The Report Committee for Aline Pinto Orr
Certifies that this is the approved version of the following report:

Effects of sample size, ability distribution, and the length of Markov Chain Monte Carlo burn-in chains on the estimation of item and testlet parameters

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor: _____
Barbara G. Dodd

_____
Youngsuk Suh

Effects of sample size, ability distribution, and the length of Markov Chain Monte Carlo burn-in chains on the estimation of item and testlet parameters

by

Aline Pinto Orr, B.A., M.S., Ph.D.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

May 2011

Abstract

Effects of sample size, ability distribution, and the length of Markov Chain Monte Carlo burn-in chains on the estimation of item and testlet parameters

Aline Pinto Orr, M.A.

The University of Texas at Austin, 2011

Supervisor: Barbara G. Dodd

Item Response Theory (IRT) models are the basis of modern educational measurement. In order to increase testing efficiency, modern tests make ample use of groups of questions associated with a single stimulus (testlets). This violates the IRT assumption of local independence. However, a set of measurement models, testlet response theory (TRT), has been developed to address such dependency issues. This study investigates the effects of varying sample sizes and Markov Chain Monte Carlo burn-in chain lengths on the accuracy of estimation of a TRT model's item and testlet parameters. The following outcome measures are examined: Descriptive statistics, Pearson product-moment correlations between known and estimated parameters, and indices of measurement effectiveness for final parameter estimates.

Table of Contents

Introduction

This study focuses on the methods used to estimate item parameters when groups of multiple-choice questions are associated to a single stimulus. This set up has appeared in standardized tests for decades and has been commonly referred to as context-dependent item sets (Haladyna, 1992a), item bundles (Rosenbaum, 1988), and testlets (Wainer & Kiely, 1987). One of the main reasons for the use of testlets is the efficiency that such groups of items provide in the testing context (Wainer & Kiely, 1987) and the increase in validity that can be achieved with this set up (Haladyna, 1992b). However, when items are associated with a single stimulus, a subject's response to one item may not be independent of his/her responses to the other items in the testlet. In other words, a dependency effect may arise that violates the Item Response Theory (IRT) assumption of local independence between items (Bradlow, Wainer, & Wang, 1999).

IRT models provide a method for obtaining student ability estimates that are independent of the items being used, and test statistics (such as item difficulty) that are independent of the sample of examinees used to calibrate the test (Hambleton & Cook, 1977). However, most IRT models make two fundamental assumptions, the assumption that all items on a test are locally independent and the assumption that there is only one latent trait being measured. The violation of these assumptions can interfere with estimation of the item characteristics and affect the estimation of the latent trait being measured.

In educational measurement, IRT defines the probability of a correct response as a function of the item statistics and the ability level of an examinee (Embretson & Reise,

2000). In addition, based on this probability, IRT provides a method for estimating how much information a specific item contributes to the overall information a test provides about an examinee's ability level (Lord, 1980). However, optimal estimation of item information functions depends on accurate item parameter estimation (Hambleton, 1994).

A consequence of ignoring the violation or local independency is the inaccurate estimation of item parameters and information functions which results in erroneous estimation of the precision of measurement (Hambleton, 1994). In order to address this dependency issue, testlet response theory (TRT) models were developed as modifications of IRT models, and include a person parameter (Wainer, Bradlow, & Wang, 2007) that addresses the dependency between items. Hence the TRT models not only include the parameters present in the IRT models but also an extra parameter to be estimated per testlet for each examinee. The estimation of item and person parameters for the TRT models is conducted within a Bayesian framework, and prior information about the items and the examinees' ability is combined with the observed data to obtain a joint posterior probability distribution for the parameters of interest. A Markov Chain Monte Carlo (MCMC) algorithm is then used to randomly sample from this posterior distribution and to estimate item and person parameters (Novick & Jackson, 1974).

However, several aspects of the MCMC method can affect the accuracy of the estimation. The selection of the probability distributions (representing our prior knowledge of the students' ability and of characteristics of the items), the size of the calibration sample, and the length of the Markov chain (Gilks, Richardson, & Spiegelhalter, 1998) are important for the successful construction of the posterior

distribution and for accurate estimation of item and person parameters. A prior distribution that does not correctly reflect the distribution of item characteristics (such as difficulty and discrimination) or the examinees' ability will hinder the accuracy of item and person parameter estimation. This is especially true when the calibration sample size is small and the prior distribution is extreme (Gao & Chen, 2005; Gifford & Swaminathan, 1990).

In addition, in the early iterations of the Markov chain the item and person parameter values sampled in consecutive iterations suffer from high auto correlation. The high correlation may result in the initial values of the chain having a large influence on the final parameter estimates, and in biased Monte Carlo standard error estimates (Gilks, et al., 1998). Researchers have suggested two main ways of addressing the initial autocorrelation issue. The first recommendation has been to obtain a large number of iterations (what ensures that the posterior distribution of interest has been thoroughly sampled from) and to keep every n^{th} iteration of the chain. This procedure would simultaneously reduce the autocorrelation between initial chain values and reduce the final amount of data to be stored and processed. However, evidence indicates that using every n^{th} iteration of the chain may negatively affect the precision of the parameter being estimated (MacEachern & Berliner, 1994).

The second recommendation involves discarding the set of iterations prior to chain convergence (referred to as the burn-in phase of the MCMC chain) before final estimation of parameters. However, the identification of chain convergence and the choice of the number of iterations to discard are not always easily discerned. In addition,

the specific effect that non-convergent Bayesian chains have on item parameter estimation has not yet been investigated.

The main purpose of this study is to investigate the effects of different sample sizes, examinee ability distributions, and MCMC chain lengths on the estimation of item parameters for the 3PL TRT model.

Integrative Analysis and Interpretation

The purpose of this integrative analysis and interpretation is to provide the theoretical framework for investigating the effects of several conditions on the accuracy of item and testlet parameters estimation for the 3PL TRT model. This section is divided into four parts: item response theory (IRT), testlet response theory (TRT), an overview of parameter estimation using a Bayesian method, and the statement of research purpose.

Item Response Theory

In classical test theory, a measurement of ability is dependent on the instrument used for measuring. Hence, a harder set of questions may result in lower scores whereas an easier set of questions may result in higher scores. In addition, the determination of whether a test is hard or easy depends on the sample of examinees taking the test and their ability level. For example, for a sample of examinees with higher average ability, a test may be characterized as easy, whereas the same test being administered to a sample of examinees with lower average ability may be characterized as harder. IRT on the other hand, provides us with parameter invariance. This means that, within a linear transformation, the same estimate of an item's parameters will be obtained regardless of the examinee's ability level, and the same ability estimate will be obtained regardless the item's parameters (Lord, 1980).

IRT is a collection of mathematical models that outline the relationship between a person's probability of a given response and the trait level assumed to underlie that performance (Hambleton & Cook, 1977). This relationship can be represented

graphically in the form of item characteristic curves (ICC). In the context of educational tests, an ICC plots the probability of responding correctly to an item as a function of the latent trait level (in this case ability). When IRT is applied to responses that can only be scored as right or wrong, ICCs tend to assume an S shape, corresponding to a monotonically increasing probability function

IRT Models

IRT models can be grouped into two families, defined by the way in which items are scored. The first family consists of dichotomous models, which treat item responses as binary possibilities and estimate the probability of a response in one of the two categories. For example, the response to an item can be correct or incorrect, agree or disagree, or success or failure. The second family consists of polytomous models, which allow scoring of items with multiple response categories. For example, in attitude surveys a subject can select one of several response options. Another example would be the scoring of essays, where the rubric often allows for partial credit and more points are awarded to better responses. These models take into consideration the probability of an examinee responding in each category of an item when calculating the total item response probability. This study focuses on dichotomous items. Therefore, while the dichotomous models are discussed in detail in this document, only a brief and superficial discussion of polytomous models is provided in the context of polytomous IRT scoring of testlets.

IRT Assumptions

There are three fundamental assumptions made by dichotomous IRT models. The first one is the assumption of unidimensionality. This occurs when only one construct or factor is expected to account for the variance in the responses to items in a test.

The second assumption is that a mathematical function can be derived to model the probability of a given response to an item conditional on trait level (Hambleton & Swaminathan, 1985). In other words, it assumes that an ICC describes the true relationship between the latent trait (ability) and the item responses, that is, as the ability level increases the probability of a correct response increases monotonically.

