IMPACT OF AND CORRECTION FOR ITEM RESPONSE THEORY (IRT) SCORE ESTIMATION ERROR ON A MULTILEVEL VALUE-ADDED MODEL

By

Changhui Zhang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods

2012

ABSTRACT

IMPACT OF AND CORRECTION FOR ITEM RESPONSE THEORY (IRT) SCORE ESTIMATION ERROR ON A MULTILEVEL VALUE-ADDED MODEL

By

Changhui Zhang

In common educational research settings, a latent achievement construct for the student is measured at both the beginning and the end of a learning program so that the added value of the instructional program can be quantified. Explanatory variables at both student level and school level are used to help explain the change in levels of performance on the construct. Test scores, which serve as a proxy for students' true achievement, contain estimation error. Some of the explanatory variables may also be latent variables and they are subject to error as well. Traditionally, the item response theory (IRT) estimates of the latent variables are obtained before they are entered into linear regression models. The problem with this two-step approach is that the relationship between dependent and independent variables could be distorted due to error in IRT score estimates. To address this problem, this dissertation proposes a combined IRT and multilevel model which estimates achievement and the added value simultaneously. A Bayesian MCMC (Monte Carlo Markov Chain) method is used to fit the combined model. The performance of the combined model and the one-step Bayesian approach are compared with the traditional two-step approach on simulated data. The simulations are carried out on a simple linear model and a multilevel model as well. The results indicate that this one-step approach recovers the true relationships among latent variables with less bias and more accuracy. Following the simulation, the study applied the new approach to an empirical dataset obtained

from the Mathematical Education for Elementary Teachers (ME.ET) project. Special attention is given to the missing data issue in the application. The final chapter of the dissertation discusses the sources of IRT score error and their implications.

ACKNOWLEDGEMENTS

My first thought is always to recognize my committee: Dr. Mark Reckase, Dr. Raven McCrory, Dr. Tenko Raykov and Dr. Cassandra Guarino, for their kind guidance, inspirational encouragement, and generous support. The completion of this dissertation would not be possible without the time and effort they spent with me. I can easily recognize their brilliant ideas and insights throughout this dissertation. I am also grateful to Dr. Kimberly Maier because my Bayesian experience originated from her classes and workshops.

The High Performance Computer Center (HPCC) at Michigan State University (MSU) was a primary resource of the computational power that I needed to crunch the numbers. I admit to having been impressed by the state-of-the-art hardware as well as the knowledgeable people working there such as Dr. Dirk Colbry.

Many people lent their helping hand along this journey. Dr. Chueh-An Hsieh helped improving the quality of the Bayesian code. Dr. Wei He and Dr. Xuechun Zhou offered valuable suggestions on writing and formatting. For many others whose name I did not mention here, I am also thankful to you and my best wishes are with you.

Finally, it is a duty and pleasure to thank my family for their patient understanding during those exacting hours that somehow results in this dissertation.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES x

CHAPTER 1
INTRODUCTION 1
Background 1
Literature review 2
Homogeneous measurement error is a known problem 2
IRT score error is more complicated 4
Purposes of this study 6
The impact of IRT score error 6
Evaluation of one-step Bayesian approach 8
General organization of this dissertation 10

CHAPER 2
IRT SCORE ERROR IMPACT IN A SIMPLE LINEAR REGRESSION MODEL 12
Model and presence of IRT score error 12
Model: simple regression model with latent variables 12
Presence of IRT score error 13
Data generation 14
Parameter recovery 15
Parameter β recovery: 1) two-step approach 15
Parameter β recovery: 2) one-step Bayesian MCMC approach 16
Results 18
1) IRT score error in the dependent variable 19
2) IRT score error in the independent variable 26
3) IRT score error in both dependent and independent variables 33
More about Bayesian: comparison of priors 39
More about Bayesian: convergence test results 40

CHAPTER 3
IRT SCORE ERROR IMPACT IN A MULTILEVEL VALUE-ADDED MODEL 41
Model and presence of IRT score error 41
Model: value-added models (gain score models) 41
Presence of IRT score error 47
Data generation 48
Parameter Recovery 50
Parameters recovery: 1) two-step approach 50
Parameters recovery: 2) one-step Bayesian MCMC approach 51
Results 54
More about Bayesian: setting priors 70
More about Bayesian: convergence test results 73

CHAPTER 4
ONE-STEP BAYESIAN APPLICATION ON ME.ET PROJECT DATA 76
The Mathematical Education of Elementary Teachers (ME.ET) Project 76
Model 76
Data 77
Estimation 78
Results 79
Results if using full data with imputation 82
Comparison of the listwise deleted sample and full sample 86

CHAPTER 5
CONCLUSIONS AND DISCUSSION 88
Conclusions 88
Discussion 89
The recovery of latent varialbe correlation coefficient 89
Sources of IRT score error 92
Bayesian approach 93

APPENDICES 95
Appendix A Item parameters in multilevel simulation 96
Appendix B BUGS code for the simple linear model in chapter 2 100
Appendix C BUGS code for the multilevel value-added model in chapter 3 105
Appendix D BUGS code for ME.ET project model in chapter 5 121

REFERENCES 127

LIST OF TABLES

Table 2.1 β estimation using the two-step approach when IRT score error is present in the dependent variable. 20

Table 2.2 β estimation using the one-step approach when IRT score error is present in the dependent variable. 22

Table 2.3 β estimation using the two-step approach when IRT score error is present in the independent variable. 27

Table 2.4 β estimation using the one-step approach when IRT score error is present in the independent variable. 29

Table 2.5 β estimation using the two-step approach when IRT score error is present in both dependent and independent variables. 34

Table 2.6 β estimation using the one-step Bayesian approach when IRT score error is present in both dependent and independent variables. 35

Table 2.7 The burn-in length for the Markov chains when passing the convergence test....40

Table 3.1 Coefficients recovery when no error present in latent variables 55

Table 3.2 Coefficients recovery when errors in dependent variables (student scores)....56

Table 3.3 Coefficients recovery when errors in independent variables (*attitude* and *method*) 57

Table 3.4 Coefficients recovery when errors in both dependent and independent variables 58

Table 3.5 Estimated coefficients with seven different priors 71

Table 3.6 Convergence of Markov chains: coefficient of *attitude* (= 0.14) 74

Table 3.7 Convergence of Markov chains: coefficient of *female* (= 0.00)....74

Table 3.8 Convergence of Markov chains: coefficient of *method* (= 0.30)....75

Table 3.9 Convergence of Markov chains: coefficient of *textbook* (= 0.48) 75

Table 4.1 Heidelberger and Welch diagnostic of ME.ET project model....81

Table 4.2 Covariates effects estimated from Bayesian approach. 81

Table 4.3 Heidelberger and Welch diagnosis of ME.ET project model, full dataset used....85

Table 4.4 Covariates effects estimated from Bayesian approach, full dataset used 85

Table 4.5 Statistics of the listwise deleted data and full data 87

Table 5.1 The standard deviation of IRT scores and their correlation with true θ 90

Table 5.2 The bias of correlation coefficient recovery 91

Table 6.1 Teaching method questionnaire item parameters 97

Table 6.2 Student attitude questionnaire item parameters 98

Table 6.3 Student pretest and posttest item parameters 99

LIST OF FIGURES

Figure 2.1 Comparison of β estimation methods when IRT score error is present in the dependent variable and ρ = 0.2. 23

Figure 2.2 Comparison of β estimation methods when IRT score error is present in the dependent variable and ρ = 0.5. 24

Figure 2.3 Comparison of β estimation methods when IRT score error is present in the dependent variable and ρ = 0.8. 25

Figure 2.4 Comparison of β estimation methods when IRT score error is present in the independent variable and ρ = 0.2. 30

Figure 2.5 Comparison of β estimation methods when IRT score error is present in the independent variable and ρ = 0.5. 31

Figure 2.6 Comparison of β estimation methods when IRT score error is present in the independent variable and ρ = 0.8. 32

Figure 2.7 Comparison of β estimation methods when IRT score error is present in both variables and ρ = 0.2. 37

Figure 2.8 Comparison of β estimation methods when IRT score error is present in both variables and ρ = 0.5. 38

Figure 2.9 Comparison of β estimation methods when IRT score error is present in both variables and ρ = 0.8. 39

Figure 3.1 Coefficient of attitude (= 0.14) recovery. 59

Figure 3.2 Coefficient of female (= 0.00) recovery. 60

Figure 3.3 Coefficient of method (= 0.30) recovery. 61

Figure 3.4 Coefficient of textbook (= 0.48) recovery. 62

Figure 3.5 Recovery of correlation between student pretest and gain (= -0.4). 64

Figure 3.6 Recovery of correlation between student pretest and gain (= 0.0). 65

Figure 3.7 Recovery of correlation between student pretest and gain (= 0.4). 66

Figure 3.8 Recovery of correlation between school mean pretest and gain (= -0.38). 68

Figure 3.9 Recovery of variance of school gain (= 0.08). 69

Figure 3.10 Estimated coefficients with seven different priors 72

Figure 4.1 Trace plot and density plot of ME.ET project parameters. Missing data were deleted listwise.. 80

Figure 4.2 Trace plot and density plot of ME.ET project parameters. Full dataset with imputation was used. 84

CHAPTER 1 INTRODUCTION

Background

One important goal of educational research is to gauge the effectiveness of teaching programs. To achieve this, researchers measure students' achievement as a latent construct before and after a learning program to quantify its quality. Often the pretest and posttest consist of items with discrete scores that can be described with item response theory (IRT) models. Gain score, defined as the difference between student initial achievement and posterior achievement, is an indicator of added value from the educational program.

In addition to the magnitude of the added value a teaching program can bring, researchers are often interested in the factors that cause it. During the intricate process of learning, numerous factors at both student level and school level could have contributed to the student's achievement. At the student level, commonly used explanatory variables include gender, social-economical status (SES), attitudes and beliefs etc. At the school level the variables include the textbook or materials used (as a proxy for content) and teaching methods. Some of these variables, such as gender and textbook, can be observed and recorded accurately. But other variables such as the student's attitude and the teacher's teaching methodology are latent variables that have to be measured indirectly by questionnaires and estimated within an IRT framework.

Once the values of these latent variables are obtained, they are entered into linear models so that the relationship between variables can be explored. This two-step approach, estimating latent variables first and fitting the linear models afterward, is straightforward but with a potential problem. The problem lies in the error contained in IRT scores that may lead to biased estimation of regression coefficients. Measurement error, especially homogeneous measurement

error, is a known issue in regression models, but the consequences of tolerating the error in IRT scores are yet to be fully investigated.

This measurement error issue may become even more complicated when the multilevel linear regression models (or hierarchical linear models, HLM) are involved. Since the 1990s, researchers have paid close attention to the data structure of students nested in schools. In order to better estimate the influence of student and school factors, researchers have developed the multilevel models to account for the correlation within schools (2002). In HLM, the regression coefficients are often called *effects* of those covariates. They are further divided into *fixed effects* and *random effects* depending on whether they vary in the model. Both HLM and IRT models are commonly used in educational research, but it is not clear what the impact of the IRT score error is on the estimation of regression coefficients in multilevel models.

Literature review

Homogeneous measurement error is a known problem

In the field of statistics, measurement error has long been a popular topic. Researchers have realized that the regression coefficient tends to be underestimated due to measurement error and have developed various remedies to correct such attenuation. Since Cochran (1968; 1970) called attention to the measurement error impact on statistical models, a rich and elegant literature on measurement error has been developed, and it was well summarized in the books by Fuller (1987) and Buonaccorsi et al (2010).

In educational research, minimizing measurement error impact is as important a task as in any other fields. Sutcliffe (1958) elaborated on the implications of measurement error on tests of significance and demonstrated how measurement error decreases the power of the *F* test for

differences among means. Lord and Novick (1968) proposed a test theory that describes the effects of measurement error on scores. Dunivant (1981) discussed the problems caused by measurement error in linear models for assessing change and expressed the bias in ordinary least squares estimators as a function of covariance among true scores, among the measurement errors, and sample size. Rogers et al. (1988) explored the complex nature of the relationship between power and covariance, which often increases power, and measurement error, which reduces power.

To address the impact of measurement error, researchers have developed various remedies to correct such attenuation. Stroud (1973) demonstrated a correction method for problems of measurement error in independent variables when measurement error variances are known. Daniel (1996) advocated likelihood analysis for regression models with measurement errors in explanatory variables and presented an EM algorithm as a straightforward approach for likelihood analysis of normal linear regression with normal explanatory variables, and normal replicate measurements. Bartlett et al. (2009) compared two correction for measurement error methods: regression calibration (RC) and maximum likelihood (ML) and found ML is better than RC in dealing with covariate measurement error. Zimmerman (2007) pointed out the important conceptual difference between population and sample with regard to correlation and investigated different research settings under which the correction for attenuation can be useful in data analysis and those under which it is inaccurate. His simulation experiments proved the advantages and general superiority of estimators proposed by Fuller (1987).

However, some of the correction methods have practical problems. Findings of Schmidt et al. (1996) revealed the importance of eliminating bias caused by measurement error since research in psychology is becoming more sophisticated and more oriented toward the

development and testing of theory. Their paper illustrated appropriate and inappropriate instances of correction for measurement error in commonly seen research situations. Fan (2003) highlighted the importance of reporting measurement reliability information in substantive research and suggested that correction for attenuation should be considered when information about score reliability is available. However, the coefficient alpha is a not a perfect measure of reliability. Osburn (2000) found alpha, when used in corrections for attenuation, can result in nontrivial overestimation of the corrected correlation. Woodhouse et al. (1996) found that in multilevel models, level 1 residual variance decreases when correction for measurement error in an independent variable at level 1. This effort increases the interschool correlation, which is further increased when correction for measurement error is made in the dependent variable. While measurement error research has been expanded into nonlinear models (Carroll, 2006), the error correction effort becomes more challenging.

It should be noticed that the majority of research adopts the homogeneous assumption of error due to its better compatibility with the Classical Test Theory (CTT) framework and easier computation in practice. When IRT became popular in educational and psychological research, not much special attention was given to its special heterogeneous error structure.

IRT score error is more complicated

In contrast with the rich literature of homogeneous measurement error, heterogeneous estimation error in IRT scores has had much less attention. The main difficulty is the heterogeneous nature of IRT estimation error that is quite different from the assumption of constant error variance in CTT framework. In CTT framework, the error structure of an observed score can be described as

$$\boldsymbol{Var(X|T) = c} \tag{1.1}$$

where *c* is a constant denoting the variance of the observed score, $\boldsymbol{X}$, given a known true score, $\boldsymbol{T}$. Larger *c* means less precision with the measurement instrument.

In IRT, however, the variance of estimated score $\hat{\theta}$ given the true θ is not a constant, but a function of θ. The function used to describe the precision of estimated IRT score is called test information function I(θ):

$$I(\theta) \equiv I\{\theta, \hat{\theta}\} = \sum_{k=1}^{K} \frac{P_k'^2}{P_k Q_k} \tag{1.2}$$

where $P_k(\theta)$ is the probability of student with proficiency θ to answer item *k* (k = 1, … , K) correctly and $Q_k(\theta) = 1 - P_k(\theta)$. Lord (1980) pointed out that the asymptotic variance of the maximum likelihood estimator of proficiency is the reciprocal of the information function:

$$Var(\hat{\theta}|\theta) = \frac{1}{I\{\theta, \hat{\theta}\}} = \frac{1}{\sum_{k=1}^{K} \frac{P_k'^2}{P_k Q_k}} \tag{1.3}$$

The information function is the upper bound to the information that can be obtained from the test (Lord, 1980). That is to say, any estimated IRT score cannot have a smaller error than the standard error of measurement (SEM) given by

$$SEM(\hat{\theta}|\theta) = \sqrt{\frac{1}{I\{\theta, \hat{\theta}\}}} \tag{1.4}$$

That limit represents the precision of the measurement instrument and the ambiguousness of the estimated IRT scores.

Purposes of this study

The research literature contains a lot of information on IRT modeling methods. However, the nature and characteristics of IRT score estimation error and its impact on structural explanations has not been thoroughly discussed. It is not clear whether an HLM and IRT combination can correct for estimation error. Many studies have demonstrated how their estimation approaches work, but few discussed the consistency and efficiency of the estimators in these one-step estimation methods. This dissertation study intends to answer these questions.

This dissertation aims to investigate the consequences of ignoring IRT score errors in a simple linear model as well as in a multilevel linear model, and to evaluate a potential solution: a combined HLM and IRT model solved by a Bayesian MCMC approach. To be specific, the research questions of this study include:

1) To what extent could the IRT score estimation error lead to biased estimation of the regression coefficients in simple linear models and multilevel models?

2) To what extent could the Bayesian MCMC approach, which estimates latent variables and regression coefficients simultaneously, correct for the impact of IRT score estimation error?

The impact of IRT score error

The error of an estimator is the difference between the estimated value and the true value of the same parameter. If the expected value of an error is zero, the estimator is called unbiased.

The bias of an estimator is the difference between the expected value and the true value of the parameter being estimated. There is not a known method to calculate the bias of linear regression coefficient estimator when IRT scores are used as dependent or independent variables. So this study will use simulation to illustrate the magnitude of IRT score error impact, or whether there is any impact on the regression models.

Multiple replications (*N* = 100 for each condition) will be run to control for the influence of sampling fluctuation. The mean of the *N* estimates in the replications will be used as an approximation of the expected value of the estimator. The bias value ($Bias_{\hat{\beta}}$) and empirical sampling standard deviation ($SD_{\hat{\beta}}$) across *N* replications are computed as

$$Bias_{\hat{\beta}} = \frac{1}{N}\sum_{d=1}^{N}\left(\hat{\beta}_d - \beta\right) \tag{1.5}$$

$$SD_{\hat{\beta}} = \sqrt{\sum_{d=1}^{N}\left(\hat{\beta}_d - \bar{\hat{\beta}}\right)^2 /(N-1)} \tag{1.6}$$

where β denotes the generating values of linear regression coefficients, $\hat{\beta}_d$ denotes the estimator of β in replication *d (d = 1, … ,N)*, and $\bar{\hat{\beta}}$ denotes the mean estimates across *N* (*N* = 100) replications.

It is known that the OLS estimator is unbiased when no measurement error is in the dependent and independent variables. So the observed bias of β estimation, if there is any, will be attributed to the IRT score error.

Homogeneous measurement error literature suggests that the impact of error depends on whether the variable is a dependent variable or an independent variable. There is no bias in estimating β if the dependent variable contains homogeneous measurement error. The regression coefficient β will be underestimated if the independent variable contains homogeneous measurement error. To fully understand the impact of error in IRT scores, this study will explore three different scenarios: 1) error in dependent variables, i.e., the dependent variable is a latent variable and it uses estimated IRT scores in the regression model; 2) error in independent variables, i.e., the independent variable is a latent variable and it uses estimated IRT scores in the regression model; 3) error in both dependent and independent variables, i.e. estimated IRT scores of both independent and dependent variables will be used in regression models; and 4) they will be compared with a reference scenario in which no variable has error in it.

IRT scores can be estimated in different ways. This study also compare the adoption of both Maximum Likelihood (ML) estimator and Bayesian Model (BM) estimator to determine if different IRT scores have different influences on regression coefficient estimation.

Evaluation of one-step Bayesian approach

The second purpose is to evaluate a combination model of IRT, HLM and Value Added Models (VAM). There has been efforts (Kolen, Zeng, & Hanson, 1996; Lee, Brennan, & Kolen, 2000) to study scale score measurement using IRT. However research that directly addresses IRT estimation error is less known until the combination of multilevel models, IRT models and

Bayesian approach. Adams et al. (1997) proposed model where an IRT model and single level linear model are combined. An explanatory item response modeling (EIRM) approach elaborated by de Boeck and Wilson (2004) is presented within the statistical framework of generalized linear mixed models. This approach provides a powerful framework for both a psychometric and statistical analysis of value-added models. Briggs (2008) found that the EIRM approach results in estimated racial/ethnic achievement gaps that are larger than those found in the two-step approach. Later research (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2008; Hsieh, von Eye, & Maier, 2010; Kamata, 2001; Maier, 2001, 2002) merged an IRT model with a multilevel linear models. The connection between hierarchical modeling and the structural measurement models has been discussed and the combination model of IRT and HLM have been investigated (Fox, 2003, 2005; Fox & Glas, 2001). In Hsieh (2010), an item response theory (IRT) model was incorporated into a latent variable model (LVM) by using a commonly used link function and Bayesian estimation. The generalized linear latent and mixed models (GLLAMM), combining features of generalized linear mixed models (GLMM) and structural equation models (SEM), was introduced by Rabe-Hesketh (2004). The idea was further explored in Rabe-Hesketh and Skrondal's later paper (Rabe-Hesketh & Skrondal, 2007).

The model in this study is an expansion of the IRT and HLM models (HMM) and a multilevel value-added model. This model tries to account for all the elements in one model. It has latent variables as a dependent variable and an independent variable as well. For measurement IRT models, not only the two-parameter logistic (2PL) model but also the graded response model (Samejima, 1969; 1997) are included. The response matrix instead of the IRT scores will be analyzed. This is contrasted with the typical two-step approach, in which psychometric analysis (i.e., measurement) and statistical analysis (i.e., explanation) occur

separately. This study also includes a special but frequently encountered case of a value-added model, in which only two test occasions (pretest and posttest) are presented. The complicity of the combination models cannot be estimated easily and a powerful tool, Bayesian MCMC approach, will be used to solve it.

It is expected that such a combination model and Bayesian MCMC solution will work. However it is not known how well the combination will work. This study intends to evaluate the performance of the proposed combination model with Bayesian approach, compared with the traditional two-step approach. For every scenario and condition listed above, the one-step Bayesian approach and two-step approach will be compared based on *N* replications. Two criteria will be used to evaluate the two approaches. The first one is bias. Bias of an estimator of a parameter is defined as the difference between the estimator's expected value and the true value of the parameter. It is used to describe how systematically an estimator is different from the true parameter. In this simulation study, the mean of *N* one-step Bayesian estimates and the mean of *N* estimates using the two-step approach will be calculated to see how far they are from the true value. The second criterion is efficiency. Efficiency is defined as the standard deviation of an estimator, and it is an indicator of the accuracy of an estimator. It is approximated by the standard deviation of the *N* estimates in the simulation recovered by those two different approaches.

General organization of this dissertation

To achieve the above goals, this study starts with the simple regression model in chapter 2, where only one dependent variable and one independent variable are presented in a single level linear regression:

$$\theta_y = \beta\theta_x + \varepsilon. \tag{1.7}$$

Different scenarios, such as 1) error in the dependent variable, 2) error in the independent variable, and 3) error in both dependent and independent variables, will be explored because error in the dependent or independent variables can have a different influence on coefficient estimation. For each scenario, simulated data will be generated and parameters will be recovered by the two approaches: one-step Bayesian approach and two-step approach. The two approaches are compared based on *N* replicated simulations.

Then the research moves to a more complex multilevel value-added model in chapter 3, where a combination model including HLM, VAM and IRT elements are presented. The error impact and correction are considered for different scenarios in the same way as they are in chapter 2.

Following the simulation, the Bayesian approach is applied on an empirical study, the Mathematical Education of Elementary Teachers (ME.ET) Project, in chapter 4. Special attention is given to the commonly encountered missing value problem in real data. Further discussions on IRT measurement error and improvement of the Bayesian approach will be included in chapter 5.

CHAPER 2 IRT SCORE ERROR IMPACT IN A SIMPLE LINEAR REGRESSION MODEL

Model and presence of IRT score error

Model: simple regression model with latent variables

To better understand the IRT score error issue, exploratory simulations will start with a basic simple linear regression model

$$\theta_y = \beta\theta_x + \varepsilon \quad (2.1)$$

where there are only one dependent latent variable θ_y, one independent latent variable θ_x and a random error term ε. Assuming that both latent variables are latent IRT constructs that are centered on 0, there is no intercept term in the equation.

Let $Y_{I\times K}$ denote the student (0, 1) response matrix where 1 indicates a correct answer and 0 a wrong one when measuring θ_y of *I* students with a *K*-item test, assuming that each response Y_{ik} (*i = 1, ... , I; k = 1, ... , K*) fits a two-parameter logistic (2PL) IRT model :

$$P(Y_{ik} = 1) = \frac{e^{a_{yk}(\theta_{yi} - b_{yk})}}{1 + e^{a_{yk}(\theta_{yi} - b_{yk})}} \quad (2.2)$$

where b_{yk} is the item difficulty parameter and a_{yk} is the discrimination parameter of item *k* that measures latent construct θ_y.

Similarly, $X_{I\times K}$ is the student (0, 1) response matrix when measuring θ_x and each X_{ik} (*i = 1,…, I; k = 1,…, K*) fits a two-parameter logistic (2PL) IRT model :

$$P(X_{ik}=1)=\frac{e^{a_{xk}(\theta_{xi}-b_{xk})}}{1+e^{a_{xk}(\theta_{xi}-b_{xk})}} \tag{2.3}$$

where b_{xk} is the item difficulty parameter and a_{xk} is the discrimination parameter of item *k* that measures latent construct θ_x.

