12-1-2011

The Exploration of the Relationship Between Guessing and Latent Ability in IRT Models

Song Gao
Southern Illinois University Carbondale, songgao@siu.edu

THE EXPLORATION OF THE RELATIONSHIP BETWEEN GUESSING AND LATENT ABILITY IN IRT MODELS

by

Song Gao

A Dissertation

Submitted to the Faculty of the Graduate School of Southern Illinois University at Carbondale

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Department of Educational Psychology and Special Education
in the Graduate School
Southern Illinois University at Carbondale

December 2011

DISSERTATION APPROVAL

THE EXPLORATION OF THE RELATIONSHIP BETWEEN GUESSING AND LATENT ABILITY IN IRT MODELS

BY

Song Gao

A Dissertation Submitted in Partial

Fulfillment of the Requirements

for the Degree of Philosophy

in the field of Educational Statistics and Measurement

Approved by

Todd Headrick, Chair

Yanyan Sheng

Earnie Lewis

Keith Waugh

Mike Grey

Graduate School

Southern Illinois University Carbondale

21st October 2011

AN ABSTRACT OF THE DISSERTATION OF

Song Gao, for the Doctor of Philosophy degree in Educational Statistics and Measurement, presented on 21, October 2011, at the Southern Illinois University Carbondale.

TITLE: THE EXPLORATION OF THE RELATIONSHIP BETWEEN GUESSING AND LATENT ABILITY IN IRT MODELS

MAJOR PROFESSOR: Dr. Todd Headrick

This study explored the relationship between successful guessing and latent ability in dichotomous IRT models. Two new IRT models, the Rasch-Guessing model and the 2PL-Guessing model were developed with guessing functions integrating probability of guessing an item correctly with the examinee's ability and the item parameters. The conventional 3PL IRT model was compared with the new 2PL-Guessing model on parameter estimation using the Monte Carlo method. SAS program was used to implement the data simulation and the maximum likelihood estimation.

Compared with the traditional 3PL model, the new model should reflect: a) the maximum probability of guessing should not be more than 0.5, even for the highest ability examinees; b) different ability of examinees should have different probability of successful guessing because a basic assumption for the new models is that higher ability examinees have a higher probability of successful guessing than lower ability examinees; c) smaller standard error in estimating parameters; d) better AIC for goodness of fit; and e) faster running time. Three criteria were used to compare parameter estimates: correlation, RMSD (root mean squared

deviation), and bias.

Two item response data sets on 20 items from 100, 200, 500, and 1000 examinees using the 3PL model and the 2PL-Guessing model with 10 replications were simulated. Each data set was used by both models to recover parameters to compare the accuracy of parameter recovery between these two models in terms of three aforementioned criteria.

The new 2PL-Guessing model can control the probability of the successful guessing between the probability of random guessing and 0.5 by applying logistic function to the successful guessing probability, successfully reflecting different probability of successful guessing with different ability. The parameter estimate results illustrated that the new 2PL-Guessing model produced higher correlations between true parameter values and estimated parameter values, smaller RMSD, smaller bias, and better AIC for goodness of fit using the dataset generated by the new model. When using the dataset generated by the conventional 3PL model, the new model produced better results for ability and discrimination parameter estimates and smaller average AIC indices across all sample sizes compared than the 3PL model, but the 3PL model produced better difficulty parameter estimates.

ACKNOWLEDGEMENT

The accomplishment of this study could not fulfilled without assistance from my teachers and my family. I would like to express my deep appreciation to all those who have given me huge support in writing my dissertation.

First, I want to give many thanks to my advisor, Dr. Headrick who patiently helped me with ideas, my writing, and organization of this study. I also would like to thank Dr. Sheng who helped me with methodology of parameter estimation and designs of my dissertation study. I also would like to extend my gratitude to other members of my committee, Mike Grey, Keith Waugh, and Ernie Lewis for their constructive suggestions.

Second, I want to express my appreciation to Dr. Reynolds for her measurement class to motivate me to do item response theory study on this topic and her article to guide me through this study.

Third, I would like to thank my wife and my kids for their devoted support in my tough time and my colleagues for their patient help in the computer lab.

TABLE OF CONTENT

LIST OF TABLES

TABLES PAGE

LIST OF FIGURES

FIGURES PAGE

CHAPTER ONE

INTRODUCTION

Guessing and Multiple Choice Tests

Multiple-choice format questions are most frequently used in educational testing, in market research, and in elections. Multiple-choice items consist of a stem and a set of options which are the possible answers from which the examinees can choose. Because only one answer can be correct, when unanswered questions are counted as incorrect for many multiple-choice tests, it makes sense to guess when all else fails. Therefore, most often examinees taking a multiple-choice test may make a guess at the answers when they are not sure which alternative is correct to improve their test scores. This kind of behavior is especially prevalent when there is no penalty for guessing wrong.

Wright (1991) stated that guessing, which can increase opportunities for unqualified individuals, is considered to be a construct-irrelevant response. It is necessary to reevaluate those misfitting persons caused by guessing after they are identified by using the error estimates (also see Pelton, 2002).

Generally, there are two forms of guessing: “blind guessing” or “informed guessing”. Blind guessing occurs when the examinee has no idea of the correct answer and responds randomly while informed guessing occurs when the examinee responds to an item on the basis of partial knowledge. Guessing of one form or another can especially occur on multiple-choice test items and it can increase error variance of test scores, thereby damaging their reliability and validity (Rogers, 1999).

Random guessing, however, provides no information about ability. Correct

responses due to random guessing are quite different from correct responses by guessing when examinees can eliminate some options by partial information (Smith, 1993). Some researchers have tried to distinguish random guessing which contains no information at all from informed guessing which contains some information. Birnbaum (1968) introduced the 3 parameter logistic (PL) item response theory (IRT) model which integrated a guessing parameter reflecting the possibility of a correct guess. The pseudo-guessing item parameter, however, in the three-parameter IRT model mistakes guessing as the only function of the item properties, when, in fact, the guessing is an interaction between item properties and person ability.

There are two good reasons to believe that the success of guessing is related to ability. The first reason is that for a certain item, only some examinees exhibit guessing behavior, especially low ability examinees. The more difficult the item is, the more guessing behavior is exhibited; the easier the item is, the less guessing behavior is exhibited, especially for high-ability examinees. This may explain why there is such a big difference in item parameter estimates between capable and weak students. The second reason is that 3PL IRT model guessing parameter is sometimes more than 1/N (N is the number of options), so a plausible explanation for this issue is that some respondents can eliminate one or more of the options and then guess among the non-eliminated options. Partial knowledge may be reflected in this guessing parameter, so it is ability related (Martin, del Pino, & De Boeck, 2006).

The most important purpose of an exam is to estimate the examinees' ability or academic achievement and to make decisions on the basis of test scores; therefore, it is

very important to consider the effect of guessing on multiple-choice tests because guessing behavior either increases the measured error or can be held accountable for construct-irrelevant variance (Messick, 1995). If guessing behavior is not considered in the IRT parameter estimation, the standard IRT models will misestimate the true levels of the examinees' ability and will cause to make wrong decisions.

Statement of the Problem

Some researchers (e.g. Cao & Stokes, 2008) have engaged to integrate IRT models with guessing. The 3PL model developed by Birnbaum (1968) assumes that the examinees would make a guess if he does not know the correct answer and the probability of guessing correctly will be 1/N (N is the number of options). The model applies this guessing behavior as an item parameter to all examinees assuming that the probability of successful guessing is entirely a quality of the item which has the same fixed effect on all examinees. Wietzman (1996) combined the Rasch model with guessing for a fixed-length, multiple-choice test with the requirement that all multiple-choice items must have equally guessworthy options. That is to say, if an item has 4 options, the guessing parameter c should be equal to 1/4 for all test items regardless of examinees' ability.

However, this pseudo-guessing parameter has stirred a lot of concerns. De Ayala (2008) expressed his concerns on the guessing parameter: a) the difference between the guessing parameter and the random guessing probability occurs all the time and the random guessing assumption for guessing parameter is not reflected in the observed data; b) the responses from low ability individuals demonstrate the interaction between the person's ability and the item characteristics; and c) the assumption for the guessing

parameter that every examinee has the same probability to guess an item correctly may not reflect the real guessing situation.

In addition, the uniform guessing parameter cannot distinguish random guessing from informed guessing, therefore, this guessing parameter revealing nothing about the examinee's partial knowledge. However, Hutchinson (1991) demonstrated that examinees getting a high proportion of items correct at their first attempt tend to get a high proportion of items correct at their second attempt, thus, showing some form of partial information was operating. With partial information, the examinee can eliminate one or more of the distracters as being obviously wrong, and then he guesses randomly among the remainder.

Furthermore, the inaccuracy of estimating guessing parameter causes another concern. Renolds (1986) indicated that the guessing parameter could not be precisely estimated in her simulation study although she increased the sample size and test length and changed the distribution of ability. A research study conducted by Ree (1979) concerning the accuracy of the guessing parameter estimate revealed that the accuracy of the guessing parameter estimate was still poor even with 2000 subjects and 80 test items being simulated. There must be some other factors affecting the accuracy of estimation since sample size and test length are not the primary factors to influence the accuracy of guessing parameter estimate.

Pelton (2002) concluded from his empirical study of the accuracy and stability of estimates on 1PL, 2PL, and 3PL models that the estimation of the guessing parameter is likely to fluctuate substantially with different guessing information. The 3PL model can

produce the best estimates only if a moderate amount of guessing was assumed.

Martin, del Pino, and Boeck (2006) described that the guessing parameter for different ability levels of students may have a substantial impact on item parameter estimates. “In fact, most information about the lower asymptote in the item characteristic curve is obtained for relatively easy items, while the discrepancy between capable and less capable persons may also come from the probability of a correct guessing being dependent on ability” (p.185).

Martin, del Pino, and De Boek (2006) developed a model to integrate guessing behavior with individual latent ability by putting the guessing parameter into a function of the ability of the examinee. However, their models failed to control guessing probability under 0.5 and an extra parameter of the weight of ability in the guessing function had to be estimated for all test items; therefore, the presence of such a parameter increased the complexity of parameter estimation.

Bock (1997) proposed a nominal response model (NRM) to collect more information from incorrect answers and improve the accuracy of ability estimation for multiple-choice items (see also Verstralen, 1997); however, the model can increase accuracy mainly for low test scores. Nedelsky (1954) developed a model based on the idea that the borderline test-taker responds to a multiple-choice question by first eliminating the incorrect options, then guesses randomly from the remaining options. Nedelsky (1954) then generalized this method to all levels of ability. However, the model requires an assumption that the correct answer is never rejected, or the test taker will never think that the correct answer is wrong and a very large sample size is required to get

reliable estimates. Further, Farr, Pritchard, and Smitten (1990) found no evidence to support the assumption of the Nedelsky (1954) model in terms of reading comprehension tests.

Cao and Stokes (2008) proposed three different models based on three different guessing behaviors by using Bayesian estimation methods: a) the IRT threshold guessing model; b) the IRT difficulty-based guessing model; and c) the IRT continuous guessing model.

However, there are some limitations associated with the Cao and Stokes (2008) models. Cao and Stokes (2008) used only low-stake tests to apply three models with assumptions that 60% of the examinees were guessers and the guessing parameter for all three models was equal to the reciprocal of the number of options of the test item. This fails to reflect the relationship between the probability of correct guessing and the examinee's ability. Furthermore, Cao and Stokes (2008) provided little or no discussion in terms of when or in what situation, or which model should be applied.

The Purpose of the Study

The purpose of this study was to: a) analyze and determine the relationship between the probability of guessing a test item correctly and the examinee's ability and item parameters so that different ability examinees have different guessing probability of success; b) propose new IRT models with a guessing function related to the examinee's ability and item parameters; c) use the Monte Carlo method to generate the proposed models' and 3PL model's response data, estimate item parameters, and compare the new 2PL-Guessing model with the 3PL IRT models in item parameter recovery; and d)

compare goodness-of-fit using the real data between the 2PL-Guessing model and the 3PL model. To do parameter estimation for the new models, the Monte Carlo approach will be used to generate simulation data and the marginal maximum likelihood estimation method will be used to estimate model parameters.

The Limitations of the Study

First, the response data were generated under the assumptions of unidimensionality and the normal distribution of ability; therefore, the models may not be appropriate to be applied to multidimensionality tests or polytomous assessments. Second, the models were also developed under the strong assumption that examinees have a high motivation to guess if they do not have the knowledge for the answer because guessing can increase their performance, so the best situations for the application of the models are high-stake tests, achievement tests or licensure tests. Third, the proposed new model may not be appropriate for classroom exams because sometimes instructors want all students to answer some items correctly for classroom tests.

The Significance of the Study

The issue of guessing is important to multiple-choice assessments because guessing behavior can be a source of construct-irrelevant variance, posing a major threat to construct validity; furthermore, the use of guessing strategies not only increases error, but also weakens the relationships among test items. Therefore, it is imperative to account for guessing in the evaluation of multiple choice tests (Messick, 1995).

A multiple-choice question always provides opportunity for successful guessing. This kind of systematic error may increase the probability of success for the lower ability

examinees. If this bias or systematic error is not handled appropriately by the model, it may have a negative effect on the precision of the item difficulty and discrimination parameter estimates (Pelton, 2002).

Multiple-choice test items are subject to guessing, so answering an item correctly and knowing the correct answer to the item are not equivalent. Birnbaum (1968) introduced the 3PL model with a controversial guessing parameter which completely depends on item property and has nothing to do with the examinee's ability. Furthermore, estimation of the guessing parameter in the 3PL model is the most unstable and Ree (1979) found that even large samples and long tests could not improve the accuracy of the guessing parameter estimate.

The proposed new models should reflect: a) the maximum probability of guessing should not be more than 0.5, even for the highest ability examinees; b) different ability of examinees should have different probability of successful guessing, because a basic assumption for the new models is that higher ability examinees have a higher probability of successful guessing than lower ability examinees; and c) because the new 2PL-Guessing model has only two item parameters, the running time for estimation of item parameters is much shorter than the 3PL IRT model.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction of Item Response Theory

Gulliksen (1950) indicated that an important contribution to the theory and practice of item analysis would be the discovery of item parameters that are relatively invariant to different examinee groups on which item analysis is based. Although classical test theory (CTT) has been widely used in the measurement field for a long time because of its simplicity and relatively weak assumption requirements which make CTT easily applied to many test situations, IRT has experienced tremendous growth in recent decades since IRT overcomes the circular dependency , the major weakness associated with CTT (Hambleton & Jones, 1993). IRT, also known as latent trait theory, is a model-based measurement in which ability estimates depend on both examinees' responses and on the properties of the administered items (Embreston & Reise, 2000).

Compared to CTT, IRT is more theory-driven and models the probability of examinees' successful responses by the item statistics independent of examinee samples, the individual latent ability, and the particular set of items administered. That is to say, when the IRT model fits the data, the same item characteristic curve (ICC) is obtained for the test item regardless of the distribution of ability in the group of examinees used to estimate the item parameter. The chief advantage of IRT is the properties of item and ability parameter invariance which is crucial for inferences to be equally valid for different populations of examinees or different measurement conditions. "The importance of the property of invariance of item and ability parameter cannot be overstated. This

property is the cornerstone of item response theory and makes possible such important applications as equating, item banking, investigation of item bias, and adaptive testing." (Hambleton, Swaminathan & Rogers, 1991, p. 25)

More and more test developers are using IRT to design standardized tests due to IRT's potential to solve practical issues and its theoretical invariance advantage. IRT now are applied to several major tests such as the Armed Services Vocational Aptitude Battery, SAT, and GRE. The early IRT applications involved mainly unidimensional IRT models (Embreston, 2000). Since Bock, Gibbons, and Muraki (1988) developed a multidimendional IRT model, IRT applications to personality, attitude, and behavioral self-reports have become possible as well. IRT has increasingly become the mainstream in the measurement field.

IRT Assumptions

Even though IRT has many advantages over CTT, these advantages can only take effect when its assumptions are met. There are two important assumptions for IRT models: unidimensionality and local independence.

A common assumption of IRT models is that only one latent trait (or ability) is measured by a set of items in an exam. However, it is impossible to meet this assumption because other factors such as personality, motivation, anxiety, and guessing always affect test performance to some extent. Therefore, if there is the presence of a dominant component or factor in a set of test data, we would say, the unidimensionality assumption is met and this dominant factor is referred to as the latent trait measured by the test.

Local independence means that the probability of answering any test item is

independent of the probability of answering any other test item when the abilities are held constant (Hambleton, Swaminathan & Rogers, 1991). The property of local independence, for a given examinee, means that the probability of a response pattern on a test is equal to the product of each test item probability. The assumptions of unidimensionality and local independence are equivalent because local independence can be obtained if unidimensionality is met. The property of local independence can be expressed mathematically in the following way:

$$\Pr ob(U_1, U_2, ..., U_n \mid \theta) = P(U_1 \mid \theta)P(U_2 \mid \theta)...P(U_n \mid \theta) = \prod_{i=1}^{n} P(U_i \mid \theta), \qquad (2.1)$$

where θ represents the examinee's ability level; U_i represents the response of a randomly chosen examinee to item $i\,(i = 1, 2, \ldots, n)$; $P(U_i \mid \theta)$ denotes the probability of the response of a randomly chosen examinee with ability θ; $P(U_i = 1 \mid \theta)$ denotes the probability of a correct response, and $P(U_i = 0 \mid \theta)$ denotes the probability of an incorrect response.

The final assumption for any selected IRT model is that the model must fit the data. That assumes that the ICC of chosen IRT model must be able to provide an accurate reflection of the relationship between examinees' ability and item response (Davis, 2002). The advantages of IRT models can be achieved only if there is a satisfactory goodness-of-fit between the model and test data. If the model fits the data poorly, the invariance of the examinee's latent ability and item parameters will be compromised (Hambleton, Swaminathan & Rogers, 1991).

There is no an absolute statistical method to determine a particular model fit or not overall, but Embreston and Reise (2000) suggested two approaches to evaluating

goodness-of-fit for IRT models: item fit and person fit.

Dichotomous IRT Models

The three most popular unidimensional IRT models are the one-, two-, and three-parameter logistic models named because of the number of item parameters each model has also, these models are appropriate for dichotomous item response data. A primary distinction among these three models is the number of parameters used to describe items.

One-Parameter Logistic Model.

The one-parameter logistics model, which is often called the Rasch model, is one of the most widely used IRT models. The probability of answering an item correctly is given by the equation

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad i = 1, 2, ..., n, \qquad (2.2)$$

where $P_i(\theta)$ is the probability that a randomly chosen examinee with ability θ answer item i correctly, b_i is the item i difficulty parameter, n is the number of items in the test, and e is a transcendental number whose value is 2.718. In the one-parameter model, it is assumed that item difficulty is the only item property that affects examinee performance. This is equivalent to the assumption that all item discrimination indices are equal. The lower asymptote of the ICC is zero which means examinees of very low ability have zero probability of answering the item correctly. Thus, there is no allowance for guessing in this model (Hambleton, Swaminathan & Rogers, 1991).

Two-Parameter Logistic Model.

Birnbaum (1968) substituted the two-parameter logistic function for the two-parameter normal ogive function developed by Lord (1952) because logistic functions have the

important advantage of being more convenient to work with than normal ogive functions. The probability of answering an item correctly is expressed by two-parameter model developed by Birnbaum (1968) as:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad i = 1, 2, ..., n \qquad (2.3),$$

where a_i represents the discrimination parameter of item *i*. The item discrimination is proportional to the slope of ICC at the point b_i on the ability scale. The steeper the slope is, the more useful the item is to separate examinees into different ability levels (Hambleton, Swaminathan & Rogers, 1991). *D* is a scaling factor developed to make the logistic function as close as possible to normal ogive function and it is a constant and is equal to 1.7.

Three-Parameter Logistic Model.

The mathematical expression for the three-parameter logistic model is

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad i = 1, 2, ..., n, \qquad (3.5)$$

where c_i is called the guessing parameter which provides nonzero lower asymptote for the ICC and represents the probability of examinees with low ability answering the item correctly. It is important to note that by definition, the value of *c* does not vary as a function of ability level in this equation. Thus, the lowest and highest ability examinees have the same probability of answering the item correctly by guessing.