The third one is the assumption of local independence. According to this assumption, conditional on the ability level, the probability of responding to an item is statistically independent of the probability of responding to any other item (Embretson & Reise, 2000). For example, the content provided in one item should not aide an examinee in answering any other items. In other words, taking an examinee's ability level into account, the examinee's response to one item should not influence his/her responses to any other items in the test. When items are locally independent, the probability of a response pattern for a set of items (at a given ability level) is equal to the product of the probabilities of the examinee's response to each item (Hambleton & Swaminathan, 1985). Violation of this assumption occurs when examinees' responses to test items are conditionally correlated (Wainer, Bradlow, & Du, 2000). In which case, the probability of a response pattern is less than the product of the probabilities of responses to the individual items (Devore, 2007).

Dichotomous IRT Models

Dichotomous IRT models are appropriate for questions whose responses can be classified as either correct or incorrect, resulting in a binary scoring of 1 (for a correct response) and 0 for an incorrect response. The dichotomous IRT models are the one parameter logistic (1PL), the two parameters logistic (2PL), and the three parameters logistic (3PL), each being a generalization of the previous one.

The 1PL model (Rasch, 1960) takes into consideration the ability level of the examinee and the difficulty of the item when estimating the probability of a correct response to that item. For this model it is assumed that all items in a test have the same level of discrimination and that no guessing occurs when an examinee responds to the items. For the 1PL model, the probability that an examinee *i*, with a certain ability level will respond correctly to an item *j* of a certain difficulty, $P_{ij}(y_j = 1 \mid \theta_i)$, is given by the following expression,

$$P_{ij}(y_j = 1 | \theta_i) = \frac{e^{(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}} \quad (1)$$

Where y_j=1 corresponds to a correct response to item j, θ represents the ability level of a given examinee and b_j represents the item difficulty. The difficulty parameter indentifies the relative easiness of an item and places it on the same scale as ability. An item's difficulty is defined as the point along the ability scale at which the slope of the probability function (the ICC) reaches its maximum (the point of inflection). For the 1PL model this corresponds to a .50 probability of giving a correct response.

The 2PL model (Birnbaum, 1958, 1968) is an extension of the 1PL model that takes into account the examinee's ability level, the difficulty of the items, and the discrimination capacity of each item. The discrimination parameter indicates how well an item distinguishes lower ability examinees from higher ability examinees. In addition, this model assumes that no guessing occurred when the examinees provided their answers. For the 2PL model, the probability that an examinee *i*, with a certain ability level will respond correctly to an item *j* of a certain difficulty, $P_{ij}(y_j = 1 \mid \theta_i)$, is given by the following expression,

$$P_{ij}(y_j = 1 \mid \theta_i) = \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} \quad (2)$$

Where y_j=1 corresponds to a correct response to item j, θ_i represents the ability level of a given examinee, b_j represents the item difficulty, and a_j represents the discrimination capacity of the item. For the 2PL, an items' difficulty is defined in the same way as for the 1PL model. The item discrimination is proportional to the slope of ICC at the point of inflection, consequently the higher the slope of the probability function the higher the item discrimination.

Finally, the 3PL model (Birnbaum, 1968) is an extension of the 2PL model. In addition to the examinee' ability level, the item difficulty, and the discrimination of the item, this model includes a pseudo-guessing parameter that accounts for the possibility that examinees will answer an item correctly by guessing. This is likely to occur when an individual encounters an item that is more difficult than his/her ability level, a scenario

that is more probable for individuals at the low end of the ability scale. Consequently, by introducing a guessing parameter, the 3PL model accounts for the performance of individuals in the low end of the ability scale. For this model, the probability that an examinee *i*, with a certain ability level will respond correctly to an item *j* of a certain difficulty, $P_{ij}(y_j = 1 | \theta_i)$, is given by the following expression,

$$P_{ij}(y_j = 1 | \theta_i) = c_j + (1 - c_j) \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} \quad (3)$$

Where y_j=1 corresponds to a correct answer to item *j*, θ_i represents the ability level of a given examinee *i*, b_j represents the difficulty of item *j*, a_j represents the item discrimination capacity of item *j*, and c_j represents the pseudo guessing parameter for item *j*. The item discrimination (a_j) is defined in the same way as for the 2PL model. i.e., it is proportional to the slope of the item characteristic curve at the point of inflection. For the 3PL model, the item difficulty parameter corresponds to the ability level at which the probability of a correct response equals $(1 + c_j)/2$. The pseudo guessing parameter represents a non-zero probability of success for examinees with low ability and is indicated by the lower asymptote of the item characteristic curve.

Item Information Function

The item information function quantifies the precision of measurement for θ at each level on the ability scale. It is denoted $I_j(\theta)$ and is expressed as

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)} \qquad \text{for } i = 1, 2, \ldots, n$$

(4)

Where $P_i(\theta)$ is the probability of a correct response by examinee i given an ability level θ. The term $Q_i(\theta)$ corresponds to the probability of an incorrect response $(1 - P(\theta))$ by examinee i and $P_i'(\theta)^2$ is the first derivative of the item response curve evaluated at a particular θ level squared. The higher the information function at a particular θ value, the more precisely the item can measure examinees at that θ level. The first derivative of a function corresponds to the slope of that function, what indicates that the precision contributed by the item is closely related to the slope of the item characteristic function at a particular θ level. This is important because the discrimination of an item is proportional to the slope of the item characteristic function at the point of inflection, indicating a close relationship between the amount of information an item can provide and the discrimination parameter for that item (Embretson & Reise, 2000).

Item information functions can be added to create a test information function, $TI(\theta)$,

$$TI(\theta) = \sum I(\theta)$$

(5)

The test information function can be used to evaluate the measurement precision of a test at different levels of the ability scale. This is done by calculating the standard error of the

ability estimate, *SE(θ)*, at each ability level. The standard error of θ estimate is related to the test information function by the following formula:

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}} \quad (6)$$

This knowledge can be used in the construction of tests. For example, if we assume the examinees' ability will be normally distributed, we might desire a test that provides most information, and highest precision of measurement, at the center of the ability scale, rather than on the lower or upper ends of the scale. The test items would then be selected based on the items' information functions to provide most precision of measurement at the center of the θ scale.

The use of item information functions in test development relies on the appropriate estimation of the item parameters and on the fit of the IRT model to the data. When item calibration and model fit are poor, the item information functions will be misleading and will result in inappropriate test construction (Hambleton, Swaminathan, & Rogers, 1991).

Person Parameter Estimation

IRT scoring methods attempt to estimate an examinee's ability based on that examinee's pattern of responses to items on a test and based on the characteristics of those items. In such cases, the item parameters are assumed to be known and the estimation error is ignored when calculating the examinees' ability. Three popular IRT

strategies for scoring dichotomous or polytomous items are the maximum likelihood estimation (MLE), the maximum a posteriori (MAP), and the expected a posteriori (EAP).

The MLE procedure finds the examinee's ability level that maximizes the likelihood of an examinee's response pattern. In other words, if finds which ability level has the highest likelihood of producing a specific response pattern. This can be thought of as a graph of a likelihood function, with the ability scale represented in the x-axis and the likelihood of a particular response pattern on the y-axis. For each ability value, we can calculate the likelihood that a specific response pattern would occur, and the maximum likelihood of this function is the ability value (the value on the x-axis) for which the graphed function corresponds to the highest y value (the highest likelihood).

The likelihood function of an examinee's response pattern corresponds to the product of the response functions for each individual item. For example, if a subject responds correctly to the first 2 items on a sequence a wrong to the third item (1,1,0), we can find the likelihood function of this pattern by multiplying the individual item response curves P1(θ), P2(θ), and Q3(θ). Or the natural logarithm of the item response curves can be taken and instead of being multiplied, the item response curves are added to each other. One common way of finding the maximum point in a log-likelihood function is by using an iterative procedure called the Newton-Raphson procedure.

This algorithm finds the mode of each examinee's log-likelihood function. It starts at an arbitrary trait level and calculates the first and second derivatives of the log-likelihood function at this θ value. The first derivative of the log-likelihood function

represents the slope of the function, whereas the second derivative corresponds to the rate of change in the first derivative (i.e., the rate of change in the slope). The ratio of the first derivative to the second derivative is calculated and subtracted from the initial ability estimate to generate a new updated ability estimate. Since the first derivative is the slope of the function, and we are interested in the highest point of this function (i.e. the point at which the slope is zero), we will find the final ability estimate when the change between the current and the new estimated theta is negligible (such as a difference of less than 0.001).

