statistic (Brooks & Gelman, 1998) shows whether MCMC simulations reach stability by using multiple chains with different starting points. A value of R near 1 for all parameters of interest indicates that MCMC has converged. The green and blue lines reflect the pooled and within-chain posterior variance, respectively.

The ratio of these two variances is represented by the red line. The Gelman-Rubin convergence statistic plot of the first item was shown in Figure 4.27. The red line (Gelman-Rubin statistic) converged to 1, indicating equality between the pooled and within-chain variances. Thus, the Gelman-Rubin plot demonstrates that the convergence seems to occur around 4,200 iterations. However, the fluctuating red line might indicate the necessity of longer iterations.

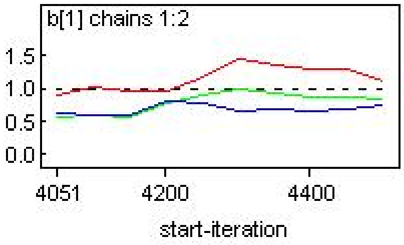

Figure 4.27: The Gelman-Rubin Convergence Statistic of the First Item with 5,000 Iterations under Normal Distribution

A smoothed kernel density is estimated for the posterior distributions. Figure 4.28 showed the density plots of the first and the sixth items based on the initial 5,000 iterations from two chains for the difficulty parameter, b. The density plot for b of the first item showed unimodal distribution which is nearly symmetric and close to the normal distribution. However, other items does not show approximate symmetric density plots. Item 6, for instance, showed bi-modal density distribution which might suggest the necessity of longer chains.

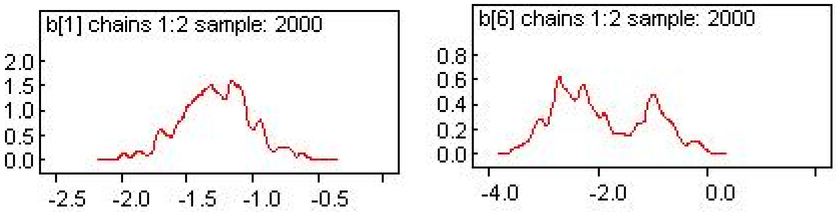

Figure 4.28: A Kernel Density of the First Item with 5,000 Iterations under Normal Distribution

4.3.2 ESTIMATION

Prior to data analysis, the first 5,000 iterations of each chain were discarded as burn-in periods for the initial 5,000 iterations. Point estimates of the model parameters and standard errors were computed, respectively, of 20,000 iterations (10,000 iterations for each chain) sampled from each parameter's marginal posterior distribution after burn-in periods. Once the convergence of the model was checked, parameters obtained from SCORIGHT, Gibbs, and WinBUGS were compared to one another. For instance, correlations between item parameter estimates from SCORIGHT and WinBUGS were .94 for item discrimination, .99 for item difficulty, and .33 for pseudo-guessing. The results in Table A.7, Table A.8, and Table A.9 showed that the estimated discrimination, difficulty, and pseudo-guessing parameters for all three methods were slightly divergent: Values of original pseudo-guessing parameters in the WinBUGS program were used instead of transformed pseudo-guessing parameters. As a number of iterations increased, correlations among item parameters from the three programs were getting higher. The summary statistics of the estimates from the SCORIGHT, Gibbs, and WinBUGS runs of the real data also presented in the Table 4.7 and Table 4.8.

An individual's response to the items not only depend on an individual's ability and item difficulty, but also on additional random testlet effects, which are assumed to be normally distributed with a mean of zero and a variance, σ^2_γ. A testlet effect with a similar magnitude to the variance of the corresponding latent variable means that the variance associated with LD is of the same order of magnitude as the variance of individuals.

In the line of recommendation for testlet effect (Bradlow et al., 1999), testlet effects among six testlets were moderately significant, which confirmed that items on a test violated the local independence assumption. However, slightly different testlet effects obtained from the three programs were shown in Table A.10. Results obtained from SCORIGHT and Gibbs showed that the third testlet had the largest testlet effect, whereas the largest testlet effect existed at the last testlet in WinBUGS.

Noninformative prior distributions specified as $\mu_a \sim N(0, 100^2)$, $\mu_b \sim N(0, 100^2)$, and $\mu_q \sim N(0, 100^2)$ for item discrimination, item difficulty, and pseudo-guessing parameters, respectively appeared to perform well in the 3PLT model reflecting a half normal distribution on item discrimination parameter, $a \sim N(\mu_a, \sigma_a)I(0,)$. In the WinBUGS code, the variance, $\sigma^2 = 100^2$, designate $\tau = 0.0001$. As the sample size of this preliminary study was relatively small (N=1,000), informative prior distributions seemed to be imposed for the item discrimination parameters. It accelerated the WinBUS runs when informative long-normal distribution for item discrimination, $a \sim LN(\mu_a, \sigma_a^2)$.