Due to each model's different properties and assumptions, the selection of the model should be determined by the primary purpose. The Rasch model has the advantage of estimating the fewest parameters (Davis, 2002). In addition, the Rasch model is robust in

that it is capable of calibrating data containing substantial variations to the item discrimination parameters (Linacre, 2002) and other deviations from model assumptions (Fisher, 1993; Linacre, 1995). On the other hand, the 3PL model which includes the possible potential for guessing on multiple-choice questions with guessing parameter *c* requires the most parameter estimation and "the pseudo-guessing parameter is especially difficult to estimate because of sparse data conditions at low ability level" (Davis, 2002, p14).

One of the important advantages of IRT models over CTT is that each item has an item information function *I(θ)* that can be transformed into an item information curve (IIC) which reflects the amount of information an item provides for different latent trait level (Embreston & Reise, 2000). The item information function can be expressed as:

$$I_i(\theta) = \frac{P_i^{'}(\theta)^2}{P_i(\theta)Q(\theta)} \quad i = 1, 2, ..., n, \qquad (2.5)$$

where $P_i(\theta)$ equals the probability of correctly responding to item i given θ, $P_i^{'}(\theta)$ is the first derivative of $P_i(\theta)$ with respect to θ, and $Q_i(\theta)$ is equal to $(1 - P_i(\theta))$. The information functions can be used to select test items on the basis of ability level.

Hambleton, Swaminathan and Rogers (1991) stated the amount of information provided by a test at θ is inversely correlated to the precision with which ability is estimated at that point:

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}, \qquad (2.6)$$

where $SE(\hat{\theta})$ is called the standard error of estimation and $I(\theta)$ represents the test information function (which is the sum of all item functions) at θ.

How much information an item can provide completely depends on the item parameters. In the Rasch and the 2PL models, the item provides the maximum information at b_i , so those examinees whose ability is equal to the item difficulty parameter receive most information from the item. In the 3PL model, because of the effect of the guessing parameter, the maximum amount of item information occurs slightly to the right of b_i depending on the value of the guessing parameter (Embreston & Reise, 2000). The amount of information an item can provide is associated with item discrimination parameter. The higher the discrimination, and the more information the item provides. In the 3PL model, the guessing parameter has a negative effect on the information the item gives (Davis, 2002).

2.2 Methods of Parameter Estimation

There are two main techniques of estimating parameters for binary response IRT models: the maximum likelihood estimation and the Bayesian estimation. The maximum likelihood is the most popular method to estimate item parameters for IRT models, while Bayesian estimation method can be more effective if prior information for item parameters is available (Embreston & Reise, 2000). A review of three maximum likelihood methods: a) joint maximum likelihood (JML); b) conditional maximum likelihood (CML); c) marginal maximum likelihood (MML); and d) Bayesian estimation methods in the literature will be discussed in this section.

Maximum Likelihood Function

Maximum Likelihood estimation is a popular statistical method used to estimate the model's parameters through a joint probabilistic function of observed data. The likelihood function is equal to the product of probabilities associated with each item response if the local independence is true (Si & Schumacker, 2004). Let us use 2-PL IRT as an example and $y_{i1}, y_{i2}, ...y_{ik}$ denote the binary responses of the ith individual to k test items, $\mathbf{a} = (a_1, a_2 ... a_k)$ and $\mathbf{b} = (b_1, b_2 ... b_k)$ be the vector of item discrimination and difficulty parameters respectively. The probability of obtaining a response vector $\mathbf{y}$ given θ_j, $\mathbf{a}$ and $\mathbf{b}$ for ith individual is given by

$$\Pr(Y_{i1} = y_{i1}, ..., Y_{ik} | \theta_i, \mathbf{a}, \mathbf{b}) = \prod_{j=1}^{k} \Pr(Y_{ij} = y_{ij} | \theta_i, \mathbf{a}, \mathbf{b})$$

$$= \prod_{j=1}^{k} f(a_j \theta_i - b_j)^{y_{ij}} \left[1 - f(a_j \theta_i - b_j)\right]^{(1-y_{ij})}. \qquad (2.7)$$

If the responses of each of the n individuals to the test items are assumed to be independent, then the likelihood function for all individuals will be

$$L(\theta, \mathbf{a}, \mathbf{b}) = \prod_{i=1}^{n} \prod_{j=1}^{k} f(a_j \theta_i - b_j)^{y_{ij}} \left[1 - f(a_j \theta_i - b_j)\right]^{(1-y_{ij})}. \qquad (2.8)$$

This function represents the likelihood of obtaining the observed data as a function of the model parameters. By applying a logarithm to L, maximum likelihood estimation is used to calculate the value of the parameters that maximize the value of L by solving the first derivative likelihood equation = 0 (Si & Schumacker, 2004).

The Newton-Raphson algorithm is an iterative process used to find a solution for likelihood equation. Si and Schumacker (2004) described:

The Newton-Raphson algorithm starts with an initial value for the estimate of the

parameter in the model. The number of items correct is typically used for the ability estimates and item statistics are used for item estimates. In each iteration, a new estimate for the parameter is generated based on the estimate obtained from the previous iteration. The difference between the new and old estimates are calculated for each iteration. The iterations continue until the difference is smaller than a pre-set minimal value, then the estimates has converged and is maximum likelihood estimate of the parameter. (p.154-155).

Joint Maximum Likelihood Estimation.

JML is one of the most widely used parameter estimation methods and both item and ability parameters are estimated simultaneously in this method (Lim & Drasgow, 1990). There are two steps for JML. Initial values for ability parameters must be selected on the basis of examinees' test scores and used as known ability values to estimate item parameters in the first step; and then in the second step, item parameters are treated as known to estimate ability parameters. This two-step process is stopped until there is no difference between two-step estimations (Hambleton, Swaminathan & Rogers, 1991).

Even though JML is easily programmable, applicable to many IRT models, and efficient in computation (Embreston & Reise, 2000), there are several disadvantages to JML. First, it does not produce consistent item and ability parameter estimation for the 2PL or the 3PL IRT models because an increase in sample size does not result in improved estimations. Second, ability parameters cannot be estimated for perfect or zero scores. Third, parameter estimates for items answered correctly by all examinees are not available. Fourth, JML is very sensitive to item parameter fixed initial values when

applied to the 3PL model (Hambleton, Swaminathan & Rogers, 1991).

Conditional Maximum Likelihood Estimation.

The CML estimation method, compared with JML, produced more consistent and efficient parameter estimates by removing the trait level parameters from the likelihood equations (Si & Schumacker, 2004). CML can be implemented only if a sufficient statistic is available in the data for ability and item parameters. Embreston and Reise (2000) explained that "A sufficient statistic means that no other information is needed from the data for estimating the parameter" (p. 215). In the 1-PL Rasch model, the item total scores are sufficient statistics for the ability parameter and the number of correct responses to an item is a sufficient statistic for the item difficulty parameter (Si & Schumacker, 2004). CML can be only applicable to the Rasch model because of this sufficient statistic condition.

While CML has the advantages of no requirement for ability distribution, more reliable parameter estimations compared with JML, its several disadvantages are discussed here. First, CML cannot be applied to the 2PL and 3PL models and this limits its applications. Second, estimations for examinees with extreme scores (zero or perfect) and these scores have to be removed prior to estimation. Third, CML loses its accuracy in estimating parameters for a long test (Embreston & Reise, 2000).

Marginal Maximum Likelihood Estimation.

Due to CML's limitation that can be only applied to the Rasch model, an alternative estimation method for the 3PL and 2PL models developed by Bock and Lieberman (1970) is marginal maximum likelihood estimation (MMLE). The most important advantage of

MML over CML and JML is that the ability parameter is treated as a random nuisance parameter can be removed by integrating over ability distribution (Lord, 1986; Bock & Aitkin, 1981; Harwell, Baker, & Zwarts, 1988). More formally, by defining θ to represent ability level and y_i to denote the ith examinee's responses to test items, then the likelihood function for individual *i* is:

$$L_i(y_i) = \int \Pr(y_i / \theta) f(\theta) d\theta,$$

$L_i(y_i)$ is a function of the item parameters because the ability parameter θ has been integrated out.

Although MML method by Bock and Lieberman (1970) can be used to estimate the 2PL and 3PL model parameters, this approach is computationally expensive and it was not feasible for very long tests (Si & Schumacker, 2004). Bock and Aitken (1981) used the EM algorithm for MML to estimate item parameters. The EM algorithm involves an iterative two-stage procedure for finding maximum likelihood estimates (Harwell, Baker , & Zwarts, 1988): an expectation (E) stage and a maximization (M) stage. Embreston and Reise(2000) explained that in the expectation stage, the expected numbers of the examinees at each quadrature point and the expected numbers of examinees passing each single item are computed, and then these expected values are used to execute the regular maximum parameter estimations in the maximization stage. These parameters are then used to determine the distribution of latent variables in the next expectation step. This repetition stops until the estimates converge. The Newton-Gauss method is used to solve the maximum likelihood equation and find the standard errors(Si & Schumacker, 2004).

Compared with other maximum likelihood estimation methods, MML has several advantages (Embreston & Reise, 2000). First, it can be applied to all types of IRT models and any length of tests. Second, estimates for perfect and zero scores are available and thus no loss of information. Third, the estimates of item standard error in MML are good approximations of expected sampling variance of the estimates. Fourth, the item parameter estimate is completely independent of the ability distribution, so MML can obtain reliable estimates even for small sample sizes and short tests (Si & Schumacker, 2004; Harwell, Baker , & Zwarts, 1988). However, the main disadvantage associated with MML estimation are its complicated computational process. MML computational process has created a huge problem for computer programming; another disadvantage of MML is that ability distribution has to be assumed normal if there is no prior ability distribution information available. (Embreston & Reise, 2000; Si & Schumacker, 2004; Baker, 1992).

Bayesian Estimation

Bayesian model parameter estimation for IRT models is similar to marginal maximum likelihood estimation, however, Bayesian method requires prior information of item parameters (Johnson, 2007). The posterior distribution is obtained through the product of the likelihood function and prior distribution (Lim & Drasgow, 1990). It can be expressed for the 2-PL model as:

$$P(\boldsymbol{\theta},\mathbf{a},\mathbf{b}/\mathbf{y}) \propto L(\mathbf{y}/\boldsymbol{\theta},\mathbf{a},\mathbf{b})P(\boldsymbol{\theta},\mathbf{a},\mathbf{b}), \qquad (2.9)$$

$P(\boldsymbol{\theta},\mathbf{a},\mathbf{b}/\mathbf{y})$ represents the distribution of parameter estimates based on the item response vector $\mathbf{y}$ (the posterior distribution). The $P(\boldsymbol{\theta},\mathbf{a},\mathbf{b})$ is the prior distribution of parameter estimates. $L(\mathbf{y}/\boldsymbol{\theta},\mathbf{a},\mathbf{b})$ is the likelihood function. Bayesian methodology uses equation

(2.9) to estimate parameters (Si & Schumacker, 2004).

There are two types of priors in terms of their distributions: noninformative priors and informative priors. A noninformative prior distribution has a large variance and has little effect on the parameter estimates, while an informative prior distribution has a small variance and can estimate parameters close to the mean of the prior distribution and this is the main reason why informative priors are favored in some cases (Sheng, 2008; Si and Schumacker, 2004).

The most important advantage of Bayesian estimation is that the parameter estimates can be controlled in a reasonable range because item parameter prior information is used (Lim & Drasgow, 1990). On the other hand, the major problem associated with Bayesian model estimation occurs when prior information is incorrect and this may cause systematic bias to item parameter estimates (Baker, 1987; Lim & Drasgow, 1990; Mislevy, 1986).

2.3 Guessing Parameter and Latent Ability

Approaches to the Guessing Effect in CTT

Ever since multiple-choice tests became popular, there has been concern over the guessing effect on test scores. In the beginning, score increases due to guessing were deemed as being dishonest even though these score components usually reflect partial knowledge-the ability to eliminate some wrong options before guessing. Some educators even think guessing on test items has caused the primary psychometric problem since it increases the error variance of test scores, thereby reducing their reliability and validity. Hambleton and his colleagues (1992) indicated:

The inclination to guess is an idiosyncratic characteristic of particular low ability examinees. Lucky guessing is a random event. Neither feature contributes to valid measurement of a latent trait. Parameterizing guessing penalizes the low performer with advanced special knowledge and also the non-guesser. Rasch flags lucky guesses as unexpected responses. They can either be left intact which inflates the ability estimates of the guessers, or removed which provides a better estimate of the guessers' abilities on the intended latent trait. In practice, 3-P guessing parameter estimation is so awkward that values are either pre-set or pre-constrained to a narrow range (p.215).

As a result, many educators try to avoid the use of multiple choice tests and some educators admonish students against guessing. However, multiple-choice tests have become inevitably dominant in mass testing because of their advantages of broader coverage of instructional content, reliable scoring, and easily calculated item statistics. Hence, neither admonishment against guessing nor avoidance of multiple-choice tests was an effective approach to the guessing problem.

Since the 1920s, when multiple-choice tests came into widespread use, there has been considerable research conducted to reduce the effects of guessing on test scores. Since guessing is not directly measured in CTT, much of the research in CTT on corrections for guessing has focused on correction formulas scoring. The most widely used correction formula is based on the assumption that the examinee either has the complete knowledge or know nothing about the item, and he either skips the item or makes a random guessing. Therefore, wrong answers are deemed as the result of unlucky guessing; and then the number of wrong answers can be used to predict the number of

lucky guesses, which need to be deducted from the examinee's score (Rogers, 1999).

The standard correction for guessing is given by the formula:

$$F = R - \frac{W}{(A-1)},$$

where R is the number of right answers, W is the number of wrong answers, and A is the number of alternatives for each item.

Thorndike (1982) developed a corresponding correction that can be applied to the item difficulty index or p – value (the proportion of examinees answering the item correctly). The corrected p -value is given by the formula

$$p_c = p - \frac{p_w}{(A-1)},$$

where p is the item difficulty index and p_w the proportion of examinees attempting the item who answered it incorrectly. A problem with this correction is that when the proportion of correct answers falls below the chance level, the corrected difficulty index can be less than zero.

Rowley and Traub (1977) criticized the formula because it ignored the possibility that the examinee can use partial knowledge to eliminate some distracters and is more likely to get an item right than if the examinee guesses randomly, so the formula scoring discriminates against the examinee who omits items. It has been discussed that informed guessing increases true score variance rather than error variance and thus increases the validity of scores (Mehrens & Lehman, 1987). Moreover, when examinees respond to an item on the basis of partial knowledge, their test scores are based on a greater sample of content, and hence may have better validity.

Rogers (1999) indicated that another criticism about the appropriateness and effectiveness of formula scoring was that: (a) it is based on false assumptions about examinee behavior, and (b) it disadvantages examinees who exhibit the reluctance to take a risk to guess. With respect to the first point, critics argue that there are no such ignorant examinees that they will not attempt or be completely unable to rule out a single distracter on a large number of questions, that is, examinees who have no knowledge to answer the question rarely guess randomly. Thorndike (1982) demonstrated this point by the example of a set of verbal analogy items from a published test, where the most popular distracter was chosen by about 20 per cent of examinees and the least popular by about 4 per cent of examinees, therefore, the effort of "correcting for guessing" is largely useless.

With respect to the second point, there is a considerable body of research which shows that the extent to which examinees comply with the instructions associated with formula scoring (i.e., the instruction to omit rather than guessing randomly). This reflects a personality trait which may bias against some examinees (Diamond & Evans, 1973; Rowley & Traub, 1977). Examinees who are more willing to take risks will not be penalized on average, because at most they will lose the points gained by randomly guessing. The research in this area indicates that the tendency to omit items under formula scoring directions is personality trait which is more reliably measured by multiple choice tests than the cognitive trait of interest (Rogers, 1999).

Even though the formula scoring has its assumption problems, some educators are in favor of it because it increases the reliability and validity of test scores (Mattson, 1965; Lord, 1975). Prihoda and Pinckard (2006) compared uncorrected and corrected for

guessing scores on multiple-choice examinations with scores on short-answer examinations for dental students; they found that students guessed at a level close to random guessing and correcting for guessing increased the validity in multiple-choice tests and they suggested that instructors using multiple-choice tests should either correct for guessing or take the effect of guessing into account when establishing the criterion for passing grades at different levels.

A second argument in favor of formula scoring is based on empirical studies that formula scoring has an advantage of equating the mean scores of randomly equivalent groups of examinees who have been given different instructions regarding guessing. Angoff and Schrader (1984) compared the mean number-right and formula scores of groups of examinees and found that while the means for two groups of number-right scores were significantly different, but there was no difference between the means for two groups of formula scores.

IRT Approaches to the Guessing Effect

Item response theory provides an alternative approach to the problem of guessing. Under IRT, an examinee's observed performance on a test item is assumed to depend on the latent trait level and properties of the test item and the probability of a correct response to an item as a function of person and item parameters; the examinee ability estimate is not a simple transformation of number-correct test score; it is estimated in the presence of item parameters, thus taking into account the properties of the item.

The most commonly used IRT model integrated with guessing is 3-paramter model assuming that examinee's probability of a correct response on a test item is

affected by three characteristics of the item: its difficulty, discrimination, and a guessing factor which reflects the probability that a very low ability examinee will answer the item correctly. Items differ in their c-parameters due to their difficulty and the attractiveness of the distracters.

Although the c-parameter takes into account the ability of the examinee for the adjustment of probability of guessing (that is why c-parameter estimate is usually higher than the probability of random guessing), a uniform nonzero value of guessing parameter applied to all examinees should be the biggest concern for educators because the precision of estimation of ability is reduced and error variance is increased. As a matter of fact, the c-parameter is always poorly estimated even though the data for three-parameter model have large samples of examinees and long tests. For this reason, many practitioners choose less-restricted one-parameter or two-parameter model which is easier to fit to test data, making no allowance for guessing behavior. If guessing behavior is a factor in test scores, the ability of the examinee will be overestimated (Rogers, 1999).

Multiple choice items are subject to guessing which can cause irrelevant variance and increases measurement error, so some researchers have engaged to solve this problem by two different methods: one is to get rid of random guessing effect on multiple choice items; the other is to integrate guessing parameter with latent ability into IRT models.

Waller (1973) introduced the Ability Removing Random Guessing (ARRG) model to deal with the problem by focusing on the interaction between the person and the item. He simply omitted those item-person interaction for estimation of θ to eliminate the effect of random guessing on any particular item by including only items for which essentially

random guessing is unlikely to occur. The following is his model:

$$P_{ij} = \frac{1}{1+\exp\left[a_j(b_j-\theta_i)\right]} \quad P_{ij} > P_c,$$

where the ARRG cutoff value P_c is less than or equal to $1/A_j$, A_j is the number of alternatives for item j. The model divides the items into two groups for each person: those items whose P_{ij} is greater than P_c, and those items whose P_{ij} is less than or equal to P_c. The ARRG model uses only those items from the first group to estimate a person's ability.

Even though the ARRG model estimates a person's ability on the basis of fewer items than two-parameter model does, the resulting estimated precision has been found to increase (Waller, 1973) because the noise caused by random guessing is removed for the estimation of item parameters. However, when the ARRG model was compared with three-parameter model, the ARRG model failed to produce better fit to empirical data. He also indicated that the three-parameter model using the individualized method to estimate guessing parameter produced a better fit to the data than did the three-parameter model using fixed value for guessing parameter.

Cao and Stokes (2008) developed three Bayesian IRT guessing models to accommodate different guessing behavior: the threshold guessing model, the difficulty-based guessing model, and the continuous guessing model. The threshold guessing model assumes that some examinees answer questions on the basis of their knowledge up to a certain test item, and guess randomly thereafter. An item location threshold for each examinee has to be specified for this model. The difficulty-based model assumes that some examinees answer relatively easy items on the basis of their knowledge and guess

randomly on relatively difficulty items. The continuous guessing model is constructed under the assumption that low-motivated examinees use less effort to answer test items than motivated examinees, and thus they are more likely to answer questions wrong.

A few critical limitations for Cao and Stokes' guessing models must be highlighted here: 60% of the examinees are guessers under the threshold guessing model; the probability of guessing is equal to .25 (assuming each question has 4 options), which means once examinees guess on test items, they guess randomly; these models can be applied to only low-stake tests.

Martin, del Pino, and De Boeck (2006) developed more reasonable models to integrate ability with guessing parameter. They ended up with only one reasonable model:

$$P(Y_{ij} = 1) \mid \theta) = p_{ij} + (1 - p_{ij}) g_j$$

The first is that the p -process comes first and that, depending on the result, the g – process follows. This would mean that the examinee first works on the question with a probability of p_{ij} to answer it correctly; if the examinee could not find correct answer, he or she would make a guess with a success probability of g_j. The guessing probability of g_j is defined by:

$$g_j(Y_{ij} = 1 \mid \theta_i) = \frac{e^{(\alpha\theta_i + \gamma_j)}}{1 + e^{(\alpha\theta_i + \gamma_j)}},$$

where $\theta_i \sim N(0, \sigma^2)$ is a latent ability of the examinee; γ_i is the guessing parameter of item j on the logistic scale, corresponding to a person with average ability; and α is the weight of the ability in the guessing component.