The ML method has the advantage of not being biased, and consequently the estimated value of θ is a close reflection of the true θ. In addition, this is an efficient estimator with errors that are normally distributed. However, these qualities are true for large samples (reflecting longer tests), and depend on the assumption that the examinees' responses fit the model (Bock & Mislevy, 1982). In addition, a problem with MLE is that a trait level can not be estimated for examinees with all-correct or all wrong response patterns. In such cases the likelihood function tends to infinity and the function does not have a mode (Baker & Kim, 2004). One way of addressing this issue is by incorporating prior information about the examinees' ability distribution into the estimation algorithm. Two methods that use prior information in the estimation of students' ability, and are considered Bayesian estimation procedures, are the Maximum a Posteriori (MAP) and the Expected a Posteriori (EAP).

The MAP method addresses this issue by incorporating prior information about the examinees ability into the likelihood function. The prior distribution is a hypothetical

probability distribution from which it is assumed that the examinees are a random sample. It is commonly assumed that examinees are sampled from a normal distribution with a mean ability of zero and a variance of 1.0. In MAP the prior distribution is multiplied by the likelihood function to create a posterior distribution, and the mode of this distribution is used as the ability estimate. In MAP, because the prior distribution is incorporated into the likelihood function, the shorter the test the more influence the prior will have on the final ability estimate. Two advantages of MAP are the fact that this method can produce ability estimates for extreme responses (such as all right or all wrong) and the fact that, because there is more information about examinees, the estimates have lower standard errors. However, because a prior distribution is incorporated into the likelihood function, the MAP estimates are biased towards the mean of the prior distribution. This is especially problematic in short tests and when the prior distribution assumed does not correspond to the real ability distribution of the examinees.

The EAP method is similar to MAP in the sense that it uses a prior distribution in estimating examinees' ability levels. However, EAP is a non-iterative method that finds the mean of the posterior distribution (instead of the median). In EAP, for each set of test items, a fixed number of θ values are specified (these θ levels are referred to as quadrature nodes) and a probability or weight is computed at each of these pre-specified levels. These weights serve as a discrete (as opposed to continuous) prior distribution. Once the quadrature nodes and the weights are established, a trait level estimate is calculated (and corresponds to the mean of the posterior distribution).

Both MAP and EAP produce θ level estimates and standard errors that are similar because both make use of prior information in their θ estimation. In addition, both MAP and EAP can estimate extreme response strings such as all correct or all wrong. However, both methods have a tendency for the ability estimate to be biased towards the mean of the prior distribution, which is referred to as regression towards the mean (Lord, 1986). In addition, in order for both methods to produce accurate estimates, it is assumed that the prior distribution being used is an accurate representation of the examinees ability distribution. This is especially true for short tests.

Item Parameter Estimation

Item parameters are estimated from the data during test standardization and when new tests are being implemented. Two commonly used MLE methods are the joint maximum likelihood estimation (JMLE) and the marginal maximum likelihood estimation (MMLE). The two methods differ in the way the probability of the observed response patterns is conceptualized.

The JMLE method maximizes the joint likelihood function of both person and item parameters in order to simultaneously estimate the trait level and the item parameters. The likelihood function is the probability of a person's responses to a test. In other words, it is the product of an examinee's response probabilities (conditional on the examinee's ability level and item parameters) across all items on a test. The joint likelihood function is the product of likelihood functions across all examinees, and represents the encounter of each examinee with each particular item.

In the first step, the algorithm starts with arbitrary provisional values for the examinees' θ, such as the mean of an assumed ability distribution, and uses these provisional values to estimate the item parameters. The estimation of item parameters is done with the MLE method. It takes place first because typically there are a larger number of examinees than items and therefore there is more information for estimating the item parameters. In the next step, the item parameters are treated as known and are used in the MLE procedure to estimate the examinees' θ level. In the next iteration of the procedure, the item parameters are re-estimated using the newly estimated person parameters followed the person parameters being re-estimated using the new item parameter estimates. This back and forth process continues until successive improvements in the examinee's ability estimates and the item parameter estimates are less than a pre-established convergence criterion.

The JMLE is applicable to several IRT dichotomous models and is computationally efficient. However, the JMLE item parameter estimates tend to be biased for short length tests, and consequently, the standard errors are difficult to interpret. In addition, the item parameter estimates are inconsistent for fixed length tests. This means that the addition of more examinees to the procedure does not result in improved estimates because as we add more examinees we also add more parameters to be estimated. Last, the JMLE procedure (as well as all other MLE procedures) can not estimate item or person parameters for extreme response patterns, such as all right or all wrong answers(Embretson & Reise, 2000).

The MMLE method separates the process of estimating item parameters from the process of estimating person parameters. It first estimates only item parameters and once a satisfactory model-data fit is obtained, it proceeds to estimate the person parameters. The MMLE procedure can be divided into two stages, the expectation and the maximization stages. In the expectation stage, the algorithm handles the unknown person parameters by expressing the probability of response patterns as expectations (or means) from a population trait distribution. In other words, the probability of the examinees' trait level is specified by a probability distribution that is based on knowledge of the population or on the test data. For example, for each response pattern, the number of persons with the pattern is noted. Then for each particular ability level, the probability of a response pattern can be computed from the IRT model being used. Hence, the algorithm assumes an ability distribution for the population, and the expected number of persons passing each particular item is computed for each trait level.

In the maximization stage, these expectations are used to obtain item parameters estimates that maximize the likelihood functions. A second expectation step uses the item parameters to re-calculate the expectations that are then used in a second maximization stage to re-estimate the item parameters. This back and forth procedure is employed until changes in the item parameter estimates are smaller than a pre-specified convergence criterion. After the final item parameter estimates are obtained, a separate algorithm such as the MLE, the MAP, or the EAP can be employed to estimate the examinees' ability level.

Some of the advantages of the MMLE method are that it can be applied to all types of IRT models, it is efficient for both long and short tests, and estimates can be obtained for extreme response strings. However, this method must assume an ability level distribution and consequently, the item parameter estimates are contingent on the appropriateness of this assumption.

In order to estimate item parameters, both methods make the assumption of local independence. Local independence means that given an ability level, the response to an item is unrelated to the responses to any other item in the test. Hence, after controlling for the trait level, local independence implies that no relationships remain between the test items. However, this assumption may be violated when several test items refer to one common stimulus. For example, when a passage is presented in a reading and comprehension test and the following questions refer to the passage. In such situations an alternative estimation method has been suggested. The next few sections of this document will discuss the use of testlets and the methods that have been employed in estimating item parameters when a test is composed of testlets.

Testlets

The term testlet refers to a group of items that is developed as a unit, relates to a single concept area, and contains a fixed and predetermined number of paths that an examinee can follow. For example, a testlet may consist of a group of multiple choice questions referring back to a single passage, or a group of math questions referring to a single diagram (Wainer & Kiely, 1987). One advantage of using testlets is that they

increase efficiency in situations where the ability to understand a stimulus (such as a passage or a diagram) is being examined and a substantial amount of time is necessary for processing the stimulus. In such cases, it seems inefficient to use a large amount of testing time for a student to process a stimulus and to ask only one question about it (Wainer, et al., 2007). In addition, testlets can be used to examine multistep problem solving behaviors, which permits a more meaningful and valid interpretation of higher order thinking (Haladyna, 1992b). For example, context dependent item sets have been used for examining and scoring mathematical problem solving and for relating the observed item responses to stages of cognitive development (Biggs & Collis, 1982).

However, grouping questions around a single stimulus may result in dependency between the items. This is a possibility because all items within the set require appropriate analysis and interpretation of a single stimulus in order for the correct answer to be selected. Hence, a misinterpretation can result in more than one incorrect response. This characteristic of testlets allows for the possibility of correlated errors of measurement within the testlet (Crehan, Sireci, Haladyna, & Henderson, 1993), which would violate the IRT assumption of local independence. It has been demonstrated that ignoring the dependency between items, and scoring the test according to a dichotomous IRT model, may result in overestimation of the precision of proficiency estimates and in a bias in the estimation the difficulty and discrimination parameters (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989).

The next session discusses two methods that have been used to address the dependency effect in testlets. The first method uses polytomous IRT models to score the

items in a testlet as a single unit, whereas the second method implements a derivation of the 3PL IRT model and takes into account the dependency between the items in a testlet.

Polytomous IRT Scoring of Testlets

Previous studies have shown that applying dichotomous IRT models to data where local dependency is present results in overestimation of the precision of measurement (Sireci, et al., 1991). This issue arises because the probability of two or more dependent events occurring is less than the probability of two or more independent events occurring. One approach to managing local dependence proposed by Thissen and coleagues (1989) has been to consider the whole testlet as the unit of measurement and to apply a polytomous IRT response model. Several polytomous IRT response models have been developed and studied, however only a few of these models have been used in testing programs. This section of the document gives a general description of polytomous models and how these models can be used to score testlets.