The results of the analyses were as expected. The selection of prior distributions in 3PLT model affected the estimation of item parameters as well as model convergence. It was noteworthy that the 3PLT model in the WinBUGS 1.4 program needed either several chains or relatively longer iterations as Sinharay (2004) suggested.

Table 4.7: Summary Statistics of the SCORIGHT, Gibbs, WinBUGS Estimates with 5,000 iterations

									WinBUGS							
	Gibbs				SCORIGHT				Normal				Lognormal			
	a_i	b_i	c_i	$\gamma_{jt(i)}$	a_i	b_i	c_i	$\gamma_{jt(i)}$	a_i	b_i	c_i	$\gamma_{jt(i)}$	a_i	b_i	c_i	$\gamma_{jt(i)}$
Min	.459	−1.609	.109	.120	.594	−1.736	.153	.173	.538	−1.618	.106	.146	.677	−1.620	.105	.138
Max	2.382	2.142	.329	.289	2.712	2.261	.216	.344	2.460	2.105	.317	.297	2.706	2.173	.315	.295
Mean	1.436	−.058	.201	.186	1.354	−.099	.176	.244	1.420	−.072	.198	.210	1.464	−.028	.196	.209
SD	.454	.891	.051	.078	.468	.931	.015	.070	.442	.883	.054	.067	.475	.897	.050	.068

Table 4.8: Summary Statistics of the SCORIGHT, Gibbs, WinBUGS Estimates with 15,000 iterations

									WinBUGS							
	Gibbs				SCORIGHT				Normal				Lognormal			
	a_i	b_i	c_i	$\gamma_{jt(i)}$	a_i	b_i	c_i	$\gamma_{jt(i)}$	a_i	b_i	c_i	$\gamma_{jt(i)}$	a_i	b_i	c_i	$\gamma_{jt(i)}$
Min	.631	−1.650	.098	.140	.563	−1.778	.164	.151	.539	−1.612	.109	.129	.636	−1.644	.114	.143
Max	2.457	2.093	.366	.281	2.593	2.292	.209	.313	2.418	2.126	.328	.294	2.753	2.117	.324	.289
Mean	1.446	−.051	.195	.197	1.336	−.109	.181	.226	1.431	−.050	.193	.207	1.433	−.066	.205	.207
SD	.433	.885	.064	.062	.461	.954	.012	.060	.439	.884	.056	.063	.479	.889	.050	.058

Chapter 5

DISCUSSION

The previous chapter reported the results of the studies conducted in this dissertation. This chapter included a summary of the findings from these studies and a discussion of their significance. The chapter closed by presenting some of the study limitations and suggesting directions for future research.

The two primary goals of this research are: First, to conduct simulation studies to investigate the impacts of means of particular prior distributions and different degrees of freedom under the context of the 3PLT model under different number of sample sizes; and second, to apply the 3PLT model to the empirical data sets.

5.1 Summaries

Bayesian estimation using MCMC methods offer lots of potential for estimation of complex IRT models such as testlet models (Bradlow et al., 1999; Li et al., 2005; Wang, 2002). The advantage of the testlet model over the standard IRT models is that the former can provide a quantitative idea about the dependence of the response of an examinee to the items within the testlets. The 3PLT in the WinBUGS runs requires either several chains or relatively longer iterations.

Based on the simulation study, convergence was slow for conditions in which relatively small degrees of freedom was placed on testlets containing 10 items. It might be due to the fact that there was not much information in estimating testlet effect. In addition, the findings indicated that when a small number of items were nested in testlets in a test (large number of testlets), convergence rate was relatively faster compared to one containing large number of items within small number of testlets.

The 3PLT model assumes that parameters for testlet follow a normal distribution $\mathrm{N}(0, \gamma(g))$. The magnitudes of the testlet effects are determined by the variance of the testlet parameters. In the simulation study, the true testlet parameter were specified to be normal distributed. Estimates of the testlet variance were overestimated when there was no testlet effect. However, Estimates of the testlet tended to be an underestimates if there were testlet effects. The results showed that the same testlets in different means of prior distributions exhibited different magnitudes of testlet effects. While it is hard to illuminate the change of testlet effect across different means of prior distributions by studying item responses only, analyzing the content of these testlets may shed some light on these shifts.