Martin, del Pino, and De Boeck (2006) showed, by using the previous ability-based guessing model, that the ability contributed the chances of correct guess. Even

though the number of the replications was very small, the results of parameter recovery from the simulation study were very consistent and accurate. When this ability-based model was applied to two real tests, language and mathematics, ability played much more important role in making a correct guess on the language test than on the mathematics test. Martin, del Pino, and de Boeck (2006) explained the difference by giving two reasons: a) mathematics is perhaps more like an know all or know nothing matter; and b) the examinees were not motivated to guess because of higher non-response rate for the mathematics test.

CHAPTER THREE

METHODOLOGY

3.1 The Proposed Models

The guessing parameter of a conventional 3-PL IRT model has the same value for every examinee regardless of their ability, which means all examinees have the same probability of guessing the same item correctly. The purpose of this study is to develop new IRT models in which the guessing function is associated with the examinee's latent trait and item characteristics and the proposed IRT models based on the conventional 3-PL IRT model (Birnbaum, 1968) should integrate the examinees' ability and item parameters into the guessing function.

The proposed models associate guessing with examinees' abilities and the number of multiple choice alternatives and item parameters; it can reflect that the higher ability examinees have higher probability of guessing the same item correctly.

Assumptions for the Proposed Models

The proposed models were developed on the basis of the following assumptions: First, these new models are applied to achievement, high-stake, and licensure multiple choice tests with no penalty for wrong answers; if the examinee cannot find correct answers to test items, he/she will guess at those items. Second, all examinees try to use knowledge to answer test questions first and if they cannot find correct answers they will apply guessing strategies to test items, meaning they will use their partial knowledge to eliminate some alternatives to increase the probability of guessing items correctly. Third, the probability of guessing correctly is related only to item difficulty and discrimination

and the examinee's ability; higher ability examinees have higher probability of guessing correctly the same item correctly than lower ability examinees. Fourth, the highest ability groups are engaged in some level of guessing, no matter how small. Freedle (2006) examined hundreds of test items for low ability students and found that none of the data fit the classic definition of truly random guessing data; he also found that 6% of the students who earned 600 on SAT verbal part were engaged in guessing. Fifth, unidimensionality and local independence assumptions are also applied to the new models. Sixth, the guessing probability cannot be greater than 0.5 because the highest probability of successful guessing is between two options if the examinee does not know the correct answer after eliminating other options, he or she has to make a random guess between the remaining two options. The following equation is the general equation for all IRT models

$$P_{ij}(Y_{ij}=1\,|\,\theta_i, a_j, b_j) = P_{ij} + (1-P_{ij})g(\theta_i\,|\,a_{j,}b_j) \quad (3.1),$$

where the guessing probability distribution$g(\theta_i|a_j, b_j)$ is associated with ability and item parameters.

The Proposed Models with Guessing Function

Let $\theta_i = \theta$ (for convenience reason only) and *M* is equal to the number of options for a multiple choice item. To control the probability of successful guessing between the probability of successful random guessing (1/*M*) and the highest probability of guessing correctly (0.5), a logistic function $g(\theta) = \frac{e^{(\theta)}}{1+e^{(\theta)}}$ was used to start to develop the guessing function first.

If $g(\theta)$ need to be under 0.5, then

Let $g(\theta) = \frac{e^{(\theta)}}{1+2e^{(\theta)}}$.

We know the probability of successful guessing is inversely apportioned to item difficulty and its relationship with discrimination parameter should be just like the probability of answering correctly in 2PL IRT model, so we can change the guessing function into

$$g(\theta) = \frac{e^{a_j(\theta-b_j)}}{1+2e^{a_j(\theta-b_j)}}. \qquad (3.2)$$

To make the random guessing probability equal to 1/*M*, 1/*M* constant should be added to the guessing function 3.2. To make the highest probability of guessing correctly equal to 0.5, let $\lim_{\theta\to\infty} g(\theta) = 0.5 - 1/M = 1/2 - 1/M = \frac{M-2}{2M}$, the coefficient for $e^{a_j(\theta-b_j)}$ in the denominator has to change into $\frac{2M}{M-2}$ $(M \geq 3)$.

Model 1. The proposed Rasch model with guessing function

$$P_{ij}(Y_{ij}=1 \mid \theta, b_j) = \frac{e^{(\theta-b_j)}}{1+e^{(\theta-b_j)}} + (\frac{1}{1+e^{(\theta-b_j)}})(\frac{1}{M} + \frac{e^{(\theta-b_j)}}{1+(\frac{2M}{M-2})e^{(\theta-b_j)}}), \qquad (3.3)$$

$$g(\theta) = \frac{1}{M} + \frac{e^{(\theta-b_j)}}{1+(\frac{2M}{M-2})e^{(\theta-b_j)}}, \qquad (3.4)$$

where *M* is the number of options in any multiple choice item and b_j is the *jth* item difficulty parameter.

Model 2. The 2PL IRT model with guessing function

$$P_{ij}(Y_{ij}=1 \mid \theta, a_j, b_j) = \frac{e^{a_j(\theta-b_j)}}{1+e^{a_j(\theta-b_j)}} + (\frac{1}{1+e^{a_j(\theta-b_j)}})(\frac{1}{M} + \frac{e^{a_j(\theta-b_j)}}{1+(\frac{2M}{M-2})e^{a_j(\theta-b_j)}}), \qquad (3.5)$$

where a_j is the *jth* item discrimination parameter and its the guessing function is:

$$g(\theta) = \frac{1}{M} + \frac{e^{a_j(\theta - b_j)}}{1 + (\frac{2M}{M-2})e^{a_j(\theta - b_j)}}. \qquad (3.6)$$

Properties of Guessing Function.

Property 1. The probability of guessing an item correctly is associated with the examinee's ability (θ) and the item difficulty (b) and discrimination (a) parameters. For the Rasch model, the probability of guessing is associated with the examinee's ability and item difficulty only; for the 2PL model, the probability of guessing is associated with the examinee's ability and item difficulty and discrimination. Figure 1 is 3D graph for guessing function 3.4 with M=4, the graph shows that when the difficulty b increases the probability of guessing correctly at the same ability level will decrease; while the probability of guessing correctly will increase at the same difficulty value when the ability level θ increases.

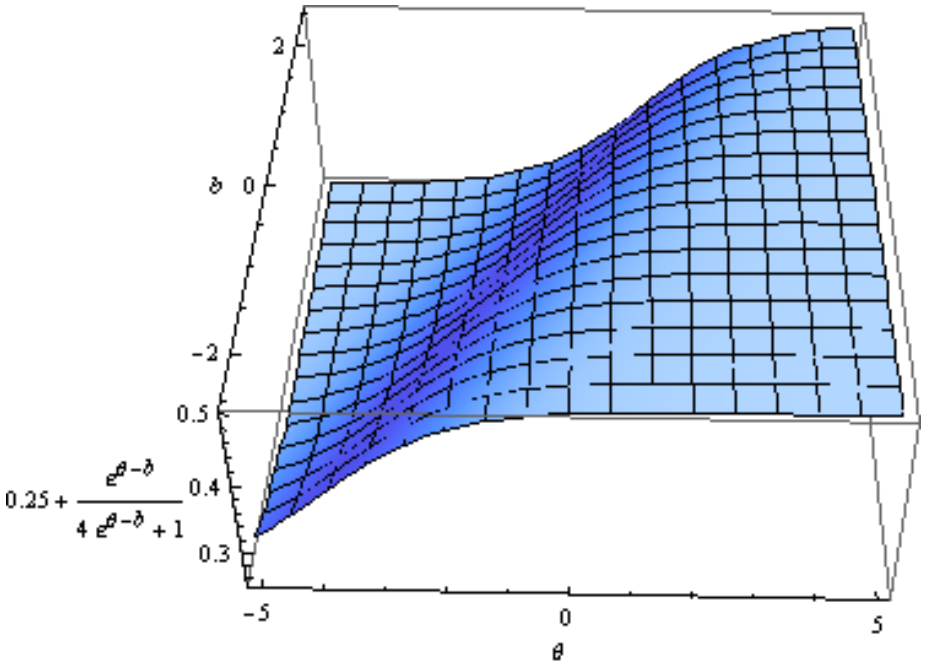

Figure 1. Guessing function 3D graph (M=4) for the Rasch-GuessingModel.

Property 2. The minimum of probability of guessing is equal to 1/M (M is the

number of options) and the maximum of probability of guessing can go up to 0.5 as the following equation and figure 2. (M=4):

$$\lim_{\theta \to \infty} g(\theta) = 0.5 \qquad (3.7)$$

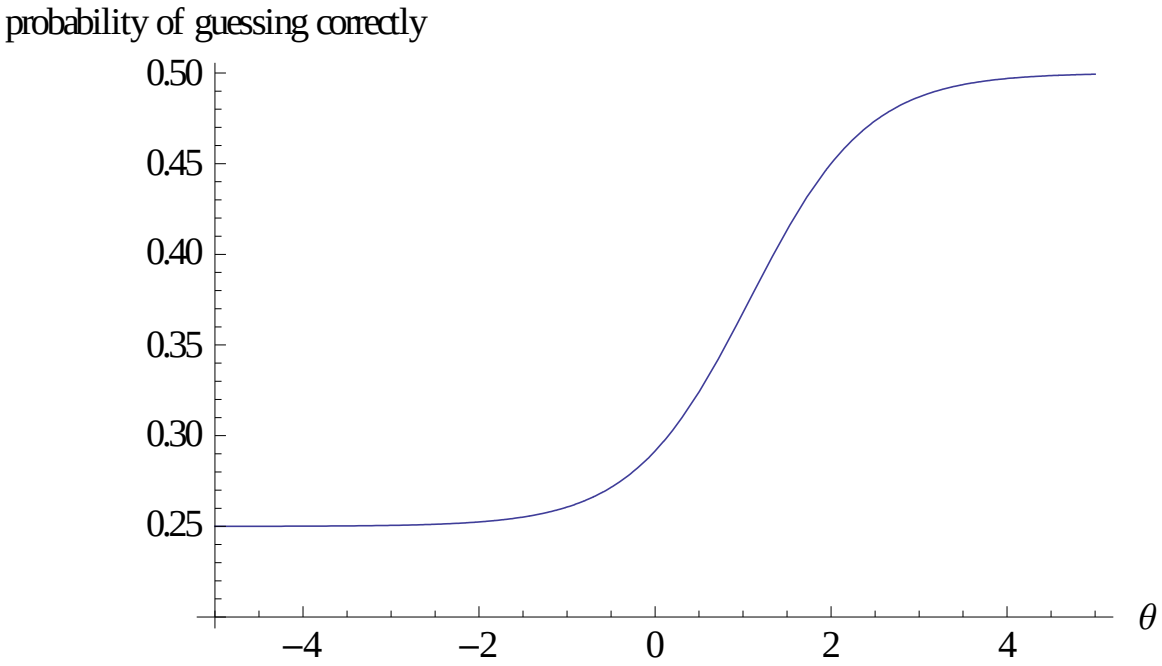

Figure 2. The guessing function for 2PL IRT model graph (a=1.5, b=2.0).

Comparison of Proposed Models with 1PL and 3PL model

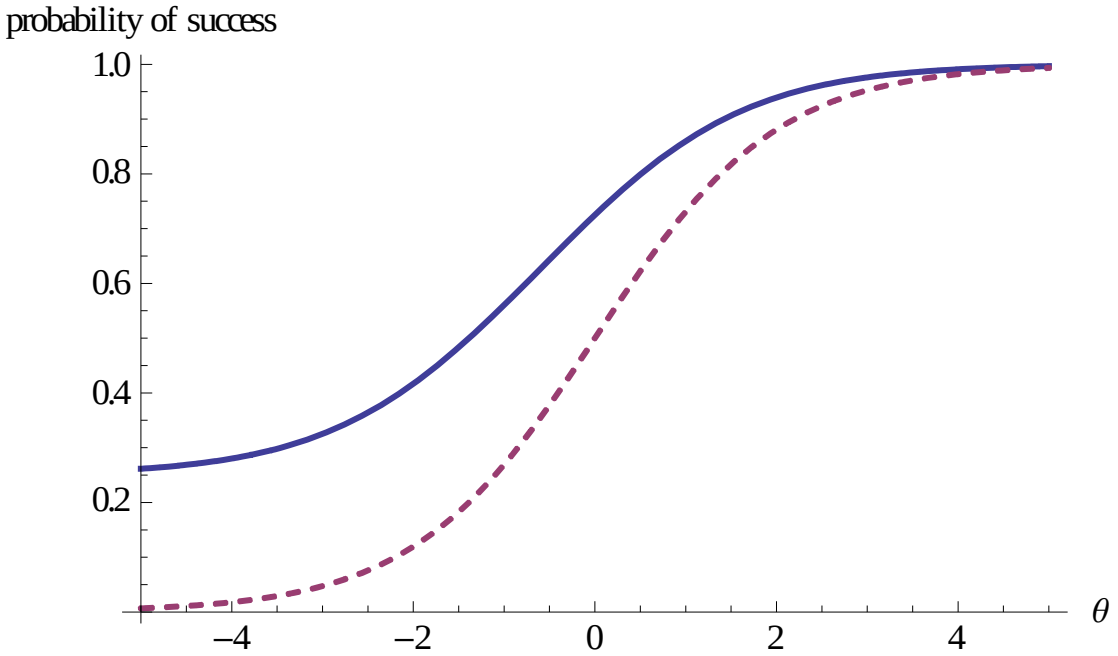

Figure 3. ICC for the Rasch model (dashed line) and the Rasch-Guessing model (continuous line).

For the Rasch model, the probability of success is always higher for the Rasch-Guessing than the Rasch model for any ability θ because of guessing function. This can be seen in Figure 3.

Compared with the 3PL model, the 2PL-Guessing model always has higher probability of success than the 3PL model, independent of θ, because the proposed model guessing can reflect that higher ability examinees have higher probability of successful guessing. Two ICC merge at two ends (extremely high ability and low ability), this indicates that when extremely low ability examinees guess, they make random guessing and the probability of success is equal to 1/*M*; while as extremely high ability examinees almost don't guess, so two ICC merge at high end. The maximum contribution of guessing to the success probability is to be found somewhere between two extremes of ability scale. For high ability examinees, the knowledge exclusively contributes to success, but for low ability examinees, the guessing does not help too much to success. This can be seen in Figure 4.

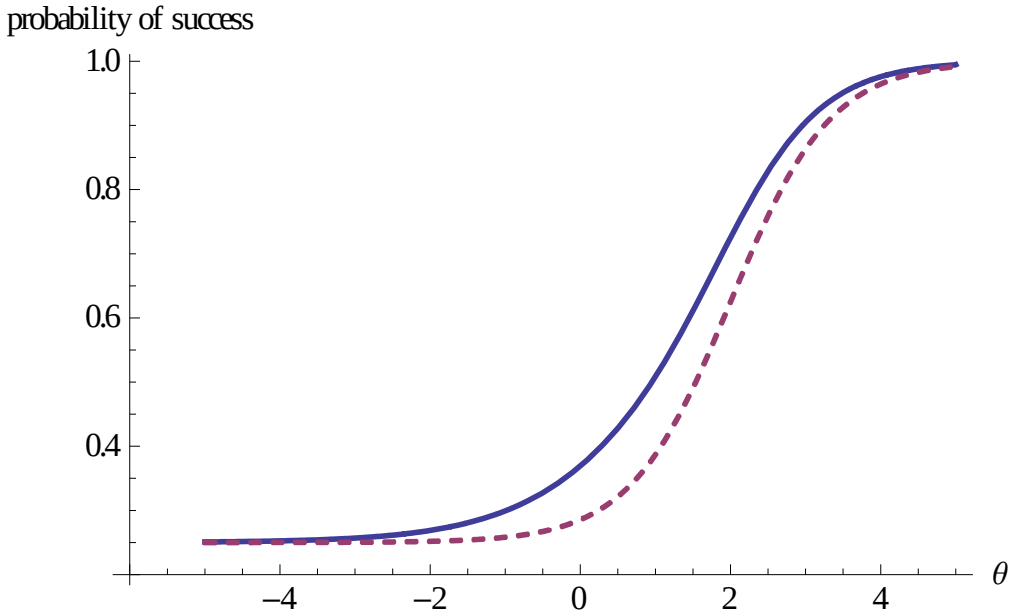

Figure 4. ICC of the 2PL-Guessing model (continuous line) and 3PL model (dashed line).

3.2 Generating IRT Parameters

In this study, In order to make the model more general to the real achievement tests, high-stake tests, or licensure tests (examinees will guessing questions if they do not know correct answers without being penalized). Item difficulty and discrimination parameters were generated on the basis of previous empirical studies and real-world test parameter ranges.

Ability and Difficulty Distribution

To avoid the deviation from the unidemensionality assumption, item difficulty parameter or person ability parameter distributions are expected to be standard normal. IRT programs like BILOG or SAS require person ability distribution to be standard normal (Misvey & Bock, 1990; Pelton, 2002).

Allen and Yen (1979) suggested that item difficulty between 0.3 and 0.7 can provide the maximum information to distinguish examinees in CTT. This difficulty range will be -0.52 to 0.52 if converted to normal standardized score. A high-stake test used to select graduate students for a university that admits only 10% of applicants should include extremely difficult items such as difficulty value is greater than 1.7 (or 0.05 in CTT).

In this study, in order to make the models more general to the real tests containing both easy items and extremely difficult items, the difficulty parameter values were focused on a range from -0.7 to +2.0 with normal distribution.

Discrimination Parameter Distribution

The Rasch model assumes equal item discrimination parameter while the 2PL and 3PL models assume discrimination parameter varies, so it is appropriate to assume that

discrimination parameter might be truncated normal. However, those items with negative discrimination are always removed from ability tests because if the probability of answering an item correctly decreases as examinee ability increases, there must be something wrong with the item. It is hardly to see the discrimination parameter is greater than 2, so the normal range for discrimination parameters is usually (0, 2) (Hambleton, Swaminathan & Rogers, 1991).

In this study, the discrimination parameter values were narrowed in a range of 0.4 to 2.0, because too low or too high discrimination values are either not practical or not stable to estimate. The discrimination parameters were generated from (0, 1) uniform distribution.

Pseudo-guessing Function

The guessing probability can go as high as 0.5 because some examinees can rule out some distracters from partial information (Kubinger & Gottschall, 2007). The probability of guessing an item correctly is determined by the examinee's ability and item parameters only. The guessing function should reflect the assumption that the higher ability persons have the higher probability of guessing the same item correctly. The highest probability of guessing cannot be greater than 0.5; otherwise, it would not be called knowledge-based answer instead of guessing. The relationship between probability of guessing correctly and the examinee's ability is logistic.

3.3 Data Simulation Design and Computer Program

A SAS program was used to generate simulation data for this study. The first step of the simulation used the random number seed RANNOR to generate item parameters for

20-item test, 30-item test, and 40-item test. These item parameters were treated as independent variables and parameter values were fixed for different lengths of tests. The second step of the program started with generating ability parameters which were normally distributed with mean equal to zero and standard deviation equal to 1. The probability of success for each examinee on each item was calculated in terms of proposed model functions and 3PL model with previously fixed item parameters and ability parameters. The calculated probability was compared with a random number drawn from uniform distribution (0, 1) produced by RANUNI to generate dichotomous response data sets. If the probability calculated was greater than the randomly drawn number from the uniform distribution, the response was assigned 1 as a correct answer; if the probability calculated was less than the randomly drawn number from the uniform distribution, the response was assigned 0 as an incorrect answer.

This study employed a design of one, two, and three item parameterization models (the Rasch-Guessing model, the 2PL-Guessing model, and the 3PL model) with normal ability distributions. 10 sets of dichotomous item responses of 1,000, 500, 200, and 100 subjects for 20 items , 1,000, 500, 300, and 200 subjects for 30 items, and 2,000, 1,000, 500, 300 subjects for 40 items were simulated using SAS computer program. Therefore, the total 3×12×10 different sets of dichotomous item responses were generated. These sets of item response data were used to estimate item parameters given the Rasch and the 2-PL IRT models with guessing function using SAS NLMIXED computer program described in the next section. The 36 combinations for two models are shown in Table 3.1 and 3.2.