Polytomous IRT models allow for scoring of partially correct answers and are used when item responses are allowed to have multiple categories. In such cases, the model represents the nonlinear relationship between an examinee's trait level and the probability of responding in a particular category. When used to score testlets, polytomous models consider the testlet as the unit of measurement. In other words, instead of calibrating each item that composes the testlet individually using a dichotomous IRT model, the items in the testlet are combined and calibrated as a single polytomous item. The testlet, represented as a polytomous item, is scored from zero to the total number of items

associated with the common stimulus. This approach is considered a practical method for accounting for the dependency across the items in the testltet (Wainer, 1995) but it doesn't provide information about examinees' response patterns within the testlets. Because the testlet is scored as the total number of items correct, there is no distinction about which items were answered correct.

Testlet Response Theory

Testlet response theory (TRT) was designed as an extension of IRT dichotomous models that accounts for the local dependency between items within a testlet. This method treats the items in a testlet as independent units of measurement (Wainer, et al., 2007) and estimates a testlet effect for each examinee. The testlet effect, measured by the gamma parameter, represents the interaction of person *i* with item *j* that is nested within a specific testlet. The dependency between items in a testlet (for a given examinee) is modeled by the gamma parameter, and all items in a testlet share the same gamma parameter in their scoring equation.

Dichotomous TRT models

Two dichotomous models have been created as generalizations of the 2PL and the 3PL IRT models, the two parameter and three parameter TRT models (Bradlow, et al., 1999; Wainer, et al., 2000). The study proposed in this document involves the 3PL TRT model and consequently only this model will be discussed in detail.

In order to account for the local dependency between items in a testlet, the 3PL TRT model includes as random effect parameter to account for the shared variance among items within a testlet. This is referred to as the testlet effect parameter. The model includes the 3 item parameters present in the 3PL IRT model, difficulty (b), discrimination (a), and pseudo-guessing (c), and two person-specific parameters, ability (θ) and the testlet effect ($\gamma_{id(j)}$).

The testlet effect parameter models the local dependency between testlet items by including the same random effect for each item within a testlet. This common random effect across items accounts for the communality created by the items' association with the same stimulus (Wainer, et al., 2000). When items are calibrated for the TRT model, the gamma parameter retrieved for each testlet corresponds to the estimated variance of the testlet effect and can be used as a measure of local item dependence. The testlet effect parameter used in the TRT model is a random variable selected from a normal distribution with mean of zero and standard deviation equal to the square root of the variance of the testlet effect (for a given testlet).

The probability of person *i*, with ability level θ_i, correctly answering item *j* within testlet $d_{(j)}$ was denoted as

$$P_{ij}(y_j = 1|\theta_i) = c_j + (1 - c_j)\frac{\exp(a_j(\theta_i - b_j - \gamma_{id(j)})}{1 + \exp(a_j(\theta_i - b_j - \gamma_{id(j)})} \quad (7)$$

where the parameters a_j, b_j, and c_j represented the item discrimination, item difficulty, and pseudo-guessing parameter respectively for item *j*. The additional parameter $\gamma_{id(j)}$ modeled the dependency for person *i* responding to item *j* nested within testlet *d(j)*.

Wainer et al. (2000) compared the performance of the 3PL IRT and the 3PL TRT models when applied to Graduate Record Examination (GRE) data. The authors found evidence of substantial testlet effect variance that was not accounted for by the 3PL IRT model. Consequently, the IRT model produced significantly larger discrimination estimates for the discrimination parameter (α_j) than the 3PL TRT mode,l which resulted in inflation of item and test information functions and in the under estimation of the standard error of ability estimation (Wainer, et al., 2007).

The 3PL TRT model was implemented within a Bayesian probability model that provides a joint probability distribution for all observable and unobservable quantities. This means that prior information about the test items and the characteristics of the examinees can be incorporated into the test calibration and into the examinees' ability estimation by assigning a probability distribution for each of the parameters of interest. In TRT calibration, a Markov Chain Monte Carlo algorithm combined with Gibbs Sampling is used to produce a posterior distribution for each of the parameters of interest. The posterior distributions can then be used to compute summary statistics for the parameters such as mean and standard deviation.

Probabilities

Before discussing the MCMC method for estimating the TRT item parameters, some consideration is given to probability theory and how it fits in the Bayesian methodology.

A probability is defined as the numerical likelihood, measured between 0 and 1, that an uncertain event will occur. In order to determine the probability of an event, we must define the sample space that includes the outcome of interest. For example, the outcome of interest might be obtaining a head after a flip of a coin. So, if we have a course of action (referred to as a random experiment) whose outcome is uncertain but includes the event of interest, we can create a list of all possible and mutually exclusive outcomes for that course of action. This list is the sample space of our random experiment.

Once the sample space is determined, the probabilities for each possible outcome are typically assigned based on certain assumptions (such as independence) and based on previously observed data. For example, considering the flip of a coin, our sample space is head or tail. After repeating the coin flip an infinite number of times we come upon the observation that we get heads ½ of the times and tails ½ of the times. Hence, we assign a probability of .5 to heads, and the same probability to tails. In this example, each coin toss is an independent event (because obtaining a head in one coin toss, does not affect the likelihood of obtaining a head on the next coin toss) and heads and tails have equal likelihood.

It is important to note that an event does not have to correspond to a single outcome, an event can be defined as a collection of one or more individual outcomes. For example, an event can correspond to obtaining three heads out of five coin-tosses, and the likelihood of different numbers of heads out of n coin-tosses can be described by a probability distribution. For continuous variables, the likelihood of the variable assuming certain values is described by a density function, and is calculated by integrating the area under the density curve.

Bayesian Theory

We are typically interested in the probability of outcomes that result from combining various events in various ways. For example, the joint probability of event A and B, correspond to the probability of the event of *A* and *B* occurring together, and is denoted $P(A \cap B)$. Whereas the conditional probability is used to determine how two events are related, or in other words, the probability that an event will occur, given that a related event has occurred. For example, if we are interested in the probability of event A happening given event B has occurred, this is denoted by *P(A|B)* and is calculated as:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{8}$$

By moving the terms of the equation above, we also have:

$$P(A \cap B) = P(A \mid B)P(B) \tag{9}$$

Whereas the probability of event B occurring given A has taken place is calculated as:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{10}$$

By moving the terms of the equations above we obtain the following:

$$P(A \cap B) = P(B \mid A)P(A) \tag{11}$$

Since the right side of equations 9 and 11 are both equal to $P(A \cap B)$, we can combine both equations to obtain:

$$P(A \mid B)P(B) = P(B \mid A)P(A) \tag{12}$$

Equation 12 is important because it give us Bayes' rule, which constitutes the core of Bayesian inference. Bayes' rule indicates how probabilities change in the light of new data. In practice, it states that the probability of event *A* given event *B*, is equal to the conditional probability of event *B* given event *A* multiplied by the prior probability of event *A*, and divided by the sum of the conditional probability of *B* under all possible events of *A*. This is expressed as:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{\sum P(B \mid A)P(A)} \tag{13}$$

The denominator in the equation above corresponds to the marginal probability of the observed data, and acts as a normalizing constant that allows the posterior distribution to integrate to one. Because the denominator is simply scaling the posterior distribution, Bayes' Theorem can be re-stated as a proportionality, where we say that the posterior distribution of interest is proportional to the likelihood of the data times the prior distribution.

In IRT, the goal is to estimate examinees' ability level given the examinees' responses to the items on a test (the ability being estimated is referred to as the posterior distribution of interest). Following Bayes' Theorem, we can multiply the sampling distribution of the observed responses (which represents the likelihood of the data given the model's parameters) by the probability distribution assigned to the parameters (corresponding to our prior knowledge or belief about the students' ability distribution) to obtain a posterior distribution.

As described earlier, IRT item parameters are usually calibrated with an MMLE approach. Whereas the MLE, EAP or MAP methods are used to estimate the students' ability. In EAP and MAP, the likelihood of the data given the model's parameters, is multiplied by a probability distribution representing prior knowledge or belief about the students ability. The multiplication of these two distributions gives us a posterior distribution, and the mean or mode of this posterior distribution is used to estimate the students' ability level.

For TRT models, a different method is used for the estimation of parameters. The TRT models were designed within a Bayesian probability frame that provides a joint probability distribution for the data, the person parameters (θ, γ), and the item parameters (a, b, c). This method uses a Markov Chain Monte Carlo (MCMC) algorithm combined with Gibbs sampling to produce a posterior distribution for each of the parameters being estimated. These posterior distributions are used to compute summary statistics for each of the parameters, such as mean and standard deviation. The summary statistics are then used as estimates for the parameters of interest.