Before examining the success of the 3PLT model in recovering of the true parameters, this study examined correlations between the generating parameter (true) and the estimated parameters. The average values of correlations across 25 replications are .92 for θ, .94 for a, .93 for b, and 0.45 for q, which indicates that estimation process had added small amount of error to the estimates. The results showed slightly lower correlations between the estimated parameters and the true parameters.

Additionally, RMSE for each condition was computed for 25 replications to evaluate the success of recovering the true parameters. In general, the variances of testlet effect tended to be underestimated across all conditions, yielding smaller testlet effects than it should have. However, it should be noted that the tendency of underestimating testlet effects reversed for the condition, which assumed no testlet effects. The tendency displayed minor testlet effects when there was no testlet effect.

It is common to use reasonably non-informative prior distributions about the mean and the variance of the random effects. However, the important thing is to distinguish primary parameters of interest in which one may want minimal influence of priors from the secondary structure used for smoothing in which either moderate or strong informative priors may be more acceptable. However, great caution should be considered in complex models.

5.2 Limitations of the Study

This study explored the impacts of means of prior distributions, different degrees of freedom, different number of items nested in testlets on testlet effects within the 3PLT model. Though the conditions were carefully designed and the factors were fixed at realistic values, the results obtained from this study cannot be generalized beyond the conditions studied here. For example, this study was limited in terms of the means of prior distributions, degrees of freedom, and number of items nested in testlets because all those conditions were applied across all parameters under the same number of sample sizes ($N = 1000$ and $N = 2000$). The relative differences among means of prior distributions and different degrees of freedom could vary more drastically depending on conditions such as sample size and number of items. A lack of information on σ_a^2, σ_b^2, and σ_q^2 had been imposed with inverse-chi square distribution. However, it is necessary to clarify difference between inverse-gamma distribution and gamma distribution used in other contexts (Spiegelhalter et al., 2003) .

By including a set of person-testlet interaction parameters in addition to the usual item and person parameters, the testlet models are able to account for the testlet effects which have been ignored by the traditional unidimensional IRT. However, the performance of the particular means of prior distributions and different degrees of freedom for the 3PLT model requires further study. For example, the effect of factors such as different prior distributions and different degrees could be further explored. Regarding the testlet effect, the practice of assuming normal distributions for testlet parameters has almost been exclusively applied by researchers in their specification and estimation of testlet models (Bradlow et al., 1999; Wainer et al., 2000; Wainer et al., 2007; Wainer & Wang, 2000). Although this is a generally accepted practice, there is no guarantee that the true testlet parameters are normally distributed universally for different tests that target different content domains and examinees in real life. The discrepancies between the assumed testlet parameters distribution and the true testlet parameter distribution can lead to inaccuracies in model. Therefore, it is recommended to study the behavior of the testlet effects parameters and investigate the appropriateness testlet response models that employ different testlet parameter distributions.

In this study, the group invariance property with the real data was not evaluated. It should be noted that test equating results may be different for various populations. In this study, the examinee population taking the FCAT was the same population but abilities of those examinees were estimated by using different programs (e.g., Gibbs, SCORIGHT, WinBUGS).

Another limitation of the current study lies in the data generation method for the simulation study. In order to keep the generated discrimination and difficulty parameters within the range of the FCAT items, samples were discarded until all the obtained parameters fell within the rage of real item parameters. As a result, the final data samples were not randomly generated in a strict sense.

Running the WinBUGS program is highly computation-intensive. Due to to computing constraints of the WinBUGS 1.4 program (Spiegelhalter et al., 2003), only 25 replications were implemented. Since an average run took about 6 hours with the sample size of 1000 and 14 hours with the sample size of 2000 under the conditions studied here, simulation research, which typically requires large number of replications, faces even greater computing challenges. Though it was smaller than which is typical for other Monte Carlo research, it was larger compared to previous research involving the 3PLT model.