The tests measuring θ_x and θ_y don't have to be the same length *K*, but they are fixed as so to simplify the notation.

Presence of IRT score error

Different scenarios of error presence are considered in the simple linear regression models:

1) IRT score estimation error in dependent variable $\hat{\theta}_y$ only:

$$\hat{\theta}_y=\beta\theta_x+\varepsilon \tag{2.4}$$

where estimated $\hat{\theta}_y$ and true θ_x will be entered into the regression model to estimate β.

2) IRT score estimation error in independent variable $\hat{\theta}_x$ only:

$$\theta_y=\beta\hat{\theta}_x+\varepsilon \tag{2.5}$$

where true θ_y and estimated $\hat{\theta}_x$ will be entered into the regression model.

3) IRT score estimation error in both independent $\hat{\theta}_y$ and dependent variables and $\hat{\theta}_x$:

$$\hat{\theta}_y = \beta\hat{\theta}_x + \varepsilon \quad (2.6)$$

where both estimated $\hat{\theta}_y$ and $\hat{\theta}_x$ will be entered into the regression.

For the reference scenario 4) in which there is no error in either variable, it is well known and will not be included for this simple linear regression model. However, it will be considered later in the multilevel regression simulation in chapter 3.

Data generation

Both θ_y and θ_x are latent variables, e.g. student reading achievement and attitudes towards learning. A number *I* students with latent construct θ_y and θ_x are generated first. They are assumed to be obtained from a multivariate normal distribution $\begin{pmatrix}\theta_y \\ \theta_x\end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}\right)$, where ρ is the correlation of the two latent variables.

In each dataset, the *a* item parameters are generated from a lognormal distribution $logN(0, 0.25)$, and the *b* item parameters are generated from a normal distribution $N(0, 0.5)$. The item parameter distributions are adopted from an earlier research by Fox (2005).

Several factors are expected to have an impact on the parameter recovery and they are explored in the simulation. The first factor is the correlation between the two latent variables. Three levels of correlation (ρ = 0.2, 0.5, 0.8) will be considered for data generation. The second

factor is test length *K*, i.e. the number of items in a test. Generally speaking, a longer test is a more accurate measurement instrument than a shorter one. Shorter tests yield more error in IRT score estimation and thus have a more prominent impact on the regression coefficient recovery. So a longer test ($K = 25$) and a shorter test ($K = 15$) are considered and compared. The third influencing factor is the number of students, which will affect the accuracy of item parameter estimation and in turn affect the IRT score estimation. Both a larger student sample ($I = 1000$) and a smaller sample ($I = 500$) are considered in the simulation.

A number *N* ($N = 100$) replications for each case were carried out. In total there are 3 levels of correlation × 2 levels of test length × 2 levels of student sample size × 100 replications = 1200 datasets for analysis. Each dataset was analyzed under the three scenarios.

Software R (R Development Core Team, 2008) was used for data generation. The data generation of student responses to test items and the recovery of latent construct θ_x and θ_y were made easier by using *irtoys* package (Partchev, 2006) in R.

Parameter recovery

Parameter β recovery: 1) two-step approach

The two-step approach is the traditional straightforward approach. At step one, the student latent constructs θ_y and θ_x are estimated under IRT models based on their responses to test items. At step two, the estimated IRT score $\hat{\theta}_y$ and $\hat{\theta}_x$ are used to fit the simple linear regression model $\theta_y = \beta\theta_x + \varepsilon$. And β is estimated with the ordinary least square (OLS) method.

There are different options when choosing IRT score estimates, among which includes maximum likelihood (ML) IRT and Bayesian Model (BM) IRT estimation. Obviously different IRT estimates are not identical and they will probably lead to different estimates of the regression parameter β. To explore the influence of different IRT estimates, both maximum likelihood (ML) estimates $\hat{\theta}_{y.ML}$, $\hat{\theta}_{x.ML}$ and Bayesian Model (BM) estimates $\hat{\theta}_{y.BM}$, $\hat{\theta}_{x.BM}$ are entered into the linear regression model to show how the estimates affect the results.

At step one, the student achievement recovery process was carried out in software R using *irtoys* package (Partchev, 2006). At step two, the regression coefficient is estimated in software R with *lm* function using OLS method.

Parameter β recovery: 2) one-step Bayesian MCMC approach

With the one-step approach, the student response matrix $Y_{I\times K}$ and $X_{I\times K}$ will be used to estimate the student latent achievement θ_y and θ_x and the regression coefficient β simultaneously by using Bayesian MCMC approach. The one-step Bayesian approach is implemented in WinBUGS (D. J. Lunn, Thomas, A., Best, N., and Spiegelhalter, D., 2000) /OpenBUGS (D. Lunn, Spiegelhalter, D., Thomas, A., Best, N., 2009)[1], which allows researchers to use the Markov Chain Monte Carlo (MCMC) sampling method to fit the combination model without complex analytic or numerical integrations. The BUGS code for each scenario is included in Appendix B.

Bayesian approach details: setting priors

[1] WinBUGS runs on Windows platform whereas OpeBUGS runs on Linux platform. The model codes are identical.

For the one-step Bayesian approach, the priors of θ_y and θ_x were all set to the standard normal distribution $N(0, 1)$. It is equivalent to the distribution assumption of the latent variables within IRT framework. The prior of β is set to a uniform distribution U [-1, 1]. This prior is a noninformative one and its boundary settings of -1 and 1 are based on the knowledge that the regression coefficient β in the simple regression between two standardized variables is equal to their correlation coefficient, which falls between -1 and 1.

The setting of the error variance prior takes advantage of the known relationship between the variance of error and regression coefficient: $Var(\varepsilon) = 1- \beta^2$. There is only one degree of freedom for the two variables β and $Var(\varepsilon)$. The prior settings of the above parameters are pretty much standard and non-informative.

There is more flexibility when specifying the priors of item parameters. Two sets of priors are used and compared in this study. One set of priors is $a \sim \log N(0,0.25)$ and $b \sim N(0,0.5)$, which was used by Fox (2005). In fact, it is the same as the data generation scheme. The other set of priors is $a \sim \log N(0,0.25)$ and $b \sim N(0,4)$, which was used by Patz and Junker (1999). These two settings are common in similar studies. The comparison of the recovery of regression coefficient β will show whether Bayesian estimates are sensitive to the item parameter prior setting. Generally speaking, the influence of priors gets smaller as the number of students increases. And with a student sample size of 500 or 1000, it is not expected to see much difference in the estimates of β when different item parameter priors are used.

Bayesian approach details: convergence diagnosis

The convergence of Markov chains will be evaluated by Heidelberger and Welch Diagnostic Test (Heidelberger & Welch, 1983). Three Markov chains were run at the same time

for each model. The starting point of each Markov chain was randomly generated according to the prior. The initial 2000 iterations were discarded as the burn-in, and the following 2000 iterations were used to generate the estimates of the parameters. If a MCMC chain does not pass the Heidelberger and Welch Diagnostic test, a longer chain will be run and the length of the burn-in period will be increased by 4000. The above step will be repeated until the chain passes the convergence test or the length of the chain reaches the upper limit of 20,000, where the computing resources with the default configuration are nearly exhausted. Trial runs indicate that if a chain does not converge at the length of 20,000, running even longer chains can be of little help. On the other hand, the estimates are close to the true parameter at the length of 20,000, even though occasionally the chain does not pass the convergence test during the trial.

Result

The simulation results are presented by three scenarios: 1) IRT score error in dependent variable; 2) IRT score error in independent variable; and 3) IRT score error in both dependent and independent variable.

In general, the simulation outcome shows that error in IRT score estimation leads to biased regression coefficient β estimates, while the one-step Bayesian MCMC approach handled the problem well. With the two-step approach, IRT score from different procedure (ML or BM) can bias the estimate in different directions. As the correlation of θ_y and θ_x, i.e. the regression coefficient β, gets stronger, the two-step approach works worse while the one-step Bayesian MCMC approach takes advantage of the correlation and yields even better results. Overall, the impact of error gets smaller when a longer test and/or a larger number of students are involved.

The results are summarized below by scenario.

1) IRT score error in the dependent variable

In the literature on homogeneous measurement error, the presence of measurement error in dependent variables causes no bias in the regression coefficient estimation. However, it is not the case when IRT scores are used as the dependent variable. The regression coefficient tends to be biased downward, as found in the study of Adams, et al. (1997). This dissertation study found that the conclusion cannot be generalized because it depends on what kind IRT score estimate is used. If a IRT score from Bayesian model (BM) procedure is chosen for analysis, as is the case with Adams, et al. (1997), the regression coefficient tends to be attenuated. On the other hand, if a IRT score from maximum likelihood (ML) procedure is used, the regression coefficient tends to biased upward (see Table 2.1). The bias is more prominent as the correlation between the two variables becomes stronger.

Table 2.1 β estimation using the two-step approach when IRT score error is present in the dependent variable

Generating value of parameter ρ	Number of items	Number of students	$\hat{\beta}_{2.ML}$: Two-step approach with IRT score from Maximum Likelihood (ML) estimate (N = 100)			$\hat{\beta}_{2.BM}$: Two-step approach with IRT score from Bayesian Model (BM) estimate (N = 100)		
			Bias	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*
0.2	15	500	0.029	0.054	0.061	-0.053	0.034	0.063
		1000	0.026	0.041	0.049	-0.053	0.026	0.059
	25	500	0.021	0.053	0.057	-0.034	0.038	0.051
		1000	0.018	0.037	0.041	-0.035	0.027	0.044
0.5	15	500	0.057	0.042	0.071	-0.141	0.026	0.143
		1000	0.066	0.036	0.075	-0.136	0.023	0.138
	25	500	0.042	0.044	0.061	-0.093	0.035	0.099
		1000	0.042	0.030	0.052	-0.093	0.023	0.096
0.8	15	500	0.105	0.038	0.112	-0.220	0.028	0.222
		1000	0.108	0.030	0.112	-0.215	0.025	0.216
	25	500	0.065	0.034	0.073	-0.149	0.025	0.151
		1000	0.065	0.022	0.069	-0.148	0.017	0.149

While the two-step approach can be biased both ways, the one-step Bayesian approach, which combines IRT and regression, yields satisfactory results regardless of the prior that is specified (see Table 2.2).

Table 2.2 β estimation using the one-step approach when IRT score error is present in the dependent variable

Generating value of parameter ρ	Number of items	Number of students	$\hat{\beta}_{1.prior1}$: one-step Bayesian MCMC approach prior 1 (N = 100)			$\hat{\beta}_{1,prior2}$: one-step Bayesian MCMC approach prior 2 (N = 100)		
			Bias	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*
0.2	15	500	0.000	0.047	0.047	-0.001	0.047	0.047
		1000	0.001	0.035	0.035	0.000	0.035	0.035
	25	500	0.002	0.046	0.046	-0.001	0.046	0.046
		1000	0.002	0.033	0.033	0.000	0.033	0.033
0.5	15	500	-0.011	0.032	0.034	-0.013	0.032	0.035
		1000	0.000	0.030	0.030	-0.002	0.030	0.030
	25	500	-0.002	0.040	0.040	-0.006	0.040	0.040
		1000	0.000	0.026	0.026	-0.002	0.026	0.026
0.8	15	500	-0.003	0.020	0.020	-0.004	0.021	0.021
		1000	0.001	0.016	0.016	0.001	0.016	0.016
	25	500	-0.003	0.021	0.021	-0.004	0.021	0.021
		1000	-0.001	0.013	0.013	-0.002	0.014	0.014

The illustrated comparison of simple linear regression coefficient β recovery is shown by the boxplot from Figure 2.1 to Figure 2.3. When IRT score estimation error is present in a dependent variable, the one-step approach is a much better choice than two-step approach.

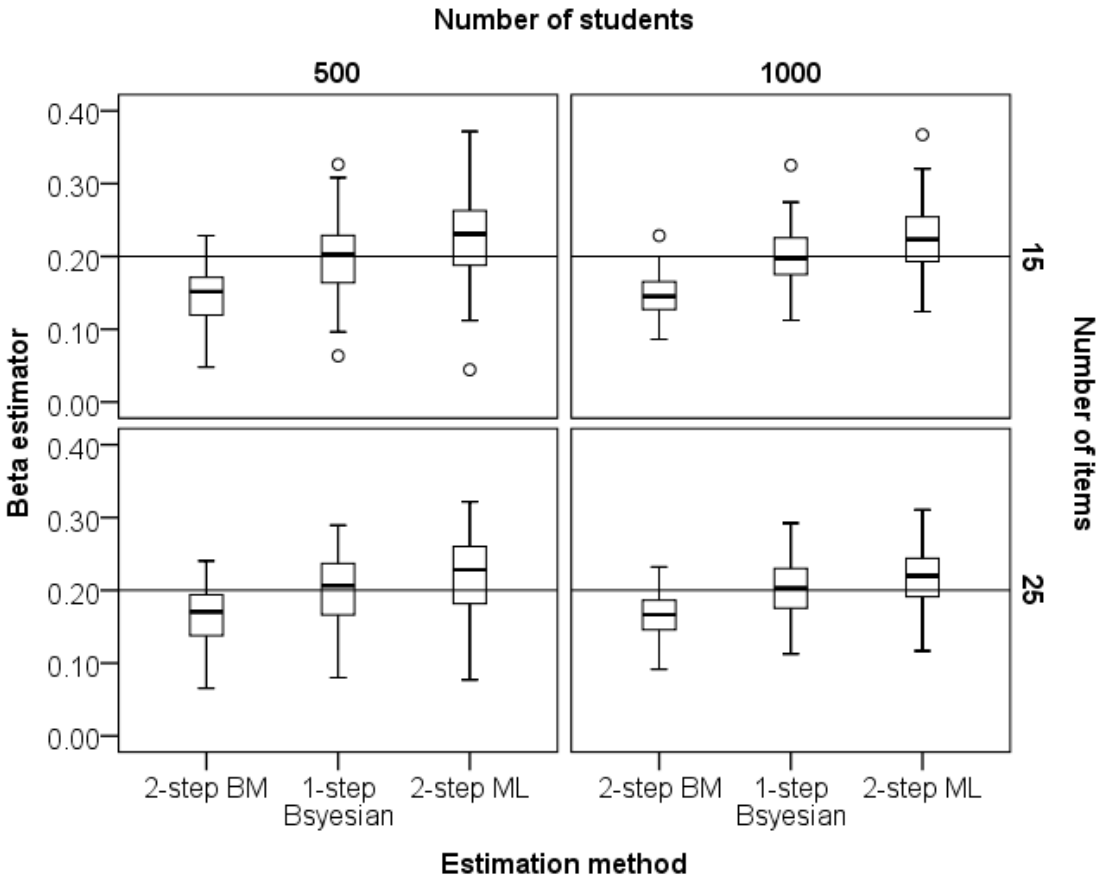

Figure 2.1 Comparison of β estimation methods when IRT score error is present in the dependent variable and ρ = 0.2.

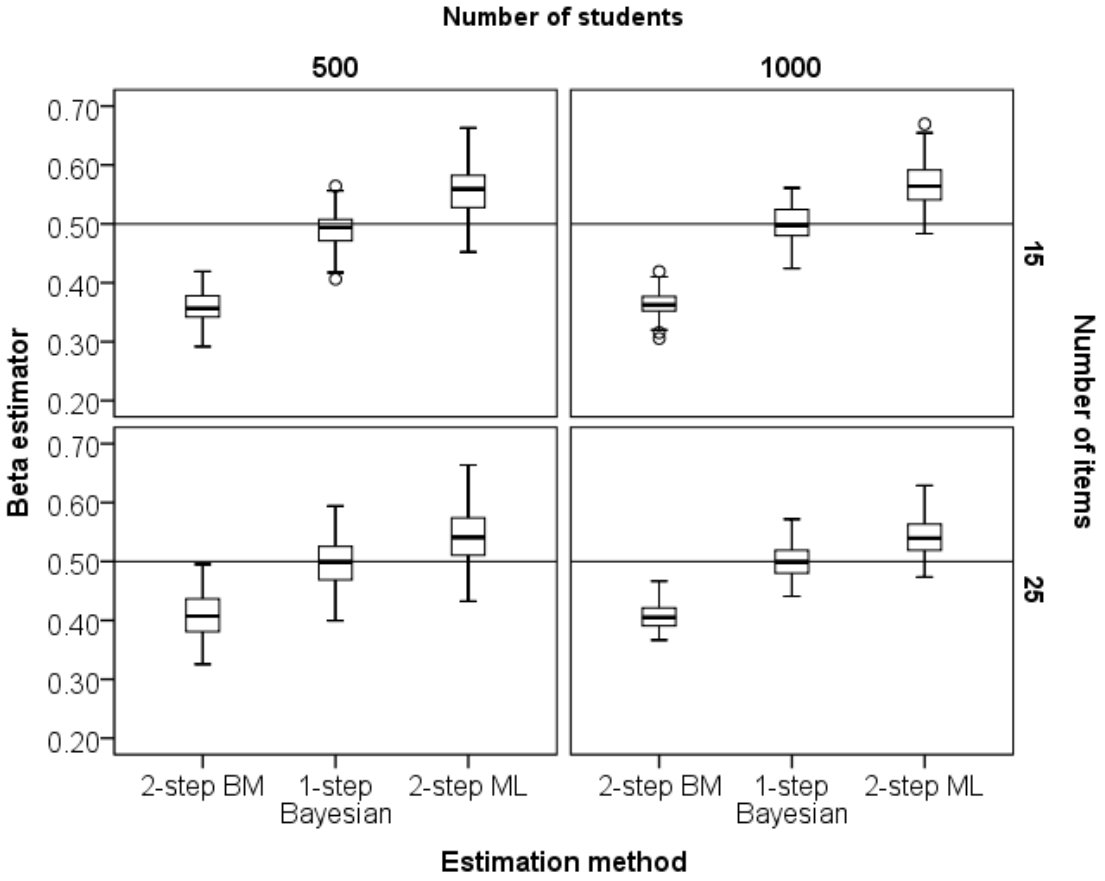

Figure 2.2 Comparison of β estimation methods when IRT score error is present in the dependent variable and ρ = 0.5.

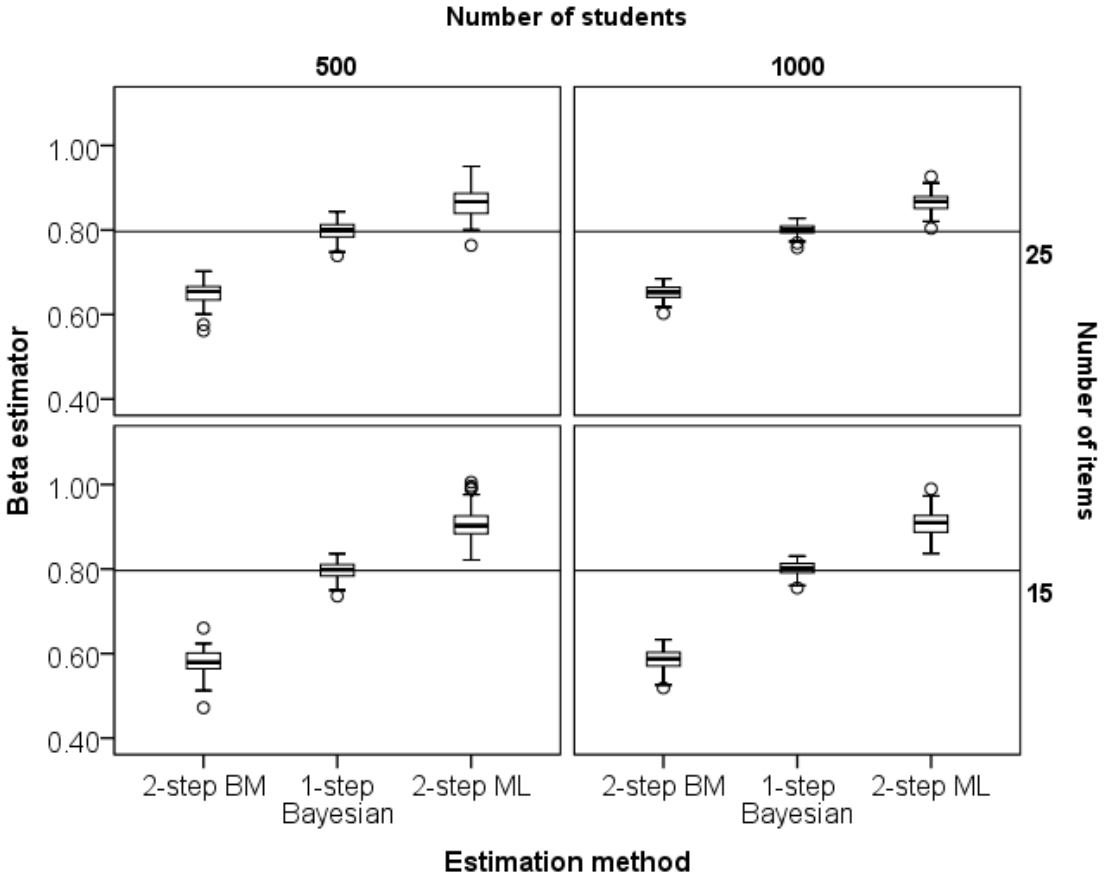

Figure 2.3 Comparison of β estimation methods when IRT score error is present in the dependent variable and ρ = 0.8.

2) IRT score error in the independent variable

In the homogeneous measurement error literature, errors in the independent variables lead to attenuated regression coefficient estimation. This study shows that using a maximum likelihood (ML) estimate does attenuate the β estimation. But it is not the case when a Bayesian model (BM) estimate is used. On the contrary, using Bayesian model (BM) estimate overestimates the coefficient β (see Table 2.3) and biases the estimation of β in the opposite direction.

Table 2.3 β estimation using the two-step approach when IRT score error is present in the independent variable

Generating value of parameter ρ	Number of items	Number of students	$\hat{\beta}_{2.ML}$: two-step approach with IRT score from Maximum Likelihood (ML) estimate (N = 100)			$\hat{\beta}_{2.BM}$: two-step approach with IRT score from Bayesian Model (BM) estimate (N = 100)		
			Bias	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*
0.2	15	500	-0.068	0.036	0.077	0.014	0.055	0.057
		1000	-0.068	0.027	0.073	0.012	0.040	0.042
	25	500	-0.042	0.037	0.056	0.013	0.050	0.052
		1000	-0.041	0.024	0.048	0.015	0.033	0.036
0.5	15	500	-0.177	0.029	0.179	0.019	0.046	0.050
		1000	-0.173	0.028	0.175	0.026	0.039	0.047
	25	500	-0.115	0.039	0.121	0.021	0.049	0.053
		1000	-0.118	0.024	0.120	0.019	0.030	0.036
0.8	15	500	-0.282	0.032	0.284	0.041	0.042	0.059
		1000	-0.273	0.029	0.275	0.046	0.034	0.057
	25	500	-0.195	0.033	0.198	0.026	0.044	0.051
		1000	-0.188	0.026	0.190	0.032	0.032	0.045

While the two-step approach can be biased in both directions, the one-step Bayesian approach, which combines IRT and regression, yields satisfactory results regardless of the prior (see Table 2.4Table 2.4).

Table 2.4 β estimation using the one-step approach when IRT score error is present in the independent variable

Generating value of parameter ρ	Number of items	Number of students	$\hat{\beta}_{1.prior1}$: One-step Bayesian MCMC approach with prior 1 (N = 100)			$\hat{\beta}_{1,prior2}$: One-step Bayesian MCMC approach with prior 2 (N = 100)		
			Bias	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*
0.2	15	500	0.001	0.051	0.051	0.000	0.052	0.052
		1000	0.001	0.038	0.038	0.000	0.038	0.038
	25	500	0.003	0.047	0.047	0.000	0.047	0.047
		1000	0.006	0.030	0.031	0.004	0.030	0.030
0.5	15	500	-0.011	0.035	0.037	-0.014	0.035	0.038
		1000	0.000	0.031	0.031	-0.002	0.031	0.031
	25	500	-0.003	0.038	0.038	-0.007	0.038	0.039
		1000	-0.002	0.024	0.024	-0.005	0.024	0.025
0.8	15	500	-0.003	0.023	0.023	-0.004	0.023	0.023
		1000	0.001	0.016	0.016	0.001	0.016	0.016
	25	500	-0.002	0.017	0.017	-0.004	0.018	0.018
		1000	-0.002	0.012	0.012	-0.002	0.012	0.012

Comparisons of simple linear regression coefficient β recovery when error is present in the independent variable are shown in the boxplot from Figure 2.4 to Figure 2.6. When estimation error is present in the IRT score, the one-step approach is a much better choice than the two-step approach.