Markov Chain Monte Carlo

The Markov Chain Monte Carlo (MCMC) method provides a practical way of estimating multivariate posterior distributions. The Monte Carlo algorithm evaluates the true mean of random variables by drawing samples from a proposed posterior distribution and approximating the expectation by adding the obtained values and dividing by the number of draws. Thus, the means of a large number of samples are used to approximate the population mean. According to the law of large numbers, when the samples are independent and the number of draws is large, the mean of the sample means approximates the true population value. The Monte Carlo algorithm uses a Markov chain (ran for a long series of iterations) and a sampling mechanism based on the Metropolis-Hasting algorithm to draw values from the posterior distribution (Gilks, et al., 1998).

The Markov Chain performs its sampling by drawing random values throughout the space of the posterior distribution. In Markov Chains, once the procedure is initialized, the transition probabilities between sample values depend only on the most recent value of the chain, and this quality is what characterizes the sampling chain as a Markov Chain (Gilks, et al., 1998).

Metropolis-Hastings and Gibbs Sampling

The Metropolis-Hastings algorithm provides a transition kernel that evaluates the probability of the newly drawn value given the immediately previous one, and decides whether the chain will assume the new state or keep the previous one (Chib & Greenberg, 1995). In simple terms, if the jump from the present state x to the new state y goes

"uphill" on the density function, i.e., the probability of y occurring is higher than the probability of x occurring, the jump is always accepted, and the Markov chain assumes a new value for the parameters being estimated. If it goes "downhill", the jump is accepted with some nonzero probability.

The Gibbs sampler is a special case of the Metropolis-Hasting method where the chain updating consists of sampling from a fully conditional distribution (Gilks, et al., 1998). This means that the Gibbs sampler iteratively draws values from the conditional distribution of one component of a vector of parameters given the current values of all other parameters (Casella & George, 1992).

Chain Convergence

After a sufficiently large number of draws, the chain will gradually assume a unique stationary distribution, where the probability values are independent of the actual starting value. A stationary distribution is one in which the unconditional probability of moving from a present state x to a new state y is the same as the unconditional probability of moving from y to x (Chib & Greenberg, 1995).

The process of achieving a stationary distribution is referred to as chain convergence, and the set of iterations from the beginning of the chain to this point is sometimes called the burn-in portion of the chain. The term "burn-in" refers to the practice of discarding an initial portion of a Markov chain sample so that the effect of initial values on the posterior inference is minimized.

Issues affecting the MCMC estimates

Autocorrelation

In the early stages of the Markov chain, parameter values being sampled in consecutive iterations are highly correlated (typically referred to as auto correlation). This can result in the initial values of the chain having a large influence on the final parameter estimates and in the underestimation of the Monte Carlo standard error (Gilks, et al., 1998).

Researchers have suggested two main ways of addressing this autocorrelation issue: One advice has been to run the MCMC algorithm for a very large number of iterations (in order to ensure that the entire posterior space has been sampled from) while only saving every n^{th} iteration of the chain (where n is a value larger than 1). This process is sometimes referred to as thinning the chain. However, evidence indicates that thinning a Markov chain can reduce the precision of the parameter estimation (MacEachern & Berliner, 1994). Another suggestion has been to discard the set of iterations prior to chain convergence before obtaining final parameter estimates. Neither discarding the burn-in phase, nor thinning the chain are mandatory practices, however both are capable of drastically reducing the amount of data saved from a MCMC run.

The number of iterations to be discarded as burn-in is not always easy to identify. It typically depends on the starting point of the chain and the rate of convergence of the sampling distribution into the stationary distribution (Gilks, et al., 1998). This is important because the reliability and accuracy of the MCMC estimation depends of the certainty that the algorithm is drawing samples from the stationary posterior distribution.

Several methods have been proposed in order to assess chain convergence, the following are the most commonly used methods: Graphical methods include history plots, which chart parameter value at a time t against the iteration number, and autocorrelation plots, which show the correlation between parameter values over a range of chain iterations. The history plot allows the visual inspection of whether the chain is sampling around the mode of the distribution (an indication of convergence). The autocorrelation plot allows the visual inspection of the serial correlation present in the sampling chain, and a decrease in autocorrelation indicates that the chain may be closer to convergence.

Another commonly used method is to run and monitor multiple parallel chains, and use some diagnostic procedure to compare the parallel chains. Two well know procedures are the Brooks, Gelman and Rubin method, and the Raftery and Lewis indices (Cowles & Carlin, 1996). According to the Brooks, Gelman and Rubin method, the variance within the chains should be very similar to the variance across the chains. Hence, if the chains have converged, the two variances should be equal and an analysis of variance should be used to test this equality. The Raftery and Lewis method takes into consideration the autocorrelation present in the chains and provides an index that indicates the increase in the number of iterations necessary to reach convergence. In addition, some authors suggest that researchers create a habit of discarding the first 2% of the full chain (Geyer, 1992), and run some parallel estimation method (other than the Bayesian one) that can be used as a comparison base (Sinharay, 2004). The techniques mentioned above are typically used in combination and tend to provide some guidance

for the number of iterations to run, however a fail proof method to diagnose convergence has not yet been identified.

Prior characteristics, sample size, test length

Issues associated with the characteristics of priors, the examinee sample size, and the length of the test being administered will be discussed together in this session. The effects of these three variables are related to each other and apply to MCMC estimation as well as other Bayesian estimation methods, and consequently can be better understood when presented together.

Several studies have indicated that when prior distributions are a good representation of the real parameters, Bayesian methods produce estimates that are more accurate than the estimates produced by MLE based methods. This is the case because Bayesian methods add prior knowledge about the parameters of interest to the observed data, which results in more information being available for the estimation of parameters. However, a loss in accuracy can be observed when prior distributions do not match the real distribution of the parameters being estimated.

For example, in a simulation study using the 3PL model, Gao and Chen (2005) compared item parameter estimates obtained by MMLE (described earlier in this document) with the ones obtained by marginal Bayes modal estimation (MBME). In general terms, MBME is a Bayesian method that adds prior information about the item parameters to the likelihood function that the MMLE estimator maximizes. The results of their study indicated that when a prior distribution matched (or closely resembled) the

distribution of the true parameters, the Bayesian method provided more accurate estimates than the MMLE method for all three item parameters. But when priors were not well matched, the Bayesian parameter estimates were less accurate than the MMLE estimates. This decrease in estimation accuracy is due to regression towards the mean (also referred to as bias or shrinkage). Regression towards the mean corresponds to the tendency of parameter estimates to be pulled towards the mean of the prior distribution. In other words, when the prior distribution doesn't reflect the true distribution of the parameter, the mean of the prior distribution will not reflect the true mean of the parameter being estimated, and regression towards the mean will pull the estimates away from their true value.

Similar results were observed by Sheng (2010) in a simulation study that investigated the effect of sample sizes, test length, prior variances, and the extent of match between priors and true item parameter distributions on the accuracy of parameter estimation for the 3 parameter normal-ogive model (3PNO). The 3PNO and the 3PL models produce near equivalent values and interpretations of item parameters when the 3PL model is scaled by a constant. The study combined a MCMC algorithm with Gibbs sampling to estimate item parameters. The sample sizes ranged from 100 to 1000 examinees and the item parameter priors had means between 1 and 6 standard deviations from the true item parameter means. This choice of prior distributions ranged from being appropriate to being severely mismatched to the true item parameters. The study demonstrated that well matched priors (with means no more than two standard deviations

from the true parameter means), combined with larger sample sizes and/or longer tests resulted in more accurate estimates and faster Markov chain convergence.

Sheng's study (2010) also indicated that when priors were well matched to the true parameters, the use of more informative priors contributed to the increase in estimation accuracy. This indicates that the contribution of a prior to parameter estimation also depends on the variance of the prior distribution (larger variances result in less informative priors). This is the case because when the variance of a prior distribution is small, that prior will be narrowly distributed around its mean. In such case, the peak of the prior distribution will have a larger influence in the posterior distribution and consequently on the estimation of item parameters. In addition, final parameter estimation is also influenced by the sample size (i.e., the number of examinees), and by the length of the test being administered. When a large sample and/or a long test is used, the observed data (the examinees' responses) will have more influence on item parameter estimation and prior distributions will be less influential. Conversely, in the presence of small samples and/or short tests, the prior distributions attributed to the model parameters will have more influence on the final parameter estimates.