Appendix A

Table A.1: Average RMSE of Variances of Testlet

		$N = 1000$			$N = 2000$		
$\sigma^2_{\mu.} = 0.01$							
$d.f.$	Items	0	0.5	1.0	0	0.5	1.0
0.4	3	0.146	0.147	0.452	0.124	0.128	0.257
	6	0.156	0.154	0.291	0.143	0.117	0.254
	10	0.196	0.136	0.271	0.184	0.097	0.238
0.5	3	0.178	0.135	0.292	0.177	0.113	0.238
	6	0.187	0.132	0.252	0.177	0.112	0.244
	10	0.195	0.097	0.234	0.186	0.093	0.229
2.0	3	0.177	0.230	0.292	0.148	0.114	0.269
	6	0.175	0.210	0.252	0.175	0.098	0.215
	10	0.198	0.097	0.242	0.179	0.082	0.210
$\sigma^2_{\mu.} = 0.001$							
0.4	3	0.165	0.263	0.322	0.142	0.131	0.273
	6	0.153	0.262	0.245	0.103	0.118	0.245
	10	0.137	0.265	0.271	0.101	0.117	0.225
0.5	3	0.298	0.266	0.249	0.176	0.119	0.249
	6	0.285	0.266	0.231	0.173	0.118	0.230
	10	0.223	0.253	0.228	0.171	0.114	0.197
2.0	3	0.201	0.255	0.267	0.167	0.118	0.248
	6	0.167	0.243	0.245	0.154	0.117	0.218
	10	0.164	0.158	0.215	0.151	0.119	0.152

Table A.2: Average RMSE Item Discrimination Parameter

		$N = 1000$			$N = 2000$		
$\sigma^2_{\mu.} = 0.01$							
$d.f.$	Items	0	0.5	1.0	0	0.5	1.0
0.4	3	0.210	0.168	0.165	0.056	0.056	0.057
	6	0.179	0.142	0.138	0.049	0.046	0.057
	10	0.166	0.137	0.163	0.049	0.049	0.061
0.5	3	0.218	0.164	0.163	0.054	0.056	0.056
	6	0.184	0.151	0.148	0.048	0.041	0.057
	10	0.118	0.148	0.162	0.052	0.054	0.060
2.0	3	0.217	0.172	0.168	0.053	0.058	0.059
	6	0.184	0.140	0.145	0.048	0.050	0.057
	10	0.165	0.132	0.153	0.047	0.047	0.060
$\sigma^2_{\mu.} = 0.001$							
0.4	3	0.170	0.183	0.211	0.050	0.056	0.060
	6	0.108	0.184	0.205	0.049	0.056	0.058
	10	0.107	0.185	0.196	0.049	0.059	0.059
0.5	3	0.166	0.174	0.230	0.055	0.058	0.064
	6	0.159	0.183	0.229	0.048	0.058	0.076
	10	0.138	0.174	0.184	0.049	0.049	0.060
2.0	3	0.166	0.182	0.211	0.054	0.059	0.060
	6	0.157	0.174	0.186	0.048	0.058	0.063
	10	0.155	0.182	0.179	0.048	0.058	0.058

Table A.3: Average RMSE Item Difficulty Parameter

			$N = 1000$			$N = 2000$	
$\sigma^2_{\mu .} = 0.01$							
$d.f.$	Items	0	0.5	1.0	0	0.5	1.0
0.4	3	0.392	0.493	0.472	0.303	0.372	0.370
	6	0.391	0.494	0.492	0.300	0.343	0.360
	10	0.393	0.389	0.390	0.306	0.369	0.361
0.5	3	0.394	0.393	0.393	0.302	0.357	0.395
	6	0.394	0.408	0.509	0.288	0.358	0.395
	10	0.421	0.421	0.497	0.287	0.353	0.343
2.0	3	0.388	0.447	0.446	0.298	0.395	0.395
	6	0.394	0.399	0.433	0.288	0.372	0.395
	10	0.395	0.421	0.387	0.302	0.361	0.372
$\sigma^2_{\mu .} = 0.001$							
0.4	3	0.380	0.386	0.385	0.305	0.370	0.370
	6	0.365	0.342	0.383	0.304	0.308	0.360
	10	0.364	0.337	0.363	0.308	0.280	0.361
0.5	3	0.315	0.346	0.363	0.314	0.342	0.360
	6	0.375	0.381	0.348	0.283	0.370	0.360
	10	0.353	0.348	0.362	0.305	0.280	0.360
2.0	3	0.348	0.372	0.368	0.294	0.300	0.361
	6	0.377	0.340	0.375	0.284	0.287	0.360
	10	0.372	0.332	0.393	0.300	0.277	0.359