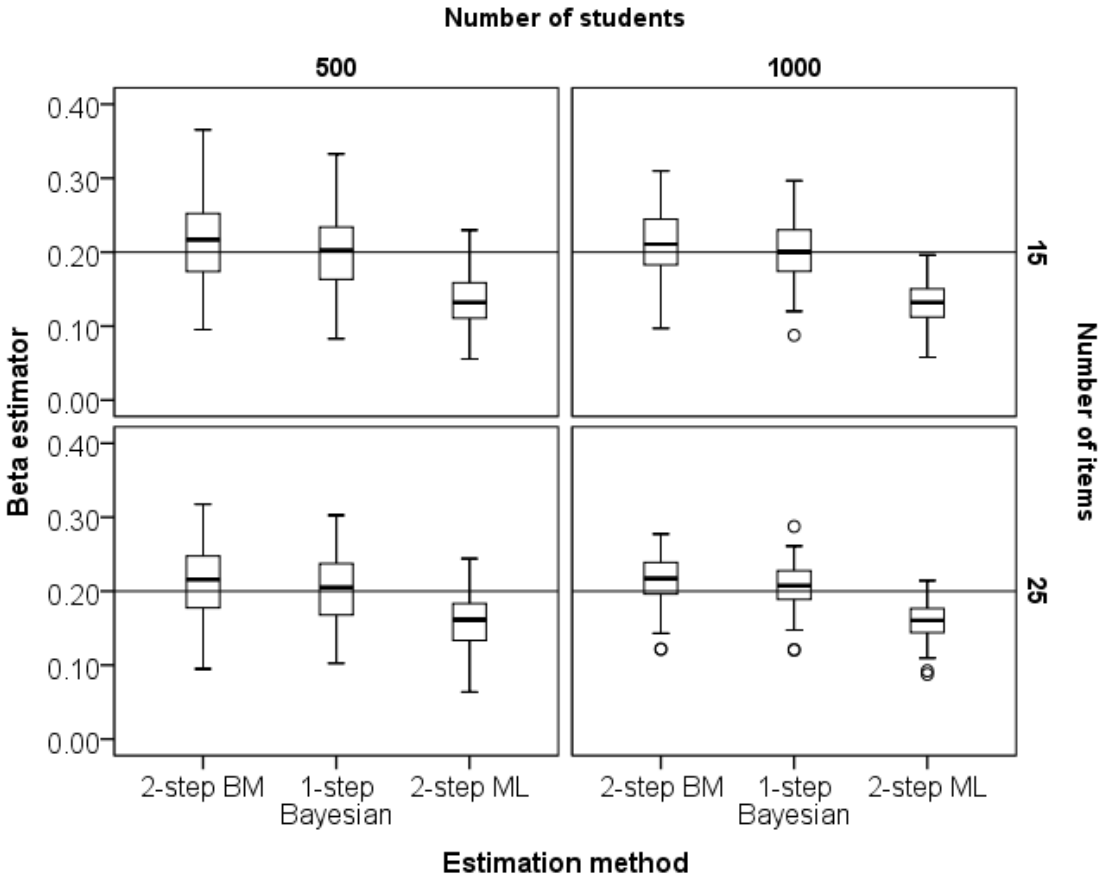

Figure 2.4 Comparison of β estimation methods when IRT score error is present in the independent variable and ρ = 0.2.

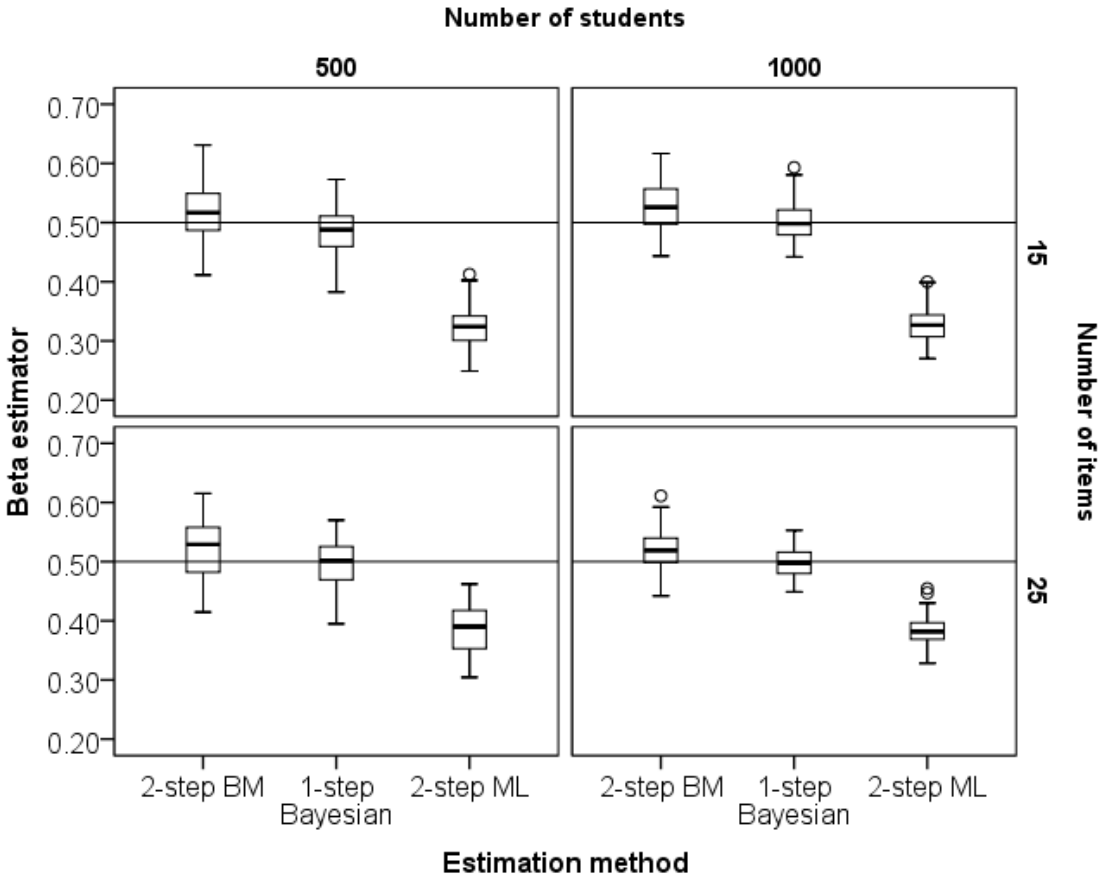

Figure 2.5 Comparison of β estimation methods when IRT score error is present in the independent variable and ρ = 0.5.

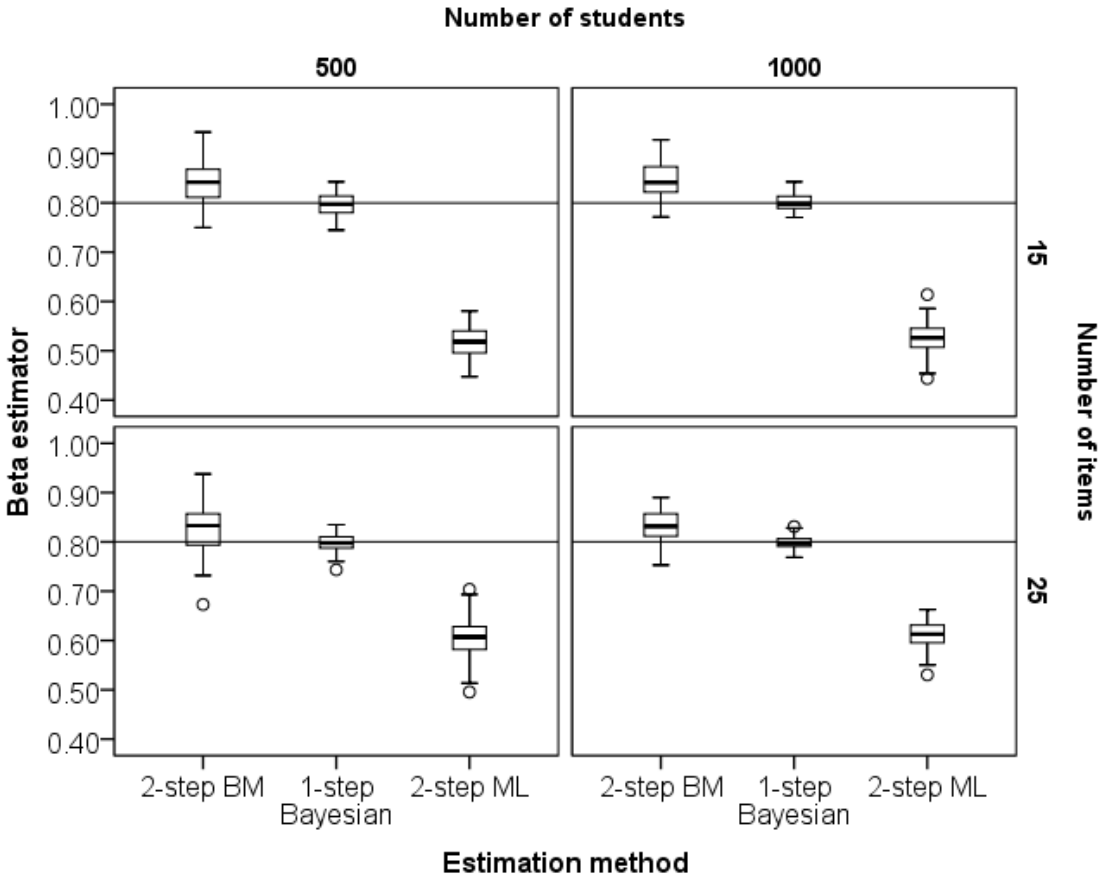

Figure 2.6 Comparison of β estimation methods when IRT score error is present in the independent variable and ρ = 0.8.

3) IRT score error in both dependent and independent variables

From scenarios 1) and 2) we have learned that the maximum likelihood (ML) IRT score estimate in the dependent variable leads to overestimation of β while in the independent variable it leads to underestimation. When IRT score error is presented in both dependent and independent variables, the direction of bias is unpredictable with the combination of the two opposite effects. The same thing happens when IRT BM estimate is used. Whether the β estimation is overestimated or underestimated depends on which effect is dominant, but some bias is expected in most of the cases. Although β estimates are biased downward given the settings in this particular simulation (see Table 2.5), the direction of bias should be regarded as a coincidence rather than a conclusion.

On the other hand, the one-step Bayesian approach, which combines IRT and regression, yields satisfactory results regardless of the prior (see Table 2.6).

Table 2.5 β estimation using the two-step approach when IRT score error is present in both dependent and independent variables

Generating value of parameter ρ	Number of items	Number of students	$\hat{\beta}_{2.ML}$: two-step approach with IRT score from Maximum Likelihood (ML) estimate (N = 100)			$\hat{\beta}_{2.BM}$: two-step approach with IRT score from Bayesian Model (BM) estimate (N = 100)		
			Bias	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*
0.2	15	500	-0.051	0.047	0.069	-0.044	0.045	0.063
		1000	-0.052	0.034	0.062	-0.046	0.034	0.057
	25	500	-0.027	0.043	0.051	-0.025	0.043	0.050
		1000	-0.030	0.030	0.042	-0.026	0.030	0.040
0.5	15	500	-0.137	0.035	0.141	-0.124	0.033	0.128
		1000	-0.130	0.032	0.134	-0.116	0.030	0.120
	25	500	-0.086	0.043	0.096	-0.077	0.043	0.088
		1000	-0.084	0.028	0.089	-0.074	0.026	0.078
0.8	15	500	-0.213	0.035	0.216	-0.185	0.036	0.188
		1000	-0.203	0.032	0.206	-0.178	0.032	0.181
	25	500	-0.142	0.034	0.146	-0.122	0.030	0.126
		1000	-0.138	0.025	0.140	-0.119	0.021	0.121

Table 2.6 β estimation using the one-step Bayesian approach when IRT score error is present in both dependent and independent variables

Generating value of parameter ρ	Number of items	Number of students	$\hat{\beta}_{1.prior1}$: one-step Bayesian MCMC approach with prior 1 (N = 100)			$\hat{\beta}_{1,prior2}$: one-step Bayesian MCMC approach with prior 2 (N = 100)		
			Bias	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*
0.2	15	500	0.002	0.058	0.058	-0.002	0.057	0.057
		1000	0.001	0.044	0.044	-0.001	0.044	0.044
	25	500	0.005	0.051	0.051	-0.001	0.049	0.049
		1000	0.005	0.036	0.036	0.001	0.035	0.035
0.5	15	500	-0.014	0.041	0.043	-0.020	0.041	0.046
		1000	-0.001	0.040	0.040	-0.005	0.040	0.040
	25	500	-0.004	0.046	0.046	-0.016	0.046	0.049
		1000	0.002	0.030	0.030	-0.005	0.030	0.030
0.8	15	500	-0.004	0.030	0.030	-0.008	0.032	0.033
		1000	0.002	0.023	0.023	-0.001	0.024	0.024
	25	500	-0.005	0.027	0.027	-0.014	0.028	0.031
		1000	-0.001	0.017	0.017	-0.006	0.017	0.018

Comparisons of simple linear regression coefficient β recovery when errors are present in both independent and dependent variables can be found from Figure 2.7 to Figure 2.9. When estimation error is presented in the IRT score, the one-step approach is a much better choice than the two-step approach. The one-step approach yields an estimator very close to the true parameter, which means the error impact has been reduced in this approach. Also one-step approach is more efficient than the two-step approach because the standard deviations of the simulated estimates are smaller. The larger the correlation, the less spread for the estimates from Bayesian approach. So the one-step Bayesian estimator is more efficient than the two-step estimators.

Also it should be mentioned that the scale for a correlation is restricted and the spread of β estimates for all approaches will get smaller when ρ is getting closer to 1 or -1 (Fisher, 1915). The smaller spread is an artifact of the scale for both two-step approach and one-step Bayesian approach. But for each fixed level of ρ in this study, β estimates from one-step Bayesian approach have a smaller spread than those from two-step approach.

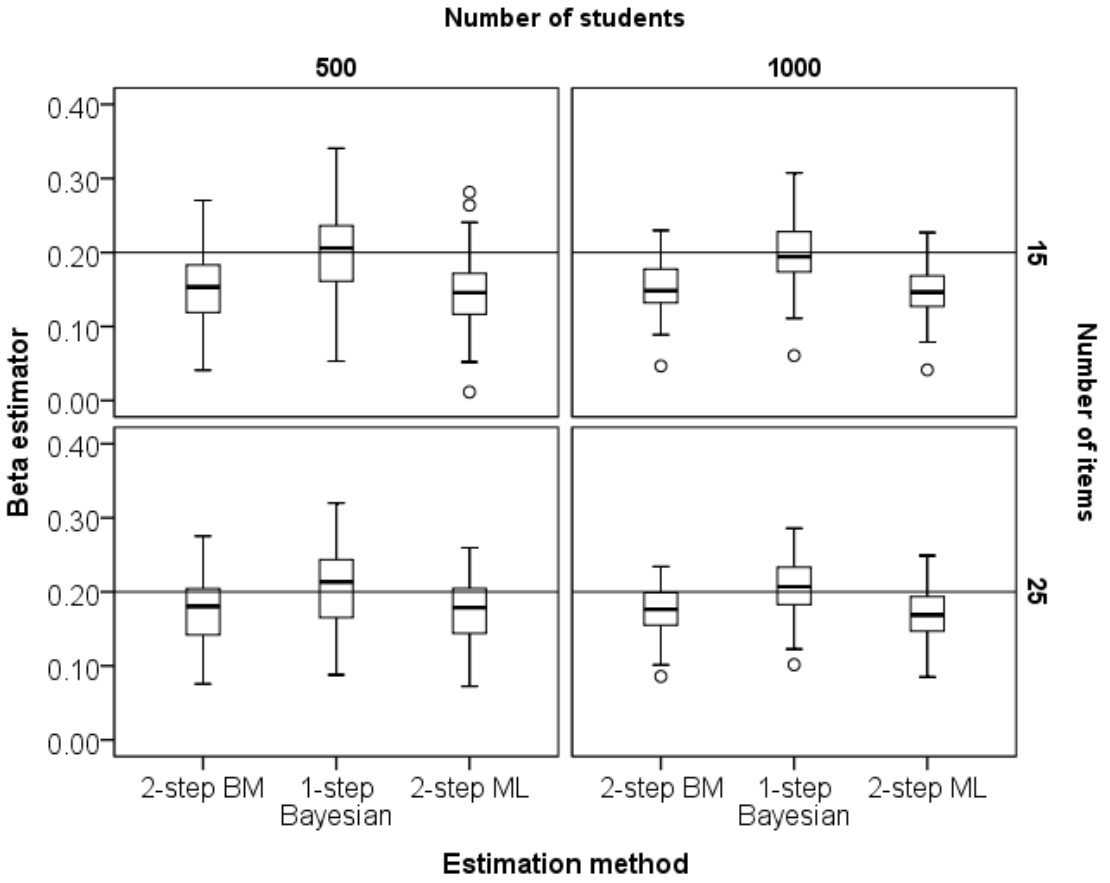

Figure 2.7 Comparison of β estimation methods when IRT score error is present in both variables and ρ = 0.2

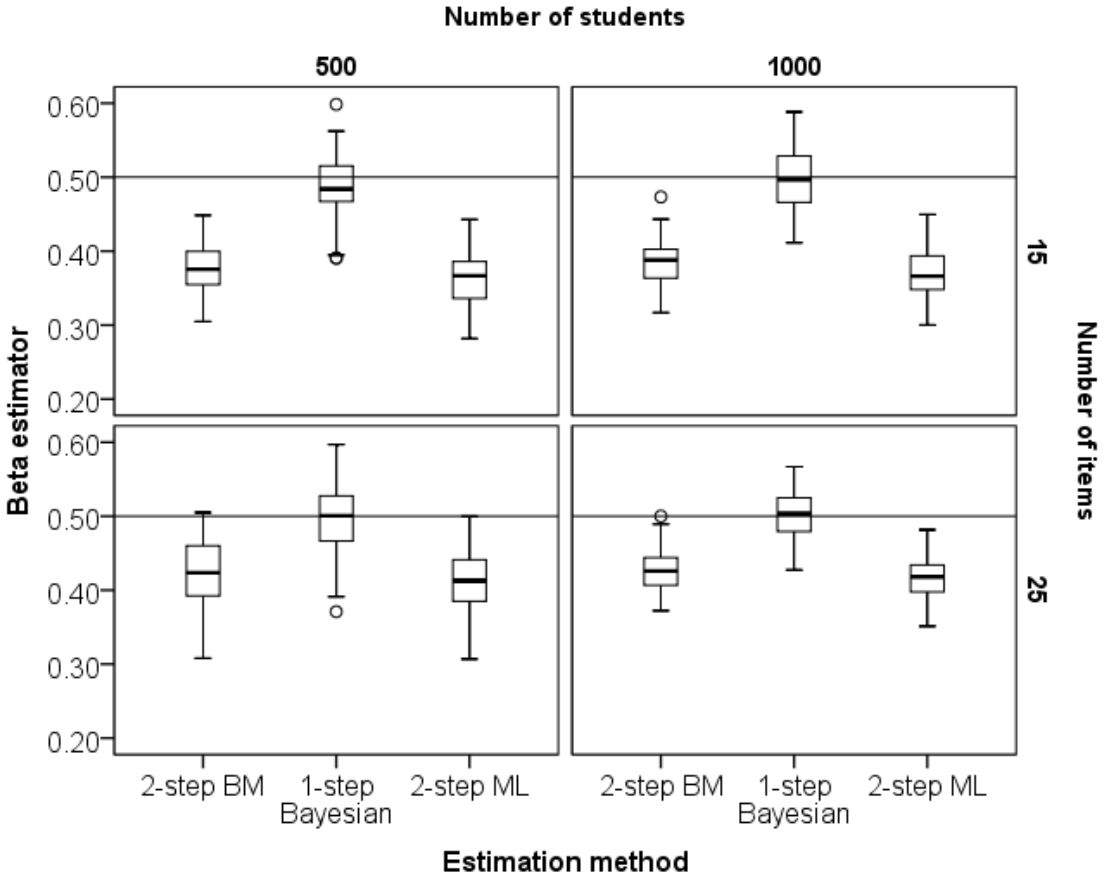

Figure 2.8 Comparison of β estimation methods when IRT score error is present in both variables and ρ = 0.5

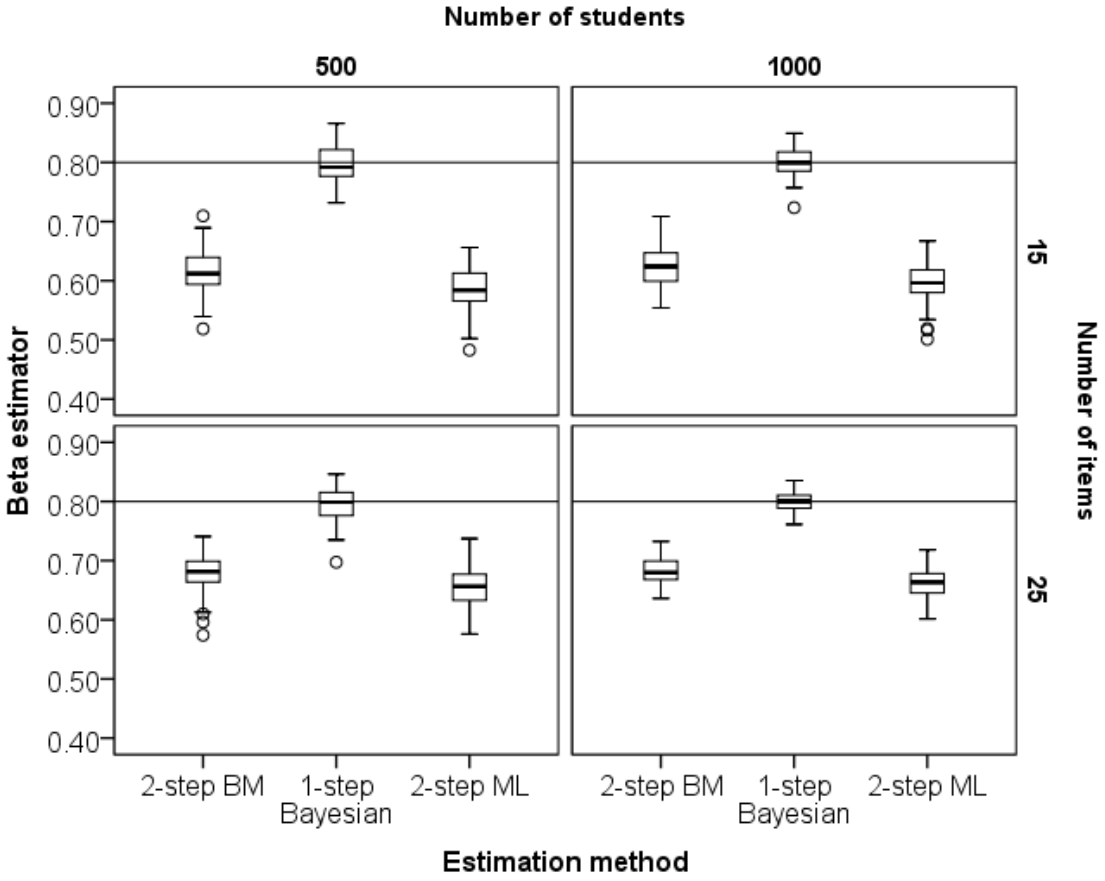

Figure 2.9 Comparison of β estimation methods when IRT score error is present in both variables and ρ = 0.8

More about Bayesian: comparison of priors

One thing that should be considered when using a Bayesian method is selecting priors for the parameters. This study used two sets of priors to see if the results are heavily influenced by the priors. One set of priors is $a \sim \log N(0, 0.25)$ and $b \sim N(0, 0.5)$, and the other is $a \sim \log N(0, 0.25)$ and $b \sim N(0, 4)$. The details of the priors setting can be found in previous section. The results are quite similar which means the results are rather stable and not very sensitive to these two commonly used priors. But it should be cautioned that the misspecified priors could badly

influence the results. Researchers should carefully select those reasonable priors and compare the results during their own study.

More about Bayesian: convergence test results

The majority of the MCMC chains passed the convergence test when the first 2,000 were discarded as burn-in and the total length was 4,000. Each of the 7200 (100 replications × 3 levels of β × 2 levels of student size × 2 levels of test length × 3 scenarios × 2 sets of priors) simulations achieved stationarity before reaching the 20,000 arbitrary maximum limits (see Table 2.7).