The literature on the consequences of choosing inappropriate priors and of the effect of sample size and test length on item parameter calibration spans several decades. Earlier studies using other Bayesian methods provide similar evidence. Gifford and Swaminathan (1990) used a Bayesian procedure that obtained joint modal estimates of the posterior distribution and used the Newton-Raphson procedure to solve the modal equations. The authors observed that bias in item parameter estimates was stronger for

priors that were more informative and where more mismatched to the real parameter distributions. In addition, Gifford and Saminathan's study showed that the detected bias could be attenuated or reinforced by the sample size and by the length of the test (Gifford & Swaminathan, 1990). Together, the studies mentioned above indicate that several factors can affect the accuracy and efficiency of the MCMC algorithm and should be taken into consideration when calibrating item parameters.

Summary and statement of problem

The main focus of this study is to investigate the impact that different sample sizes, mismatched prior distributions for the ability parameter, and the number of iterations discarded from the Markov chain will have on the 3PL TRT person parameter (γ) and item parameters *(a, b, c)* estimation.

Accurate item parameter estimation plays a fundamental role in test item analysis, test construction, and ability estimation. In fact, successful application of IRT methods depends on the availability of reliable and accurate methods for estimating item parameters. Previous simulation studies that have compared maximum likelihood and Bayesian methods indicated that the Bayesian methods produced item parameter estimates that had a higher correlation with and lower root mean squared deviation from the true parameters (Gao & Chen, 2005; Sheng, 2010; Swaminathan & Gifford, 1986; Tsutakawa, 1990). However, the accuracy of Bayesian item parameter estimates can be impacted by the characteristics of the priors chosen to represent knowledge about the parameters. Moreover, test length and sample size have also been shown to affect the

accuracy of item and person parameter estimates.

Several studies have demonstrated that informative priors that do not correctly match the true parameter distribution result in regression towards the mean (Gao & Chen, 2005; Gifford & Swaminathan, 1990). This shrinkage of the estimates corresponds to a tendency of parameter estimates to be pulled towards the mean of the prior distribution. Gao and Chen (2005) investigated the 3PLitem parameter estimation with the MBME method. In this study, the authors compared the accuracy of estimation when priors were non informative, informative and matched, and informative but mismatched. In addition, the authors compared the accuracy of estimation for various sample sizes (100, 500, 2000) and various test lengths (10, 30, 60). Gao and Chen's (2005) study demonstrated that mismatched priors negatively impacted the estimation of all item parameters but were most detrimental to the estimation of the pseudo guessing parameter. In addition, this effect was lessened when the sample size being used for the calibration and the test length increased.

These findings were confirmed by another study that investigated the MBME Bayesian method. Tsutakawa and Johnson's (1990) study indicated that when sample sizes used for calibrating the 3PL item parameters are smaller than 1,000, the accuracy of the estimates decreased noticeably. Most importantly, when the inadequate item parameter values where used to estimate examinees' ability levels, the variance of the estimated theta values was increased. This indicated a lower level of precision in the ability parameter estimation (Tsutakawa & Johnson, 1990). Furthermore, a study by Sheng (2010) confirmed the effect of mismatched priors on the MCMC estimation of the

3PL item parameter estimate. More interestingly, Sheng's study demonstrated that informative priors that are mismatched to the true item parameter distributions can decrease the rate of convergence of the Markov Chain and render the MCMC estimation procedure less efficient.

Another set of concerns are related to the MCMC method. The accuracy of the MCMC estimation relies on the assumption that the Markov Chain has converged onto a stationary distribution (Gilks, et al., 1998). Some researchers suggest that anywhere between the first 300 iterations to the first 20,000 iterations should be eliminated because the chain has not yet converged (Bradlow, et al., 1999; Sinharay, 2004). While other researchers suggest running several parallel chains, and using a combination of convergence detection methods (Cowles & Carlin, 1996; Kim & Bolt, 2007) in order to decide on a number of burn-in iterations to discard. In reality, the questions of how to detect chain convergence and how to decide on a correct burn-in length are closely tied to the question of efficiency. The longer the burn-in chain, the longer it will take for the algorithm to produce an estimate. Hence, from the perspective of computer processing time, it seems advantageous to discard fewer iterations. However, the effect of varying the length of burn-in has not yet been investigated.

Together, these studies indicate that several factors can affect the Bayesian estimation of item parameters when using the MCMC method, and consequently can impact inferences made about the examinees taking a test. However, no study has investigated the impact of these factors when using TRT models. This study will investigate the effects of different sample sizes, examinee ability distributions, and

MCMC burn-in chain lengths on the estimation of item parameters for the 3PL TRT model.

Proposed Research Study

The following proposed research study section contains an overview of the research focus followed by a proposed methods section that details the design of the study and a discussion of anticipated outcomes.

The 3PL TRT model will be used to evaluate the impact of sample size, examinee ability distribution, and the MCMC burn-in chain length on the accuracy of testlet item parameters estimation. Two sample sizes, two ability distributions, and 3 MCMC burn-in chain lengths were combined to create a total of 12 study conditions.

Examinee Ability Distribution

Two underlying simulated examinee ability distribution will be used in this study: normal (with a mean of zero and standard deviation of one) and a skewed, beta distribution (with an α parameter of 5.0 and β parameter of 1.8). The normal distribution represents the ability distribution that is typically assumed for the examinee population in most educational measurement settings and research studies. A beta distribution with these specific α and β parameter values has been used previously (Gorin, Dodd, Fitzpatrick, & Shieh, 2005) and results in a negatively skewed distribution with a mean of .74, a standard deviation of .16, a skew of -.73, and a kurtosis of zero. The sampled ability values will then be transformed to center the distribution on zero and produce a standard deviation of 1, resulting in a mean ability of 1.5 (Gorin, et al., 2005). This skewed distribution represents a test-taking population that is, on average, more proficient in the trait being measured by the test. One possible scenario in which this may

occur is when, after several years in an assessment program, instructional improvements and familiarity with the test format leads to growth in students' achievement on the test Gorin (2005).

This condition is important because several studies have demonstrated that in Bayesian estimation, the adequacy of the prior distributions assigned to the parameters of interest influences the accuracy of the estimation (Gao & Chen, 2005; Gifford & Swaminathan, 1990). SCORIGHT, the software that will be used in this study for parameter estimation, automatically assigns normal prior distributions to ability and item parameters (Wang, 2004). Therefore, the MCMC algorithm should handle estimation from simulated samples coming from a normal distribution better than from a skewed distribution. In addition, the effect of inaccurate priors seems to interact with the effect of sample size on parameter estimation (Gao & Chen, 2005). Consequently, the conditions with inadequate prior and smaller examinee sample size will be of particular interest.

Examinee Sample Size

Two sample sizes, one composed of 1,000 and another of 7,000 simulated examinees will be used in this study. According to previous research, a sample of at least 1000 examinees is necessary to estimate item parameters with the 3PL model (Hulin, Lissak, & Drasgow, 1982) and an even larger sample is sometimes recommended (Gao & Chen, 2005). In addition, Gao and Chen (2005) have demonstrated that when examinee sample sizes are not large and the prior distribution being used is not appropriate, parameter estimation tends to be less accurate when compared with the real parameter

values. A group of 1,000 simulees will serve as the small sample size condition. Whereas a group of 7,000 simulees will serve as the large calibration sample. This specific number of simulees was selected in order to match the original study that provided item parameter values for the present study. The item parameters were originally calibrated from a sample of close to 7,000 examinees taking a national test (Boyd, 2003).

Burn-in Length

The choices of burn-in chain lengths were based of previous studies. Wainer et. al. (2000) reported chain convergence for a 3PLTRT model in 4,000 iterations, and indicated that even shorter chains might have been enough for the MCMC algorithm to reach a stationary distribution (Patz & Junker, 1999; Wainer, et al., 2000). However, burn-in chains of 7,000 iterations were used in Keng's (2008) and Boyd's (2003) studies and even longer burn-in chains of 20,000 iterations were recommended in Sinharay's study (Sinharay, 2004). Hence, the following burn-in chain lengths will be used: 4000, 7000, and 20,000 iterations.

Known Item Parameters

The item parameter values used to simulate examinee responses in this study will come from a previous study by Boyd (2003). In this study, item parameter values were calibrated from examinee responses to 22 forms of a nationally administered test. The average number of examinees per form was 7,234 examinees with a minimum of 2,510 and a maximum of 14,439 examinees. Each form contained 55 multiple-choice items

distributed across 8 reading passages. The passages differed in the number of associated items (6, 7, 8, or 10 items per passage) and in the content of the passages. The testlet parameters for each test form were calibrated using the SCORIGHT software (Wang, Bradlow, & Wainer, 2001).

In order to simulate a test form used by Boyd (2003), this study will use a test containing eight testlets and their calibrated parameters from the data set used by Boyd. The test will be composed of 55 questions distributed into 8 testlets. The test content will correspond to approximately 37.5% Area I, 37.5% Area II, and 25% Area III. In terms of the number of items per passage, five of the passages will have 6 items, one will have 7 items, one will have 8 items, and one will have 10 items associated with it.