Table A.4: Average RMSE Item Guessing Parameter

		$N = 1000$			$N = 2000$		
$\sigma^2_{\mu.} = 0.01$							
$d.f.$	Items	0	0.5	1.0	0	0.5	1.0
0.4	3	0.312	0.365	0.365	0.293	0.255	0.245
	6	0.311	0.494	0.497	0.288	0.317	0.282
	10	0.309	0.460	0.458	0.291	0.329	0.290
0.5	3	0.354	0.465	0.465	0.302	0.314	0.305
	6	0.352	0.490	0.458	0.291	0.321	0.298
	10	0.484	0.480	0.312	0.360	0.342	0.300
2.0	3	0.402	0.500	0.501	0.338	0.324	0.289
	6	0.296	0.470	0.345	0.291	0.346	0.285
	10	0.390	0.498	0.320	0.291	0.365	0.283
$\sigma^2_{\mu.} = 0.001$							
0.4	3	0.315	0.370	0.370	0.305	0.280	0.300
	6	0.352	0.309	0.301	0.309	0.289	0.298
	10	0.355	0.300	0.300	0.310	0.275	0.274
0.5	3	0.305	0.368	0.309	0.300	0.290	0.298
	6	0.314	0.295	0.300	0.300	0.288	0.273
	10	0.352	0.315	0.310	0.309	0.274	0.293
2.0	3	0.307	0.309	0.460	0.297	0.275	0.301
	6	0.313	0.370	0.355	0.294	0.280	0.296
	10	0.315	0.314	0.385	0.308	0.270	0.281

Table A.5: Average RMSE of Item Estimates for Items Within Testlet When $\sigma^2_{\mu.} = 0.01$

Item	♯ of	Var$\gamma = 0$			Var$\gamma = 0.5$			Var$\gamma = 1.0$		
Parameter	Items	$d.f. = 0.4$	$d.f. = 0.5$	$d.f. = 2.0$	$d.f. = 0.4$	$d.f. = 0.5$	$d.f. = 2.0$	$d.f. = 0.4$	$d.f. = 0.5$	$d.f. = 2.0$
N= 1000										
α	3	0.302	0.255	0.252	0.285	0.291	0.075	0.273	0.215	0.262
	6	0.200	0.202	0.202	0.271	0.280	0.045	0.179	0.181	0.186
	10	0.176	0.176	0.176	0.037	0.049	0.032	0.171	0.171	0.160
β	3	0.302	0.307	0.337	0.689	0.690	0.723	0.689	0.690	0.591
	6	0.302	0.304	0.302	0.688	0.707	0.536	0.683	0.656	0.579
	10	0.301	0.300	0.299	0.541	0.579	0.499	0.630	0.642	0.632
q	3	0.335	0.345	0.376	0.269	0.319	0.392	0.332	0.319	0.402
	6	0.336	0.337	0.337	0.332	0.331	0.384	0.269	0.273	0.393
	10	0.331	0.321	0.335	0.410	0.412	0.421	0.271	0.308	0.245
N= 2000										
α	3	0.066	0.064	0.062	0.063	0.065	0.067	0.072	0.074	0.070
	6	0.055	0.052	0.052	0.050	0.055	0.052	0.066	0.072	0.065
	10	0.054	0.055	0.050	0.048	0.052	0.049	0.062	0.066	0.064
β	3	0.302	0.301	0.302	0.227	0.253	0.251	0.228	0.250	0.248
	6	0.297	0.287	0.289	0.216	0.212	0.207	0.257	0.242	0.238
	10	0.290	0.301	0.279	0.187	0.208	0.199	0.239	0.240	0.225
q	3	0.310	0.329	0.290	0.320	0.345	0.332	0.265	0.283	0.265
	6	0.302	0.290	0.275	0.318	0.322	0.324	0.260	0.273	0.256
	10	0.294	0.275	0.270	0.314	0.313	0.313	0.253	0.261	0.250

Table A.6: Average RMSE of Item Estimates for Items Within Testlet When $\sigma^2_{\mu.} = 0.001$