Table 2.7 The burn-in length for the Markov chains when passing the convergence test

Burn-in length	Frequency	Percent
2,000	7149	99.3
6,000	50	0.7
10,000	1	0.0
Total	7,200	100.0

CHAPTER 3 IRT SCORE ERROR IMPACT IN A MULTILEVEL VALUE-ADDED MODEL

Model and presence of IRT score error

Model: value-added models (gain score models)

Value-added models (McCaffrey et al. 2004) have been a focus of educational research. The issue of measurement error in gain scores is well known. Linn (1977) reviewed some of the major issues that arise in the measurement of change. Lee et al. (2000) found 84 percent of the variance in gain scores is attributable to measurement error. They concluded that measurement error will only affect precision of estimates and it can be compensated for with sufficiently large numbers of observations. Fischer (2003) investigated the precision of gain scores and simple difference scores, and compared the asymptotic and exact conditional inference about change. He proposed an IRT framework for the measurement of change. Wang et al. (2004) proposed a procedure to obtain the IRT-based effect size measure. That procedure they used to correct for measurement error is based on IRT-based test reliability. Also there have been significant advances into longitudinal study designs (Hsieh, et al., 2010; Lockwood, McCaffrey, Mariano, & Setodji, 2007; Natesan, Limbers, & Varni, 2010; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003; Seltzer, Choi, & Thum, 2003).

It is beyond the scope of this study to explore the IRT measurement error impact on general value-added models. The focus will be put on a special yet typical case in which only two test occasions, pretest and posttest, are presented. General value-added models that have more than two tests will be left for future studies. The model of interest here was inspired by a research question in ME.ET project (McCrory, Zhang, Francis, & Young, 2009). It is a

combination of multilevel model, value-added model and item response model. There are two components to the combination model: a multilevel linear regression model to describe the relationship among variables, and a measurement part which includes several IRT models to describe the relationship between latent variables and their indicators.

It should be clarified that in most schools in this study there is only one class in each school and one teacher in each class. The class level is actually the second level used for HLM analysis. But to simply the notation, the terms school, class and teacher are used interchangeably when referring to the second level variables.

1) Structural component: multilevel linear model

The structural model is a multilevel linear latent growth model with only two occasions: pretest and posttest. It can be written in different formats. Readers from different trainings may prefer one over another. Two different formats are all listed here to ease the burden of switching from one to the other.

Format 1: structural model written in compact form

The structural model can be written in the format used in McCaffrey et.al. (2004) as the following:

$$\theta_{ijt} = \gamma_{000} + u_{00} + \gamma_{010}Attitude + \gamma_{020}Gender + r_{0ij} + t * \left(\gamma_{100} + u_{10} + \gamma_{101}Method + \gamma_{102}Textbook + r_{1ij}\right) \quad (3.1)$$

Or it can be written separately as at pretest (t = 0) and posttest (t = 1)

$$\theta_{ij0} = \gamma_{000} + \gamma_{010} * Attitude + \gamma_{020} * Gender + u_{00} + r_{oij} \quad (3.2)$$

$$\theta_{ij1} = \theta_{ij0} + \gamma_{100} + \gamma_{101} * Method + \gamma_{102} * Textbook + u_{10} + r_{1ij} \quad (3.3)$$

where θ_{ijt} is the latent achievement of student *i* in school *j* at time t; γ_{000} is the grand mean of all schools at pretest and μ_{00} is the variation of school means. The residual term r_{0ij} is the student initial achievement (pretest) variation within school. They are independently and identically distributed (*i.i.d.*) errors that follow a Normal distribution *N*(0, σ_0^2). γ_{100} is the mean school added value and μ_{10} is the variation of school value added. r_{1ij} is the residual error term with a distribution *N*(0, σ_1^2) that represents the variation of student gain within school. There are two explanatory variables, *Attitude* and *Gender*, for the student's initial status, θ_{ij0}. There are another two explanatory variables, *Method* and *Textbook*, for the added value (gain score). The coefficients of the independent variables, *Attitude, Gender, Method* and *Textbook*, are denoted as γ_{010}, γ_{020}, γ_{101}, and γ_{102} respectively.

Format 2: structural model in multilevel form

The above structural model can also be written in the following format that is popular among HLM software users.

a. Repeated test level / Value-added Model / Growth model

$$\theta_{ijt} = \pi_{ij} + t * \delta_{ij} \qquad (i = 1, \dots I; j = 1, \dots, J; t = 0,1) \quad (3.4)$$

where θ_{ijt} is the achievement of student *i* in school *j* at pretest (*t* = 0) or posttest (*t* = 1). π_{ij} is the initial status of student *i* in school *j* and δ_{ij} is the gain of that student. This is a special case of a latent growth model. There is no error term or residual term in the model because the measurements are carried out only twice, one at pretest and the other at posttest, thus the line (3.4) is defined by those two points. In case the students are measured at more than two occasions, an error term should be added to the equation. More discussions on latent growth models can be found in Raykov and Marcoulides (2006).

b. Student level

$$\pi_{ij} = \beta_{00j} + \beta_{01j}\, Attitude + \beta_{02j} Gender + r_{0ij} \quad (3.5)$$

$$\delta_{ij} = \beta_{10j} + r_{1ij} \quad (3.6)$$

At student level, there are two explanatory variables, *Attitude* and *Gender*, for the student's initial status π_{ij}. The coefficient of those two variables are denoted as β_{01j} and β_{02j}. The conditional mean initial status of a student in school *j* is denoted as β_{00j}. The student's initial status π_{ij} is a random variable varies around the predicted value with variation of r_{0ij}. And these is no independent variable at student level to predict student gain δ_{ij}. The student's gain δ_{ij} is a random variable varies around β_{10j}, the mean added value of school *j*, with variation of r_{1ij}.

c. School level

$$\beta_{00j} = \gamma_{000} + \mu_{00} \quad (3.7)$$

$$\beta_{01j} = \gamma_{010} \quad (3.8)$$

$$\beta_{02j} = \gamma_{020} \quad (3.9)$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101}Method + \gamma_{102}Textbook + \mu_{10} \quad (3.10)$$

At school level, there are two explanatory variables, *Method* and *Textbook*, to predict the average school gain β_{10j}. In this particular case, it is assumed that no school level variable can explain the school initial status. The *Attitude* and *Gender* effects β_{01j} and β_{02j} are assumed to be fixed across schools as γ_{010} and γ_{020}. The school initial status β_{00j} is a random variable varies around mean initial status γ_{000} of all schools with variation μ_{00}. The conditional school gain is denoted as γ_{100}. The added value of school *j* β_{10j} is a random variable that varies around the predicted value with variation μ_{10}.

2) Measurement components: IRT models

The above structural model in section 1) is a typical multilevel linear model or hierarchical linear model (HLM). What makes the model in educational settings unique is that it has latent variables and those variables can be further described with two IRT models. Particularly in the above example, three latent variables are of interest: student achievement, student attitude and teacher's teaching method.

a. Student achievement measurement model

Student achievement is a latent variable. It cannot be observed directly but it can be estimated from students' responses to test items. The probability function for a student's response to a test item can be written as

$$P\left(Y_{ijkt}=1\right)=\frac{e^{a_k(\theta_{ij}+\delta_{ij*}t-b_k)}}{1+e^{a_k(\theta_{ij}+\delta_{ij*}t-b_k)}} \tag{3.11}$$

where Y_{ijkt} is the score of student i in school j answering item k at time t. θ is the student initial achievement and δ is the added value, i.e. the increased achievement between posttest and pretest. Item characteristics are represented by a_k and b_k, as they are in a standard 2PL IRT model. To be specific, a_k is the discrimination parameter and b_k is the difficulty parameter of item *k*.

In fact, the model above is a standard two-parameter logistic (2PL) IRT model (Lord, 1980) with two additional parameters *t* and δ. Parameter *t* is the time indicator with which *t* = 0 represents pretest and *t* = 1 posttest. Parameter δ is the increased student achievement between posttest and pretest. It is the added value from the teaching program point of view.

b. Student attitude measurement model (Graded Response Model)

Student's attitude ω is another latent variable. It is measured by a questionnaire consisting of nine five-point Likert-type items. A student's respond to each attitude question fall into one of the five categories from strongly disagree to strongly agree. The responses are coded 1 – 5 respectively. An IRT Graded Response Model (GRM, Samejima 1969) was used to describe the relationship between the latent variable ω_{ij} and the response X_{ijm}.

$$P\left(X_{ijm} \geq x\right) = \frac{e^{a_m(\omega_{ij} - b_{mx})}}{1 + e^{a_m(\omega_{ij} - b_{mx})}}, (x = 2, \dots, 5) \tag{3.12}$$

where P is the probability of student i in school j scoring x (x = 1, …, 5) or above on question m (m = 1, …, M) in the student questionnaire.

c. Teaching method measurement model (Graded Response Model)

Teacher's *teaching method* ζ is another latent variable. It is measured by a questionnaire consisting of 11 four-point Likert-type questions. A teacher's respond to each method question falls into one of four categories from strongly disagree to strongly agree. The responses are coded 1 – 4 respectively. An IRT Graded Response Model (GRM) was used to describe the relationship between the latent variable, ζ_j , and the response, Z_{jn}.

$$P\left(Z_{jn} \geq x\right) = \frac{e^{a_n(\zeta_j - b_{nx})}}{1 + e^{a_n(\zeta_j - b_{nx})}}, (x = 2, \dots, 4) \tag{3.13}$$

where P is the probability of teacher j scoring x (x = 1, …, 4) or above on question n (n = 1, …, N) in the teaching method questionnaire.

Presence of IRT score error

The impact of measurement error is pertinent to whether the variable is a dependent variable or independent variable. To take this into consideration, different scenarios will be explored: 1) IRT score error in dependent variables only, i.e. the student pretest and posttest abilities are unknown and their scores are estimated based on their answers to the test questions.

At the same time, the true values of values, of student attitude and teacher teaching method are entered into the model assuming they are known; 2) IRT score error in independent variables only, i.e. estimated student attitude score and teacher teaching method score are used while true student achievement are treated as known; 3) IRT score error in both independent and dependent variables. All the true values of those latent variables are unknown and their estimated values are used, and 4) no IRT score error in neither variable. All true values of those latent variables are used to estimate the coefficients.

Data generation

Data are generated following the multilevel value-added measurement model above. It is assumed that there are 40 schools and 30 students in each of them so that the whole dataset has a magnitude similar to the empirical data.

First, the values of latent teacher variable *method* ζ are randomly drawn from standard normal distribution $N(0, 1)$. For each school, the values of variable *textbook*, which is an indicator of whether use a certain type of textbook, were independently generated from a Bernoulli distribution $B(0.65)$. The parameter was set to 0.65 to mimic the fact that 65% of the schools use the same type of textbook in the empirical study. Since there is no indication of textbook dependence on teaching method ($p = 0.09$), values for the indicator variable *textbook* were generated independently.

Second, the values of student latent variable *attitude* ω are randomly drawn from standard normal distribution $N(0,1)$. For each student, the gender variable *female* is drawn from a Bernoulli distribution $B(0.9)$ since 90% of the students are female in ME.ET project. As there is

no indication of dependence between *female* and *attitude*, the two variables are generated independently.

Third, the generating values of fixed effects of *attitude* (= $\beta_{01j} = \gamma_{010}$), *female* (= $\beta_{02j} = \gamma_{020}$), *method* (= γ_{101}) and *textbook* (= γ_{102}) are set to their estimated values 0.14, 0, 0.3 and 0.48, which were obtained from traditional two-step approach.

The values of conditional school *initial status* u_{00} and *gain* u_{10} are generated from a multivariate normal distribution $N\left(\begin{bmatrix}-0.4\\0.5\end{bmatrix}, \begin{bmatrix}0.08 & -0.03\\-0.03 & 0.08\end{bmatrix}\right)$. The values of the generating parameters are obtained from the empirical study too. The study showed that higher school pretest score is negatively associated with the school gain. The covariance of the two variables is -0.03, which is equivalent to a correlation of -0.38.

The values of conditional student initial status r_{0ij} and gain r_{1ij} are generated from a multivariate normal distribution $N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0.23 & r\\r & 0.03\end{bmatrix}\right)$. The variances of r_{0ij} are estimated values from ME.ET study, but the covariance of r_{0ij} and r_{1ij} were NOT. The reason for ignoring the estimated covariance is that the number is equivalent to a correlation as extreme as -0.98. It is a known inaccurate estimator in pre-post studies (Raudenbush, 2002). The Werts et al. (1977) study found that the correlations of initial status and gain vary within a moderate range from -0.45 to 0.45 in practice. To explore whether this correlation can be successfully recovered by the Bayesian approach, three levels of correlation, 0.4, 0, and -0.4, have been used to generate the simulation data.

Finally, pretest and posttest response patterns were generated according to the measurement model (3.11), for a pretest of 26 items and a posttest of 25 items. Student attitude response and teaching method response patterns were generated according to the GRM model (3.12) and (3.13). The generating values of the item parameters are shown in the Table 6.1 to Table 6.3 in the Appendix A.

A number N (N = 100) of replications for each case were carried out. In total there were three levels of correlation × 100 replications = 300 datasets for later analysis.

Software R (R Development Core Team, 2008) was used for data generation. The data generation of student responses to test items and the recovery of latent construct θ_x and θ_y have been made easier by using *irtoys* package (Partchev, 2006) in R.

Parameter Recovery

Parameters recovery: 1) two-step approach

At step one, student pretest score and posttest score θ_{ijt}, are estimated using two parameter logistic (2PL) model, and only maximum likelihood (ML) IRT estimates are obtained for student achievement. Again, the student achievement estimates were obtained using *irtoys* package (Partchev, 2006) in software R.

The item parameters of pretest and posttest were treated as known, as is the case in ME.ET project. Another reason to use known item parameters is to avoid the equating issue which is not a focus of this study.

Also at step one, student attitude ω and teacher teaching method ζ are estimated using Graded Response Model (GRM). The values of these latent variables are estimated using the

ltm package (Rizopoulos,2006) in software R. Parameters of student *attitude* and *teaching method* items will still be treated as unknown, as is the case in ME.ET project.

At step two, the estimated IRT score $\hat{\theta}$, $\hat{\delta}$, $\hat{\omega}$ and $\hat{\zeta}$ are entered into the structural model. The regression coefficients are calculated with HLM6 software (S. Raudenbush, Bryk, Cheong, & Congdon, 2004) .

Parameters recovery: 2) one-step Bayesian MCMC approach

For the one-step Bayesian approach MCMC, the response matrix of pretest, posttest, attitude and teaching method are used to estimate the student latent variable and regression coefficient γ simultaneously. The model codes were written ad hoc. Even though the model seems as simple as shown above, the code can be tricky (see Appendix C). They are implemented with software WinBUGS/OpenBUGS (D. Lunn, Spiegelhalter, D., Thomas, A., Best, N., 2009).

The Bayesian MCMC iteration is time consuming, especially when the 2PL and GRM model are involved simultaneously. It usually takes hours, sometimes even nearly 10 hours depending on the length of Markov chain, to finish a single run. The 300 simulations in design couldn't have been completed within a reasonable period of time if the enormous computing power had not been employed with the support from the experts in MSU High Performance Computing Center (HPCC).

Bayesian approach details: setting priors

The prior settings of the combination model are listed below.

All the latent covariate (student attitude and teacher teaching method) are set to be from a standard Normal distribution $N(0, 1)$, which is the common basic assumption of IRT models.

The fixed effects of the four independent variables (student attitude, student gender, teacher teaching method and textbook usage) were assumed to follow independent Normal distributions with zero mean and precision = 10 which is equivalent to a variance of 1/10. The priors are centered to 0 since there is no assumption on whether an independent variable has positive or negative impact on the dependent variable. The precision at first glance seems informative for a regression model, but it is rather vague, or moderate at most when the scale of the latent variables are taken into consideration. The variances of the priors are big enough for the Markov chain to explore the plausible area between -1 and 1.

The study assumes a priori that the initial student achievement and the gain of student achievement are correlated both at school and student level. At school level, the initial status and gain were assumed to arise from a multivariate normal population distribution with unknown mean and covariance matrix Σ . The same noninformative independent Normal distributions with zero mean and precision = 10 was then specified for the population means, whilst the inverse covariance matrix $T = \Sigma^{-1}$ was assumed to follow a Wishart distribution. To represent vague prior knowledge, the scale matrix R was specified as $R = \begin{bmatrix} \hat{r}_{11} & 0 \\ 0 & \hat{r}_{22} \end{bmatrix}$, where $\hat{r}_{11}$ and $\hat{r}_{22}$ are estimated values from the output of HLM6 software. It is so chosen because there is no other clue on the magnitude of the variances in the real dataset and it is the best approximation one can get so far. Positive correlation would imply that students/schools with higher pretest scores tend

to gain more rapidly than those with lower pretest scores. However, the direction and magnitude of the correlation is not known so they are centered on 0 in the priori setting.

At the student level, the initial status and gain were assumed to come from a multivariate normal population distribution with known mean and covariance matrix Σ_X. The means are known and determined by the structural model, while the inverse covariance matrix $T_X = \sum_X^{-1}$ was assumed to follow a Wishart distribution. Again, the scale matrix R_x was specified as $R_x = \begin{bmatrix} \hat{r}_{x,11} & 0 \\ 0 & \hat{r}_{x,22} \end{bmatrix}$, where $\hat{r}_{x,11}$ and $\hat{r}_{x,22}$ were estimated values from the output of HLM6 software.

The BUGS code is listed in the Appendix C.

Bayesian approach details: convergence diagnosis

The convergence of Markov chains was evaluated using the Heidelberger and Welch Diagnostic Test (Heidelberger & Welch, 1983).

The starting point of each Markov chain was randomly generated according to the prior. The initial 3000 iterations are discarded as the burn-in, and the following 1000 iterations were used to generate the estimates of the parameters.

The length of the Markov chain is arbitrary. The decision was based on the observation of the trial runs. The outcomes became stable after 3,000 iterations. For certain parameters, their Markov chains reach a stationary distribution quickly and a longer chain makes little difference. Because of constraints on computing resources, the length of 4,000 was deemed to be adequate

for the purpose of evaluating the performance of the Bayesian method. The results are based on the last 1,000 iterations while the first 3000 were discarded as burn-in.

Results

The simulation outcomes show that error in IRT scores lead to biased regression coefficient estimates for the two-step approach, while the one-step Bayesian MCMC approach handles the problem better, albeit not as perfectly as was the case for simple regression. Generally speaking, the Bayesian approach yields less biased estimates (see Table 3.1 to Table 3.4) and more efficient (less variance) estimates for recovery of the true value of the coefficients (see Figure 3.1 to Figure 3.4). In other words, the Bayesian approach is more powerful for detecting the true relationship between those variables.

One thing the Bayesian approach does not handle perfectly is the recovery of the coefficient of the school level latent variable method (teaching method, see Figure 3.3). When this latent variable needs to be estimated, the coefficient is underestimated a little by the Bayesian approach. Most likely it is due to the small sample size of 40 at the school level.

Table 3.1 Coefficients recovery when no error present in latent variables

Correlation of student initial and gain		β_attitude (= 0.14)			β_female (= 0.00)			β_method (= 0.30)			β_textbook (= 0.48)		
		Bias	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*
-0.40	Bayesian	0.000	0.015	0.015	0.001	0.045	0.045	0.005	0.040	0.040	0.002	0.093	0.093
	two-step	-0.010	0.073	0.074	0.000	0.231	0.231	0.005	0.054	0.054	0.009	0.118	0.118
0.00	Bayesian	0.000	0.015	0.015	0.000	0.046	0.046	0.000	0.042	0.042	0.000	0.087	0.087
	two-step	0.006	0.061	0.061	0.042	0.198	0.202	0.002	0.049	0.049	-0.010	0.105	0.105
0.40	Bayesian	0.000	0.015	0.015	0.000	0.046	0.046	0.000	0.037	0.037	0.002	0.085	0.085
	two-step	0.008	0.068	0.068	-0.020	0.245	0.246	0.000	0.049	0.049	-0.010	0.117	0.117

Table 3.2 Coefficients recovery when errors in dependent variables (student scores)

Correlation of student initial and gain		β_attitude (= 0.14)			β_female (= 0.00)			β_method (= 0.30)			β_textbook (= 0.48)		
		Bias	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*
-0.40	Bayesian	0.002	0.021	0.021	0.000	0.060	0.060	0.008	0.048	0.049	-0.010	0.112	0.112
	two-step	0.000	0.095	0.095	-0.030	0.348	0.349	0.025	0.172	0.174	0.020	0.329	0.330
0.00	Bayesian	0.000	0.019	0.019	-0.010	0.057	0.058	-0.020	0.052	0.056	-0.030	0.110	0.114
	two-step	0.014	0.085	0.086	0.024	0.284	0.285	-0.010	0.163	0.163	0.019	0.330	0.331
0.40	Bayesian	0.000	0.018	0.018	0.000	0.068	0.068	-0.010	0.047	0.048	-0.030	0.100	0.104
	two-step	0.020	0.095	0.097	-0.060	0.345	0.350	0.035	0.163	0.167	-0.010	0.298	0.298

Table 3.3 Coefficients recovery when errors in independent variables (*attitude* and *method*)

Correlation of student initial and gain		β_attitude (= 0.14)			β_female (= 0.00)			β_method (= 0.30)			β_textbook (= 0.48)		
		Bias	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*
-0.40	Bayesian	0.000	0.016	0.016	0.003	0.046	0.046	-0.040	0.039	0.056	-0.010	0.099	0.100
	two-step	0.000	0.081	0.081	-0.010	0.235	0.235	0.011	0.076	0.077	0.007	0.125	0.125
0.00	Bayesian	0.000	0.016	0.016	0.000	0.047	0.047	-0.050	0.040	0.064	-0.010	0.092	0.093
	two-step	0.008	0.069	0.069	0.040	0.202	0.206	0.008	0.071	0.071	-0.010	0.111	0.111
0.40	Bayesian	0.000	0.018	0.018	0.002	0.047	0.047	-0.050	0.040	0.064	-0.010	0.091	0.092
	two-step	0.012	0.071	0.072	-0.010	0.245	0.245	0.006	0.072	0.072	-0.020	0.121	0.123

Table 3.4 Coefficients recovery when errors in both dependent and independent variables

Correlation of student initial and gain		β_attitude (= 0.14)			β_female (= 0.00)			β_method (= 0.30)			β_textbook (= 0.48)		
		Bias	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*
-0.40	Bayesian	0.001	0.021	0.021	0.000	0.062	0.062	-0.050	0.045	0.067	-0.010	0.106	0.106
	two-step	0.003	0.106	0.106	-0.030	0.345	0.346	0.028	0.175	0.177	0.015	0.344	0.344
0.00	Bayesian	0.000	0.021	0.021	-0.010	0.060	0.061	-0.060	0.044	0.074	0.006	0.116	0.116
	two-step	0.009	0.099	0.099	0.024	0.289	0.290	0.000	0.191	0.191	0.022	0.334	0.335
0.40	Bayesian	0.000	0.021	0.021	0.000	0.068	0.068	-0.050	0.045	0.067	0.000	0.102	0.102
	two-step	0.018	0.107	0.109	-0.060	0.349	0.354	0.043	0.182	0.187	-0.010	0.308	0.308