Data Simulation

The testlet response data will be generated using the 3PL-TRT SAS data generation program developed by Boyd (2003). Response data will be generated for forty samples, twenty with 1,000 simulees each, and another twenty with 7,000 simulees each. For each of the twenty samples, half will correspond to a sample of examinees from a normal ability distribution, and the other half will correspond to a sample of examinees from a negatively skewed ability distribution. In summary, there will be ten samples of 1,000 subjects from a normal θ distribution, 10 samples of 1,000 subjects from a negatively skewed θ distribution, 10 samples of 7,000 subjects from the same normal θ distribution, and 10 samples of 7,000 subjects the same negatively skewed θ distribution.

Sets of 10 samples per condition have been used in previous studies (Boyd, 2003; Gorin, et al., 2005; Keng, 2008).

For the purposes of generating data, each simulee will be assigned a theta value by randomly selecting a number from a normal distribution (with mean of zero and standard deviation of one, as described above) or from a negatively skewed beta distribution (with parameters $\alpha = 5.0$ and $\beta = 1.8$, as described above). The probability of an examinee responding to an item is based on the randomly selected theta value, the item parameters obtained from Boyd's study, and a generated person specific testlet effect parameter, γ, for each testlet. For each examinee, the γ parameter will be randomly generated from a normal distribution with mean of zero and a variance equal to the variance of the testlet effect ($\sigma^2_{jd(j)}$). In order to introduce random error, the simulee's response will be compared to a randomly selected number form a uniform distribution ranging from zero to one. The simulee will receive a correct response (assigned a value of one in the response string) if the random number is less than the simulee's estimated probability and an incorrect response otherwise (assigned a value of zero). This process will be repeated for each item and every examinee in each of the forty samples.

Parameter Estimation Procedures

All parameters will be estimated from the simulated data using the SCORIGHT software (Wang, et al., 2001). This process will generate three parameter estimates per item (difficulty (b), discrimination (a), guessing (c)), and an effect parameter ($\gamma_{jd(i)}$),

representing the interaction of an examinee with the stimulus in a testlet, and an ability parameter (θ) for each examinee. The testlet dependency effect will be allowed to vary from testlet to testlet because in real test data it is likely that some passages exhibit more context effects than others (Wainer, Bradlow, & Du, 2000).

The parameters will be estimated using a MCMC algorithm with Gibbs sampling to draw random values from the posterior distribution of the model parameters and obtain final estimates for item and testlet parameters, and ability estimates for the examinees. There are 3 different conditions for the length of the burn-in chain: 4,000, 7,000, and 20,000 iterations, and therefore the model and ability parameters will be estimated three times for each examinee sample. For each estimation run, after discarding the burn-in iterations, every 5^{th} iteration of the following1,000 iterations will be used to generate the posterior distribution of the model parameters. Hence, the three Markov chains will be run for a total of 5,000, 8,000, and 21,000 iterations. For each estimation procedure, two parallel chains will be produced and will allow SCORIGHT to calculate chain convergence diagnostics.

The SCORIGHT program has a built in, non changeable, set of priors for the parameters being estimated. Accordingly, it assigns the following prior distributions to the examinee ability parameter and the item and model parameters:

$$\theta_i \sim N(0,1)$$

$$a_j \sim N(\mu_a, \sigma_a^2)$$

$$b_j \sim N(\mu_b, \sigma_b^2)$$

$$\log(\frac{c_j}{1-c_j}) \sim N(\mu_c, \sigma_c^2)$$

$$\gamma_{id(j)} \sim N(0, \sigma_{d(j)}^2)$$

Where the prior for θ_i is normally distributed, has a mean of zero, and a standard deviation of 1. The discrimination parameter, a_j, was assigned a prior distribution that is normally distributed with a mean equal to the mean of the a (μ_a) and a standard deviation equal to the variance of a. (σ_a^2). The prior for the difficulty parameter, is a normal distribution, also with mean equal to the mean of the b parameter (μ_b) and standard deviation equal to the variance of the b parameter (σ_b^2). To better approximate normality, the c_j parameter was re-parameterized to $c_j = (c_j/(1-c_j))$ and was assigned a prior that is normally distributed, has a mean equal to the mean of the c_j parameter (μ_c), and standard deviation equal to the variance of the c parameter (σ_c^2). Finally, the testlet effect parameter for each testlet and each person, $\gamma_{d(j)}$, was assigned a normally distributed prior with mean of zero and standard deviation equal to the variance of the γ parameter ($\sigma_{\gamma_{d(j)}}^2$).

In order to express the uncertainty about the means and variances in the prior distributions above ($\mu_a, \mu_b, \mu_c, \sigma_a^2, \sigma_b^2, \sigma_c^2, \sigma_{d(j)}^2$), these parameters were also assigned distributions (or hyperpriors). The means were assigned non-informative normal

distributions, all with mean of zero and very large standard deviations. Whereas the variances were all assigned slightly informative inverse chi-square distributions with degrees of freedom equal to 0.5.

Data Analysis

The goal of this study is to examine the accuracy of estimation of the item difficulty, discrimination, and pseudo-guessing parameters, and the testlets' effect parameter, and the examinees' ability across several conditions. The data analysis will focus on comparing the estimated parameters to the true parameters used to simulate the data. For each sample it will also focus on examination of the Pearson product-moment correlation between the estimated and the true parameters, the root mean squared error (RMSE), and bias of the estimated parameters. These outcome measures have been used in previous studies of parameter estimation using the 3PL TRT model (Boyd, 2003; Keng, 2008; Wainer, et al., 2007). All measures will be averaged across the 10 sample replications for each of the study conditions.

The RMSE measures the difference between the parameters estimated by the model and the original parameters used to simulated the data. The formula for the RMSE is as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{a}_i - a_i)^2}{n}}$$

Where $\hat{a}$ symbolizes an estimated parameter, a symbolizes the original parameter value used for simulating the data, and n corresponds to the sample size.

Bias corresponds to the average difference between the known and estimated parameters. The equation for Bias is as follows:

$$Bias = \frac{\sum_{k}^{n}(\hat{a}_k - a_k)}{n}$$

Where $\hat{a}$ corresponds to an estimated parameter, *a* corresponds to the original parameter value used for simulating the data, and *n* corresponds to the sample size.

Expected Results

Sample size

A previous study that investigated the effect of sample sizes on the accuracy of the 3PNO item calibration, using the MCMC estimation method, indicated that samples smaller than 1,000 examinees may result in diminished estimation accuracy (Sheng, 2010) Similar results were observed when the MBME estimation method was used to investigate the 3PL item parameter calibration (Gao & Chen, 2005; Tsutakawa & Johnson, 1990). The MCMC process for calibrating the 3PL TRT model parameters involves the combined estimation of 2 person parameters (θ, γ) and 3 item parameters (a, b, c). Therefore, as the sample size increases the number of parameters associated with the MCMC estimation process also increases. Despite this increased complexity, larger sample sizes represent more observed data and result in the likelihood of the data having a stronger influence on the final estimates. Consequently, it is anticipated that an increase in sample size will improve the calibrations accuracy.

Skewed Ability distribution

A previous study has investigated the effect of priors that were mismatched to the true item parameter distributions on the final item parameter estimation. Sheng (2010) used MCMC to estimate item parameters for the 3PNO model and demonstrated that mismatched priors can negatively impact the accuracy of item parameter estimation. However, the mentioned study did not investigate the effect of having a mismatched prior on the ability parameter In MCMC, each iteration of the Markov Chain involves the

estimation of each parameter conditional on all other parameters' values in that iteration. Consequently, it is expected that an inappropriate prior for the ability distribution will affect the efficiency of the Markov Chain. In other words, it is expected that it will take the MCMC algorithm a larger number of iterations to converge onto a stationary distribution.

Length of Burn-in

No previous studies have examined the influence of varying burn-in lengths on parameter estimation. However, a study by Sinharay (2004) investigated the rate of chain convergence for the 3PL TRT model when using a sample size of 1600 examinees and a test with 60 items divided among 7 testlets. Sinharay's study indicated that more than 10,000 iterations were needed before the TRT γ parameter distribution stabilized. Based on this evidence, it is anticipated that conditions with the shortest burn-in (4,000 iterations) will produce parameter estimates that will have smaller correlation with the real parameters and larger RMSE estimates. Studies examining item and/or ability parameter estimation for the 3PL TRT model have used different values for burn-in. Keng (2008) and Boyd (2003) both used 7,000 iterations, whereas Sinharay (2004) suggested the need to discard at least the first 20,000 iterations. Based on these studies, it is anticipated that the conditions with 7,000 and the 20,000 iterations will produce better estimates than the condition with the smaller burn-in chain. In addition, the condition with a burn-in of 20,000 iterations is anticipated to provide better estimates than the

condition with 7,000 iterations. However, the differences in parameter estimates between these two conditions are expected to be small.