Item	♯ of	Varγ = 0			Varγ = 0.5			Varγ = 1.0		
Parameter	Items	$d.f. = 0.4$	$d.f. = 0.5$	$d.f. = 2.0$	$d.f. = 0.4$	$d.f. = 0.5$	$d.f. = 2.0$	$d.f. = 0.4$	$d.f. = 0.5$	$d.f. = 2.0$
N= 1000										
α	3	0.172	0.185	0.182	0.298	0.300	0.185	0.283	0.285	0.264
	6	0.153	0.162	0.162	0.209	0.291	0.182	0.189	0.200	0.190
	10	0.103	0.103	0.103	0.194	0.149	0.182	0.181	0.181	0.173
β	3	0.172	0.198	0.200	0.389	0.407	0.423	0.389	0.396	0.391
	6	0.169	0.180	0.202	0.388	0.379	0.336	0.383	0.369	0.379
	10	0.162	0.161	0.199	0.341	0.315	0.399	0.330	0.342	0.332
q	3	0.351	0.462	0.366	0.320	0.339	0.302	0.356	0.337	0.356
	6	0.337	0.338	0.337	0.319	0.321	0.305	0.350	0.373	0.345
	10	0.332	0.325	0.334	0.317	0.312	0.302	0.371	0.358	0.334
N= 2000										
α	3	0.076	0.074	0.064	0.065	0.067	0.069	0.082	0.075	0.073
	6	0.065	0.062	0.055	0.055	0.056	0.066	0.068	0.073	0.068
	10	0.064	0.065	0.054	0.052	0.053	0.052	0.064	0.069	0.065
β	3	0.313	0.326	0.303	0.237	0.273	0.261	0.239	0.255	0.258
	6	0.312	0.282	0.294	0.226	0.243	0.247	0.267	0.243	0.248
	10	0.311	0.308	0.308	0.197	0.226	0.248	0.240	0.242	0.237
q	3	0.312	0.330	0.296	0.330	0.355	0.334	0.267	0.293	0.277
	6	0.302	0.295	0.285	0.325	0.342	0.325	0.265	0.281	0.266
	10	0.302	0.300	0.286	0.310	0.328	0.323	0.253	0.275	0.253

Table A.7: Correlation of Item Discrimination Parameter

Condition		Du	SC	N	LN
	Du	1.000	.898	.987	.978
5,000	SC		1.00	.939	.936
iterations	N			1.000	.989
	LN				1.000
2	Du	1.000	.947	.998	.993
15,000	SC		1.00	.947	.958
iterations	N			1.000	.991
	LN				1.000

Table A.8: Correlation of Item Difficulty Parameter

Condition		Du	SC	N	LN
	Du	1.000	.984	.995	.994
5,000	SC		1.00	.990	.992
iterations	N			1.000	.998
	LN				1.000
2	Du	1.000	.991	1.000	1.000
15,000	SC		1.00	.947	.992
iterations	N			1.000	.999
	LN				1.000

Table A.9: Correlation of guessing parameter

Condition		Du	SC	*N*	*LN*
5,000 iterations	Du	1.000	.213	.915	.896
	SC		1.00	.218	.326
	N			1.000	.959
	LN				1.000
2 15,000 iterations	Du	1.000	.324	.989	.987
	SC		1.00	.242	.314
	N			1.000	.979
	LN				1.000

Table A.10: Estimated Variance of Testlet

Cond	Testlet	Du	SC	*N*	*LN*
5,000 iterations	1	.144	.211	.174	.193
	2	.140	.151	.129	.140
	3	.281	.311	.272	.264
	4	.162	.207	.194	.183
	5	.185	.199	.180	.170
	6	.268	.286	.294	.290
15,000 iterations	1	.143	.213	.175	.178
	2	.120	.173	.146	.137
	3	.290	.344	.297	.294
	4	.133	.209	.169	.167
	5	.150	.206	.179	.181
	6	.281	.321	.291	.295

Appendix B

3PL TESTLET N = 2000 55 ITEMS DATA GENERATION T 0.5 SEED:123456

```
    MODEL {
    a[1] ← 0.521172928
    a[2] ← 0.077579014
    a[3] ← 0.886715500
    a[4] ← 1.676318913
    a[5] ← 2.805824457
             ⋮
    a[51] ← 1.58147946
    a[52] ← 1.64617332
    a[53] ← 2.15516394
    a[54] ← 1.44434500
    a[55] ← 1.57411228
    b[1] ← -1.13956809
    b[2] ← -0.54981831
    b[3] ← 0.41384308
    b[4] ← 1.23229356
    b[5] ← 0.06482972
            ⋮
    b[51] ← -0.96948734
    b[52] ← 0.96107453
```

```
b[53] ← -2.45063650
b[54] ← 0.32210101
b[55] ← -0.10069723
q[1] ← -1.7513594335
q[2] ← -0.9109805725
q[3] ← -1.0475347050
q[4] ← -2.3619323768
q[5] ← -1.8765800430
        ⋮
q[51] ← -1.4719871227
q[52] ← -1.9722319279
q[53] ← -1.7289024947
q[54] ← -0.9255847931
q[55] ← -1.7116853925
test[1] ← 0
test[2] ← 0
test[3] ← 0
test[4] ← 0
test[5] ← 0
        ⋮
test[51] ← 1.0
test[52] ← 1.0
test[53] ← 1.0
test[54] ← 1.0
test[55] ← 1.0
for (j in 1:N) {
for (k in 1:T) {
    p[j,k] ← (exp(q[k])/(1+exp(q[k]))+exp(a[k]*(theta[j] - b[k]-test[k])))/(1+exp(a[k]*(theta[j]
- b[k]- test[k])))
    resp[j,k] ∼ dbern(p[j,k])
    } } }
    list(N=2000, T=55,
    theta=c(
    -1.1632050534, -0.9143888319, -0.4015921904, -1.0573467692, 0.2015239017,
    -0.6371621926, -0.3196387516, -0.5640802674, 1.7068394363, 0.5722573115,
    ............
    0.8271898127, 0.9437303027, -0.6265469574, -0.4280939117, -0.0143332587 ))
```