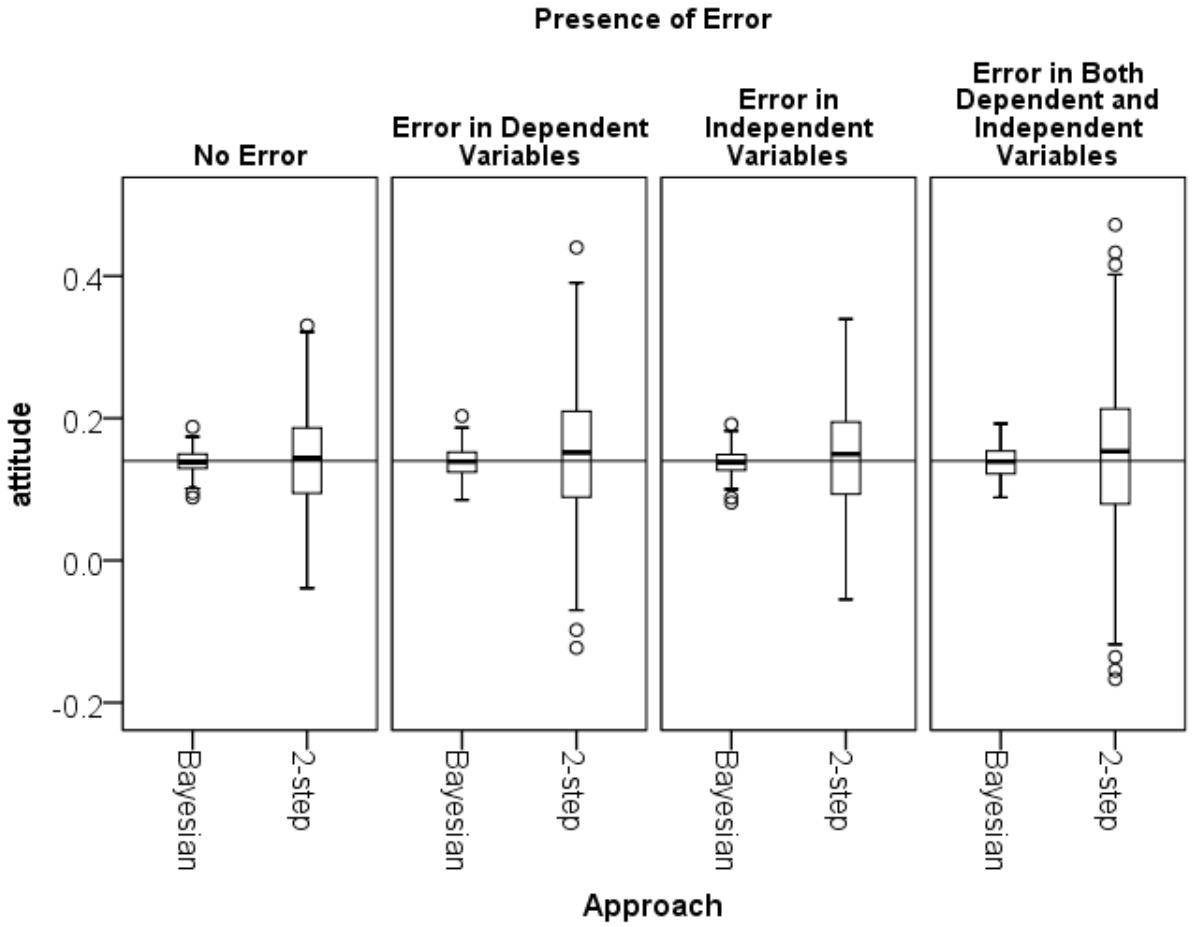

Figure 3.1 Coefficient of attitude (= 0.14) recovery

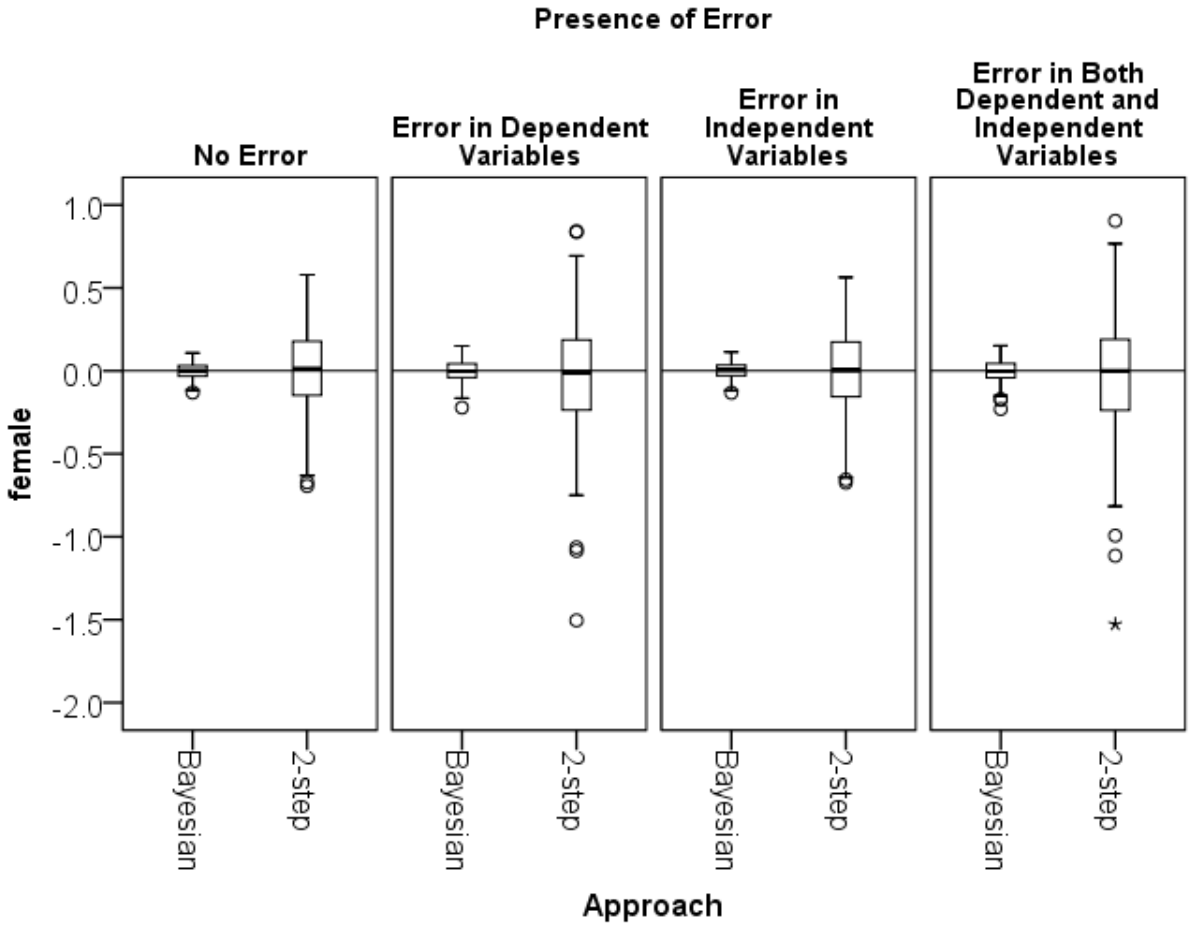

Figure 3.2 Coefficient of female (= 0.00) recovery

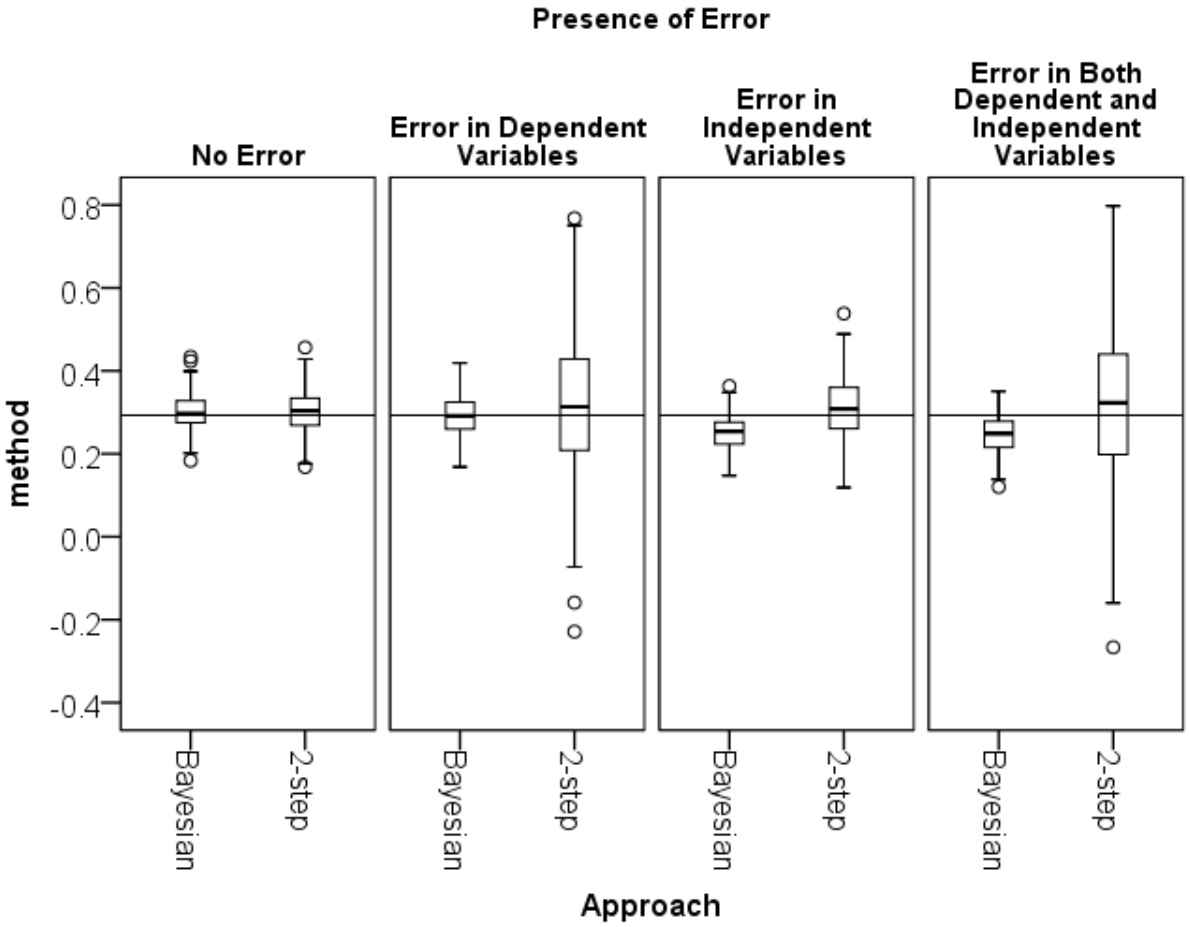

Figure 3.3 Coefficient of method (= 0.30) recovery

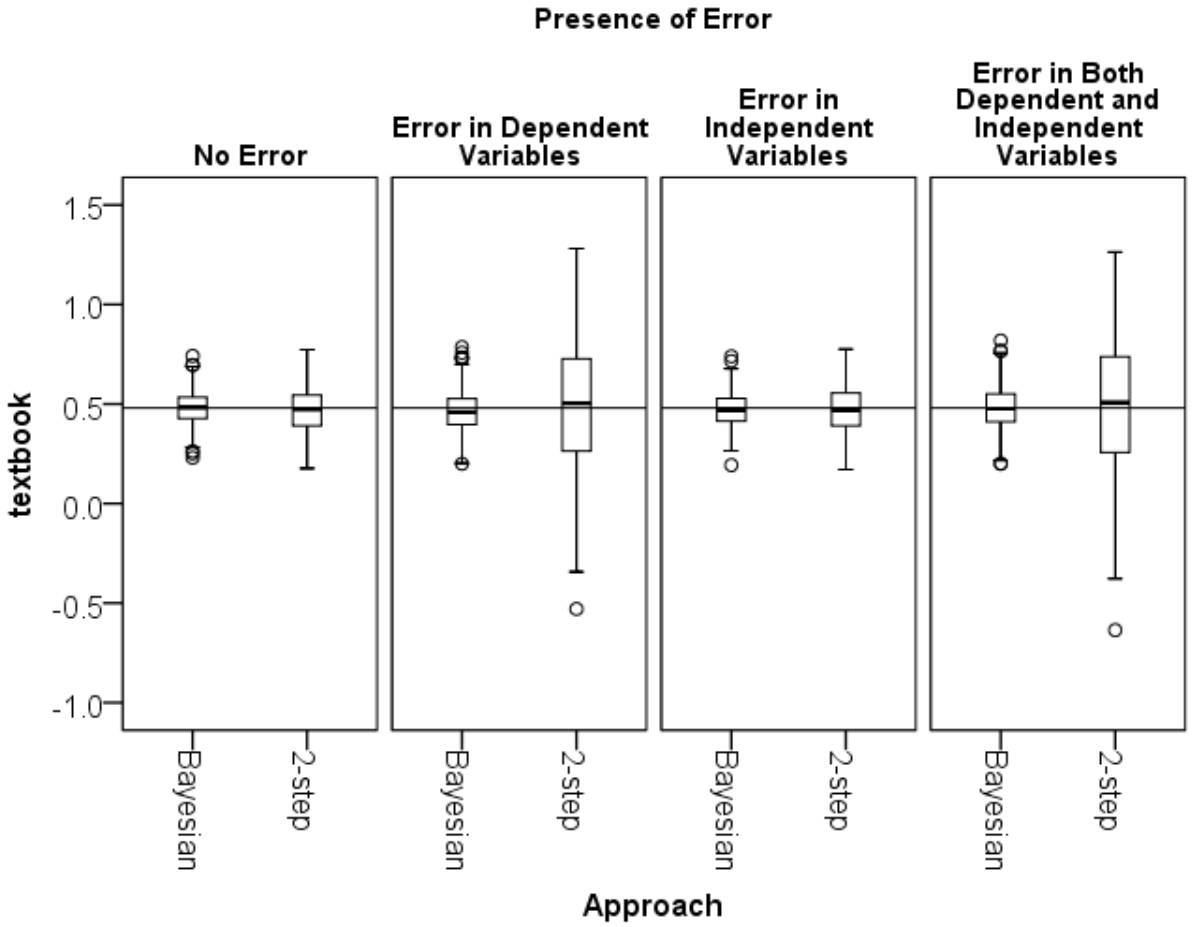

Figure 3.4 Coefficient of textbook (= 0.48) recovery

Although the focus of this study is on the recovery of regression coefficients, it will be interesting to take a look at the recovery of other parameters.

As mentioned earlier, the estimation of the correlation between student initial status (pretest) and gain is a known problem. To explore how the Bayesian approach can recover the correlation, three different levels of correlation (r = -0.4, 0, 0.4) between student initial status and gain were built into the simulation data. The result shows that neither Bayesian nor two-step approach can recover this correlation well, with Bayesian tending to generate a zero value while two-step a negative value (See Figure 3.5 to Figure 3.7). From the simulation we can conclude that those numbers should not be used in practice if they are obtained from the above two approaches.

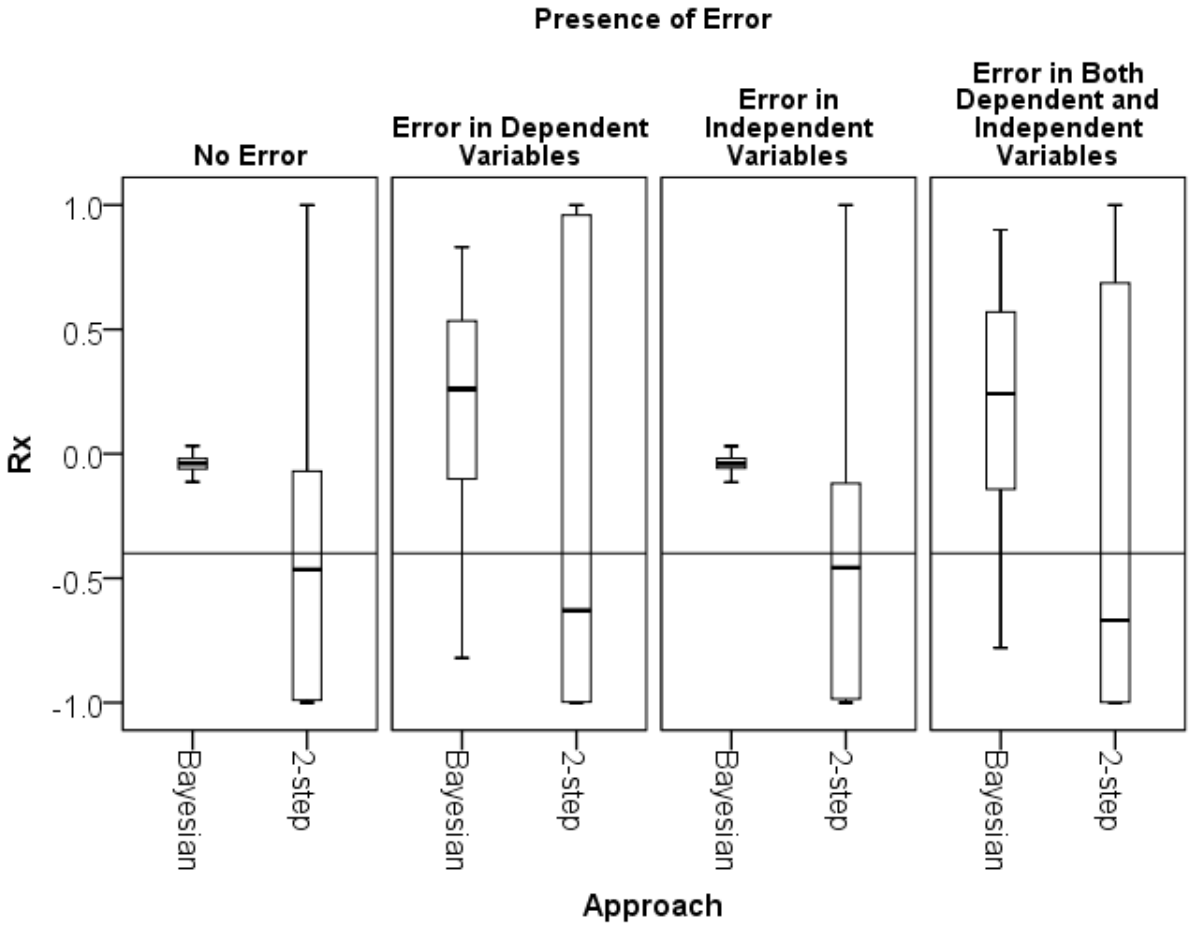

Figure 3.5 Recovery of correlation between student pretest and gain (= -0.4)

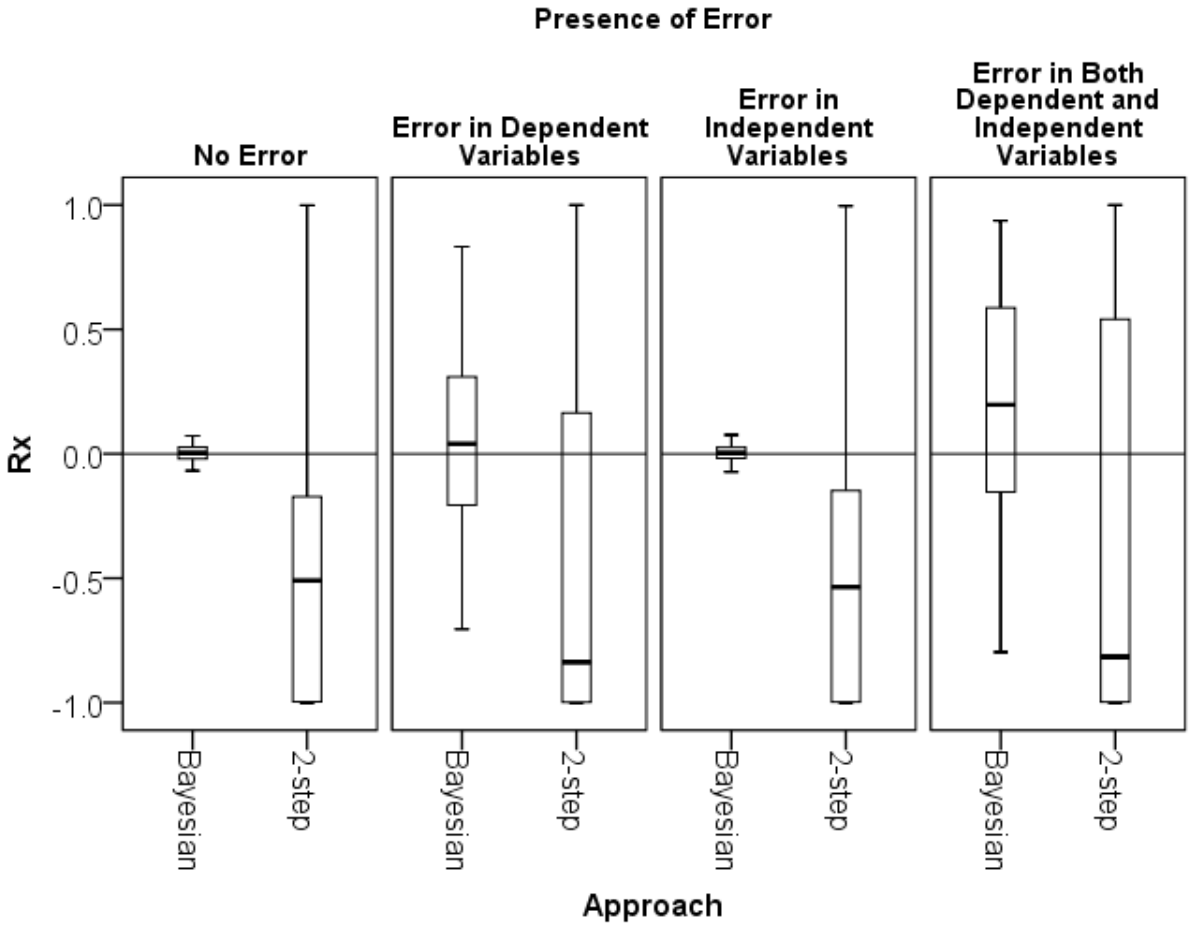

Figure 3.6 Recovery of correlation between student pretest and gain (= 0.0)

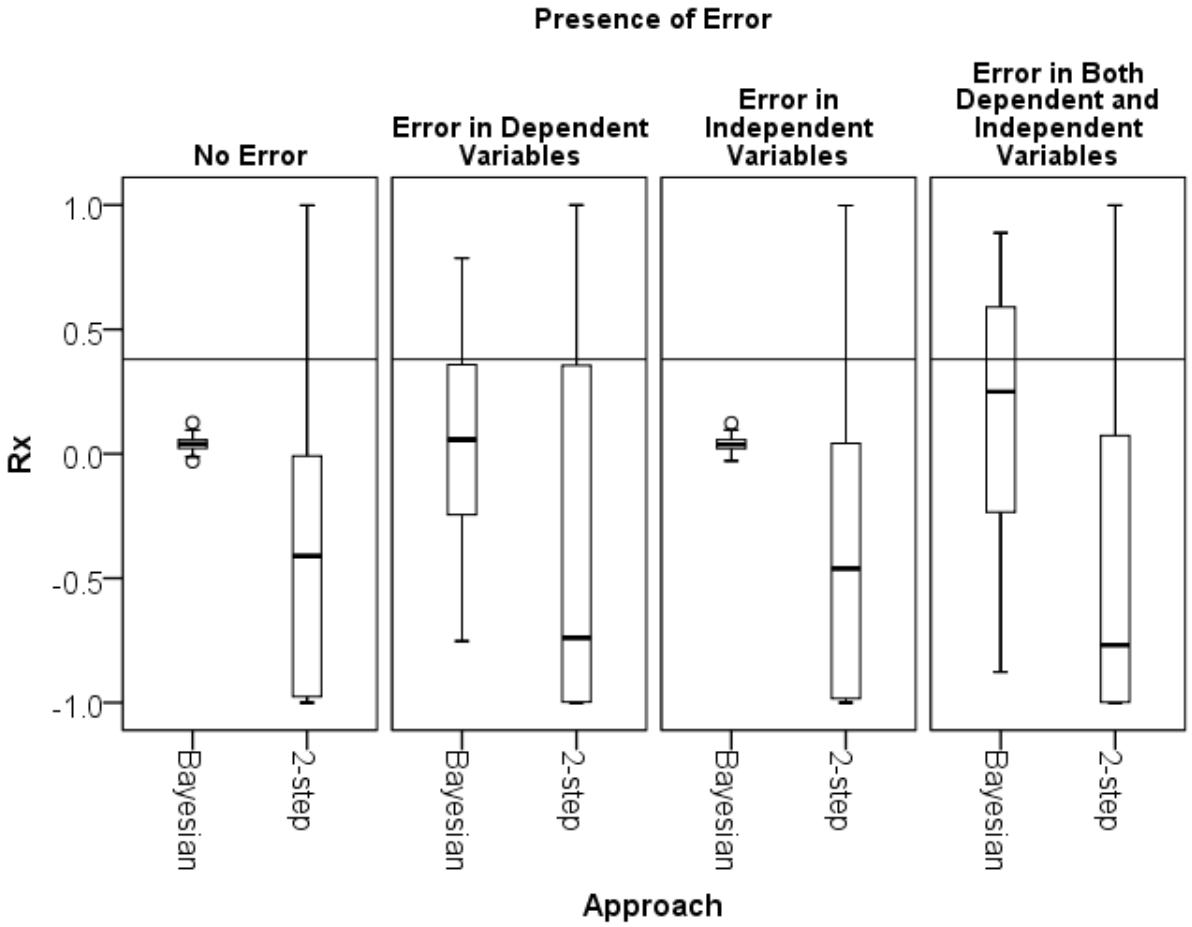

Figure 3.7 Recovery of correlation between student pretest and gain (= 0.4)

However, the Bayesian approach is much better than the two-step approach when estimating school level correlation between initial status and gain (see Figure 3.8). The Bayesian approach consistently gives the right answer while the two-step approach does not. Even when there is no error present in the variables, the two-step approach tends to yield zero correlation between pretest and gain at school level. Although the two-step approach did get close to the true correlation in some other scenarios, it is not clear whether it is a coincidence, given its performance in the no error present scenario.

In value-added models, one important interest is the performance of schools. In this model, it is captured by the school gain parameter r_{22}. From Figure 3.9 we can tell that the Bayesian approach can generally recover this parameter well in all scenarios. At the same time, the estimate from the two-step approach is not a reliable representation of the true school gain variance, thus any interpretation based on it will be called into question.