Table 3.1

Data Simulation Design for the Rasch-Guessing Model

	The Number of Test Items											
	20				30				40			
	The Number of Examinees											
	100	200	500	1000	100	200	500	1000	100	200	500	1000
Rasch-Guessing												

Table 3.2

Data Simulation Design for the 2PL-Guessing Model and the 3PL Model

	The Number of Test Items			
	20			
	The Number of Examinees			
	100	200	500	1000
2PL-Guessing				
3PL				

3.4 Number of Replications in Monte Carlo Estimation

In IRT Monte Carlo research, the number of replications is driven by the purpose of research (Harwell, Stone, Hsu, & Kirisci, 1996). If a significance test for a parameter recovery study is necessary, at least 500 replications are needed. If the purpose of study is to compare different methodologies, a small number of replications such as 10 are sufficient. The ultimate purpose of this study is to compare the 2PL-Guessing model with the 3PL model on parameter recovery and goodness-of-fit for observed data, too many

replications are not necessary.

This study employed 10 replications for each combination of conditions based on suggestions from these Monte Carlo studies because of slow computer running time for SAS program. In the each of the 10 replications of data simulation, the same random seeds were used to generate the random normal distribution ability parameters for 1000, 500, 200, 100 examinees was kept constant and the 10 random seeds that was used to generate the item response data was changed in each replication so that 20 item response data were different but with the same sample of examinees (Si & Schumacker, 2004).

3.5 Criteria to Evaluate the Proposed Model

The parameter recovery comparison between the 2PL-Guessing model and the 3PL model was evaluated by three criteria. First, averaged estimated parameter values across 10 replications were correlated with true parameter values to determine how well the proposed models recovered those parameters. However, the correlation served as a relative indicator of accuracy because it only reflects the rank ordering of variables correlated (Harwell, Stone, Hsu, & Kirisci, 1996). The higher the correlation is; the better the parameter recovery will be.

Second, a root mean squared deviation (RMSD) of parameter estimate was calculated across 10 replications for each of the study design. The RMSD indicated the variance of parameter estimate across replications, and thus serves as an indicator of accuracy; the smaller the RMSD is, the more accurate the estimate will be. The RMSD was calculated by the following formula:

$$RMSD = \sqrt{\sum_{j=1}^{n} \frac{(\hat{a}_j - a_j)^2}{n}}, \quad (3.8)$$

$$RMSD = \sqrt{\sum_{j=1}^{n} \frac{(\hat{b}_j - b_j)^2}{n}}, \quad (3.9)$$

where n = number of replications

a_j = the true discrimination parameter value of the jth item

$\hat{a}_j$ = the discrimination parameter estimates of jth item from n replications

b_j = the true difficulty parameter value of the jth item

$\hat{b}_j$ = the discrimination parameter estimates of jth item from n replications

Third, estimate bias is the mean difference between the estimated and true parameter value for an item across all replications. The smaller bias differences are, the closer the estimates are to the true parameter values. Positive bias indicates overestimation and negative bias indicates underestimation (Dawber, Roger, & Carbonaro, 2004). Bias for a_j and b_j can be calculated in this study by the following:

$$Bias\ a_j = \bar{\hat{a}}_j - a_j, \text{ where } \bar{\hat{a}}_j = \sum_{r=1}^{10} a_{jr} / 10$$

$$Bias\ b_j = \bar{\hat{b}}_j - b_j, \text{ where } \bar{\hat{b}}_j = \sum_{r=1}^{10} (b_{jr}) / 10$$

3.6 Test Length

In psychological and educational assessments, the short (20 items) and moderate (40 items) exam lengths are most frequently used (Dawber, Roger, & Carbonaro, 2004; Seong, 1990; Yen, 1987); therefore, three test lengths were employed in this study: two short exams of 20 and 30 items, a moderate exam of 40 items. The number of alternatives in each item was set to four in this study. Only 20 item tests were simulated to compare the 2PL-Guessing model with the 3PL model.

3.7 Parameter Estimation Methods

MML estimations were used to estimate item difficulty and discrimination parameters for the Rasch-Guessing and the 2PL-Guessing models. Under the MMLE approach to item parameter estimates, the examinees are treated as a random sample which is drawn from a population with ability distributed on a density function and items are treated as a fixed effect and abilities as a random effect (Baker & Kim, 2004). The most important part for MMLE is the integration over the ability distribution and the ability parameter can be removed from the likelihood function, so item parameter estimates are independent of each examinee's ability, thus producing more reliable item parameter estimates.

According to Bock and Lieberman's (1970) solution, let item response vector= $\mathbf{Y}_j$ conditional on the examinee's ability θ_j and the item parameters in ξ, $g(\theta_j \mid \boldsymbol{\tau})$ is the probability density function of ability in the population of examinees with parameter vector $\boldsymbol{\tau}$, and $P(\mathbf{Y}_j) = \int P(\mathbf{Y}_j \mid \theta_j, \xi) g(\theta_j \mid \boldsymbol{\tau}) d\theta_j$ (Baker & Kim, 2004). Because the integration is across the ability distribution, this expression is the marginal probability of item response vector $\mathbf{Y}_j$ in terms of the item parameters and the population ability density. The marginal likelihood function is

$$L = \prod_{j=1}^{n} P(\mathbf{Y}_j), \tag{3.8}$$

so, the logarithm of L is

$$\log L = \sum_{j=1}^{N} \log P(\mathbf{Y}_j), \tag{3.9}$$

and, to find the marginal likelihood equation for the *ith* item a_i take

$$\frac{\partial}{\partial a_i}(\log L) = 0,$$

then, the marginal likelihood function for discrimination parameter can be written as the following (the detailed procedures for the deduction of marginal maximum likelihood function see Appendix A):

$$\frac{\partial}{\partial a_i}(\log L) = \sum_{j=1}^{N} \int [\frac{y_{ij} - P_i(\theta_j)}{P_i(\theta_j)Q_i(\theta_j)}][\frac{\partial P_i(\theta_j)}{\partial a_i}][P(\theta_j \mid \mathbf{Y}_j, \boldsymbol{\xi}, \boldsymbol{\tau})]d\theta_j,$$

where $Q_i(\theta_j)$ is the probability of incorrect answer and is equal to 1-$P_i(\theta_j)$ which is the probability of correct answer to the *ith* item at the ability level of θ_j. y_{ij} is the *jth* examinee's response to the *ith* item and is equal to1 for correct response or 0 for incorrect response. The $P(\theta_j|\mathbf{Y}_j,\boldsymbol{\xi},\boldsymbol{\tau})$ is the probability of an examinee having ability θ_j along the conditional on the item response vector Y_j, the item parameter in $\boldsymbol{\xi}$, and the population distribution of ability $\boldsymbol{\tau}$. It is also called the posterior ability distribution.

The marginal likelihood equation for discrimination parameter a_i is:

$$\frac{\partial}{\partial a_i}(\log L) = \sum_{j=1}^{N} \int [\frac{y_{ij} - P_i(\theta_j)}{P_i(\theta_j)Q_i(\theta_j)}][\frac{\partial P_i(\theta_j)}{\partial a_i}][P(\theta_j \mid \mathbf{Y}_j, \boldsymbol{\xi}, \boldsymbol{\tau})]d\theta_j \qquad (3.10)$$

The likelihood equation for difficulty parameter b_i is :

$$\frac{\partial}{\partial b_i}(\log L) = \sum_{j=1}^{N} \int [\frac{y_{ij} - P_i(\theta_j)}{P_i(\theta_j)Q_i(\theta_j)}][\frac{\partial P_i(\theta_j)}{\partial b_i}][P(\theta_j \mid \mathbf{Y}_j, \boldsymbol{\xi}, \boldsymbol{\tau})]d\theta_j \qquad (3.11)$$

3.8 Item Information Functions

Birnbaum (1968) has defined the test function as

$$I(\theta) = \sum_{i=1}^{n} \frac{[P_i^{'}(\theta)]^2}{P_i(\theta)Q_i(\theta)}, \qquad (3.12)$$

where $P_i(\theta)$ is obtained by the ICC model function at θ and $P_i^{'}(\theta) = \frac{\partial P_i}{\partial \theta}$. The right side of equation 3.12 can be decomposed into the contribution of each item to the entire test information, so the amount of information each item contributes to the test information is given by

$$I_i(\theta) = \frac{[P_i^{'}(\theta)]^2}{P_i(\theta)Q_i(\theta)}. \qquad (3.13)$$

Inspection of equation (3.13) indicates that the test information is simply the sum of the amount of each item information at the ability level of interest. Figure 5 shows the item information functions for three items in which continuous line represents item 1 with $a = 0.5$ and $b =$ -1, dashed line represents item 2 with $a = 1$ and $b = 0$, and dotted line represents item 3 with $a = 1.5$ and $b = 1.0$.

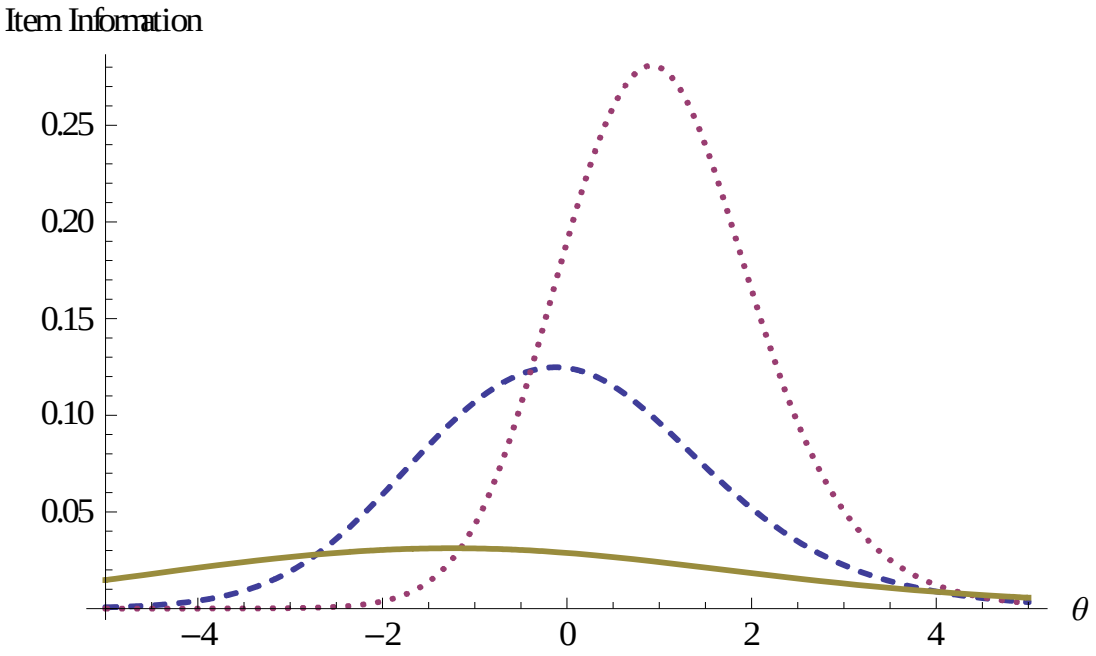

Figure 5. Item Information Functions for three items

Figure 5 highlights several important points: (a) the maximum information provided

by an item is at its difficult level equal to ability level, (b) the higher discrimination parameter is, the more information an item will provide, and (c) an item with low discrimination power is almost useless statistically in a test.

3.9 Item Parameter Estimation Computer Program

The SAS PROC NLMIXED was used to estimate item parameters on the basis of generated dichotomous response data sets. PROC NLMIXED fits nonlinear mixed models by maximizing an approximation to the likelihood integrated over the random effect. PROC NLMIXED enables you to specify a conditional distribution for your data (given the random effects) having either a standard form (normal, binomial, Poisson) or a general distribution that you code using SAS programming statements. Fixed effects were item parameters and random effect was ability in this study.

SAS PROC NLMIXED uses Gaussian quadrature to do the integral approximation and uses dual quasi-Newton algorithm as the optimization method to implement maximization. Pinheiro and Bates (1995) proved that adaptive Gaussian quadrature is the best method after they compared several different integrated likelihood approximations. Successful convergence of the optimization problem results in parameter estimates and their approximate standard errors based on the second derivative matrix of likelihood function (SAS/STAT, 2008).

CHAPTER FOUR

RESULTS

The purpose of this study was to compare the accuracy of parameter estimates for the 2PL-Guessing model with the 3PL IRT model and investigate how well the Rasch-Guessing model can recover parameters. Three criteria were used to determine how well the new models' item parameters were recovered, correlation between true parameter values and estimated parameter mean values across 10 replications for four different sample sizes, the root mean squared deviation (RMSD) and bias. The higher the correlation is and the smaller the RMSD is, the more accurate the estimate will be. Correlations and RMSD for parameter estimates were tabulated for each study design.

The Rasch-Guessing Model Parameter Recovery Results

Item difficulty parameter estimates for the Rasch-Guessing model were run via SAS PROC NLMIXED under sample sizes of 100, 200, 500, and 1000 and test length of 20, 30 and 40 items with 10 replications. Estimated difficulty parameter values for each replication were saved and then the mean of each item difficulty parameter estimated was calculated across 10 replications. The calculated means were used to correlate with their corresponding parameter true values. The results of these correlations were given in the Table 4.1. As shown in the Table 4.1, the highest correlation was 0.999 for 20 and 40 items with 1000 subjects and the lowest was 0.975 for 20 items with 100 subjects. As sample size increased, the correlation increased too; however, the number of test items had little effect on those correlations because as the test length increased from 20 to 40, the correlations didn't change too much.

Table 4.1.

Correlations for Difficulty Parameter Recovery for the Rasch-Guessing Model

No of Items	20	30	40
	Sample size (n=100)		
$r_{b\hat{b}}$	0.975	0.993	0.985
	Sample size (n=200)		
$r_{b\hat{b}}$	0.997	0.996	0.992
	Sample size (n=500)		
$r_{b\hat{b}}$	0.998	0.997	0.997
	Sample size (n=1000)		
$r_{b\hat{b}}$	0.999	0.998	0.999

Table 4.2

RMSD and Bias for The Rasch-Guessing Model Difficulty Parameter Estimates

		RMSD			Bias	
No of Items	Sample Size	Maximum	Minimum	Average	Maximum	Minimum
20	100	0.662	0.241	0.373	0.445	0.032
	200	0.371	0.116	0.232	0.165	0.004
	500	0.242	0.101	0.156	0.14	0.004
	1000	0.165	0.076	0.118	0.074	0
30	100	0.601	0.168	0.320	0.258	0.002
	200	0.337	0.123	0.233	0.251	0.005
	500	0.222	0.103	0.155	0.126	0.012
	1000	0.133	0.062	0.104	0.099	0
40	100	0.582	0.201	0.380	0.247	0
	200	0.346	0.172	0.261	0.157	0.003
	500	0.216	0.089	0.153	0.113	0.002
	1000	0.145	0.058	0.108	0.108	0

As shown in the Table 4.2, RMSD decreases as the sample size increases from 100 samples to 1000 samples for the same number of test items. The maximum RMSD was 0.662 for the estimation of 100 sample size of 20 items; the minimum is 0.058 for the estimation of 1000 sample size of 40 items. However, estimations for 30 items have the best average RMSD compared with estimations for 20 or 40 items. The estimated mean and calculated RMSD values for each item under different sample size are listed in the Appendix table…

How well the parameter estimates under different sample sizes for the Rasch-Guessing model is also illustrated by the following graphs.

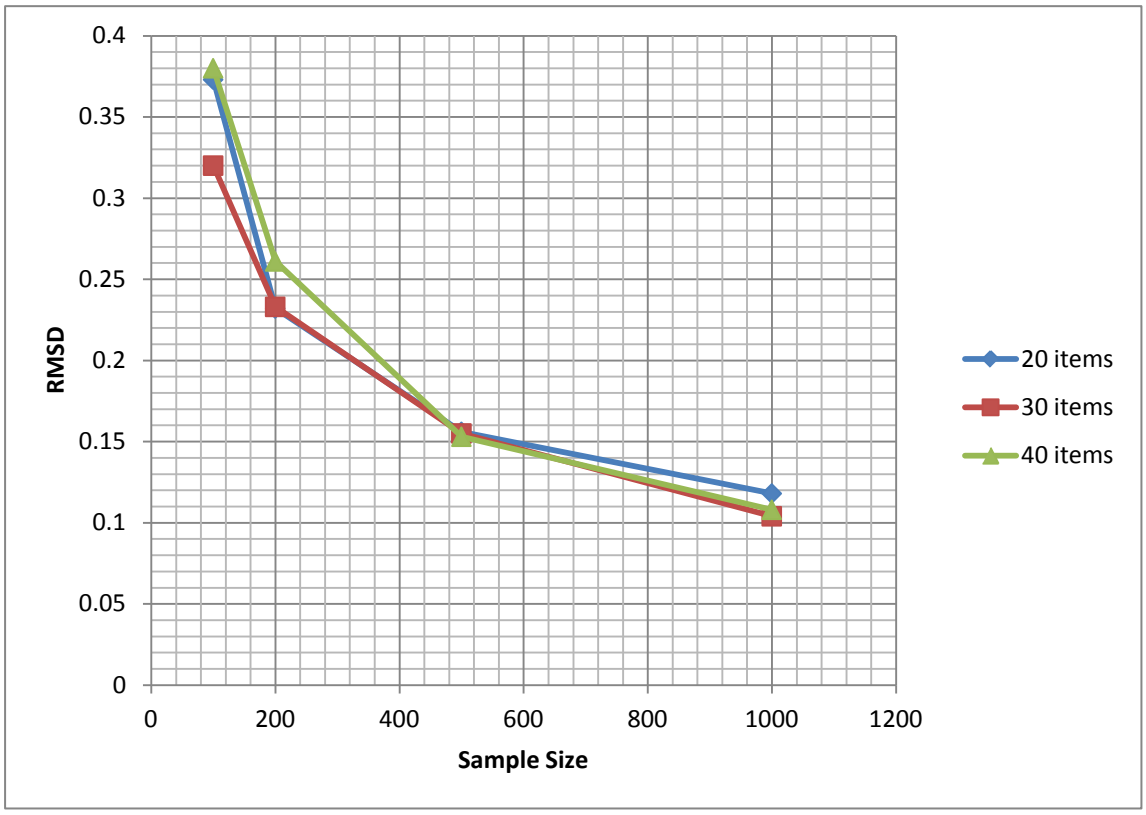

Figure.6. Average RMSD for the Rasch-Guessing model

The Comparison Between the 3PL Model and the 2PL-Guessing Model

Two designs were used to compare parameter estimates between the proposed model and 3PL model. In the first design, only the proposed model was used to generate a 20-item test for samples of 100, 200, 500, and 1000, and then parameters were estimated via the proposed model and the 3PL model. In the second design, the traditional 3PL model was used to generate a 20-item test for samples of 100, 200, 500, and 1000, and then parameters were estimated by both models.

Three criteria were used to compare the traditional 3PL model with the newly proposed 2PL-Guessing model: correlations between true parameter values and estimated values, root mean squared deviation (RMSD), and bias. Only correlations were used to compare latent ability estimates.

Ability Parameter Estimate Results

The ability parameter estimate results are shown in the Table 4.3.

Table 4.3

Correlations Between Estimated Ability Values and True Ability Values

Sample size	Estimation method			
	2PLG-2PLG	2PLG-3PL	3PL-3PL	3PL-2PLG
100	0.808	0.805	0.805	0.820
200	0.824	0.817	0.795	0.805
500	0.829	0.821	0.798	0.805
1000	0.830	0.825	0.807	0.814

Note: 2PLG-2PLG represents the estimation that the data was generated by the 2PL-Guessing model and estimated by the 2PL-Guessing model; 2PLG-3PL represents the estimation that the data was generated by the 2PL-Guessing model but estimated by the 3PL model; 3PL-3PL represents the estimation that the data

was generated by the 3PL model and estimated by the 3PL model; 3PL-2PLG represents the estimation that the data was generated by the 3PLmodel but estimated by the 2PL-Guessing model.

The correlations between true ability values and estimated values for 20 item simulated test were generally around 0.8. The highest correlation was 0.830 and the lowest was 0.795. The highest correlation occurred when the data was generated by the 2PL-Guessing model and ability parameters were estimated by the 2PL-Guessing model with 1000 sample size. The lowest correlation occurred when the data was generated by 3PL model and ability parameters were estimated by the 3PL model with 200 sample size.