Appendix C

```
♯ 3PL Testlet Model
    model
    {
    for (j in 1:N) {
    for (k in 1:T) {
    r[j,k]¡-resp[j,k]
    } }
    for (j in 1:N) {
    for (k in 6:25) {
    p[j,k] ← (exp(q[k])/(1+exp(q[k])) + exp(a[k]*(theta[j] - b[k])))/(1+exp(a[k]*(theta[j] -
b[k])))
    r[j,k] ∼ dbern(p[j,k])
    }
    for (k in 26:T) {
    p[j,k] ← (exp(q[k])/(1+exp(q[k]))+exp(a[k]*(theta[j] - b[k]-gamtes[j, test[k]])))/(1+exp(a[k]*(theta[j]
- b[k]- gamtes[j,
    test[k]])))
    r[j,k] ∼ dbern(p[j,k])
    }
    for (k in 2:M){
    gamtes[j, k] ∼ dnorm(0, siggam[k])
    } }
    for (j in 1:N) {
    theta[j] ~ dnorm(0,1)
    }
    for (k in 1:T) {
    a[k] ~ dnorm(mua, siga)I(0,)
    b[k] ~ dnorm(mub,sigb)
    q[k] ~ dnorm(muq,sigq)
    }
    a[1] ← 0.521172928
    a[2] ← 0.077579014
    a[3] ← 0.886715500
    a[4] ← 1.676318913
    a[5] ← 2.805824457
    b[1] ← -1.13956809
    b[2] ← -0.54981831
    b[3] ← 0.41384308
    b[4] ← 1.23229356
    b[5] ← 0.06482972
    q[1] ← -1.7513594335
    q[2] ← -0.9109805725
    q[3] ← -1.0475347050
    q[4] ← -2.3619323768
    q[5] ← -1.8765800430
    for (k in 2: M) { siggam[k] ~ dchisqr(.5) var[k] ← 1/siggam[k]
    }
    mua ~ dnorm(0, .01)
```

```
mub ~ dnorm(0, .01)
muq ~ dnorm(0, .01)
siga ~ dchisqr(.5)
sigb ~ dchisqr(.5)
sigq ~ dchisqr(.5)
}
```

References

Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: Dynamics of ability determinants. *Journal of Applied Psychology, 77*, 598–614.

Albert, J. H. (1992). Bayesian estimtion of normal ogive item resonse curves using Gibbs sampling. *Journal of Educational Statistics, 17*, 251–269.

Albert, J. H. & Ghosh, M. (2000). Item response modeling. In D. Dey., & S. G. Mallick (Eds.), *Generalized linear models: A Bayesian perspective*, (pp. 173–193). New York: Addison-Wesley.

Ariel, A., Veldkamp, B. P., & Breithaupt, K. (2006). Optimal testlet pool assembly for multistage testing designs. *Applied Psychological Measurement, 30*, 204–215.

Bafumi, J., Gelman, A., Park, D., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis, 13*, 171–187.

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

Bao, H. (2007). *Investigating differential item function amplification and cancellation in application of item response testlet models.* Unpublished doctoral dissertation, University of Maryland.

Bazán, J. J., Branoco, M. D., & Bolfarine, H. (2006). A skew item response model. *Bayesian Analysis, 1*, 861–892.

Bock, R. D. (1972). Estimating item parameter and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

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

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434–455.

Brooks, S. P., & Roberts, G. O. (1998). Convergence assessments of Markov chain Monte Carlo algorithms. *Statistics and Computing, 8*, 319–335.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayesian methods for data analysis.* London, UK.: Chapman & Hall/CRC.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91*, 883–904.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: The Guilford Press.

De Finetti, B. (1974), *Theory of probability.* New York: John Wiley & Sons.

Du, Z. (1998). *Modeling conditional item dependence with a three-parameter logistic testlet model.* Unpublished doctoral dissertation, Columbia University.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 269–286.