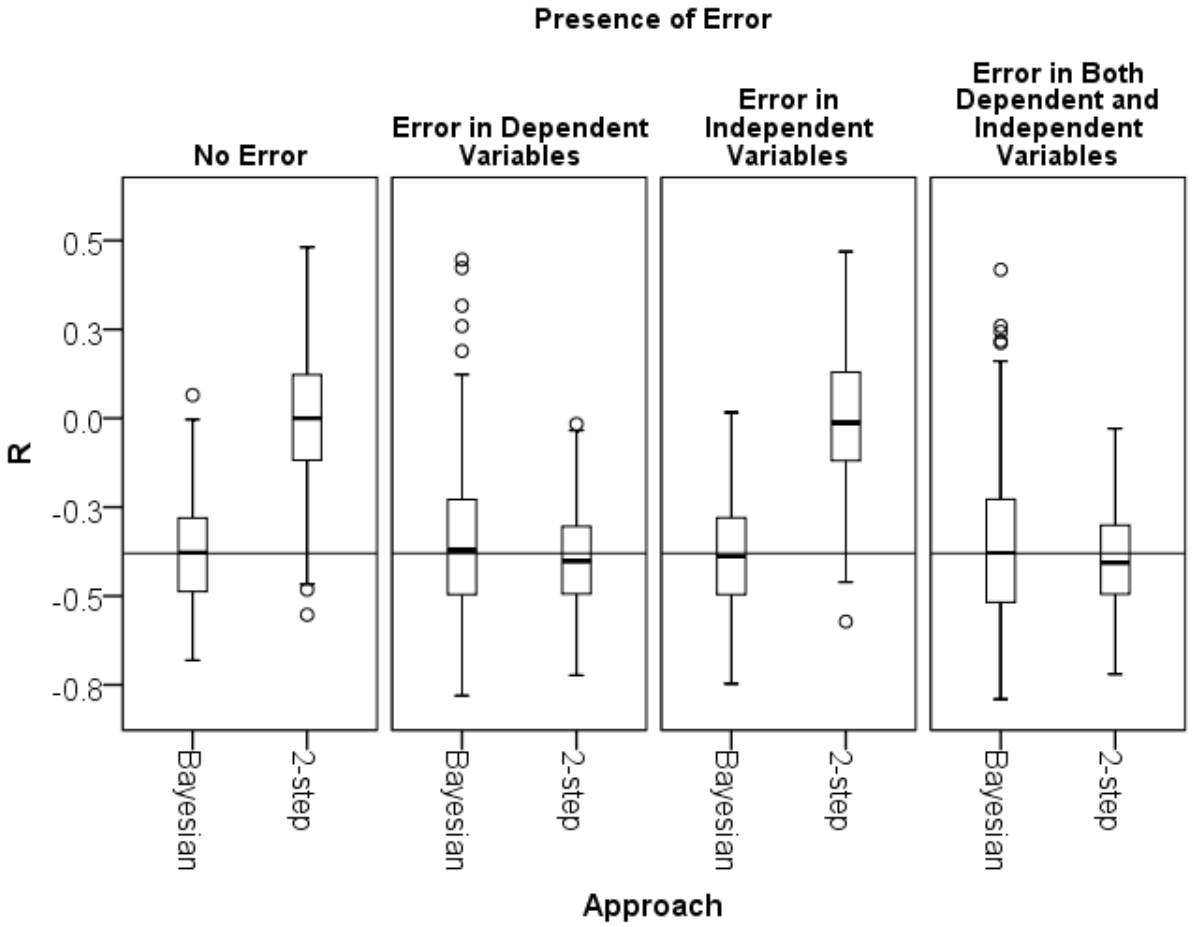

Figure 3.8 Recovery of correlation between school mean pretest and gain (= -0.38)

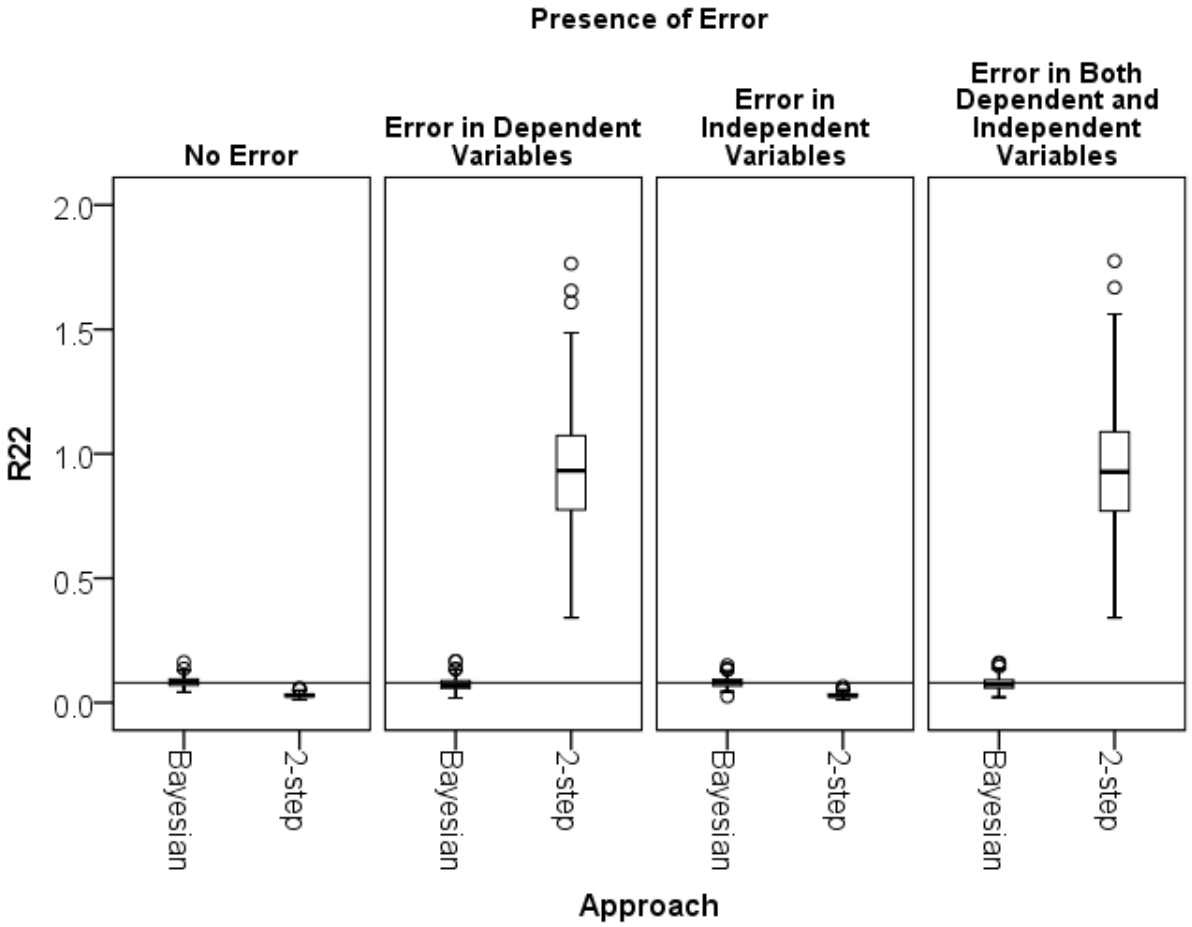

Figure 3.9 Recovery of variance of school gain (= 0.08)

More about Bayesian: setting priors

One concern in Bayesian applications is the setting of priors. Poor selection of priors could heavily distort the results. In this study, seven different priors were used to demonstrate the influence of priors. The priors for the four coefficients are all normal distributions with different standard deviations (See Table 3.5 and Figure 3.10) ranging from 0.1 to 10. From the results one can conclude that if the priors (e.g. prior 1 to prior 3) are too narrow to cover the true parameter, the results could be under the heavy influence of the prior. On the other hand, if the priors are uninformative (e.g. prior 4 to prior 7), the results are quite stable. Prior 4 was selected for the simulations in this chapter.

Table 3.5 Estimated β coefficients with seven different priors

	Priors for all four β coefficients						
	Prior 1 $N(0, 0.01)$	Prior 2 $N(0, 0.04)$	Prior 3 $N(0, 0.25)$	Prior 4 $N(0, 1)$	Prior 5 $N(0, 4)$	Prior 6 $N(0, 25)$	Prior 7 $N(0, 100)$
β_attitude	0.117	0.121	0.121	0.123	0.120	0.121	0.120
β_female	0.022	0.018	0.053	0.047	0.039	0.032	0.043
β_method	0.124	0.158	0.179	0.190	0.182	0.192	0.190
β_textbook	0.240	0.393	0.573	0.629	0.581	0.642	0.623

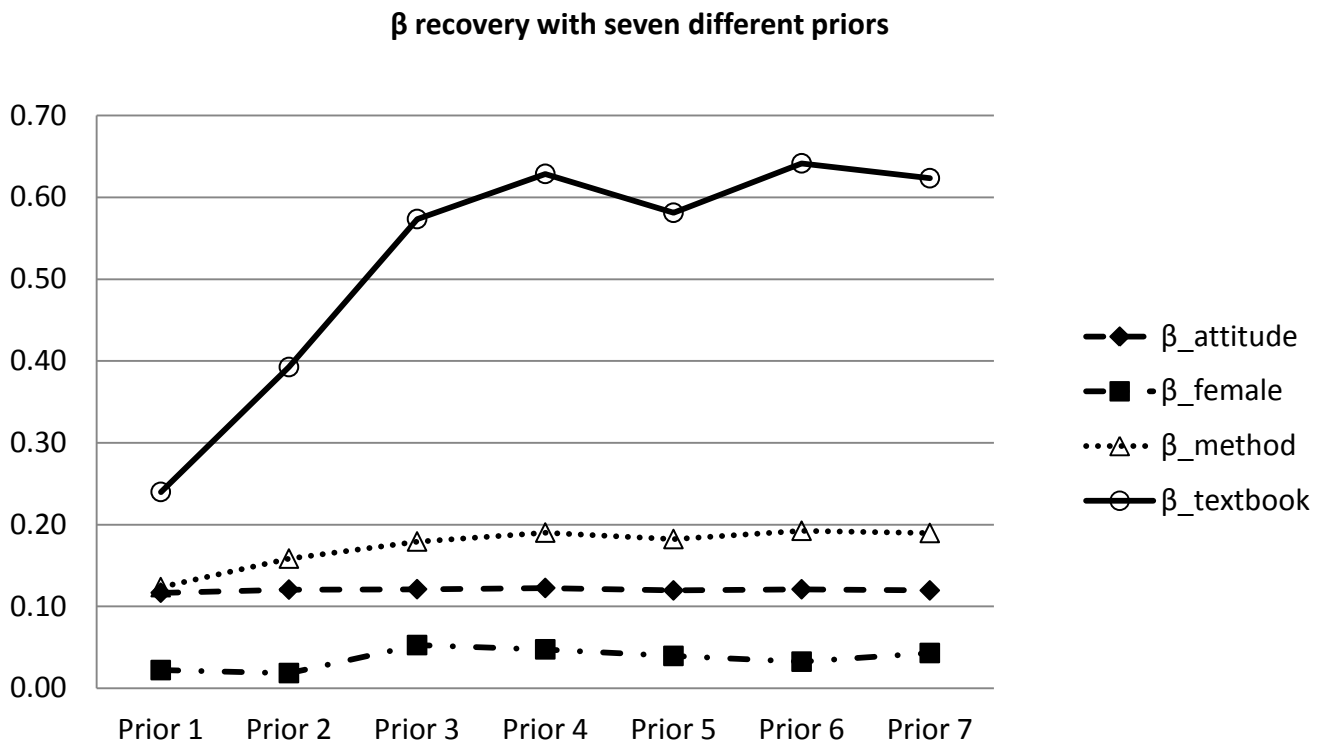

Figure 3.10 Estimated coefficients with seven different priors

More about Bayesian: convergence test results

While all the Markov chains passed convergence tests for simple regression in the previous chapter, convergence of Markov chains is a more serious concern when the model gets more complicated. It usually takes longer to reach convergence. Some chains at the fixed length (4,000) in multilevel models did not pass the convergence test. Running the chains longer might eventually solve the problem but it is beyond the time constraints of this project. On the other hand, the means of the Markov chains showed little difference (see Table 3.6 to Table 3.9) regardless of the convergence test results. So the conclusions based on the result should be maintained.

The results showed that the about 20% of the chains for the coefficient of gender did not pass the convergence test at the fixed length of 4,000 iteration. The possible reason is that the variable *gender* is extremely skewed. On average 90% of the students are female. In some of the generated data, the percentage of females could be much higher than 90%. If that is the case, the gender difference is hard to estimate and the lack of convergence reflects the situation.

Table 3.6 Convergence of Markov chains: coefficient of *attitude* (=0.14)

	β_attitude							
	Test failed		Not converged		Converged		Total	
Presence of Error	*N*	*Mean*	*N*	*Mean*	*N*	*Mean*	*N*	*Mean*
No Error	0	.	0	.	300	0.139	300	0.139
Error in Dependent Variables	0	.	3	0.125	297	0.140	300	0.139
Error in Independent Variables	0	.	3	0.159	297	0.138	300	0.138
Error in Both Variables	0	.	10	0.142	290	0.137	300	0.138
Total	0	.	16	0.142	1184	0.139	1200	0.139

Table 3.7 Convergence of Markov chains: coefficient of *female* (=0.00)

	β_female							
	Test failed		Not converged		Converged		Total	
Presence of Error	*N*	*Mean*	*N*	*Mean*	*N*	*Mean*	*N*	*Mean*
No Error	0	.	0	.	300	-0.002	300	-0.002
Error in Dependent Variables	7	-0.024	104	-0.002	189	-0.005	300	-0.004
Error in Independent Variables	0	.	18	0.020	282	-0.001	300	0.001
Error in Both Variables	7	0.021	119	-0.010	174	-0.020	300	-0.015
Total	14	-0.001	241	-0.004	945	-0.006	1200	-0.005

Table 3.8 Convergence of Markov chains: coefficient of *method* (= 0.30)

	β_method							
	Test failed		Not converged		Converged		Total	
Presence of Error	*N*	*Mean*	*N*	*Mean*	*N*	*Mean*	*N*	*Mean*
No Error	0	.	0	.	300	0.301	300	0.301
Error in Dependent Variables	5	0.264	193	0.290	102	0.296	300	0.291
Error in Independent Variables	0	.	6	0.261	294	0.242	300	0.243
Error in Both Variables	0	.	96	0.231	204	0.228	300	0.229
Total	5	0.264	295	0.270	900	0.265	1200	0.266

Table 3.9 Convergence of Markov chains: coefficient of *textbook* (= 0.48)

	β_textbook							
	Test failed		Not Converged		Converged		Total	
Presence of Error	*N*	*Mean*	*N*	*Mean*	*N*	*Mean*	*N*	*Mean*
No Error	0	.	0	.	300	0.479	300	0.479
Error in Dependent Variables	0	.	127	0.472	173	0.450	300	0.460
Error in Independent Variables	0	.	3	0.523	297	0.470	300	0.471
Error in Both Variables	0	.	64	0.460	236	0.431	300	0.437
Total	0	.	194	0.469	1006	0.460	1200	0.461

CHAPTER 4 ONE-STEP BAYESIAN APPLICATION ON ME.ET PROJECT DATA

The previous chapter evaluates the performance of the one-step Bayesian approach to the combination model. In many ways it is better than the two-step approach as the coefficient estimation is less biased and more efficient. So it is worth trying the one-step Bayesian approach on the ME.ET project data.

The Mathematical Education of Elementary Teachers (ME.ET) Project

The Mathematical Education of Elementary Teachers (ME.ET) project is a study on the preparation of prospective math teachers. It aims at depicting a picture of what is going on in math teacher training classes and identifying factors that can improve their effectiveness. It is a comprehensive project with many facets and only the information relevant to the statistical model and data will be briefly presented here. Totally there are 1706 students from 41 classes (schools) who took at least one of the pretest or posttest that measures math related achievement. The items of pretest and posttest are drawn from a pool of 51 items adopted from a previous study. The students who took the tests were also surveyed on their gender and attitude toward math. At the meantime 78 instructors, including the 41instructors whose students took part in the math pretest and posttest, were surveyed on the textbook used and their teaching method.

Model

Among many research questions that ME.ET project intended to investigate, there is one particularly suitable to be solved by the one-step Bayesian method. That is the relationship between student gain and other variables: student gender, student attitude toward math, textbook used and instructor teaching method. Obviously there are latent variables involved in both dependent and independent variables and they are ready to be described with IRT models. Gain

score is the learning outcome. The data structure includes students nested in classes. A combination model described in Chapter 3 would be appropriate to answer this research question.

Data

When the students took the pretest and posttest, they took one of the four test forms which consists of 20 to 26 items selected from a 51-item pool. In the Bayesian approach, every test form is treated as a test with full length of 51 and the student responses to those not presented items are listed as missing. This transformation has greatly eased the difficulty caused by the complexity of the student data structure.

Missing data is a common occurrence in practice and ME.ET project cannot be exempted. Not every student in the study took both the pretest and posttest. Instead, many of them took only one test for various reasons. The gain score, which is defined as the score difference between posttest and pretest, cannot be estimated if only one test is taken by a student. Meanwhile, not all of the students and teachers provided complete answer to the survey questions. The non-responses are represented as missing values in the dataset.

Naturally, a Bayesian approach can accommodate missing values conveniently by treating them as unknown parameters. This feature is usually regarded as an advantage of the Bayesian approach. But in this particular model, the drawback of using the full dataset with imputed values is that it loses the power to detect the effect of school level covariates, which is the primary concern here. In order to better understand the relationship between gain score and school level covariates, the data loaded into the Bayesian model is not the full dataset, but the one that has been listwise deleted if a missing value is found in any of the variables: any student who has omitted a pretest or posttest, or skipped a response to the gender or attitude survey

question is excluded from the analysis. Any instructor who has a no response to textbook used or teaching methods question is also removed from the dataset as well. This gives a total of 928 students from 33 schools for analysis.

For comparison purpose, results from the full dataset with imputations will be briefly presented afterward the full analysis of cases with complete data, and the pros and cons of dataset choice will be discussed in next chapter.

The data were cleaned and prepared in SAS and exported to text files. After that software R read the data from text files and transformed it into the format that can be loaded into WinBUGS program.

Estimation

In the simulation in the previous chapter the data is complete and balanced, that is to say each class has the same number of students. It is seldom the case in practice so the WinBUGS code had to be modified to accommodate this situation. An index variable for each student to identify his or her instructor was created. The details of the WinBUGS code for the combination model can be found in Appendix D.

The priors of the four coefficients are all set to a noninformative normal distribution $N(0, 1)$, which means the possible values are centered on 0 with a deviation of 1. That pretty much reflects out prior knowledge of these coefficients: they could be positive or negative and they are not far away from 0 considering the scale of IRT scores for dependent variables.

The priors for the residual variance at both student level and school level adopt Wishart distributions with 2 degree of freedom:

$$cov\begin{pmatrix} r_{0ij} \\ r_{1ij} \end{pmatrix} \sim Wishart\left(\begin{bmatrix} 0.23 & -0.08 \\ -0.08 & 0.03 \end{bmatrix}, 2\right)$$

$$cov\begin{pmatrix} \mu_{00} \\ \mu_{10} \end{pmatrix} \sim Wishart\left(\begin{bmatrix} 0.08 & -0.03 \\ -0.03 & 0.08 \end{bmatrix}, 2\right)$$

Those parameters in Wishart distribution are obtained from the traditional two-step method. It reflects the best guess we have on the distribution and turned out to be uninformative. The priors for these parameters are less important since neither the one-step Bayesian nor the two-step approach can recover these numbers well, as shown in the simulation results in the previous chapter.

Results

After the first 2000 iterations were discarded as burn-in, the rest of the 8000 iterations were used to calculate the posterior distribution of the parameters. The trace plot and the density plot are shown in Figure 4.1.

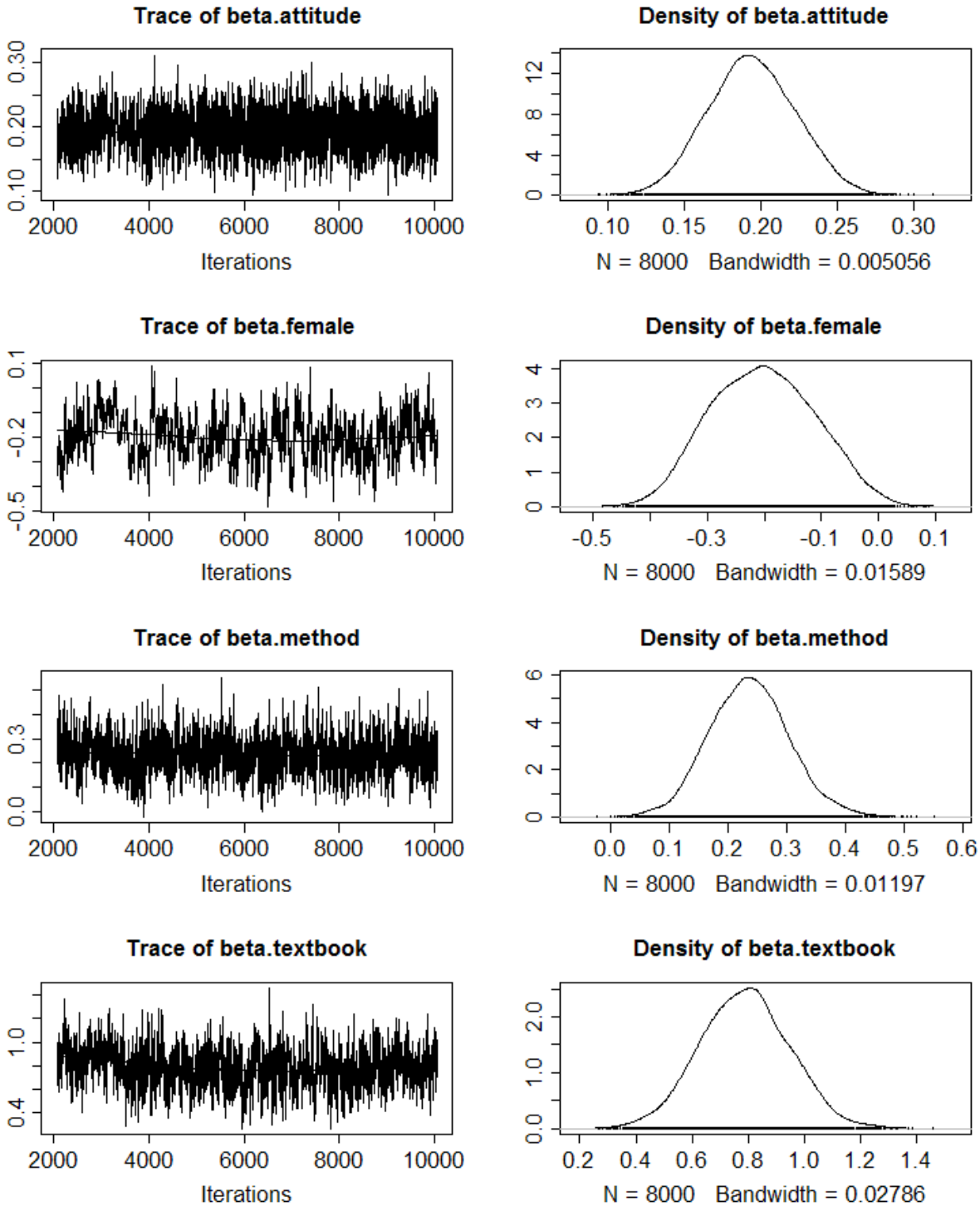

Figure 4.1 Trace plot and density plot of ME.ET parameters. Missing data were deleted listwise.

The chains passed the Heidelberger and Welch convergence diagnostic (see Table 4.1) and the results are listed in Table 4.2.

Table 4.1 Heidelberger and Welch Diagnosis of ME.ET project model

	Heidelberger and Welch Diagnosis					
	Stationary test	Start iteration	*p-value*	Halfwidth test	*Mean*	Halfwidth
β_attitude	passed	1	0.326	passed	0.194	0.002
β_female	passed	1	0.169	passed	-0.202	0.017
β_method	passed	1	0.360	passed	0.237	0.007
β_textbook	passed	801	0.136	passed	0.779	0.018

Table 4.2 Covariates Effects estimated from Bayesian approach.

	MCMC posterior			
	Mean	*SE*	2.5% CI	97.5% CI
β_attitude	0.194	0.029	0.139	0.250
β_female	-0.202	0.090	-0.368	-0.027
β_method	0.237	0.069	0.106	0.381
β_textbook	0.788	0.159	0.479	1.098

Note: 928 students from 33 schools in the dataset.