As shown in the Table 4.3, regardless of sample size, all correlations calculated were higher when the data were generated by the 2PL-Guessing model and ability parameters were estimated by the 2PL-Guessing model than those calculated when the data were generated by 2PL-Guessing model, but ability parameters were estimated by the 3PL model for corresponding sample size. Furthermore, when the 3PL model was used to simulate data and the 2PL-guessing model was used to estimate latent ability, their correlations for different sample sizes were higher than correlations calculated when the 3PL model was used to simulate data and latent ability parameters were estimated by the 3PL model.

Item Parameter Estimate Results

Item Parameter Correlations

Because we wanted to compare the accuracy of parameter estimates for the proposed model with the 3PL model under the same condition, the data was simulated by one of the two models and parameters were estimated by both two models. The correlation and RMSD were calculated for each item parameter. The Table 4.4 and 4.5 show the

correlation results for difficulty and discrimination parameter estimates (2PLG-2PLG means the data was generated by the 2PL-Guessing model and parameters were estimated by the 2PL-Guessing models too; 2PL-3PL means the data was generated by the 2PL-Guessing model , but parameters were estimated by the 3PL model. Similar explanation goes to 3PL-3PL and 3PL-2PLG).

Table 4.4

Correlations for Difficulty Parameter Estimates

Sample size	Estimation method			
	2PLG-2PLG	2PLG-3PL	3PL-3PL	3PL-2PLG
100	0.969	0.949	0.964	0.921
200	0.993	0.964	0.982	0.938
500	0.995	0.968	0.991	0.950
1000	0.998	0.943	0.981	0.958

Note: 2PLG-2PLG represents the estimation that the data was generated by the 2PL-Guessing model and estimated by the 2PL-Guessing model; 2PLG-3PL represents that the data was generated by the 2PL-Guessing model but estimated by the 3PL model; 3PL-3PL represents that the data was generated by the 3PL model and estimated by the 3PL model; 3PL-2PLG represents that the data was generated by the 3PLmodel but estimated by the 2PL-Guessing model.

The highest correlation (0.998) for difficulty parameter estimates went to the 2PLG-2PLG estimation for 1000 sample size and the lowest correlation (0.921) went to the 3PL-2PLG estimation for 100 sample size. All correlations for the 2PLG-2PLG difficulty parameter estimates were greater than those of the 2PLG-3PL estimation. All correlations for the 3PL-3PL difficulty parameter estimates were greater than those for the 3PL-2PLG estimates.

Table 4.5

Correlations for Discrimination Parameter Estimates

Sample size	Estimation method			
	2PLG-2PLG	2PLG-3PL	3PL-3PL	3PL-2PLG
100	0.890	0.787	0.782	0.856
200	0.957	0.924	0.819	0.921
500	0.980	0.909	0.842	0.972
1000	0.994	0.941	0.881	0.988

The highest correlation (0.994) for discrimination parameter estimates went to the 2PLG-2PLG estimation under the sample size of 1000 and the lowest correlation (0.782) went to the 3PL-3PL estimation under the sample size of 100. Among all the estimation methods, the 2PLG-2PLG produced the highest correlations. Even though the data were generated by the 3PL model, the discrimination parameters were estimated better by the new proposed model than by the conventional 3PL model. Generally, the correlations for both difficulty and discrimination parameter estimation tended to be enhanced as the sample size was increased.

Item Parameter Estimate RMSD

RMSD is the best indicator for the accuracy of parameter calibration. The smaller the RMSD is, the more accurate the estimates will be. The average RMSD results for item difficulty and discrimination parameter estimates with different estimation methods are presented in Table 4.6 and 4.7. The detailed RMSDs for each item is presented in the Appendix C.

Table 4.6

Average RMSD for Difficulty Parameter Estimates

Sample size	Estimation Method			
	2PLG-2PLG	2PLG-3PL	3PL-3PL	3PL-2PLG
100	0.405	0.784	0.505	0.722
200	0.273	0.826	0.462	0.666
500	0.206	0.827	0.396	0.606
1000	0.155	0.777	0 .360	0.584

Table 4.7

Average RMSD for Discrimination Parameter Estimates

Sample size	Estimation Method			
	2PLG-2PLG	2PLG-3PL	3PL-3PL	3PL-2PLG
100	0.436	0.578	0.506	0.483
200	0.322	0.477	0.477	0.391
500	0.233	0.402	0.376	0.270
1000	0.175	0.323	0.306	0.207

The 2PLG-2PLG method estimated the difficulty parameter most accurately compared with other methods as shown in Table 4.6. The least accuracy of the difficulty parameter estimates (0.827) went to the 2PLG-3PL situation for 500 sample size. The most accurate difficulty parameter estimates (0.155) went to the 2PLG-2PLG situation for 1000 sample size. When the 3PL model was used to estimate item difficulty parameter using the 2PL-Guessing model generated data, the RMSD would increase tremendously. The maximum increase was almost 400% for 1000 sample size (from 0.155 to 0.777).

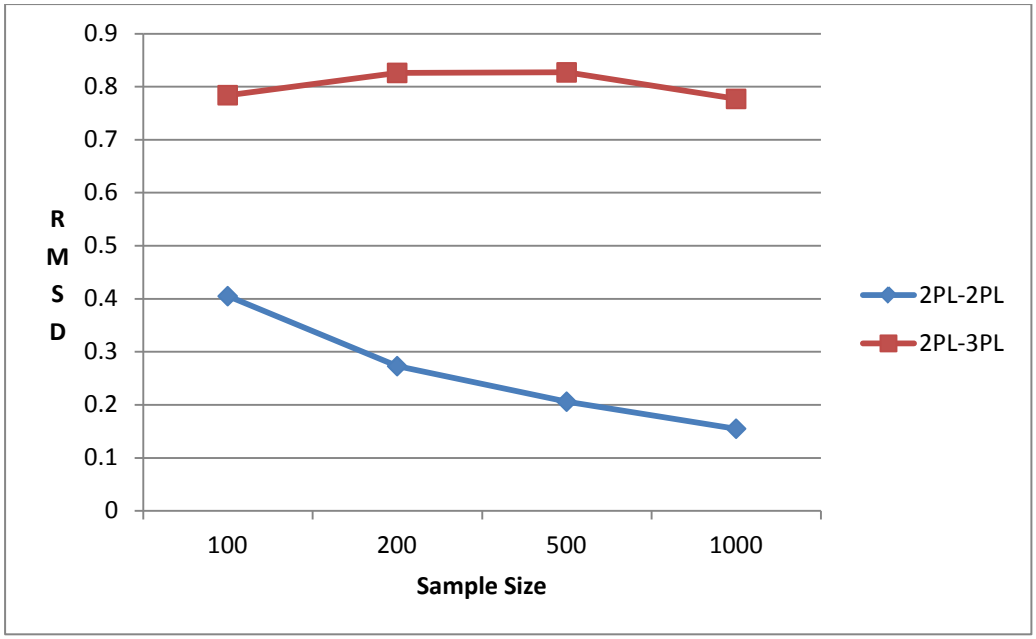

Figure 7. the 2PLG-2PLG and the 2PLG-3PL Difficulty Parameter Estimate RMSD Graph

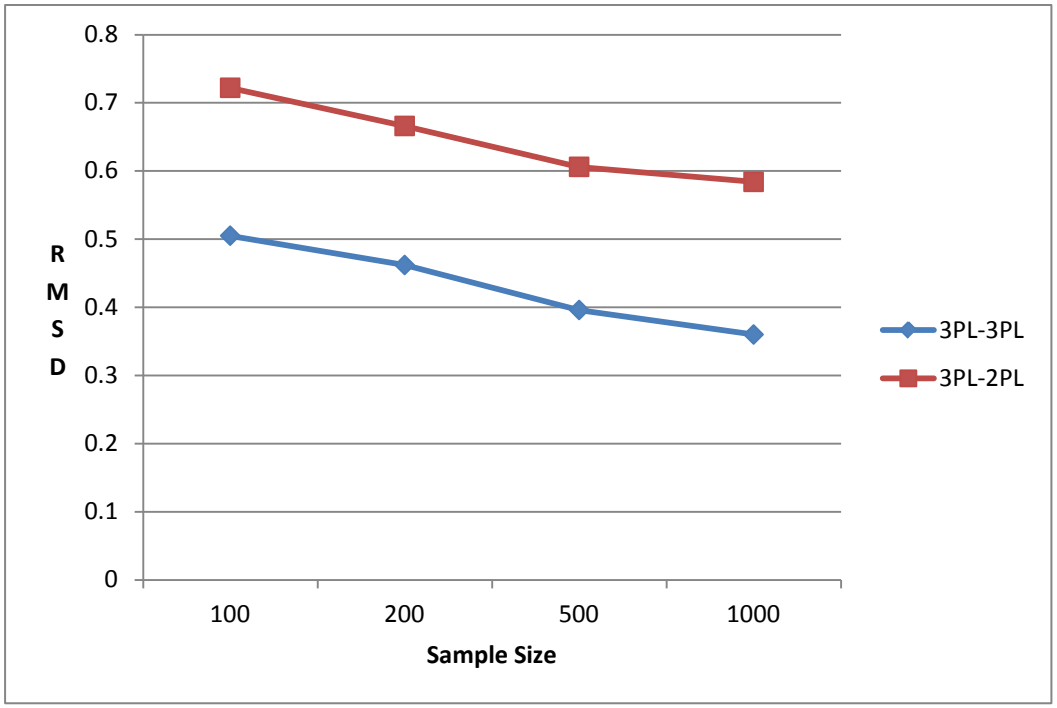

Figure 8. the 3PL-3PL and the 3PL-2PLG Difficulty Parameter Estimate RMSD Graph

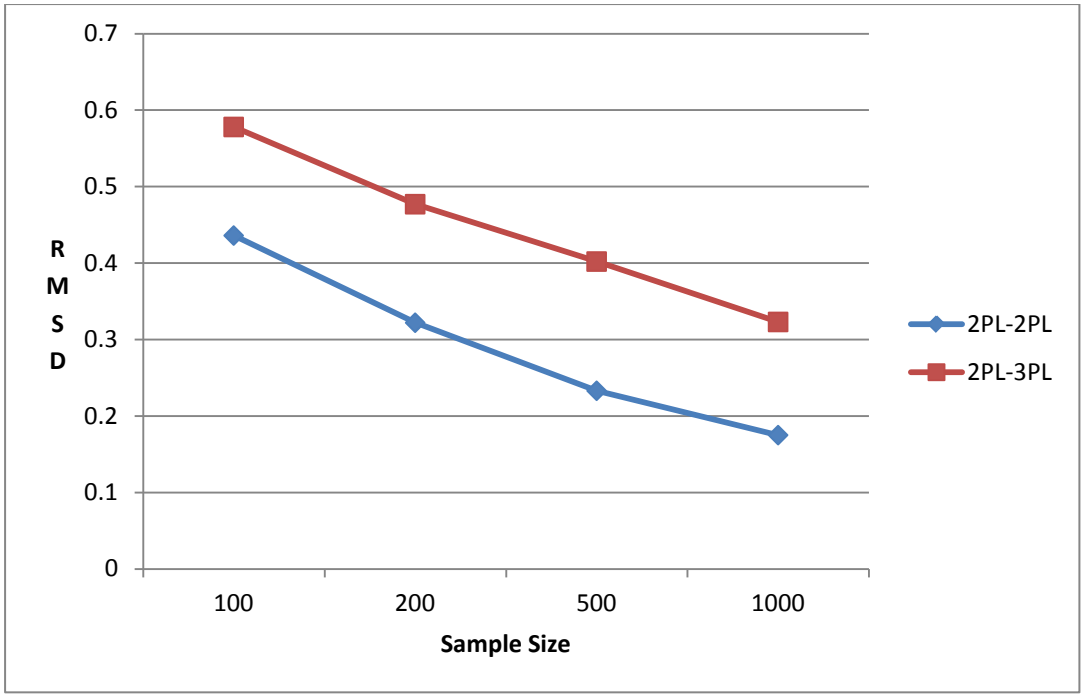

Figure 9. the 2PLG-2PLG and the 2PLG-3PL Discrimination Parameter Estimate RMSD Graph

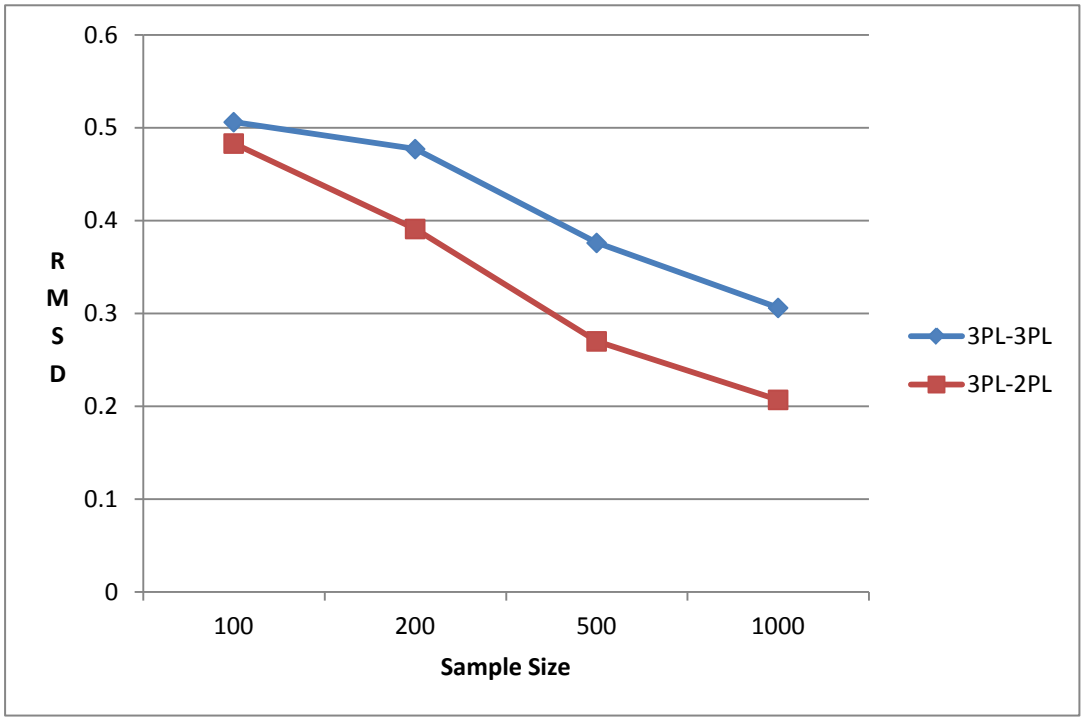

Figure 10. the 3PL-3PL and the 3PL-2PLG Discrimination Parameter Estimate RMSD Graph

When the 2PL-Guessing model was used to estimate item difficulty parameter using the data generated by the 3PL model, the RMSD would increase about 40% to 50% for all sample sizes. Figure 6 and 7 demonstrated that the average RMSD gap between the 2PLG-2PLG estimation and the 2PLG-3PL estimation for all sample sizes was much larger than the average gap between the 3PL-3PL estimation and the 3PL-2PLG estimation for all sample sizes.

The discrimination parameter was estimated most accurately by the 2PLG-2PLG estimation method for all sample sizes. The discrimination parameter was estimated most accurately with the 2PLG-2PLG estimation method for 1000 sample size (RMSD=0.175) and was estimated the least accurately with the 2PLG-3PL estimation situation for 100 sample size (RMSD=0.578). When the 3PL model was used to estimate discrimination parameter using the data generated by the 2PL-Guessing model, average RMSDs for all sample sizes were increased from 30% to 85% compared with average RMSDs estimated by the 2PL-Guessing model. When the 2PL-Guessing model was used to estimate discrimination parameter using the data generated by the 3PL model, average RMSDs for all sample sizes were decreased from 5% to 30% compared with average RMSDs estimated by the 3PL model. This can also be seen in Figure 8 and Figure 9.

Item Parameter Estimate Bias

The bias was calculated by the following:

the mean of estimated each item parameter values – the item parameter true value. Because some biases were positive and some were negative, we did not calculate the mean of the bias. The positive bias indicates an overestimated parameter and the negative bias presents an underestimated parameter. The zero mean of biases does not mean there is no bias. The absolute values of maximum and minimum biases for different sample sizes are illustrated in Table 4.8 and 4.9.

Table 4.8

Maximum and Minimum Bias for Difficulty Parameter Estimates

Bias Values	Sample Size	Estimation methods			
		2PLG-2PLG	2PLG-3PL	3PL-3PL	3PL-2PLG
Maximum	100	0.507	-0.866	-0.430	1.460
	200	0.277	-1.160	-0.458	1.332
	500	0.15	-1.026	-0.280	1.248
	1000	-0.103	-1.238	-0.571	1.159
Minimum	100	0.005	0.058	0.023	0.232
	200	-0.006	-0.220	-0.028	0.340
	500	0.001	-0.227	-0.005	0.292
	1000	0.001	-0.169	-0.009	0.299

The smallest maximum and minimum biases for difficulty parameter estimates went to the 2PLG-2PLG estimation and the biggest maximum and minimum biases for difficulty parameter estimates went to the 3PL-2PLG estimation. The smallest bias (0.001) went to the 2PLG-2PLG estimation for 1000 sample size and the highest bias was 1.460 in the 3PL-2PLG estimation for 100 sample size. The 2PLG-3PL difficulty parameter estimates had much greater biases than the 2PLG-2PLG difficulty parameter estimates and the 3PL-2PLG difficulty parameter estimates presented much higher biases then the 3PL-3PL difficulty parameter estimates.

Table 4.9

Maximum and Minimum Bias for Discrimination Parameter Estimate

Bias Values	Sample Size	Estimation methods			
		2PLG-2PLG	2PLG-3PL	3PL-3PL	3PL-2PLG
Maximum	100	0.537	0.579	-0.635	0.459
	200	0.218	0.378	-0.772	0.481
	500	0.219	-0.312	-0.754	0.320
	1000	0.096	-0.397	-0.717	0.256
Minimum	100	0.002	-0.001	0.002	0.001
	200	-0.001	0.009	0.000	-0.015
	500	0.002	0.023	-0.006	0.003
	1000	0.000	-0.023	0.003	0.001

The 2PLG-2PLG estimation had the smallest bias in estimating discrimination parameter compared with all other estimation methods, while the 3PL-3PL estimation had the largest maximum discrimination bias across all estimation methods. The 3PL-2PLG method had smaller bias in estimating discrimination parameter than the 3PL-3PL method. The smallest bias went to the 2PLG-2PLG estimation for 1000 sample size and the 3PL-3PL estimation for 200 sample size. The 2PLG-3PL discrimination parameter estimates had greater biases than the 2PLG-2PLG discrimination parameter estimates and the 3PL-2PLG discrimination parameter estimates presented smaller biases then the 3PL-3PL discrimination parameter estimates. Each item parameter estimate bias is presented in the Appendix D.

Goodness of Fit Index Results

Table 4.10

Average AIC for Goodness-of-Fit

Sample size	Estimation method					
	2PLG-2PLG	2PLG-3PL	Δ_1	3PL-3PL	3PL-2PLG	Δ_2
100	2508.9	2537.7	28.8	2659.5	2630.6	28.9
200	5018.9	5044.6	25.7	5295.1	5267.5	27.6
500	12464	12487	23	12977	12952	25
1000	24906	24925	19	25933	25918	15

As shown in the above table, the 2PLG-2PLG estimation presented the smallest average AIC for the same sample size design, while the 3PL-3PL estimation had the biggest AIC index. It is very important to point out that even though the 2PL-Guessing model was used to run the 3PL model-generated data, the average AIC for the 2PL-Guessing model was still smaller than the average AIC for the 3PL model. However, the

difference (Δ_1 and Δ_2) between two AICs decreased as the sample size increased, while Δ_1 is equal to the average AIC for the 2PLG-3PL estimation minus the average AIC for the 2PLG-2PLG estimation and Δ_2 is equal to the average AIC for the 3PL-3PL estimation minus the average AIC for the 3PL-2PLG estimation.

CHAPTER FIVE

CONCLUSIONS

Stage (2003) investigated whether the conventional 3PL model would be applicable to the Swedish Scholastic Aptitude Test (SweSAT) which is a norm-referenced and high-stake multiple choice test and Stage concluded that the 3PL model did not fit the SweSAT data even though guessing existed. Stage's study presented a big challenge to the traditional 3PL model when handling guessing. Simply assuming that every examinee has the same probability of guessing an item correctly is not appropriate for all kinds of tests. The new model in this study was developed to solve this problem.