Fox, J. P., & Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika, 68*, 169–191.

Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education, 18*, 351–380.

Gelman, A. E., Carlin, J. B., Stern, H. S., & Rubin, R. D. (2003). *Bayesian data analysis* (2nd ed). New York: Chapman & Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–472.

Ghosh, M., Ghosh, A., Chen., M.-H., & Agresti, A. (2000). Noninformative priors for one parameter item response models. *Journal of Statistical Planning and Inference, 88*, 99–115.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications theory.* Boston:Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Harwell, M. R., & Baker, F. B. (1991). The use of prior distribution in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement, 15*, 375–389.

Heikkinen, V. P., & Kanto, A. (2002). Value-at-risk estimation using non-integer degrees of freedom of Student's distribution. *Journal of Risk, 4*, 77–84.

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson, V., & Albert, J. (1999). *Ordinal data modeling.* New York: Springer-Verlag.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association, 91*, 1343–1370.

Kim, S.-H., Cohen, A. S., Baker, F. B., Subkoviak, M. j., & Leonard, T. (1999). An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika, 59*, 405–421.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*, 237–255.

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Measurement in Education, 25*, 357–372.

Li, Y. (2004). *Applications and extensions of IRT testlet models.* Unpublished doctoral dissertation, University of Wisconsin-Madison.

Li, Y., Bolt, D. M., & Fu, J. (2005). A test characteristic curve linking method for the testlet model. *Applied Psychological Measurement, 29*, 340–356.

Loevinger, J. A. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monograph, 61.*

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177–195.

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

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

Patz, R. J., & Junker, B. W. (1999b). Application and extensions of MCMC in IRT: Multiple item types, missing Data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342–366.

Raftery, A. L., & Lewis, S. (1992). How many iterations in the Gibbs sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 763-773). Oxford, UK: Oxford University Press.

Rosenbaum, P. P. (1988). Item bundles. *Psychometrika, 53*, 349–359.

Rupp, A., Dey, D., & Zumbo, B. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling, 11*, 424-451.

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation, 72*, 217–232.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplements, 17.*

Sinharay, S. (2004). Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics, 29*, 461–488.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*, 298–321.

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician, 56*, 196–201.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 3, 237–247.

Smith, B. (2001). *Bayesian output analysis program (BOA) (Version 1.0.0) [Computer software].* Iowa City, IA: University of Iowa, College of Public Health.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1996). *BUGS 0.5 Examples Volume 1.* Robinson Way, Cambridge CB2 2SR, UK: MRC Biostatistics Unit, Institute of Public Health.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. J. (2003). WinBUGS version 1.4 [Computer program]. Robinson Way, Cambridge CB2 2SR, UK: MRC Biostatistics Unit, Institute of Public Health.

"Student". (1908). The probable error of a mean. *Biometrika, 4,* 1–25.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*, 175–191.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50*, 349–364.

Swaminathan, H., & Gifford, J. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*, 581–601.

Swaminathan, H., Hambleton, R., Sireci, S., Xing, D., & Rivazi, S. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement, 27*, 27–51.

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

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in adaptive testing. In W. J. van der Linden & C. A. W. Gals (Eds.), *Computerized adaptive testing: Theroy and practice* (pp. 245–270). Boston: Kluwer-Nijhoff.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* New York: Cambridge University Press.

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. *Journal of Educational Measurement, 29*, 243–251.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–202.

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

Wainer, H., Lewis, C., & Braswell, J. (1991). Building algebra testlets: A comparison of hierarchical and linear structures. *Journal of Educational Measurement, 28*, 311–324.

Wainer, H., Lewis, C., Kaplan, B., & Braswell, J. (1991). Building algebra testlets: A comparison of hierarchical and linear structures. *Journal of Educational Measurement, 28, 4*, 311–323.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement, 57,* 749–766.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*, 197–219.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test score? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*, 22–29.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*, 203–220.

Wang, X., Bradlow, E. T., & Wainer, H. (2004). *User's guide for SCORIGHT (version 3.0): A computer program for scoring tests built of testlets including a model for covariate analysis.* Research Report, 04-49. Princeton, NJ: Educational Testing Service.

Wang, W.-C., & Wilson, M. (2005a). The Rasch testlet model. *Applied Psychological Measurement, 19*, 126–149.

Wang, W.-C., & Wilson, M. (2005b). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement, 65*, 549–576.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and application. *Applied Psychological Measurement, 26*, 109–128.

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

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependence in the Medical College Admissions Test. *Journal of Educational Measurement, 39*, 291–309.