From the results we can see that *teaching method* and *textbook* used have a positive correlation with the class gain. On average, a class using a primary textbook has a gain that is 0.79 point more than a class not doing so. The 95% credible interval (or Bayesian confidence interval) of *textbook* effect ranges from 0.48 to 1.10. An instructor who has one unit more on the *teaching method* scale is corresponding to 0.24 point more gain in his or her class. The 95% credible interval of teaching method effect is from 0.11 to 0.38.

Also the results reveal the gender difference in the student scores. On average, a female student scores 0.20 lower than a male student according to the data. That difference is bound line significant because the 95% credible interval (-0.37, -0.03) almost covers the 0. As expected, a positive attitude towards math is correlated with a higher student score with a correlation point estimate of 0.2 and a 95% credible interval of (0.14, 0.25).

Results if using full data with imputation

As mentioned earlier, the Bayesian approach can accommodate missing values in the outcome variables naturally. Missing values in pretest, posttest, attitude survey or teaching method survey can be loaded into Bayesian model directly. So no special imputation was needed beforehand except for the independent variable female. The missing values are imputed by a randomly generated number from Bernoulli distribution $B(0.9)$, where 0.9 is the proportion of females in the observed data.

The same Bayesian model and MCMC computation applying on this full dataset generated somewhat different results (see Figure 4.2, Table 4.3 and Table 4.4). The major difference is that those textbook effect and teaching method effect are close to 0 and not significant any more. While the power of detecting school level effects dwindled, the accuracy of

student level effects had improved since there are more pretest data available for their estimation. From the numbers we can say that the full dataset reveals a bigger gender difference than the one with missing values list-wise deleted.

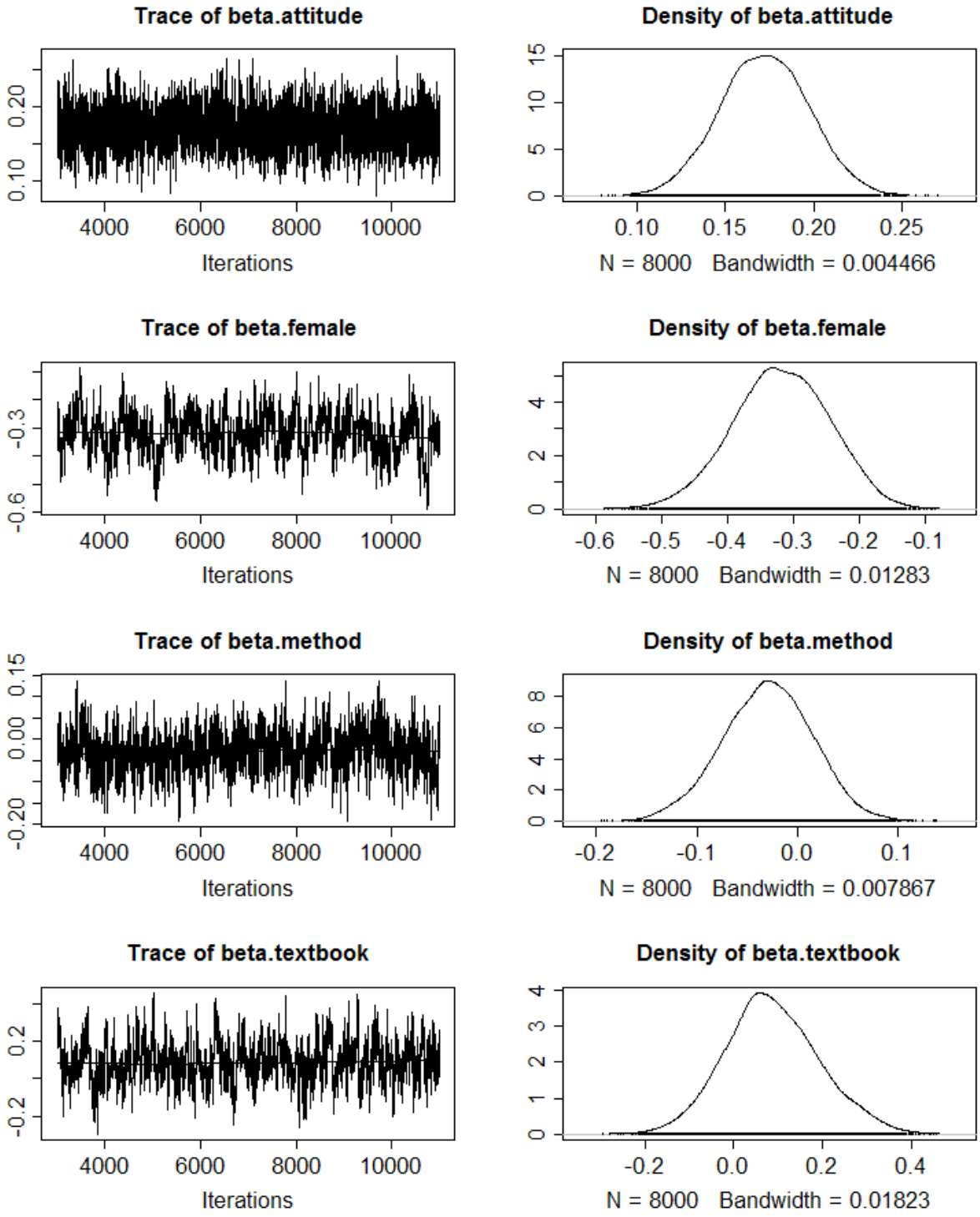

Figure 4.2 Trace plot and density plot of ME.ET project model parameters. Full dataset with imputation was used.

Table 4.3 Heidelberger and Welch diagnosis of ME.ET project model, full dataset used

	Heidelberger and Welch diagnosis					
	Stationary test	Start iteration	*p-value*	Halfwidth test	*Mean*	Halfwidth
β_attitude	passed	1	0.158	passed	0.172	0.001
β_female	passed	1	0.666	passed	-0.320	0.013
β_method	passed	1	0.140	failed	-0.030	0.005
β_textbook	passed	1	0.761	failed	0.088	0.016

Table 4.4 Covariates Effects estimated from Bayesian approach, full dataset used

	MCMC posterior			
	Mean	*SE*	2.5% CI	97.5% CI
β_attitude	0.172	0.025	0.123	0.222
β_female	-0.320	0.073	-0.467	-0.182
β_method	-0.030	0.045	-0.122	0.056
β_textbook	0.088	0.106	-0.111	0.308

Note: 1706 students and 78 schools in the dataset.

Comparison of the listwise deleted sample and full sample

There is not a huge difference in the statistics of the variables between the listwise deleted sample and full sample (see Table 5.5). For the two school level variables, *method* and *textbook*, of which the coefficient estimates are quite different in the two samples, their statistics are quite similar. So the sample bias at school level is not a major contributor to the different in the estimation of *method* effect and *textbook* effect. Imputation of the missing gain score seems to be the reason that reduces the power of detecting those two effects.

But there is a noticeable difference in the pretest scores. The pretest score is lower in the full sample. The explanation is that the students are more likely to drop out if they did not do well in pretest. There is a potential selection bias in the listwise deleted sample. The correlation between pretest score and gain score is negative (-0.56), so the overall added value of the training programs could be underestimated because the participants who only have lower pretest scores are expected to have higher gain scores. However, it is not clear yet if the selection bias affects the estimation of school level effects such as method and textbook. Future investigations should pay attention to this issue.

Table 4.5 Statistics of the listwise deleted dataset and full dataset

	Listwise deleted dataset			Full dataset		
	N	*Mean*	*SD*	*N*	*Mean*	*SD*
Method	33	2.80	0.58	78	2.69	0.64
Textbook	33	0.45	0.51	51	0.46	0.50
Female	928	0.91	0.28	1560	0.90	2.28
Attitude	928	3.70	0.54	1486	3.63	0.61
Pretest	928	-0.58	1.00	1476	-0.65	1.01
Posttest	928	0.34	1.16	1293	0.35	1.18

CHAPTER 5 CONCLUSIONS AND DISCUSSION

Conclusions

By investigating IRT score estimation error impact on multilevel value-added models, this study highlights one important but often overlooked issue. This study explored the IRT score error impact on regression models and evaluated how well the one-step Bayesian approach can solve this problem. The major findings indicate that IRT error impact is not ignorable and the one-step Bayesian approach is recommended to get a better estimation of parameters.

It can be concluded from the simulation result that IRT score estimation error can cause biased coefficient estimates with the two-step approach. Not only does it attenuate the coefficient estimation when error is present in independent variables, but it also distorts the estimation when error is in dependent variables. Different IRT score estimates, i.e. Maximum Likelihood estimate vs. Bayesian Model estimate, leads to bias in different directions. It can also be concluded that the larger the IRT score error, the more prominent the problem is. So when the test is less accurate, or there are only a small number of students to estimate the item parameters if they are unknown, the researchers should be cautious about the interpretation of relationships involving latent variables. On the other hand, the one-step Bayesian approach provides less biased and more efficient estimates of the regression coefficients. Overall one-step Bayesian has a better chance to reveal the true relationship between dependent and independent variables.

The outcomes also confirmed that the correlation of student pretest and gain is hard to estimate for both the two-step and one-step approaches. Interpretations based on the estimations should not be taken seriously because they can easily be wrong. At the school level, the correlation between school mean pretest and gain was well recovered by the one-step Bayesian

approach. Also the recovery of school gain variance is satisfactory under the one-step Bayesian, while it is not as acceptable with the two-step approach. The findings here suggest the two-step approach cannot do the job of estimating school added value well. It is possible that the one-step Bayesian approach, after further study and refinement, has potential for the task.

Discussion

The recovery of latent variable correlation coefficient

In the simple regression simulation in chapter 2, IRT score from ML estimate and BM estimate have different influences on the regression coefficient recovery. Although the assumption is that the latent variable has a standard normal distribution, the obtained IRT scores might not follow it. The standard deviation of the ML estimate is larger than the standard deviation of the true θ, while that of the BM estimate is smaller (see Table 5.1).

In chapter 2, the IRT scores are entered directly into regression model to estimate the regression coefficient. As shown in Table 5.1, the standard deviation of the obtained IRT scores is not 1, which means the regression coefficient is not a good estimator of the correlation coefficient. If it is to recover of correlation coefficient, the variables should be standardized. It turned out that the estimates of correlation coefficient are all negatively biased (see Table 5.2).

The correlation of the IRT score with true θ is an indicator of the quality of θ recovery. BM estimate has a higher correlation with true θ than ML estimate does. The IRT score obtained from one-step Bayesian approach is the most accurate among the three estimates (see Table 5.1). The difference between θ variance and $\hat{\theta}$ variance could be one of reasons that cause the biased β estimation. Future studies on correction for IRT score error impact with two-step approach should take this into consideration.

Table 5.1 The standard deviation of IRT score estimates and their correlation with true θ

Generating value of parameter ρ	Number of items	Number of students	True θ	Maximum Likelihood (ML) estimate		Bayesian Model (BM) estimate		IRT score in one-step Bayesian approach	
			SD	*SD*	*Correlation with* θ	*SD*	*Correlation with* θ	*SD*	*Correlation with* θ
0.2	15	500	0.993	1.304	0.853	0.826	0.866	0.862	0.867
		1000	0.972	1.263	0.863	0.848	0.873	0.879	0.874
	25	500	1.006	1.198	0.915	0.875	0.918	0.905	0.919
		1000	0.940	1.169	0.910	0.894	0.915	0.926	0.915
0.5	15	500	1.048	1.231	0.878	0.866	0.889	0.914	0.896
		1000	1.024	1.309	0.879	0.856	0.903	0.917	0.905
	25	500	0.989	1.150	0.924	0.893	0.929	0.935	0.931
		1000	1.025	1.184	0.909	0.891	0.925	0.948	0.928
0.8	15	500	0.986	1.313	0.854	0.834	0.867	0.911	0.888
		1000	1.004	1.298	0.867	0.831	0.884	0.912	0.909
	25	500	1.000	1.198	0.905	0.880	0.916	0.944	0.927
		1000	0.995	1.185	0.925	0.891	0.934	0.946	0.941

Table 5.2 The bias of correlation coefficient recovery (*N* = 100)

Generating value of parameter ρ	Number of items	Number of students	Error in one variable (dependent variable or independent variable)				Error in both variables			
			ML estimate		BM estimate		ML estimate		BM estimate	
			Bias	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*
0.2	15	500	-0.030	0.053	-0.027	0.052	-0.052	0.068	-0.047	0.063
		1000	-0.026	0.039	-0.023	0.036	-0.051	0.059	-0.046	0.055
	25	500	-0.021	0.061	-0.019	0.061	-0.038	0.067	-0.035	0.067
		1000	-0.011	0.034	-0.010	0.033	-0.029	0.044	-0.026	0.041
0.5	15	500	-0.082	0.089	-0.073	0.081	-0.142	0.147	-0.125	0.132
		1000	-0.071	0.076	-0.063	0.068	-0.133	0.136	-0.117	0.121
	25	500	-0.049	0.059	-0.043	0.054	-0.090	0.097	-0.081	0.088
		1000	-0.047	0.054	-0.042	0.050	-0.089	0.093	-0.079	0.084
0.8	15	500	-0.112	0.115	-0.098	0.101	-0.213	0.216	-0.184	0.187
		1000	-0.115	0.117	-0.101	0.104	-0.210	0.212	-0.182	0.184
	25	500	-0.076	0.079	-0.068	0.072	-0.143	0.146	-0.125	0.128
		1000	-0.072	0.073	-0.064	0.066	-0.136	0.138	-0.120	0.121

Sources of IRT score error

Latent variables such as student achievement are subject to error. Several sources of measurement error can be identified from the simulation study. The first one is the IRT score estimation procedure. Different procedures, such as Maximum Likelihood (ML) and Bayesian Model (BM) produce different IRT scores.

The second source of error is the sampling error with regard to subjects. When item parameters are unknown, the responses of the subjects, i.e. students etc, are used to estimate the item parameters. Whether these item parameters can be accurately estimated depends on whether the subjects sample is large enough. Generally speaking, larger samples produce better estimation of the item parameters which leads to more accurate latent trait estimation.

The third source of error is the instrument error. Assuming that the items are valid in measuring the latent traits they are intended to measure, items still differ in their capacity to reveal the individual student difference. Students with different levels of latent trait have the chance to generate identical responses to a test. It is more likely to happen when the test is shorter and the items have less discriminating values. A longer test consisting of items with larger discriminating values is more likely to tell the students apart. This characteristic of a test can be described with the test information function. In general, a longer test has more information and less error in its estimated IRT score than a shorter one.

Possibly there is a fourth source of error emerging from the latent trait instability. One subject answering one item correctly this time might give a wrong answer next time when the item discrimination parameters is not perfect. The test-retest inconsistency is the result of trait instability as well as item inaccuracy. It is necessary to have test-retest data to estimate the

stability of the latent trait. The data structure of this study does not support the investigation of this source of error and thus it was not included in the investigation. The error associated with the trait instability should not be confused with the measurement error which is thought to be associated with items.

Bayesian approach

As demonstrated in this study, the Bayesian approach is truly remarkable in its flexibility to solve complex models and its performance on the recovery of regression coefficients. The Bayesian approach is recommended for empirical study, but a few considerations should be mentioned in its application.

The first concern is the recovery of the correlation between student pretest and gain. The current Bayesian MCMC model code in this study still cannot recovery this parameter well. Further improvement should be made to better recover this parameter if it is important to answer a research question.

The second concern is the treatment of missing values. Bayesian models can handle the missing values by default, but it is not always in the best interest of the research question. As shown in chapter 5, different strategies of treating missing values generate different conclusions. One should be careful about this when working with missing data.

The third concern is model evaluation. In this study, it is assumed that the data fit the theory well. The conclusion from the study is valid on condition that the model describes the data perfectly. But in practice, whether the model fits the data is an open question. One should consider evaluating the model fit during the model building process. Model evaluation and

comparison should be carefully studied before any practical conclusion is drawn from the research. Further discussions on model evaluation can be found in Gill (2012).

APPENDICES

Appendix A

Item parameters in multilevel simulation

Table 6.1 Teaching method questionnaire item parameters

Question	b_2	b_3	b_4	a
1	-2.378	-1.113	0.036	2.227
2	-1.818	0.006	1.475	1.928
3	-1.165	-0.081	1.936	1.473
4	-2.423	-0.855	1.703	1.079
5	-1.951	-0.665	1.163	2.742
6	-2.567	-0.728	0.515	2.546
7	-1.432	-0.028	0.955	4.251
8	-1.930	-0.658	0.758	4.348
9	-1.750	-0.246	0.661	2.123
10	-1.483	0.226	1.843	1.410
11	-1.697	0.500	2.638	1.584

Note: The parameters are estimated from ME.ET project data.

Table 6.2 Student attitude questionnaire item parameters

Question	b_2	b_3	b_4	b_5	a
A1	-3.463	-1.109	-0.110	3.033	0.905
A3	-3.310	-1.242	-0.248	2.709	1.089
A9	-3.121	-1.799	-1.313	0.803	2.386
A13	-3.036	-1.760	-1.245	1.012	2.151
A14	-4.391	-2.382	-1.000	2.027	1.452
A17	-4.489	-2.569	-1.140	1.636	1.728
A18	-4.014	-2.722	-1.078	1.348	1.761
A20	-3.136	-1.670	-1.096	1.124	2.073
A27	-1.824	-0.571	0.182	2.338	1.140

Note: The item parameters are estimated based on ME.ET project data and used for the generation of simulation data.

Table 6.3 Student pretest and posttest item parameters

Item ID	*a*	*b*	Item ID	*a*	*b*
1	1.05	-1.82	27	0.91	-1.05
2	0.96	-1.70	28	0.65	0.23
3	1.00	-1.65	29	0.70	0.34
4	1.03	-2.19	30	1.12	-1.13
5	0.72	-0.31	31	0.55	0.54
6	0.98	-0.31	32	0.89	-2.54
7	1.54	-1.17	33	0.68	1.73
8	1.09	-1.15	34	0.92	-0.58
9	0.63	-1.03	35	0.88	-0.86
10	0.35	-1.03	36	0.87	-0.34
11	0.56	0.33	37	0.60	0.92
12	0.99	-1.09	38	1.50	0.28
13	1.05	0.03	39	0.84	-0.30
14	1.47	-0.63	40	0.77	-1.01
15	0.70	-0.41	41	0.25	0.23
16	0.92	-1.50	42	0.49	2.67
17	0.75	-3.08	43	0.62	-0.59
18	0.36	0.59	44	0.69	0.97
19	0.44	-0.89	45	0.34	-1.54
20	0.95	-0.16	46	0.68	-1.34
21	0.72	-0.20	47	0.62	-0.92
22	0.65	1.35	48	0.64	-0.93
23	0.60	-0.63	49	0.67	0.61
24	0.50	-1.63	50	0.60	1.80
25	0.60	-0.77	51	0.53	0.12
26	0.63	-1.16			

Note: Parameters are not estimated from ME.ET project data but from previous studies when the instrument was developed.

Appendix B

BUGS code for the simple linear model in chapter 2

1. When error present in dependent variables (dep.odc)

```
Model
{
        for  (j in 1 : n_stu) {
                for (k in 1 : n_item) {
                        rs[j,k]~dbern(prob[j,k])
                }
        }
        for (n in 1:n_stu){
                for (t in 1 : n_item) {
                        logit(prob[n,t])<-ip[t,1]*(y[n]-ip[t,2])
                }
        }
        for (n in 1:n_stu){
                y[n]~dnorm(y.hat[n],r.tau)I(-5,5)
                 y.hat[n]<-x[n]*b
        }
        r.tau<-1/(1-b*b)
        b~dunif(-1,1)
        for (t in 1 : n_item) {
                ip[t,1]~dlnorm(0,0.5)I(0,3)
                ip[t,2]~dnorm(0,1)I(-3,3)
        }
} # End of model
```

2. When error present in independent variables (pre.odc)

```
Model
{
        for  (j in 1 : n_stu) {
                for (k in 1 : n_item) {
                        rs2[j,k]~dbern(prob[j,k])
                }
        }
        for (n in 1:n_stu){
                for (t in 1 : n_item) {
                        logit(prob[n,t])<-ip[t,1]*(x[n]-ip[t,2])
                }
        }
        for (nn in 1:n_stu){
                y[nn]~dnorm(y.hat[nn],r.tau)I(-5,5)
                y.hat[nn]<-x[nn]*b
                x[nn]~dnorm(0,1)I(-5,5)
        }
         r.tau<-1/(1-b*b)
        b~dunif(-1,1)
        for (tt in 1 : n_item) {
                ip[tt,1]~dlnorm(0,0.5)I(0,3)
                ip[tt,2]~dnorm(0,1)I(-3,3)
        }
}# End of model
```

3. When error present in both variables (both.odc)

```
Model
{
    for  (j in 1 : n_stu) {
        for (k in 1 : n_item) {
            rs[j,k]~dbern(prob[j,k])
        }
    }
    for  (j in 1 : n_stu2) {
        for (k in 1 : n_item2) {
            rs2[j,k]~dbern(prob2[j,k])
        }
    }
    for (n in 1:n_stu){
        for (t in 1 : n_item) {
            logit(prob[n,t])<-ip[t,1]*(y[n]-ip[t,2])
        }
    }
    for (n in 1:n_stu2){
        for (t in 1 : n_item2) {
            logit(prob2[n,t])<-ip2[t,1]*(x[n]-ip2[t,2])
        }
    }
    for (nn in 1:n_stu){
        y[nn]~dnorm(y.hat[nn],r.tau)I(-5,5)
        y.hat[nn]<-x[nn]*b
        x[nn]~dnorm(0,1)I(-5,5)
```

```
        }
        r.tau<-1/(1-b*b)
        b~dunif(-1,1)
        for (tt in 1 : n_item) {
                ip[tt,1]~dlnorm(0,0.5)I(0,3)
                ip[tt,2]~dnorm(0,1)I(-3,3)
        }
        for (tt in 1 : n_item2) {
                ip2[tt,1]~dlnorm(0,0.5)I(0,3)
                ip2[tt,2]~dnorm(0,1)I(-3,3)
        }
}# End of model
```

Appendix C

BUGS code for the multilevel value-added model in chapter 3

1. When error not present (hlmnome.odc)

```
Model
{
        for (m in 1:n.tch) {
                for (n in 1:n.stu){
                        # student level variance
                        theta[m,n,1:2]~dmnorm(mu.x[m,n,1:2],Omega.x[,])
                        mu.x[m,n,2]<-tch[m,2]
                        mu.x[m,n,1]<-tch[m,1]+female2[m,n]*b.f+att2[m,n]*b.a
                }
 # bivariate Normal of teacher ini and gain
                tch[m,1:2]~dmnorm(mu.tch.m[m,],Omega.tch[,])
                mu.tch.m[m,1]<-mu.tch[1]
                mu.tch.m[m,2]<-mu.tch[2]+method[m]*b.m+text[m]*b.t
        }
 # class level variance
        r.tch <- Sigma2.tch[1,2] / (sqrt(Sigma2.tch[1,1])
                 *sqrt(Sigma2.tch[2,2]))
        sigma.tch[1]<-sqrt(Sigma2.tch[1,1])
        sigma.tch[2]<-sqrt(Sigma2.tch[2,2])
        Sigma2.tch[1:2,1:2] <-inverse(Omega.tch[,])
        Omega.tch[1:2,1:2]~ dwish(R[,],2)
# student level variance
        r.x <- Sigma2.x[1,2] / (sqrt(Sigma2.x[1,1])
                *sqrt(Sigma2.x[2,2]))
        sigma.x[1]<-sqrt(Sigma2.x[1,1])
        sigma.x[2]<-sqrt(Sigma2.x[2,2])
```

```
        Sigma2.x[1:2,1:2] <-inverse(Omega.x[,])
        Omega.x[1:2,1:2]~ dwish(Rx[,],2)
        b.a~dnorm(0,1)              # coefficient of attitude prior
        b.f~dnorm(0,1)              # coefficient of female  prior
        b.t~dnorm(0,1)              # coefficient of textbook prior
        b.m~dnorm(0,1)              # coefficient of method  prior
        mu.tch[1]~dnorm(0,1)        # class mean initial prior
        mu.tch[2]~dnorm(0,1)         # class mean gain prior
}# End of model
```