The primary purpose of this study is to compare the accuracy of parameter estimates via the new model with the conventional 3PL model under different situations (or designs) through the Monte Carlo method. Three criteria were used to compare how the proposed model estimated parameters more accurately than the 3PL model: correlation, RMSD, and bias. In this section, a few advantages of the new model compared with the 3PL model will be discussed.

Ability Parameter Estimate Comparison

The newly proposed model estimated ability parameter more accurately than the traditional 3PL model. Two Monte Carlo study designs were created to prove this. In the first design, the data were simulated using the new model and the ability parameters were estimated by both the new and the 3PL model. In the second design, the data was simulated by the 3PL model and the ability parameters were estimated by both the new and the 3PL model. The average correlations between true ability values and estimated ability values for each replication were used as the criterion to compare two models.

In both designs, regardless the data was generated by the new model or the 3PL

model, the proposed model produced higher correlations for all sample sizes than the 3PL model indicating that even if the real guessing situation fits the 3PL model (assuming if the examinee did not know the correct answer, he/she would guess randomly), the proposed model can estimate ability parameter more accurately than the 3PL model because the new model also takes random guessing into consideration and it can be more universally applied to multiple choice tests. If the ability estimate is the most important for those who are more interested in placement, admission, or selection, the new model can provide more accurate information than the traditional 3PL model.

Item Parameter Estimate Comparison

Three criteria were adopted to compare the accuracy of item parameter estimates for two models, correlation, root mean standard deviation, and bias. The same study design for ability estimate was used to generate the data and estimate item parameters. The means were calculated for estimated difficulty and discrimination parameters with 10 replications for sample size of 100, 200, 500 to 1000, and then those means were correlated with corresponding true values to get correlation coefficients, the equation 3.8 and 3.9 were used to calculate RMSDs, and the bias was the difference between the mean of estimated parameter and the true value.

Among all four estimation methods (2PLG-2PLG, 2PLG-3PL, 3PL-3PL, and 3PL-2PLG), the 2PLG-2PLG had the most accurate estimate for item parameters because it had the highest correlations, the smallest RMSDs, and the lowest biases, indicating that if the guessing situation is close to the assumption that examinees of different ability level have different probability of successful guessing, the proposed model will be most accurate to estimate item parameters than the 3PL model.

If we used the 3PL-model to estimate these item difficulty parameter, RMSD would increase tremendously. For 100 sample size, it would increase almost 100% (from 0.405

to 0.784), and for 1000 sample size, it would increase more than 400% (from 0.155 to 0.777). However, the 3PL model estimated the item difficulty parameter better than the new model when the data was generated by the 3PL model or in the situation of random guessing. For example, the average RMSDs for 100 and 1000 sample size were 0.505 and 0.360 respectively via the 3PL model estimation, but the average RMSDs were 0.722 and 0.584 respectively via the new model estimation. Therefore, if the guessing situation is close to random guessing and the difficulty parameter estimation is more important than any other purposes, the 3PL model should be adopted to estimate item parameters.

The new model, nonetheless, estimated the discrimination parameter more accurately for all sample sizes than the 3PL model even though the data was generated by the 3PL model. For example, the average RMSDs estimated by the 3PL model for sample sizes of 100 and 1000 were 0.506 and 0.306 respectively, however, the average RMSDs estimated by the new model for sample sizes of 100 and 1000 were 0.483 and 0.207 respectively. This indicates that even in a random guessing situation test, the new model is still better in estimating item discrimination parameter and this can also prove the huge advantage of the new model compared with the 3PL model. This is probably the main reason why the new model can estimate latent ability more accurately than the 3PL model even in random guessing situation.

Goodness of Fit Index AIC Comparison

The fit of the model to the data is very important in item response theory. Akaike's information criterion (AIC) was adopted to compare the goodness of fit between the 2PL-Guessing model and the 3PL model. The smaller the AIC is, the better fit the model will be to the data.

In this study, regardless of data generated by the 2PL-Guessing model or by the 3PL model, when the 2PL-Guessing model was used to estimate item parameters, all AIC

indices for sample size of 100, 200, and 500 were smaller than those AIC indices estimated by the 3PL model, so the 2PL-Guessing model not only fit the data generated by the 2PL-Guessing model better, but also fit the data generated by the 3PL model better for sample size of 100, 200, and 500., demonstrating that even in random guessing situation for small sample sizes, the new 2PL-Guessing model always fit the data better than the 3Pl model.

However, in random guessing situation for sample size 1000 (or the data was generated by the 3PL model), even the average AIC index estimated by the 2PL-Guessing model was smaller than the average AIC index estimated the 3PL model, not each replication's AIC estimated by the 2PL-Guessing model was smaller than the AIC estimated by the 3PL model. Some AIC indices estimated by the 3PL model were smaller than those estimated by the 2PL-Guessing model, meaning the conventional 3PL model is better applied to big sample size tests. For sample size under 1000, the 2PL-Guessing model can do better estimation than the 3PL model even in random guessing situation.

Running Time Comparison

Another big advantage of the new model was that it ran a lot faster than the 3PL model when estimating item parameters using maximum likelihood estimation method. The 3PL model has been notoriously slow in estimating item parameters because there are three parameters in the 3PL model. The new model changed the guessing parameter in the 3PL model into a function of difficulty and discrimination parameters, so the new model is still 2PL plus a guessing function model in which there are only two item parameters: difficulty and discrimination parameters. The process of estimating item parameters can be reduced tremendously because of this, for example, a laptop with 4GB ram memory was used to estimate item parameters for a 1000 sample size and 20-item test and it took 72 hours to get the results using the 3PL model, but it took only 18 hours

to get the results using the new model.

Convergence Problem for the 3PL Model

Although the optimization techniques used by SAS PROC NLMIXED are some of the best ones available, for the 3PL model, it always has difficult time in converging because the 3PL model is more complex. To achieve convergence for the 3PL model, we took some extra steps in SAS program such as changing the parameter initial values and using boundary constraints to avoid floating-point errors and overflows. The new model, however, could converge easily for all sample sizes under any condition.

In summary, the new model was a better model to estimate parameters if the assumption that different ability examinees have different probability of guessing an item correctly is viable. Even in the random guessing situation, the new model could estimate the latent ability and discrimination parameter more accurately than the 3PL model. The 3PL model performed better than the new model in estimating difficulty parameter only in the random guessing situation.

The new model successfully controlled the successful guessing probability between the probability of random guessing and 0.5, estimated parameters more accurately, ran much faster in estimating item parameters, and reflected the different probability of successful guessing for examines of different ability. However, due to the highly time-consuming estimation process of PROC NLMIXED, only 20-item short tests were simulated with minimum sample size 100, this may lead to a restriction in the generalizability of the new model.

Future Research Recommendations

Due to the limited designs in this study, there are a few directions for future research that could be considered. First, Reynolds (1986) found that the normal ability distribution estimated the difficulty parameter most accurately and the uniform ability distribution

estimated discrimination parameter most accurately. The ability distribution was normally distributed in this study, the effect of skewed ability distribution on parameter estimation for the new model should be very interesting to explore.

Second, the difficulty parameter was controlled from -0.7 to 2.0 and the item discrimination was controlled from 0.4 to 2.0 to simulate an achievement test or a high-stake test, so it may be of interest to investigate the effect of expanded range (for example, -2.0 to 2.5 for difficulty parameter and 0 to 2.5 for discrimination parameter) on the accuracy of estimating item parameters. Because we all know that some classroom test items were made very easy on teacher's purpose and students use guessing strategies to answer those items that they do not know the correct answers, the expanded item parameter range study will be crucial for the new model to apply to classroom tests.

SAS PROC NLMIXED was adopted to do marginal maximum likelihood estimation because it provides one of the best optimization techniques. It should be enlightening to use other statistical softwares such as R or Matlab to calibrate parameters using the new model, and then compare their results to find which software can produce the most accurate estimation and which one will have the worst estimation.

The number of options used in this study was 4 which is most popular in high-stake tests. We know that as the number of options increases, the probability of guessing an item correctly decreases. The number of options should have effect on guessing strategies. If the number of options, theoretically, increased up to infinite, then the probability of successful guessing would be zero. Therefore, the more the number of options is, the less motivating the examinee will be because it is too time-consuming. What will happen if the number of options is increased up to a threshold that examinees are not motivated to use partial knowledge to make a guess because there are too many

options? In this situation, examinees just make random guess to those items that they do not know the answers, thus we can simplify the guessing situation.

REFERENCES

Allen, M. J. & Yen, W. M. (1979). *Introduction to measurement theory*, Long Grove, IL: Waveland Press Inc..

Anderson, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B: Methodological, 34,* 42-54.

Angoff, W. & Schrader, W. (1984). A study of hypothesis basic to the use of rights and formula scores. *Journal of Educational Measurement, 21*, 1-17.

Baker, F. (1992). *Item response theory: Parameter estimation techniques*, New York: Marcel Dekker, Inc..

Baker, F. & Kim, S. (2004). *Item response theory: Parameter estimation techniques*, Second Edition, Revised and Expanded, New York: Marcel Dekker, Inc..

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

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practices,* Winter, 21-33.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Bock, R. D., & Lieberman M. (1970). Fitting a response curve model for dichotomously scored items. *Psychometrica, 35,* 197-198.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behavior. *Psychometrika, 73*(2), 209-230.

Davis, L. L. (2002). Strategies for controlling item exposure in computerized adaptive testing with polytomously scored items. A dissertation for PH. D. University of Texas at Autin.

Dawber, T., Rogers, W. T., & Carbonaro, M. (2004). Robustness of Lord's formula for item difficulty and discrimination conversion between classical and item response theory models. Unpublished paper presented at American Educational Research Association.

De Ayala, R. J. (2008). *The theory and practice of item response theory.* New York, NY: Guilfords Publications.

De Gruijter, D. N. M. (1990). A note on the bias of UCON item parameter estimation in the Rasch model. *Journal of Educational Measurement, 27*(3), 285-288.

Diamond, J. & Evans, W. (1973). The correction for guessing. *Review of Educational Research, 43*, 181-191.

Embreston, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum.

Fisher, W. P. (1993). Scale-free measurement revisited. *Rasch Measurement Transctions, 7*, 272-273.

Freedle, R. (2006). How and why standardized tests systematically underestimate African-Americans' true verbal ability and what to do about it: Towards the promotion of two new theories with practical applications. *St. John's Law Review, 80*, 183-226.

Gulliksen, H.O (1950). *Theory of mental tests*. New York: John Wiley and Sons, Inc.

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

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

Hambleton, R. K., et al. (1992). Hambleton 9 theses. *Rasch Measurement Transactions*, *6*(2), p.215-217.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 3847.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Education and Behavioral Statistics, 13*, 243-271.

Harwell, M., Stone, C. A., Hsu, T. & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.

Hutchinson, T. P. (1991). *Ability, partial information and guessing: Statistical modeling applied to multiple-choice tests.* Rundle Mall, South Australia: Rumsby Scientific Publishing.

Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response Models in R. *Journal of Statistical Software, 20*(10), 1-24.

Kamata, A. (1998). *Some generalizations of the Rasch model: An application of the hierarchical generalized linear model*. A dissertation for PH.D. Michigan State University.

Kubinger, K. D. & Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on different item response formats: An experiment in fundamental research on psychological assessment. *Psychology Science*, *49*, 361-374.

Lim, R. G. & Drasgow, F. (1990). Evaluation of two methods for estimating item Response theory parameters when assessing different item functioning. *Journal of Applied Psychology, 75*(2), 164-174.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, No. 7.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement*, *12*, 7-12.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale NJ: Erlbaum.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*(2), 157-162.

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

Lincare, J. M. (1995). Investigating empirical ICCs. *Rasch Measurement Transactions*, *9*(3), 449.

Lincare, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9,* 49-57.

Martin, E.S., del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing.

Applied Psychology Measurement, 30, 183-203.

Mattson, D. (1965). The effects of guessing on the standard error of measurement and reliability of test scores. *Educational & Psychological Measurement, 25*, 727-730.

Mehrens, W. A., & Lehman, I. J. (1987). *Using standardized tests in education*. (4th ed.). NY: Longman.

Messick, S. (1995). Validation of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into scoring meanings. *American Psychologist, 50,* 741-749.

Mislevy, R., & Bock, R. D. (1990). BILOG: *Item analysis and test scoring with binary logistic models* [Computer software]. Mooresville, IN: Scientific Software.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational Psychological Measurement, 16*, 159-176.

Neyman, J., Scott, E.L., (1948). Consistent estimation from partially consistent observations. *Econometrica, 16*, 1-32.

Pelton, T. W. (2002). *The accuracy of unidimensional measurement models in the presence of deviations for the underlying assumptions.* Unpublished doctoral dissertation, Brigham Young University, Department of Instructional Psychology and Technology.

Pinheiro, J. C., & Bates, D. M. (1995). Approximation to the Log-Likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics, 4*, 12-35.

Prihoda, T. J., Pinckard, R. N., McMahan, A., & Jones, A. C. (2006). Correcting

for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. Journal of Dental Education, *70*(4), 378-386.

Ree, J. M. (1979). Estimating item characteristic curve. *Applied Psychological Measurement*, *3*, 371-385.

Reynolds, T. (1986). The effects of small sample size, short test length, and ability distribution upon parameter estimation. Unpublished paper.

Rogers, H. J. (1999). Guessing in multiple choice tests. In G. N. Master & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 235-243). Amsterdam: Pergamom.

Rowley G. L., & Traub R. E.(1977). Formula scoring, number right scoring, and test-taking strategy. *Journal of Educational Measurement*, *14*(1), 15-22.

SAS Institute Inc., 2008. SAS/STAT 9.2 Users' Guide. Cary, NC.

Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*(3), 299-311.

Sheng, Y. (2008). Markov Chain Monte Carlo estimation of normal ogive IRT models in MATLAB. *Journal of Statistical Software, 25*(8), 1-15.

Si, C. F. & Schumacker, R. E. (2004). Ability estimation under different item parameterization and scoring models. *International Journal of Testing, 4*(2), 137-181.

Smith, R. M. (1993). Guessing and the Rasch model. *Rasch Measurement Transactions. 6*(4), 262-263.

Stage, C. (2003). Classical test theory or item response theory: The Swedish Experience. *Centro de Estudios Publicos, 42*.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin Company.

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

Waller, M. I. (1973). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement, 13*, 233-243.

Wietzman, R. A. (1996). The Rasch model plus guessing. *Educational and Psychological Measurement*, *56*, 779-790.

Wright, B. D. (1991). Rasch vs. Birnbaum *Rasch Measurement Transactions, 5*, 178-179.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrica*, *52*(2), 275-292.

APPENDICES

Appendix A

The Deduction of Marginal Maximum Likelihood Function for the 2PL-Guessing Model

According to Bayes' theorem, the posterior ability distribution is given as (Baker & Kim, 2004):

$$P(\theta_j \mid \mathbf{Y}_j, \boldsymbol{\tau}, \xi) = \frac{P(\mathbf{Y}_j \mid \theta_j, \xi) g(\theta_j \mid \boldsymbol{\tau})}{\int P(\mathbf{Y}_j \mid \theta_j, \xi) g(\theta_j \mid \boldsymbol{\tau}) d\theta_j}. \tag{A.1}$$

also,

$$\frac{\partial}{\partial a_i}[P(\mathbf{Y}_j \mid \theta_j, \xi)] = \frac{\partial}{\partial a_i}[\log P(\mathbf{Y}_j \mid \theta_j, \xi)] P(\mathbf{Y}_j \mid \theta_j, \xi), \tag{A.2}$$

then,

$$\frac{\partial}{\partial a_i}(\log L) = \sum_j^N \frac{\partial}{\partial a_i}[\log P(\mathbf{Y}_j)] = \sum_j^N [P(\mathbf{Y}_j)]^{-1} \frac{\partial}{\partial a_i}\left[\int P\left(\mathbf{Y}_j \mid \theta_j, \xi\right) g(\theta_j \mid \boldsymbol{\tau}) d\theta_j\right]$$

$$= \sum_j^N [P(\mathbf{Y}_j)]^{-1} \int \frac{\partial}{\partial a_i}\left[P(\mathbf{Y}_j \middle| \theta_j, \xi)\right] g(\theta_j \mid \boldsymbol{\tau}) d\theta_j,$$

because of relation in euqation A.2, we will have

$$= \sum_j^N [P(\mathbf{Y}_j)]^{-1} \int \frac{\partial}{\partial a_i}\left[\log P(\mathbf{Y}_j \middle| \theta_j, \xi)\right] P\left(\mathbf{Y}_j \mid \theta_j, \xi\right) g(\theta_j \mid \boldsymbol{\tau}) d\theta_j$$

$$= \sum_{j=1}^N \int \frac{\partial}{\partial a_i}\left[\log P(\mathbf{Y}_j \middle| \theta, \xi)\right] \left[\frac{P(\mathbf{Y}_j \middle| \theta_j, \xi) g(\theta_j \mid \boldsymbol{\tau})}{P(\mathbf{Y}_j)}\right] d\theta_j,$$

using equation A.1, result in

$$= \sum_{j=1}^N \int \frac{\partial}{\partial a_i}\left[\log P(\mathbf{Y}_j \middle| \theta, \xi)\right] \left[P(\theta \middle| \mathbf{Y}_j, \xi, \boldsymbol{\tau})\right] d\theta_j,$$

where $P(\mathbf{Y}_j \mid \theta_j, \xi) = \prod_{i=1}^n P_i(\theta_j)^{y_{ij}} Q_i(\theta_j)^{1-y_{ij}}$

so, let $\theta_j = \theta$ (for convenience)

$$\frac{\partial}{\partial a_i}(\log L) = \sum_{j=1}^{N}\int \frac{\partial}{\partial a_i}[\log \prod_{i=1}^{n} P_i(\theta)^{y_{ij}} Q_i(\theta)^{1-y_{ij}}] \times [P(\theta \mid \mathbf{Y}_j, \boldsymbol{\xi}, \boldsymbol{\tau})] d\theta$$

$$= \sum_{j=1}^{N}\int [\prod_{i=1}^{n} P_i(\theta)^{y_{ij}} Q_i(\theta)^{1-y_{ij}}]^{-1} \times \frac{\partial}{\partial a_i}[\prod_{i=1}^{n} P_i(\theta)^{y_{ij}} Q_i(\theta)^{1-y_{ij}}][P(\theta \mid \mathbf{Y}_j, \boldsymbol{\xi}, \boldsymbol{\tau})] d\theta, \quad \text{(A.3)}$$

where

$$\frac{\partial}{\partial a_i}[\prod_{i=1}^{n} P_i(\theta)^{y_{ij}} Q_i(\theta)^{1-y_{ij}}] = [\prod_{h\neq 1}^{n} P_h(\theta)^{y_{ij}} Q_h(\theta)^{1-y_{ij}}] \frac{\partial}{\partial a_i}[P_i(\theta)^{y_{ij}} Q_i(\theta)^{1-y_{ij}}],$$

where

$$\frac{\partial}{\partial a_i}[P_i(\theta)^{y_{ij}} Q_i(\theta)^{1-y_{ij}}] = \frac{\partial}{\partial a_i}[P_i(\theta)^{y_{ij}}] Q_i(\theta)^{1-y_{ij}} + P_i(\theta)^{y_{ij}} \frac{\partial}{\partial a_i}[Q_i(\theta)^{1-y_{ij}}]$$

$$= y_{ij} P_i(\theta)^{y_{ij}-1}\left[\frac{\partial P_i(\theta)}{\partial a_i}\right] Q_i(\theta)^{1-y_{ij}} + P_i(\theta)(1-y_{ij}) Q_i(\theta)^{1-y_{ij}-1}\left[\frac{\partial Q_i(\theta)}{\partial a_i}\right].$$