2. When error present in dependent variables (hlmdep.odc)

```
Model
{
        for (i in 1 : n.tch) {
                for  (j in 1 : n.stu) {
                        for (kpre in 1 : n.item.pre) {
                                rs.pre2[i,j,kpre]~dbern(prob.pre[i,j,kpre])
                        }
                        for (kpost in 1 : n.item.post) {
                                rs.post2[i,j,kpost]~dbern(prob.post[i,j,kpost])
                        }
                }
        }
        for (m in 1:n.tch) {
                for (n in 1:n.stu){
                        for (tpre in 1 : n.item.pre) {
                                logit(prob.pre[m,n,tpre])<-ip[tpre,1]*(theta[m,n,1]-ip[tpre,2])
                        }
                        for (tpost in 1 : n.item.post) {
                                logit(prob.post[m,n,tpost])<-
                                        ip2[tpost,1]*(theta[m,n,1]+theta[m,n,2]-ip2[tpost,2])
                        }
# student level variance
                        theta[m,n,1:2]~dmnorm(mu.x[m,n,1:2],Omega.x[,])
                        mu.x[m,n,2]<-tch[m,2]
                        mu.x[m,n,1]<-tch[m,1]+female2[m,n]*b.f+att2[m,n]*b.a
                }
```

```
# bivariate Normal of teacher ini and gain
                tch[m,1:2]~dmnorm(mu.tch.m[m,],Omega.tch[,])
                mu.tch.m[m,1]<-mu.tch[1]
                mu.tch.m[m,2]<-mu.tch[2]+method[m]*b.m+text[m]*b.t
        }
# class level variance
                r.tch <- Sigma2.tch[1,2] / (sqrt(Sigma2.tch[1,1])
                        *sqrt(Sigma2.tch[2,2]))
                sigma.tch[1]<-sqrt(Sigma2.tch[1,1])
                sigma.tch[2]<-sqrt(Sigma2.tch[2,2])
                Sigma2.tch[1:2,1:2] <-inverse(Omega.tch[,])
                Omega.tch[1:2,1:2]~ dwish(R[,],2)
# student level variance
                r.x <- Sigma2.x[1,2] / (sqrt(Sigma2.x[1,1])
                        *sqrt(Sigma2.x[2,2]))
                sigma.x[1]<-sqrt(Sigma2.x[1,1])
                sigma.x[2]<-sqrt(Sigma2.x[2,2])
                Sigma2.x[1:2,1:2] <-inverse(Omega.x[,])
                Omega.x[1:2,1:2]~ dwish(Rx[,],2)
                b.a~dnorm(0,1)
                b.t~dnorm(0,1)
                b.f~dnorm(0,1)
                b.m~dnorm(0,1)
                mu.tch[1]~dnorm(0,1)                # class mean initial  prior
                mu.tch[2]~dnorm(0,1)                # class mean gain   prior

}# End of Model
```

3. When errors present in independent variables (hlmpre.odc)

```
Model
{
        for (m in 1:n.tch) {
                for (n in 1:n.stu){
# student level variance
                        theta[m,n,1:2]~dmnorm(mu.x[m,n,1:2],Omega.x[,])
                        mu.x[m,n,2]<-tch[m,2]
                         mu.x[m,n,1]<-tch[m,1]+female2[m,n]*b.f+att[m,n]*b.a
                }
# bivariate Normal of teacher ini and gain
                tch[m,1:2]~dmnorm(mu.tch.m[m,],Omega.tch[,])
                mu.tch.m[m,1]<-mu.tch[1]
                mu.tch.m[m,2]<-mu.tch[2]+method[m]*b.m+text[m]*b.t
        }
# class level variance
        r.tch <- Sigma2.tch[1,2] / (sqrt(Sigma2.tch[1,1])
                 *sqrt(Sigma2.tch[2,2]))
        sigma.tch[1]<-sqrt(Sigma2.tch[1,1])
        sigma.tch[2]<-sqrt(Sigma2.tch[2,2])
        Sigma2.tch[1:2,1:2] <-inverse(Omega.tch[,])
        Omega.tch[1:2,1:2]~ dwish(R[,],2)
# student level variance
        r.x <- Sigma2.x[1,2] / (sqrt(Sigma2.x[1,1])
                *sqrt(Sigma2.x[2,2]))
        sigma.x[1]<-sqrt(Sigma2.x[1,1])
        sigma.x[2]<-sqrt(Sigma2.x[2,2])
```

```
        Sigma2.x[1:2,1:2] <-inverse(Omega.x[,])
        Omega.x[1:2,1:2]~ dwish(Rx[,],2)
        b.a~dnorm(0,1)              # coefficient of attitude prior
        b.f~dnorm(0,1)              # coefficient of female  prior
        b.t~dnorm(0,1)              # coefficient of textbook prior
        b.m~dnorm(0,1)              # coefficient of method  prior
        mu.tch[1]~dnorm(0,1)        # class mean initial prior
        mu.tch[2]~dnorm(0,1)        # class mean gain prior
# GRM attitude
        for (i in 1:n.tch) {
                for (i2 in 1:n.stu) {
                        for (j in 1:n.item.att) {
                                for (k in 1: (n.cat.att[j]-1)) {
                                        p.att[i,i2,j,k] <- 1 / (1+exp(-ip.att[j,5]*(att[i,i2]-ip.att[j,k])))
                                }
                        }
                }
        }
        for (i in 1:n.tch) {
                for (i2 in 1:n.stu) {
                        for (j in 1:n.item.att) {
                                pcat.att[i,i2,j,1] <- 1-p.att[i,i2,j,1]
                                for (k in 2: (n.cat.att[j]-1)){
                                        pcat.att[i,i2,j,k] <- p.att[i,i2,j,k-1]-p.att[i,i2,j,k]
                                }
                        pcat.att[i,i2,j,n.cat.att[j]] <- p.att[i,i2,j,(n.cat.att[j]-1)]
                        }
```

```
            }
        }
        for (i in 1:n.tch) {
            for (i2 in 1:n.stu) {
                for (j in 1:n.item.att) {
                    for (k in 1:n.cat.att[j]) {
                        pc.att[i,i2,j,k] <- pcat.att[i,i2,j,k] / sum( pcat.att[i,i2,j,
1:n.cat.att[j]] )
                    }
                    rs.att2[i,i2,j]~dcat(pc.att[i,i2,j,1:n.cat.att[j]])
                }
            }
        }
# item attitude prior
        for (j in 1:n.item.att) {
            ip.att[j,5]~dlnorm(0,0.5)I(0,5)
            ip.att[j,1]~dunif(-5,5)
            for (k in 2: (n.cat.att[j]-1)) {
                ip.att[j,k]~dunif(ip.att[j,k-1],5)
            }
        }
# student attitude prior
        for (i in 1:n.tch) {
            for (i2 in 1:n.stu) {
                att[i,i2]~dnorm(0,1)I(-4,4)
            }
        }
# GRM method
```

```
        for (i in 1:n.tch) {
                for (j in 1:n.item.method) {
                        for (k in 1: (n.cat[j]-1)) {
                                p[i,j,k] <- 1 / (1+exp(-ip.tch[j,4]*(method[i]-ip.tch[j,k])))
                        }
                }
        }
        for (i in 1:n.tch) {
                for (j in 1:n.item.method) {
                        pcat[i,j,1] <- 1-p[i,j,1]
                        for (k in 2: (n.cat[j]-1)){
                                pcat[i,j,k] <- p[i,j,k-1]-p[i,j,k]
                        }
                        pcat[i,j,n.cat[j]] <- p[i,j,(n.cat[j]-1)]
                }
        }
        for (i in 1:n.tch) {
                for (j in 1:n.item.method) {
                        for (k in 1:n.cat[j]) {
                                pc[i,j,k] <- pcat[i,j,k] / sum( pcat[i,j, 1:n.cat[j]] )
                        }
                        rs.tch[i,j]~dcat(pc[i,j,1:n.cat[j]])
                }
        }
# method items prior
        for (j in 1:n.item.method) {
                ip.tch[j,4]~dlnorm(0,0.5)I(0,5)
```

```
                ip.tch[j,1]~dunif(-5,5)
                for (k in 2: (n.cat[j]-1)) {
                        ip.tch[j,k]~dunif(ip.tch[j,k-1],5)
                }
        }
# teacher method prior
        for (i in 1:n.tch) {
                method[i]~dnorm(0,1)I(-4,4)
        }
}# End of Model
```

4. When errors present in both dependent and independent variables (hlmboth.odc)

```
Model
{
# Student Responses
        for (i in 1 : n.tch) {
                for  (j in 1 : n.stu) {
                # Pretest
                        for (kpre in 1 : n.item.pre) {
                                rs.pre2[i,j,kpre]~dbern(prob.pre[i,j,kpre])
                        }
                #Posttest
                        for (kpost in 1 : n.item.post) {
                                rs.post2[i,j,kpost]~dbern(prob.post[i,j,kpost])
                        }
                }
        }

# GRM student attitude
        for (i in 1:n.tch) {
                for (i2 in 1:n.stu) {
                        for (j in 1:n.item.att) {
                                for (k in 1: (n.cat.att[j]-1)) {
                                        p.att[i,i2,j,k] <- 1 / (1+exp(-ip.att[j,5]*(att[i,i2]-ip.att[j,k])))
                                }
                        }
                }
        }
```

```
        for (i in 1:n.tch) {
                for (i2 in 1:n.stu) {
                        for (j in 1:n.item.att) {
                                pcat.att[i,i2,j,1] <- 1-p.att[i,i2,j,1]
                                for (k in 2: (n.cat.att[j]-1)){
                                        pcat.att[i,i2,j,k] <- p.att[i,i2,j,k-1]-p.att[i,i2,j,k]
                                }
                                pcat.att[i,i2,j,n.cat.att[j]] <- p.att[i,i2,j,(n.cat.att[j]-1)]
                        }
                }
        }
        for (i in 1:n.tch) {
                for (i2 in 1:n.stu) {
                        for (j in 1:n.item.att) {
                                for (k in 1:n.cat.att[j]) {
                                        pc.att[i,i2,j,k] <- pcat.att[i,i2,j,k] / sum( pcat.att[i,i2,j,
                                                                1:n.cat.att[j]] )
                                }
                                rs.att2[i,i2,j]~dcat(pc.att[i,i2,j,1:n.cat.att[j]])
                        }
                }
        }
# item attitude prior
        for (j in 1:n.item.att) {
                ip.att[j,5]~dlnorm(0,0.5)I(0,5)
                ip.att[j,1]~dunif(-5,5)
                for (k in 2: (n.cat.att[j]-1)) {
```

```
                ip.att[j,k]~dunif(ip.att[j,k-1],5)
            }
        }
# student attitude prior
        for (i in 1:n.tch) {
            for (i2 in 1:n.stu) {
                att[i,i2]~dnorm(0,1)I(-4,4)
            }
        }
# GRM teacher method
        for (i in 1:n.tch) {
            for (j in 1:n.item.method) {
                for (k in 1: (n.cat[j]-1)) {
                    p[i,j,k] <- 1 / (1+exp(-ip.tch[j,4]*(method[i]-ip.tch[j,k])))
                }
            }
        }
        for (i in 1:n.tch) {
            for (j in 1:n.item.method) {
                pcat[i,j,1] <- 1-p[i,j,1]
                for (k in 2: (n.cat[j]-1)){
                    pcat[i,j,k] <- p[i,j,k-1]-p[i,j,k]
                }
                pcat[i,j,n.cat[j]] <- p[i,j,(n.cat[j]-1)]
            }
        }
        for (i in 1:n.tch) {
```

```
            for (j in 1:n.item.method) {
                for (k in 1:n.cat[j]) {
                    pc[i,j,k] <- pcat[i,j,k] / sum( pcat[i,j, 1:n.cat[j]] )
                }
                rs.tch[i,j]~dcat(pc[i,j,1:n.cat[j]])
            }
        }
# method items prior
        for (j in 1:n.item.method) {
            ip.tch[j,4]~dlnorm(0,0.5)I(0,5)
            ip.tch[j,1]~dunif(-5,5)
            for (k in 2: (n.cat[j]-1)) {
                ip.tch[j,k]~dunif(ip.tch[j,k-1],5)
            }
        }
# teacher method prior
        for (i in 1:n.tch) {
            method[i]~dnorm(0,1)I(-4,4)
        }
# Student latent achievement theta
        for (m in 1:n.tch) {
            for (n in 1:n.stu){
                for (tpre in 1 : n.item.pre) {
                    logit(prob.pre[m,n,tpre])<-ip[tpre,1]*(theta[m,n,1]-ip[tpre,2])
                }
                for (tpost in 1 : n.item.post) {
                    logit(prob.post[m,n,tpost])<-
```

```
                                    ip2[tpost,1]*(theta[m,n,1]+theta[m,n,2]-ip2[tpost,2])
                        }
# student level variance
                theta[m,n,1:2]~dmnorm(mu.x[m,n,1:2],Omega.x[,])
                mu.x[m,n,1]<-stu[m,n]          # Student expected initial status
                mu.x[m,n,2]<-tch[m,2]          # Student expected gain
                stu[m,n]<-tch[m,1]+female2[m,n]*b.f+att[m,n]*b.a
        }
# bivariate Normal of teacher ini and gain
                tch[m,1:2]~dmnorm(mu.tch.m[m,],Omega.tch[,])
                mu.tch.m[m,1]<-mu.tch[1]
                mu.tch.m[m,2]<-mu.tch[2]+method[m]*b.m+text[m]*b.t
         }

# class level prior and
        r.tch <- Sigma2.tch[1,2] / (sqrt(Sigma2.tch[1,1])
                  *sqrt(Sigma2.tch[2,2]))
        sigma.tch[1]<-sqrt(Sigma2.tch[1,1])
        sigma.tch[2]<-sqrt(Sigma2.tch[2,2])
        Sigma2.tch[1:2,1:2] <-inverse(Omega.tch[,])
        Omega.tch[1:2,1:2]~ dwish(R[,],2)
# student level variance
        r.x <- Sigma2.x[1,2] / (sqrt(Sigma2.x[1,1])
                  *sqrt(Sigma2.x[2,2]))
        sigma.x[1]<-sqrt(Sigma2.x[1,1])
        sigma.x[2]<-sqrt(Sigma2.x[2,2])
        Sigma2.x[1:2,1:2] <-inverse(Omega.x[,])
```

```
        Omega.x[1:2,1:2]~ dwish(Rx[,],2)
# coefficients priors  #debug priors from normal to uniform
        b.a~dnorm(0,1)              # coefficient of attitude prior
        b.f~dnorm(0,1)              # coefficient of female  prior
        b.t~dnorm(0,1)              # coefficient of textbook prior
        b.m~dnorm(0,1)              # coefficient of method  prior
        mu.tch[1]~dnorm(0,1)        # class mean initial prior
        mu.tch[2]~dnorm(0,1)        # class mean gain prior
} # End of model
```

Appendix D

BUGS code for ME.ET project model in chapter 5

```
Model
{# Class/instructor level variance
        for (m in 1:n.ins){
                # bivariate Normal of teacher ini and gain
                tch[m,1:2]~dmnorm(mu.tch.m[m,1:2],Omega.tch[,])
                mu.tch.m[m,1]<-mu.tch[1]
                #mu.tch.m is a temperary variable
                mu.tch.m[m,2]<-mu.tch[2] +text[m,1]*b.t+method[m]*b.m
        }
#Class teaching method GRM
        for (i in 1:n.tch) {
                for (j in 1:n.item.method) {
                        for (k in 1: (n.cat[j]-1)) {
                                p[i,j,k] <- 1 / (1+exp(-ip.tch[j,4]*(method[i]-ip.tch[j,k])))
                        }
                }
        }
        for (i in 1:n.tch) {
                for (j in 1:n.item.method) {
                        pcat[i,j,1] <- 1-p[i,j,1]
                        for (k in 2: (n.cat[j]-1)){
                                pcat[i,j,k] <- p[i,j,k-1]-p[i,j,k]
                        }
                        pcat[i,j,n.cat[j]] <- p[i,j,(n.cat[j]-1)]
                }
        }
```

```
        for (i in 1:n.tch) {
                for (j in 1:n.item.method) {
                        for (k in 1:n.cat[j]) {
                                pc[i,j,k] <- pcat[i,j,k] / sum( pcat[i,j, 1:n.cat[j]] )
                        }
                rs.tch[i,j]~dcat(pc[i,j,1:n.cat[j]])
                }

        }
# method items prior
        for (j in 1:n.item.method) {
                ip.tch[j,4]~dlnorm(0,0.5)I(0,5)
                ip.tch[j,1]~dunif(-5,5)
                for (k in 2: (n.cat[j]-1)) {
                        ip.tch[j,k]~dunif(ip.tch[j,k-1],5)
                }
        }
# teacher method prior
        for (i in 1:n.tch) {
                method[i]~dnorm(0,1)I(-4,4)
        }
# class level variance
        r.tch <- Sigma2.tch[1,2] / (sqrt(Sigma2.tch[1,1])
                *sqrt(Sigma2.tch[2,2]))
        sigma.tch[1]<-sqrt(Sigma2.tch[1,1])
        sigma.tch[2]<-sqrt(Sigma2.tch[2,2])
        Sigma2.tch[1:2,1:2] <-inverse(Omega.tch[,])
```

```
        Omega.tch[1:2,1:2]~ dwish(R[,],2)
# student/level pretest and posttest
        for  (j in 1 : n.stu) {
                for (k in 1 : n.item) {
                        rs.pre[j,k]~dbern(prob.pre[j,k])
                        rs.post[j,k]~dbern(prob.post[j,k])
                }
        }

        for (n in 1:n.stu){
                for (t in 1 : n.item) {
                        logit(prob.pre[n,t])<-ip[t,1]*(theta[n,1]-ip[t,2])
                        logit(prob.post[n,t])<-ip[t,1]*(theta[n,1]+theta[n,2]-ip[t,2])
                }
                theta[n,1:2]~dmnorm(mu.x[n,1:2],Omega.x[,])
                mu.x[n,1]<-tch[ins[n,1],1] + female[n,1]*b.f + att[n]*b.a
                mu.x[n,2]<-tch[ins[n,1],2]
        }
# GRM student attitude
         for (i2 in 1:n.stu) {
                for (j in 1:n.item.att) {
                        for (k in 1: (n.cat.att[j]-1)) {
                                p.att[i2,j,k] <- 1 / (1+exp(-ip.att[j,5]*(att[i2]-ip.att[j,k])))
                        }
                }
         }
         for (i2 in 1:n.stu) {
```

```
            for (j in 1:n.item.att) {
                pcat.att[i2,j,1] <- 1-p.att[i2,j,1]
                for (k in 2: (n.cat.att[j]-1)){
                    pcat.att[i2,j,k] <- p.att[i2,j,k-1]-p.att[i2,j,k]
                }
                pcat.att[i2,j,n.cat.att[j]] <- p.att[i2,j,(n.cat.att[j]-1)]
            }
        }
        for (i2 in 1:n.stu) {
            for (j in 1:n.item.att) {
                for (k in 1:n.cat.att[j]) {
                    pc.att[i2,j,k] <- pcat.att[i2,j,k] / sum( pcat.att[i2,j, 1:n.cat.att[j]] )
                }
            rs.att[i2,j]~dcat(pc.att[i2,j,1:n.cat.att[j]])
            }
        }
# item attitude prior
        for (j in 1:n.item.att) {
            ip.att[j,5]~dlnorm(0,0.5)I(0,5)
            ip.att[j,1]~dunif(-5,5)
            for (k in 2: (n.cat.att[j]-1)) {
                ip.att[j,k]~dunif(ip.att[j,k-1],5)
            }
        }
# student attitude prior
        for (i2 in 1:n.stu) {
            att[i2]~dnorm(0,1)I(-4,4)
```

```
        }
# student level variance
        r.stu <- Sigma2.x[1,2] / (sqrt(Sigma2.x[1,1])
                *sqrt(Sigma2.x[2,2]))
        sigma.x[1]<-sqrt(Sigma2.x[1,1])
        sigma.x[2]<-sqrt(Sigma2.x[2,2])
        Sigma2.x[1:2,1:2] <-inverse(Omega.x[,])
        Omega.x[1:2,1:2]~ dwish(Rx[,],2)
# coefficients
        b.f~dnorm(0,1)
        b.t~dnorm(0,1)
        b.m~dnorm(0,1)
        b.a~dnorm(0,1)
        mu.tch[1]~dnorm(0,1)          # class mean initial  prior
        mu.tch[2]~dnorm(0,1)          # class mean gain    prior
}# End of ME.ET project model
```

REFERENCES

REFERENCES

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

Bartlett, J. W., De Stavola, B. L., & Frost, C. (2009). Linear mixed models for replication data to efficiently allow for covariate measurement error. *Statistics in Medicine, 28*(25), 3158-3178.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). Overview of measuring effect sizes: the effect of measurement error. Brief 2. *National Center for Analysis of Longitudinal Data in Education Research.*

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education, 21*(2), 89-118.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. Boca Raton: CRC Press.

Carroll, R. J. (2006). *Measurement error in nonlinear models : a modern perspective* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics, 10*(4), 637-&.

Cochran, W. G. (1970). Some effects of errors of measurement on multiple correlation. *Journal of the American Statistical Association, 65*(329), 22-34.

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

Dunivant, N. (1981). The effects of measurement error on statistical models for analyzing change (pp. 185). New York: New York University.

Fan, X. T. (2003). Two approaches for correcting correlation attenuation caused by measurement error: Implications for research practice. *Educational and Psychological Measurement, 63*(6), 915-930.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population.

Fox, J. P. (2003). Stochastic EM for estimating the parameters of a multilevel IRT model. *British Journal of Mathematical & Statistical Psychology, 56*, 65-81.

Fox, J. P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical & Statistical Psychology, 58*, 145-172.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*(2), 271-288.

Fuller, W. A. (1987). *Measurement error models*. New York: Wiley.

Gill, J. (2012). *Bayesian Methods: A social and behavioral sciences approach* (3rd ed.): Chapman and Hall/CRC.

Heidelberger, P., & Welch, P. D. (1983). Simulation Run Length Control in the Presence of an Initial Transient. *Operations Research, v31*, p1109-1144.

Hsieh, C.-A., von Eye, A. A., & Maier, K. S. (2010). Using a multivariate multilevel polytomous item response theory model to study parallel processes of change: the dynamic association between adolescents' social isolation and engagement with delinquent peers in the National Youth Survey. *Multivariate Behavioral Research.*

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79-93.

Kolen, M. J., Zeng, L. J., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*(2), 129-140.

Lee, W. C., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 37*(1), 1-20.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics, 32*(2), 125-150.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, N.J.: L. Erlbaum Associates.

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

Lunn, D., Spiegelhalter, D., Thomas, A., Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine, 28*, 3049-3067.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing, 10*, 325--337.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics, 26*(3), 307-330.

Maier, K. S. (2002). Modeling incomplete scaled questionnaire data with a partial credit hierarchical measurement model. *Journal of Educational and Behavioral Statistics, 27*(3), 271-289.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67-101.

McCrory, R., Zhang, C., Francis, A. P., & Young, S. (2009). Factors in the achievement of preservice elementary teachers in mathematics classes. Paper presented at the Psychology of Mathematics Education North America Conference (PME-NA), Atlanta, GA.

Natesan, P., Limbers, C., & Varni, J. W. (2010). Bayesian Estimation of Graded Response Multilevel Models Using Gibbs Sampling: Formulation and Illustration. *Educational and Psychological Measurement, 70*(3), 420-439.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5*(3), 343-355.

Partchev, I. (2006). irtoys: Simple interface to the estimation and plotting of IRT models.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146-178.

R Development Core Team. (2008). R: A Language and Environment for Statistical Computing (Version 2.7.0). Retrieved from http://www.R-project.org

Rabe-Hesketh, S. (2004). Generalized multivevel structural equation modeling. *Psychometrika, 69*(2).

Rabe-Hesketh, S., & Skrondal, A. (2007). Multilevel and latent variable modeling with composite links and exploded likelihoods. *Psychometrika, 72*(2), 123-140.

Raudenbush, S., Bryk, A., Cheong, Y., & Congdon, R. (2004). *HLM 6: Hierarchical Linear and Nonlinear Modeling*: Scientific Software International, Inc.

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

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*(2), 185-205.

Rogers, W. T., & Hopkins, K. D. (1988). Power Estimates in the Presence of a Covariate and Measurement Error. *Educational and Psychological Measurement, 48*(3), 647-656.

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

Samejima, F. (1997). Graded response model. In v. d. Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*(2), 199-223.

Seltzer, M., Choi, K., & Thum, Y. M. (2003). Examining relationships between where students start and how rapidly they progress: Using new developments in growth modeling to gain insight into the distribution of achievement within schools. *Educational Evaluation and Policy Analysis, 25*(3), 263-286.

Stroud, T. W. F. (1973). Comparing regressions when measurement error variances are known (pp. 31). N. J.: Educational Testing Service.

Sutcliffe, J. P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika, 23*(1), 9-17.

Wang, W.-C., & Chen, H.-C. (2004). The standardized mean difference within the framework of item response theory. *Educational and Psychological Measurement, 64*(2), 201-223.

Werts, C. E., & Hilton, T. L. (1977). Intellectual status and intellectual growth, again. *American Educational Research Journal, 14*(2), 137-146.

Woodhouse, G., Yang, M., Goldstein, H., & Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society Series a-Statistics in Society, 159*, 201-212.

Zimmerman, D. W. (2007). Correction for attenuation with biased reliability estimates and correlated errors in populations and samples. *Educational and Psychological Measurement, 67*(6), 920-939.