Using the relationship $\frac{\partial Q_i(\theta)}{\partial a_i} = -\frac{\partial P_i(\theta)}{\partial a_i}$ and we can get

$$= \left[\frac{\partial P_i(\theta)}{\partial a_i}\right][P_i(\theta)^{y_{ij}} Q_i(\theta)^{1-y_{ij}}]\left[\frac{y_{ij}}{P_i(\theta)} - \frac{1-y_{ij}}{Q_i(\theta)}\right]$$

$$= \left[\frac{\partial P_i(\theta)}{\partial a_i}\right][P_i(\theta)^{y_{ij}} Q_i(\theta)^{1-y_{ij}}]\left[\frac{y_{ij} - P_i(\theta)}{P_i(\theta) Q_i(\theta)}\right],$$

then,

$$\frac{\partial}{\partial a_i}[\prod_{i=1}^{n} P_i(\theta)^{y_{ij}} Q_i(\theta)^{1-y_{ij}}] = [\prod_{i=1}^{n} P_i(\theta)^{y_{ij}} Q_i(\theta)^{1-y_{ij}}][\frac{\partial P_i(\theta)}{\partial a_i}][\frac{y_{ij} - P_i(\theta)}{P_i(\theta) Q_i(\theta)}]. \quad \text{(A.4)}$$

Put equation A.4 into equation A.3, we get the marginal likelihood equation for a_i can be

written as the follows:

$$\frac{\partial}{\partial a_i}(\log L) = \sum_{j=1}^{N}\int [\frac{y_{ij} - P_i(\theta_j)}{P_i(\theta_j) Q_i(\theta_j)}][\frac{\partial P_i(\theta_j)}{\partial a_i}][P(\theta_j \mid \mathbf{Y}_j, \boldsymbol{\xi}, \boldsymbol{\tau})] d\theta_j,$$

let $\frac{\partial P_i(\theta_j)}{\partial a_i} = Ka$ (for discrimination parameter), and

let $\frac{\partial P_i(\theta_j)}{\partial b_i} = Kb$ (for difficulty parameter), and then we have

$$\frac{\partial}{\partial a_i}(\log L) = \sum_{j=1}^{N} \int [\frac{y_{ij} - P_i(\theta_j)}{P_i(\theta_j)Q_i(\theta_j)}][Ka][P(\theta_j \mid \mathbf{Y}_j, \boldsymbol{\xi}, \boldsymbol{\tau})]d\theta_j \qquad \text{(A.5)}$$

The likelihood equation for b_i is

$$\frac{\partial}{\partial b_i}(\log L) = \sum_{j=1}^{N} \int [\frac{y_{ij} - P_i(\theta_j)}{P_i(\theta_j)Q_i(\theta_j)}][Kb][P(\theta_j \mid \mathbf{Y}_j, \boldsymbol{\xi}, \boldsymbol{\tau})]d\theta_j \qquad \text{(A.6)}$$

For Rasch-Guessing model, there is only item difficulty parameter, then (N is the number of options for the following equations)

$$Kb = \frac{e^{2(\theta_j - b_i)}}{[1+e^{(\theta_j - b_i)}]^2} - \frac{e^{(\theta_j - b_i)}}{1+e^{(\theta_j - b_i)}} + \left(1 - \frac{e^{(\theta_j - b_i)}}{1+e^{(\theta_j - b_i)}}\right)\left(\frac{2Ne^{2(\theta_j - b_i)}}{(N-2)\left(1+\frac{2Ne^{(\theta_j - b_i)}}{N-2}\right)^2} - \frac{e^{(\theta_j - b_i)}}{1+\frac{2Ne^{(\theta_j - b_i)}}{N-2}}\right)$$
$$+\left(-\frac{e^{2(\theta_j - b_i)}}{[1+e^{(\theta_j - b_i)}]^2} + \frac{e^{(\theta_j - b_i)}}{1+e^{(\theta_j - b_i)}}\right)\left(\frac{1}{N} + \frac{e^{(\theta_j - b_i)}}{1+\frac{2Ne^{(\theta_j - b_i)}}{N-2}}\right).$$

For 2PL-Guessing model, (N is the number of options)

$$Ka = (\theta_j - b_i)\left[\begin{array}{l}\dfrac{e^{2a_i(\theta_j-b_i)}}{\left(1+e^{a_i(\theta_j-b_i)}\right)^2}+\dfrac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}+\left(\dfrac{1}{N}+\dfrac{e^{a_i(\theta_j-b_i)}}{1+\dfrac{2Ne^{a_i(\theta_j-b_i)}}{N-2}}\right)\left(\dfrac{e^{2a_i(\theta_j-b_i)}}{\left(1+e^{a_i(\theta_j-b_i)}\right)^2}-\dfrac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}\right)\\ +\left(1-\dfrac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}\right)\left(-\dfrac{2Ne^{2a_i(\theta_j-b_i)}}{(N-2)\left(1+\dfrac{2Ne^{a_i(\theta_j-b_i)}}{N-2}\right)^2}+\dfrac{e^{a_i(\theta_j-b_i)}}{1+\dfrac{2Ne^{a_i(\theta_j-b_i)}}{N-2}}\right)\end{array}\right]$$

$$Kb = a_i\left[\begin{array}{l}\dfrac{e^{2a_i(\theta_j-b_i)}}{\left(1+e^{a_i(\theta_j-b_i)}\right)^2}-\dfrac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}+\left(\dfrac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}-\dfrac{e^{2a_i(\theta_j-b_i)}}{\left(1+e^{a_i(\theta_j-b_i)}\right)^2}\right)\left(\dfrac{1}{N}+\dfrac{e^{a_i(\theta_j-b_i)}}{1+\dfrac{2Ne^{a_i(\theta_j-b_i)}}{N-2}}\right)\\ +\left(1-\dfrac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}\right)\left(\dfrac{2Ne^{2a_i(\theta_j-b_i)}}{(N-2)\left(1+\dfrac{2Ne^{a_i(\theta_j-b_i)}}{N-2}\right)^2}-\dfrac{e^{a_i(\theta_j-b_i)}}{1+\dfrac{2Ne^{a_i(\theta_j-b_i)}}{N-2}}\right)\end{array}\right].$$

Appendix B

Difficulty Parameter RMSD Estimates for the Rasch-Guessing Model

Table B1

RMSD of Difficulty Parameter Estimates for the Rasch-Guessing Model (20 items)

True values	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD
	N=100		N=200		N=500		N=1000	
-0.327	-0.283	0.303	-0.307	0.224	-0.267	0.205	-0.361	0.138
0.724	0.852	0.320	0.802	0.287	0.864	0.242	0.798	0.165
1.67	1.712	0.418	1.651	0.221	1.652	0.154	1.689	0.134
0.413	0.510	0.244	0.441	0.167	0.361	0.141	0.399	0.134
0.032	-0.311	0.446	-0.048	0.245	0.020	0.101	0.028	0.094
1.113	1.173	0.370	1.136	0.230	1.127	0.109	1.154	0.109
0.202	0.293	0.341	0.157	0.243	0.192	0.117	0.182	0.103
0.986	0.984	0.385	1.036	0.239	0.954	0.184	0.983	0.103
1.417	1.862	0.662	1.495	0.244	1.378	0.168	1.357	0.108
1.843	2.058	0.397	2.008	0.260	1.895	0.156	1.846	0.116
0.567	0.546	0.317	0.597	0.142	0.571	0.105	0.547	0.105
0.106	-0.104	0.312	0.022	0.183	0.032	0.185	0.070	0.111
-0.67	-0.220	0.505	-0.526	0.371	-0.629	0.155	-0.626	0.138
2.018	1.917	0.358	1.960	0.275	2.002	0.193	1.983	0.163
-0.245	-0.279	0.337	-0.284	0.200	-0.316	0.146	-0.280	0.092
0.116	0.084	0.316	0.132	0.262	0.112	0.169	0.111	0.143
1.216	1.360	0.330	1.234	0.193	1.251	0.143	1.216	0.095
1.682	1.793	0.425	1.693	0.242	1.693	0.170	1.696	0.118
0.291	0.154	0.423	0.274	0.301	0.254	0.168	0.273	0.111
0.451	0.483	0.241	0.455	0.116	0.446	0.109	0.432	0.076

Table B2

RMSD of Difficulty Parameter Estimates for the Rasch-Guessing Model(30 Items)

True values	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD
	N=100		N=200		N=500		N=1000	
-0.327	-0.213	0.284	-0.311	0.186	-0.407	0.103	-0.394	0.108
0.724	0.785	0.429	0.687	0.244	0.710	0.158	0.734	0.102
1.67	1.751	0.276	1.716	0.179	1.705	0.165	1.638	0.155
0.413	0.502	0.387	0.565	0.323	0.539	0.204	0.451	0.119
0.032	-0.049	0.312	-0.019	0.186	-0.008	0.137	0.030	0.116
1.113	0.999	0.291	1.032	0.337	1.101	0.222	1.096	0.102
0.202	0.287	0.256	0.250	0.201	0.234	0.145	0.228	0.098
0.986	1.077	0.215	0.929	0.207	0.894	0.154	0.899	0.133
1.417	1.587	0.322	1.668	0.324	1.445	0.152	1.422	0.080
1.843	2.040	0.375	1.837	0.297	1.773	0.175	1.799	0.099
0.567	0.686	0.278	0.629	0.221	0.487	0.187	0.548	0.077
0.106	0.115	0.245	0.067	0.248	0.092	0.164	0.084	0.100
-0.67	-0.584	0.258	-0.711	0.204	-0.638	0.172	-0.670	0.062
2.018	2.276	0.579	2.064	0.237	2.059	0.173	2.039	0.089
-0.245	-0.243	0.312	-0.203	0.254	-0.294	0.142	-0.282	0.101
0.116	0.058	0.282	0.035	0.215	0.072	0.114	0.059	0.083
1.216	1.267	0.335	1.262	0.157	1.254	0.124	1.245	0.105
1.682	1.781	0.385	1.736	0.283	1.763	0.138	1.758	0.095
0.291	0.369	0.381	0.329	0.238	0.306	0.192	0.290	0.112
0.451	0.465	0.206	0.476	0.123	0.490	0.207	0.460	0.111
0.46	0.374	0.312	0.361	0.228	0.361	0.164	0.406	0.104
1.419	1.316	0.350	1.442	0.170	1.493	0.125	1.518	0.153
1.886	2.125	0.476	2.017	0.299	1.937	0.207	1.945	0.117
1.169	1.160	0.340	1.230	0.261	1.200	0.133	1.162	0.100
0.944	0.853	0.601	0.949	0.307	0.957	0.166	0.958	0.096
0.39	0.340	0.274	0.320	0.208	0.345	0.136	0.336	0.107
0.046	-0.034	0.205	0.040	0.220	-0.019	0.108	-0.007	0.083
0.858	0.898	0.168	0.911	0.186	0.842	0.148	0.862	0.095
0.734	0.760	0.200	0.740	0.193	0.721	0.103	0.699	0.100
0.258	0.320	0.255	0.285	0.254	0.310	0.141	0.291	0.126

Table B3

RMSD of Difficulty Parameter Estimates for the Rasch-Guessing Model (40 Items)

True values	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD
	N=100		N=200		N=500		N=1000	
-0.327	-0.269	0.385	-0.265	0.265	-0.334	0.098	-0.326	0.072
0.724	0.804	0.493	0.712	0.242	0.730	0.117	0.669	0.092
1.670	1.584	0.331	1.537	0.186	1.597	0.216	1.592	0.127
0.734	0.906	0.388	0.765	0.224	0.815	0.173	0.764	0.127
0.413	0.393	0.378	0.401	0.323	0.383	0.151	0.382	0.089
0.448	0.399	0.424	0.465	0.292	0.461	0.129	0.427	0.122
0.032	-0.015	0.475	-0.026	0.310	0.038	0.113	0.032	0.058
0.425	0.369	0.241	0.369	0.195	0.450	0.124	0.415	0.125
0.490	0.460	0.201	0.506	0.172	0.504	0.133	0.446	0.092
1.419	1.652	0.474	1.525	0.239	1.464	0.165	1.410	0.145
1.113	1.261	0.430	1.154	0.234	1.121	0.107	1.113	0.082
0.202	0.044	0.321	0.068	0.207	0.089	0.182	0.094	0.146
0.983	1.044	0.331	0.988	0.202	1.033	0.163	1.008	0.089
0.986	1.042	0.503	1.065	0.275	1.069	0.182	0.986	0.083
0.858	0.785	0.308	0.745	0.255	0.808	0.151	0.838	0.095
0.451	0.499	0.441	0.368	0.346	0.404	0.130	0.366	0.128
0.046	0.011	0.216	0.062	0.227	0.082	0.121	0.046	0.114
1.417	1.505	0.283	1.543	0.219	1.419	0.089	1.400	0.088
0.460	0.460	0.339	0.533	0.304	0.540	0.193	0.475	0.135
0.315	0.562	0.360	0.396	0.312	0.347	0.186	0.318	0.103
1.843	2.004	0.371	2.000	0.317	1.889	0.210	1.800	0.131
1.286	1.267	0.315	1.330	0.203	1.338	0.141	1.320	0.108
0.567	0.658	0.412	0.563	0.287	0.552	0.129	0.565	0.109
0.955	0.946	0.480	1.022	0.259	1.006	0.105	0.999	0.074
1.133	1.143	0.407	1.096	0.327	1.177	0.211	1.139	0.130
1.118	1.361	0.582	1.153	0.304	1.154	0.200	1.128	0.098
0.390	0.310	0.308	0.228	0.255	0.349	0.175	0.367	0.116
0.944	0.813	0.318	0.883	0.231	0.950	0.125	0.923	0.090
0.106	0.008	0.380	0.069	0.243	0.025	0.151	0.080	0.108
1.216	1.150	0.301	1.092	0.258	1.179	0.156	1.180	0.111
-0.670	-0.628	0.279	-0.640	0.270	-0.691	0.127	-0.692	0.081
0.116	0.050	0.349	0.131	0.206	0.140	0.144	0.111	0.059
1.169	1.283	0.379	1.224	0.242	1.130	0.181	1.152	0.136

Table B3 (continued).

RMSD of Difficulty Parameter Estimates for the Rasch-Guessing Model (40 Items)

True values	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD
	N=100		N=200		N=500		N=1000	
0.291	0.080	0.361	0.138	0.307	0.194	0.148	0.207	0.131
1.682	1.682	0.524	1.528	0.324	1.599	0.201	1.638	0.135
-0.245	-0.272	0.465	-0.311	0.287	-0.294	0.115	-0.270	0.097
0.481	0.360	0.280	0.484	0.220	0.465	0.111	0.413	0.119
0.258	0.183	0.356	0.193	0.236	0.229	0.130	0.212	0.120
1.886	1.803	0.393	1.840	0.264	1.859	0.253	1.913	0.148
2.018	1.800	0.636	1.885	0.373	1.982	0.174	1.967	0.113

Appendix C

RMSD of Each Item Parameter Estimates for All Methods

Table C1

RMSD of Difficulty Parameter Estimates for the 2PLG-2PLG (20 items)

True values	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD
	N=100		N=200		N=500		N=1000	
-0.327	-0.322	0.282	-0.309	0.225	-0.249	0.193	-0.350	0.160
0.724	0.863	0.338	0.823	0.242	0.842	0.204	0.756	0.151
1.67	1.747	0.455	1.691	0.361	1.808	0.324	1.749	0.240
0.413	0.431	0.186	0.462	0.159	0.377	0.091	0.414	0.096
0.032	-0.225	0.396	0.000	0.253	0.033	0.130	0.036	0.097
1.113	1.089	0.523	1.026	0.216	1.130	0.136	1.118	0.146
0.202	0.222	0.339	0.161	0.274	0.184	0.144	0.179	0.101
0.986	1.092	0.446	1.046	0.309	0.950	0.210	1.003	0.106
1.417	1.718	0.519	1.576	0.474	1.519	0.309	1.422	0.213
1.843	1.570	0.485	1.793	0.352	1.736	0.347	1.740	0.231
0.567	0.632	0.384	0.561	0.112	0.571	0.117	0.534	0.077
0.106	-0.029	0.506	0.036	0.354	0.011	0.419	0.063	0.257
-0.67	-0.163	0.588	-0.393	0.369	-0.587	0.123	-0.578	0.133
2.018	1.943	0.230	2.029	0.068	2.028	0.100	2.032	0.104
-0.245	-0.304	0.351	-0.304	0.229	-0.316	0.150	-0.269	0.089
0.116	0.067	0.262	0.106	0.198	0.084	0.140	0.092	0.126
1.216	1.619	0.621	1.410	0.429	1.366	0.432	1.205	0.307
1.682	1.577	0.390	1.618	0.290	1.623	0.239	1.701	0.251
0.291	0.231	0.325	0.298	0.250	0.257	0.142	0.269	0.103
0.451	0.431	0.479	0.495	0.301	0.455	0.175	0.397	0.120

Table C2

RMSD of Discrimination Parameter Estimates for the 2PLG-2PLG (20 Items)

True values	$\bar{\hat{a}}$	RMSD	$\bar{\hat{a}}$	RMSD	$\bar{\hat{a}}$	RMSD	$\bar{\hat{a}}$	RMSD
	N=100		N=200		N=500		N=1000	
0.873	1.101	0.547	0.976	0.280	0.968	0.209	0.908	0.156
1.409	1.435	0.373	1.408	0.362	1.395	0.217	1.476	0.221
0.831	0.906	0.436	0.881	0.295	0.764	0.156	0.811	0.100
1.676	1.605	0.430	1.522	0.404	1.662	0.263	1.666	0.212
1.218	1.264	0.452	1.155	0.212	1.184	0.224	1.201	0.164
1.457	1.693	0.453	1.675	0.347	1.455	0.213	1.533	0.216
1.054	1.056	0.417	1.030	0.352	1.100	0.188	1.039	0.127
1.318	1.385	0.532	1.454	0.452	1.379	0.223	1.318	0.104
1.3	1.452	0.449	1.429	0.519	1.277	0.411	1.273	0.250
0.924	1.461	0.676	1.131	0.394	1.143	0.465	1.020	0.180
1.378	1.390	0.497	1.367	0.214	1.402	0.162	1.410	0.178
0.54	0.899	0.537	0.607	0.200	0.542	0.145	0.530	0.102
0.833	0.819	0.260	0.816	0.216	0.823	0.211	0.884	0.186
1.578	1.423	0.420	1.413	0.315	1.523	0.245	1.528	0.274
0.964	0.859	0.315	0.974	0.297	0.999	0.119	1.021	0.116
1.3	1.395	0.483	1.382	0.380	1.430	0.374	1.350	0.228
0.603	0.551	0.122	0.589	0.138	0.601	0.165	0.648	0.123
1.444	1.675	0.429	1.584	0.387	1.591	0.344	1.462	0.272
1.858	1.757	0.335	1.741	0.336	1.797	0.215	1.843	0.204
0.599	0.800	0.554	0.707	0.342	0.625	0.116	0.595	0.087

Table C3

RMSD of Difficulty Parameter Estimates for the 2PLG-3PL (20 Items)

True values	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD
	N=100		N=200		N=500		N=1000	
-0.327	-0.270	0.351	-0.560	0.546	-0.554	0.780	-0.496	0.584
0.724	0.184	0.715	0.223	0.697	0.265	0.659	0.149	0.756
1.67	1.168	1.128	0.914	0.997	0.728	1.143	0.805	1.052
0.413	-0.137	0.706	0.072	0.611	-0.150	0.700	-0.184	0.672
0.032	-0.204	0.578	-0.626	0.949	-0.398	0.671	-0.247	0.394
1.113	0.487	1.018	0.484	0.752	0.555	0.617	0.645	0.512
0.202	-0.205	0.743	-0.461	0.967	-0.369	0.840	-0.626	1.026
0.986	0.315	0.908	0.229	0.923	0.293	0.808	0.590	0.501
1.417	1.277	0.769	0.861	0.788	0.965	0.583	0.750	0.781
1.843	0.977	1.079	1.017	1.040	1.059	0.960	1.207	0.711
0.567	0.064	0.698	0.141	0.587	0.145	0.612	-0.086	0.775
0.106	-0.096	0.636	-0.434	1.065	-0.382	1.129	-0.097	0.856
-0.67	-0.408	0.746	-1.011	0.761	-1.696	1.093	-1.563	1.018
2.018	1.311	0.830	1.270	0.834	1.491	0.634	1.541	0.517
-0.245	-0.345	0.366	-0.465	0.436	-0.882	0.975	-0.704	0.812
0.116	-0.031	0.520	-0.318	0.803	-0.271	0.702	-0.526	0.758
1.216	0.504	1.472	0.056	1.420	0.533	1.257	-0.022	1.401
1.682	1.122	0.846	1.030	0.805	0.918	0.818	1.005	0.731
0.291	-0.133	0.684	0.067	0.496	-0.175	0.607	-0.097	0.420
0.451	0.321	0.884	-0.107	1.039	0.028	0.943	-0.405	1.252

Table C4

RMSD of Discrimination Parameter Estimates for the 2PLG-3PL (20 Items)

True values	$\bar{\hat{a}}$	RMSD	$\bar{\hat{a}}$	RMSD	$\bar{\hat{a}}$	RMSD	$\bar{\hat{a}}$	RMSD
	N=100		N=200		N=500		N=1000	
0.873	1.346	0.713	0.998	0.378	1.096	0.400	1.001	0.271
1.409	1.331	0.529	1.435	0.441	1.373	0.462	1.378	0.418
0.831	1.120	0.724	1.209	0.795	0.875	0.537	0.731	0.268
1.676	1.574	0.506	1.632	0.425	1.559	0.443	1.384	0.399
1.218	1.571	0.541	1.060	0.417	1.183	0.214	1.301	0.228
1.457	1.375	0.429	1.529	0.449	1.278	0.473	1.412	0.348
1.054	1.152	0.426	1.151	0.580	1.081	0.227	0.964	0.266
1.318	1.388	0.650	1.327	0.537	1.220	0.374	1.245	0.386
1.3	1.267	0.644	1.283	0.395	1.343	0.459	1.341	0.484
0.924	1.234	0.720	0.911	0.517	1.113	0.666	1.037	0.268
1.378	1.372	0.560	1.456	0.485	1.529	0.429	1.274	0.395
0.54	1.119	0.803	0.610	0.360	0.563	0.153	0.517	0.126
0.833	0.832	0.238	0.787	0.209	0.723	0.256	0.747	0.219
1.578	1.222	0.740	1.294	0.604	1.266	0.523	1.395	0.560
0.964	1.006	0.429	1.149	0.415	1.059	0.350	1.046	0.246
1.3	1.609	0.559	1.425	0.536	1.567	0.569	1.248	0.329
0.603	0.995	0.762	0.641	0.473	0.707	0.248	0.565	0.146
1.444	1.562	0.539	1.313	0.633	1.194	0.568	1.047	0.570
1.858	1.785	0.336	1.792	0.323	1.685	0.362	1.787	0.303
0.599	0.921	0.711	0.853	0.578	0.727	0.338	0.623	0.226

Table C5

RMSD of Difficulty Parameter Estimates for the 3PL-3PL (20 Items)

True values	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD
	N=100		N=200		N=500		N=1000	
-0.327	-0.210	0.410	-0.423	0.338	-0.280	0.323	-0.414	0.336
0.724	0.557	0.361	0.598	0.370	0.599	0.353	0.410	0.467
1.67	1.627	0.508	1.517	0.469	1.783	0.404	1.441	0.486
0.413	-0.017	0.538	0.216	0.435	0.251	0.393	0.214	0.363
0.032	-0.092	0.562	-0.263	0.519	-0.111	0.395	-0.116	0.405
1.113	0.914	0.484	0.816	0.478	1.032	0.299	1.040	0.271
0.202	-0.170	0.571	-0.256	0.601	-0.078	0.595	-0.042	0.468
0.986	0.882	0.375	0.791	0.455	0.743	0.451	0.922	0.307
1.417	1.490	0.642	1.340	0.482	1.412	0.462	1.345	0.307
1.843	1.590	0.545	1.514	0.570	1.632	0.477	1.750	0.249
0.567	0.244	0.467	0.214	0.516	0.310	0.414	0.393	0.361
0.106	-0.186	0.472	-0.052	0.621	0.038	0.447	0.185	0.444
-0.67	-0.370	0.507	-0.582	0.271	-0.676	0.041	-0.636	0.140
2.018	1.670	0.495	1.703	0.535	1.904	0.323	2.060	0.073
-0.245	-0.394	0.362	-0.449	0.343	-0.404	0.309	-0.226	0.254
0.116	-0.015	0.465	0.088	0.263	-0.032	0.427	-0.109	0.390
1.216	1.109	0.817	1.046	0.588	1.087	0.600	0.645	0.802
1.682	1.468	0.562	1.428	0.379	1.421	0.330	1.523	0.367
0.291	0.314	0.369	0.324	0.320	0.230	0.297	0.282	0.127
0.451	0.789	0.594	0.572	0.692	0.273	0.581	0.260	0.578

Table C6

RMSD of Discrimination Parameter Estimates for the 3PL-3PL (20 Items)

True values	$\bar{\hat{a}}$	RMSD	$\bar{\hat{a}}$	RMSD	$\bar{\hat{a}}$	RMSD	$\bar{\hat{a}}$	RMSD
	N=100		N=200		N=500		N=1000	
0.873	1.121	0.578	1.002	0.483	0.920	0.177	0.882	0.184
1.409	1.345	0.499	1.345	0.418	1.403	0.450	1.281	0.428
0.831	0.796	0.499	1.026	0.623	0.838	0.379	0.731	0.321
1.676	1.526	0.487	1.676	0.406	1.555	0.417	1.501	0.358
1.218	1.220	0.563	1.153	0.426	1.189	0.365	1.122	0.181
1.457	1.539	0.414	1.340	0.544	1.449	0.393	1.460	0.359
1.054	1.070	0.445	0.849	0.382	0.969	0.307	1.003	0.256
1.318	1.443	0.593	1.317	0.438	1.339	0.432	1.445	0.395
1.3	1.060	0.663	1.382	0.694	1.274	0.518	1.227	0.381
0.924	1.174	0.691	1.033	0.676	0.898	0.467	1.031	0.321
1.378	1.269	0.501	1.160	0.439	1.229	0.340	1.356	0.297
0.54	0.832	0.505	0.498	0.141	0.508	0.124	0.530	0.084
0.833	0.703	0.281	0.790	0.257	0.793	0.172	0.859	0.155
1.578	0.943	0.913	0.806	0.952	0.824	0.940	0.861	0.904
0.964	0.837	0.270	0.943	0.402	0.995	0.232	1.000	0.178
1.3	1.574	0.488	1.562	0.518	1.459	0.408	1.244	0.289
0.603	0.622	0.228	0.772	0.435	0.543	0.155	0.527	0.155
1.444	0.933	0.772	1.178	0.767	0.902	0.704	1.129	0.598
1.858	1.896	0.213	1.781	0.285	1.725	0.393	1.827	0.174
0.599	0.691	0.506	0.590	0.257	0.574	0.155	0.577	0.106

Table C7

RMSD of Difficulty Parameter Estimates for the 3PL-2PLG (20 Items)

True values	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD	$\bar{\hat{b}}$	RMSD
	N=100		N=200		N=500		N=1000	
-0.327	0.423	0.839	0.341	0.703	0.368	0.714	0.306	0.644
0.724	1.298	0.657	1.213	0.539	1.202	0.503	1.107	0.410
1.67	2.244	0.698	2.187	0.630	2.362	0.720	2.368	0.718
0.413	0.719	0.381	0.770	0.420	0.705	0.316	0.747	0.350
0.032	0.454	0.775	0.485	0.555	0.513	0.507	0.513	0.501
1.113	1.484	0.485	1.536	0.469	1.554	0.461	1.527	0.449
0.202	0.780	0.725	0.883	0.737	0.765	0.589	0.747	0.552
0.986	1.537	0.666	1.498	0.588	1.387	0.476	1.416	0.455
1.417	2.062	0.806	1.850	0.641	1.853	0.551	1.875	0.535
1.843	2.075	0.470	2.239	0.522	2.323	0.538	2.366	0.561
0.567	0.983	0.625	1.008	0.523	0.964	0.402	0.924	0.359
0.106	0.773	0.856	1.253	1.227	1.354	1.407	1.265	1.281
-0.67	0.502	1.377	0.202	0.967	0.023	0.713	0.031	0.711
2.018	2.315	0.382	2.358	0.375	2.419	0.417	2.457	0.442
-0.245	0.293	0.662	0.321	0.618	0.245	0.512	0.321	0.575
0.116	0.445	0.415	0.492	0.437	0.489	0.387	0.522	0.416
1.216	2.036	0.894	2.015	0.909	2.107	0.981	2.084	0.928
1.682	2.167	0.636	2.109	0.543	2.147	0.539	2.129	0.510
0.291	0.596	0.484	0.654	0.441	0.608	0.329	0.590	0.307
0.451	1.911	1.606	1.783	1.485	1.458	1.052	1.404	0.976

Table C8

RMSD of Discrimination Parameter Estimates for the 3PL-2PLG (20 Items)

True values	$\bar{\hat{a}}$	RMSD	$\bar{\hat{a}}$	RMSD	$\bar{\hat{a}}$	RMSD	$\bar{\hat{a}}$	RMSD
	N=100		N=200		N=500		N=1000	
0.873	1.193	0.628	1.077	0.433	1.024	0.218	0.971	0.179
1.409	1.654	0.455	1.574	0.325	1.601	0.296	1.665	0.321
0.831	1.078	0.429	1.078	0.439	0.917	0.155	0.934	0.158
1.676	1.786	0.328	1.767	0.285	1.861	0.261	1.809	0.184
1.218	1.237	0.592	1.318	0.486	1.310	0.315	1.295	0.179
1.457	1.620	0.343	1.583	0.386	1.534	0.279	1.592	0.269
1.054	1.275	0.475	1.086	0.354	1.173	0.253	1.160	0.182
1.318	1.546	0.505	1.620	0.515	1.555	0.367	1.477	0.217
1.3	1.759	0.532	1.678	0.519	1.486	0.423	1.405	0.293
0.924	1.640	0.823	1.405	0.640	1.244	0.509	1.112	0.291
1.378	1.521	0.526	1.434	0.334	1.493	0.214	1.531	0.194
0.54	0.888	0.583	0.577	0.192	0.535	0.188	0.557	0.131
0.833	0.712	0.357	0.818	0.276	0.836	0.211	0.930	0.207
1.578	1.468	0.481	1.515	0.384	1.636	0.253	1.592	0.238
0.964	0.875	0.305	1.042	0.386	1.096	0.216	1.064	0.181
1.3	1.694	0.465	1.631	0.476	1.556	0.328	1.447	0.244
0.603	0.817	0.461	0.775	0.315	0.715	0.230	0.710	0.164
1.444	1.617	0.486	1.591	0.445	1.537	0.311	1.515	0.249
1.858	1.859	0.266	1.742	0.328	1.804	0.232	1.859	0.139
0.599	0.759	0.622	0.650	0.302	0.664	0.143	0.667	0.121

Appendix D

Biases of Each Item Parameter Estimates for All Estimation Methods

Table D1

Difficulty Parameter Estimate Biases for the 2PLG-2PLG (20 Items)

True values	Sample Size			
	N=100	N=200	N=500	N=1000
-0.327	0.005	0.018	0.078	-0.023
0.724	0.139	0.099	0.118	0.032
1.67	0.077	0.021	0.138	0.079
0.413	0.018	0.049	-0.036	0.001
0.032	-0.257	-0.032	0.001	0.004
1.113	-0.024	-0.087	0.017	0.005
0.202	0.020	-0.041	-0.018	-0.023
0.986	0.106	0.060	-0.036	0.017
1.417	0.301	0.159	0.102	0.005
1.843	-0.273	-0.050	-0.107	-0.103
0.567	0.065	-0.006	0.004	-0.033
0.106	-0.135	-0.070	-0.095	-0.043
-0.67	0.507	0.277	0.083	0.092
2.018	-0.075	0.011	0.010	0.014
-0.245	-0.059	-0.059	-0.071	-0.024
0.116	-0.049	-0.010	-0.032	-0.024
1.216	0.403	0.194	0.150	-0.011
1.682	-0.105	-0.064	-0.059	0.019
0.291	-0.060	0.007	-0.034	-0.022
0.451	-0.020	0.044	0.004	-0.054

Table D2

Discrimination Parameter Estimate Biases for the 2PLG-2PLG (20 Items)

True values	Sample Size			
	N=100	N=200	N=500	N=1000
0.873	0.228	0.103	0.095	0.035
1.409	0.026	-0.001	-0.014	0.067
0.831	0.075	0.050	-0.067	-0.020
1.676	-0.071	-0.154	-0.014	-0.010
1.218	0.046	-0.063	-0.034	-0.017
1.457	0.236	0.218	-0.002	0.076
1.054	0.002	-0.024	0.046	-0.015
1.318	0.067	0.136	0.061	0.000
1.3	0.152	0.129	-0.023	-0.027
0.924	0.537	0.207	0.219	0.096
1.378	0.012	-0.011	0.024	0.032
0.54	0.359	0.067	0.002	-0.010
0.833	-0.014	-0.017	-0.010	0.051
1.578	-0.155	-0.165	-0.055	-0.050
0.964	-0.105	0.010	0.035	0.057
1.3	0.095	0.082	0.130	0.050
0.603	-0.052	-0.014	-0.002	0.045
1.444	0.231	0.140	0.147	0.018
1.858	-0.101	-0.117	-0.061	-0.015
0.599	0.201	0.108	0.026	-0.004

Table D3

Difficulty Parameter Estimate Biases for the 2PLG-3PL (20 Items)

True values	Sample Size			
	N=100	N=200	N=500	N=1000
-0.327	0.058	-0.233	-0.227	-0.169
0.724	-0.540	-0.501	-0.459	-0.575
1.67	-0.502	-0.756	-0.942	-0.865
0.413	-0.550	-0.341	-0.563	-0.597
0.032	-0.236	-0.658	-0.430	-0.279
1.113	-0.626	-0.629	-0.558	-0.468
0.202	-0.407	-0.663	-0.571	-0.828
0.986	-0.671	-0.757	-0.693	-0.396
1.417	-0.140	-0.556	-0.452	-0.667
1.843	-0.866	-0.826	-0.784	-0.636
0.567	-0.503	-0.426	-0.422	-0.653
0.106	-0.202	-0.540	-0.488	-0.203
-0.67	0.262	-0.341	-1.026	-0.893
2.018	-0.707	-0.748	-0.527	-0.477
-0.245	-0.100	-0.220	-0.637	-0.459
0.116	-0.147	-0.434	-0.387	-0.642
1.216	-0.712	-1.160	-0.683	-1.238
1.682	-0.560	-0.652	-0.764	-0.677
0.291	-0.424	-0.224	-0.466	-0.388
0.451	-0.130	-0.558	-0.423	-0.856

Table D4

Discrimination Parameter Estimate Biases for the 2PLG-3PL (20 Items)

True values	Sample Size			
	N=100	N=200	N=500	N=1000
0.873	0.473	0.125	0.223	0.128
1.409	-0.078	0.026	-0.036	-0.031
0.831	0.289	0.378	0.044	-0.100
1.676	-0.102	-0.044	-0.117	-0.292
1.218	0.353	-0.158	-0.035	0.083
1.457	-0.082	0.072	-0.179	-0.045
1.054	0.098	0.097	0.027	-0.090
1.318	0.070	0.009	-0.098	-0.073
1.3	-0.033	-0.017	0.043	0.041
0.924	0.310	-0.013	0.189	0.113
1.378	-0.006	0.078	0.151	-0.104
0.54	0.579	0.070	0.023	-0.023
0.833	-0.001	-0.046	-0.110	-0.086
1.578	-0.356	-0.284	-0.312	-0.183
0.964	0.042	0.185	0.095	0.082
1.3	0.309	0.125	0.267	-0.052
0.603	0.392	0.038	0.104	-0.038
1.444	0.118	-0.131	-0.250	-0.397
1.858	-0.073	-0.066	-0.173	-0.071
0.599	0.322	0.254	0.128	0.024

Table D5

Difficulty Parameter Estimate Biases for the 3PL-3PL (20 Items)

True values	Sample Size			
	N=100	N=200	N=500	N=1000
-0.327	0.117	-0.096	-0.280	-0.087
0.724	-0.167	-0.126	-0.125	-0.314
1.67	-0.043	-0.153	0.113	-0.229
0.413	-0.430	-0.197	-0.162	-0.199
0.032	-0.124	-0.295	-0.143	-0.148
1.113	-0.199	-0.297	-0.081	-0.073
0.202	-0.372	-0.458	-0.280	-0.244
0.986	-0.104	-0.195	-0.243	-0.064
1.417	0.073	-0.077	-0.005	-0.072
1.843	-0.253	-0.329	-0.211	-0.093
0.567	-0.323	-0.353	-0.257	-0.174
0.106	-0.292	-0.158	-0.068	0.079
-0.67	0.300	0.088	-0.006	0.034
2.018	-0.348	-0.315	-0.114	0.042
-0.245	-0.149	-0.204	-0.159	0.019
0.116	-0.131	-0.028	-0.148	-0.225
1.216	-0.107	-0.170	-0.129	-0.571
1.682	-0.214	-0.254	-0.261	-0.159
0.291	0.023	0.033	-0.061	-0.009
0.451	0.338	0.121	-0.178	-0.191

Table D6

Discrimination Parameter Estimate Biases for the 3PL-3PL (20 Items)

True values	Sample Size			
	N=100	N=200	N=500	N=1000
0.873	0.248	0.129	0.047	0.009
1.409	-0.064	-0.064	-0.006	-0.128
0.831	-0.035	0.195	0.007	-0.100
1.676	-0.150	0.000	-0.121	-0.175
1.218	0.002	-0.065	-0.029	-0.096
1.457	0.082	-0.117	-0.008	0.003
1.054	0.016	-0.205	-0.085	-0.051
1.318	0.125	-0.001	0.021	0.127
1.3	-0.240	0.082	-0.026	-0.073
0.924	0.250	0.109	-0.026	0.107
1.378	-0.109	-0.218	-0.149	-0.022
0.54	0.292	-0.042	-0.032	-0.010
0.833	-0.130	-0.043	-0.040	0.026
1.578	-0.635	-0.772	-0.754	-0.717
0.964	-0.127	-0.021	0.031	0.036
1.3	0.274	0.262	0.159	-0.056
0.603	0.019	0.169	-0.060	-0.076
1.444	-0.511	-0.266	-0.542	-0.315
1.858	0.038	-0.077	-0.133	-0.031
0.599	0.092	-0.009	-0.025	-0.022

Table D7

Difficulty Parameter Estimate Biases for the 3PL-2PLG (20 Items)

True values	Sample Size			
	N=100	N=200	N=500	N=1000
-0.327	0.750	0.668	0.695	0.633
0.724	0.574	0.489	0.478	0.383
1.67	0.574	0.517	0.692	0.698
0.413	0.306	0.357	0.292	0.334
0.032	0.422	0.453	0.481	0.481
1.113	0.371	0.423	0.441	0.414
0.202	0.578	0.681	0.563	0.545
0.986	0.551	0.512	0.401	0.430
1.417	0.645	0.433	0.436	0.458
1.843	0.232	0.396	0.480	0.523
0.567	0.416	0.441	0.397	0.357
0.106	0.667	1.147	1.248	1.159
-0.67	1.172	0.872	0.693	0.701
2.018	0.297	0.340	0.401	0.439
-0.245	0.538	0.566	0.490	0.566
0.116	0.329	0.376	0.373	0.406
1.216	0.820	0.799	0.891	0.868
1.682	0.485	0.427	0.465	0.447
0.291	0.305	0.363	0.317	0.299
0.451	1.460	1.332	1.007	0.953

Table D8

Discrimination Parameter Estimate Biases for the 3PL-2PL (20 Items)

True values	Sample Size			
	N=100	N=200	N=500	N=1000
0.873	0.320	0.204	0.151	0.098
1.409	0.245	0.165	0.192	0.256
0.831	0.247	0.247	0.086	0.103
1.676	0.110	0.091	0.185	0.133
1.218	0.019	0.100	0.092	0.077
1.457	0.163	0.126	0.077	0.135
1.054	0.221	0.032	0.119	0.106
1.318	0.228	0.302	0.237	0.159
1.3	0.459	0.378	0.186	0.105
0.924	0.716	0.481	0.320	0.188
1.378	0.143	0.056	0.115	0.153
0.54	0.348	0.037	-0.005	0.017
0.833	-0.121	-0.015	0.003	0.097
1.578	-0.110	-0.063	0.058	0.014
0.964	-0.089	0.078	0.132	0.100
1.3	0.394	0.331	0.256	0.147
0.603	0.214	0.172	0.112	0.107
1.444	0.173	0.147	0.093	0.071
1.858	0.001	-0.116	-0.054	0.001
0.599	0.160	0.051	0.065	0.068

VITA

Graduate School
Southern Illinois University

Song Gao

gaosong0620@yahoo.com

Dalian University of Science & Technology
Bachelor of Science, Chemical Engineering, July 1990

Southern Illinois University Carbondale
Master of Science, Workforce Education & Development, May 2002

Dissertation Title:
The Exploration of the Relationship Between Guessing and Latent Ability in IRT Models

Major Professor: Todd Headrick