Sample size, skewed distribution, and length of burn-in

When all three factors are taken into account, it is anticipated that the condition with the small sample size, skewed prior, and short burn-in (4,000) will produce the least accurate parameter estimates. Whereas, the condition with the large sample size (7,000 simulees), normal prior, and longest burn-in chain (20,000) is anticipated to produce the most accurate estimates.

The contrast between the 7,000 and 20,000 burn-in conditions is anticipated to be small when the sample size is large and the prior distribution for the simulees' ability matches the true distribution. However, it is anticipated that noticeable differences in the accuracy of parameter estimation will occur between all burn in conditions when the smaller sample size and the skewed ability distribution are combined. This is because the effects of a mismatched prior and a smaller sample size on the MCMC estimation may be partially compensated for by the longer chain.

Summary and conclusions

Previous studies of Bayesian methods have stated the need for determining that the sampling chain has converged onto the posterior distribution, and have recommended that the pre-convergence iterations be discarded before proceeding to estimate the parameters of interest (Baker & Kim, 2004; Gilks, et al., 1998; Sinharay, 2004). However, no previous studies have investigated the direct effects of varying the length of burn-in on parameter estimation.

In addition, several factors may affect chain convergence. For example, the priors selected for the parameters of interest, and whether they reflect the true probability distributions of such parameters, may influence how long a Bayesian chain takes to converge onto a stationary distribution. This is important because one of the main criticisms to the MCMC method is that it is very time consuming and consequently impractical for real life applications. In TRT, a large portion of the MCMC chain corresponds to the burn-in phase. Therefore it seems interesting to investigate whether it is possible to shorten the burn-in without affecting the accuracy of the estimation.

Furthermore, the examinee sample size is another factor shown to influence parameter estimation in Bayesian methods. The sample size reflects the amount of data available to the algorithm. The more data available, the smaller the influence of the priors, and consequently the smaller the tendency for the parameter estimates to be shifted towards the mean of the prior distributions. The proposed study investigates how sample sizes, prior distributions, and the number of burn-in iterations discarded from the Bayesian estimation chain affect the accuracy of the MCMC parameter estimation.

In Boyd's study (2003), the MCMC algorithm was allowed to run for 8,000 iterations. Of those, 7,000 cycles were discarded as burn-in. Therefore, it will be interesting to examine the effects of a shorter (4,000 iterations) burn-in when the examinee's ability distribution matches the distribution assigned by SCORIGHT (i.e., the normal distribution) and the examinee sample size is large (7,000 simulees).

In addition, several studies have indicated that priors can have a biasing effect on parameter estimates. Therefore, it will be interesting to examine the accuracy of estimates when the large sample size (7,000 simulees) and the longest burn-in (20,000) chain are combined with the negatively skewed (and consequently mismatched) prior for the simulees' ability distribution. Because the longer burn-in chain will afford the MCMC algorithm more iterations to stabilize onto a stationary distribution, and the large sample size will have more weight on the estimation than the prior, it is speculated that the effect of a mismatched prior will not be as pronounced in this condition.

Limitations

There are several advantages of using an already developed, tested, and well established software for calibrating test data. However, it is important to note that SCORIGHT has a predetermined set of priors for the model parameters that can not be modified to fit the researcher's needs. This may become a problem when the person parameter distributions do not match the priors in the program. For example, the examinee ability distribution may be skewed, or the test may be targeting a population with skewed ability distribution, in which case the majority of the items on the test may

be of higher (or lower) difficulty, resulting in the difficulty parameter not following a normal distribution either. Therefore, it would be useful to ensure that the priors assigned by SCORIGHT do fit what is known about the parameters real distributions, specially when small samples are being used in the calibration.

Future directions

Several studies have indicated a positive effect of test length on the accuracy of the Bayesian item parameter estimation. For example, in Tsutakawa and Johnson's (1990) study of the 3PL model using the MBME method to estimate item parameters, the authors indicated that increasing the size of the calibrating sample alone would not increase the precision of the ability estimates, because the major component of the observed variance in estimates was the randomness of the examinees' responses to the test items. Consequently, the authors concluded that increases in test length would be necessary to improve estimation precision In addition, other studies have indicated that the sample size and test length have an interactive effect on Bayesian parameter estimation accuracy (Baker, 1998) and this is even more pronounced for the 3PL model (Hulin, et al., 1982). Consequently, in future studies, it would be interesting to investigate the effects of the test length, a mismatched prior distribution for the ability parameter, and the length of burn-in.

References

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

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

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (structure of observed learning outcomes)*. New York: Academic.

Birnbaum, A. (1958). *On the estimation of mental ability*. Randolph Air Force Base, TX: USAF School of Aviation Medicine.

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* (pp. 397-472). Reading, MA: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431-444.

Boyd, A. M. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems.* University of Texas, Austin, TX.

Bradlow, E., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. [10.1007/BF02294533]. *Psychometrika, 64*(2), 153-168.

Casella, G., & George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician, 46*(3), 167-174.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician, 49*(4), 327-335.

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

Crehan, K. D., Sireci, S. G., Haladyna, T. M., & Henderson, P. A. (1993). *A comparison of testlet reliability for polytomous scoring methods*. Paper presented at the Annual Meeting of the American Educational Research Association.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J.: L. Erlbaum Associates.

Gao, F., & Chen, L. (2005). Bayesian or non-bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education, 18*(4), 351-380.

Geyer, C. (1992). Practical Markov Chain Monte Carlo. *Statistical Science, 7*(4), 473-483.

Gifford, J. A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological Measurement, 14*, 33-43.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1998). *Markov chain Monte Carlo in practice*. Boca Raton, Fla.: Chapman & Hall/CRC.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement, 29*(6), 433-456.

Haladyna, T. M. (1992a). Context dependent item sets. *Educational Measurement: Issues and Practice, 11*, 21-25.

Haladyna, T. M. (1992b). The effectiveness of several multiple choice formats. *Applied Measurement in Education, 5*, 73-88.

Hambleton, R. K. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 7*(3), 171-186.

Hambleton, R. K., & Cook, L. L. (1977). Latent Trait Models and Their Use in the Analysis of Educational Test Data. *Journal of Educational Measurement, 14*(2), 75-96.

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

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

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Reocvery of two and three parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*(3), 249-260.

Keng, L. (2008). *A comparison of the performance of testlet-based computer adpative tests and multistage tests.* University of Texas at Austin, Austin, TX.

Kim, J.-S., & Bolt, D. M. (2007). An NCME Instructional Module on Estimating Item Response Theory Models Using Markov Chain Monte Carlo Methods. *Educational Measurement: Issues and Practice, 26*(4), 38-51.

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

Lord, F. M. (1986). Maximum Likelihood and Bayesian Parameter Estimation in Item Response Theory. *Journal of Educational Measurement, 23*(2), 157-162.

MacEachern, S. N., & Berliner, L. M. (1994). Subsampling the Gibbs sampler. *American Statistician, 48*, 188-190.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*(4), 342-366.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: The University of Chicago Press.

Rosenbaum, P. R. (1988). Item Bundles. *Psychometrika, 53*(3), 349-359.

Sheng, Y. (2010). A sensitivity analysis of Gibbs sampling for 3PNO IRT models: Effects of prior specifications on parameter estimates. *Behaviormetrika, 37*(2), 87-110.

Sinharay, S. (2004). Experiences with Markov Chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics, 29*(4), 461-488.

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

Swaminathan, H., & Gifford, J. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*(4), 589-601.

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

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*(2), 371-390.

Wainer, H. (1995). Precision and Differential Item Functioning on a Testlet-Based Test: The 1991 Law School Admissions Test as an Example. *Applied Measurement in Education, 8*(2), 157-186.

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

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

Wainer, H., & Kiely, G. L. (1987). Item Clusters and Computerized Adaptive Testing: A Case for Testlets. *Journal of Educational Measurement, 24*(3), 185-201.

Wang, X., Bradlow, E., & Wainer, H. (2001). *A user's guide for SCORIGHT (version 1.2): A computer program built for scoring tests built of testlets including a module for covariate analysis*. Princeton, NJ.

Vita

Aline Orr was born in Rio de Janeiro, Brazil. She received the degree of Bachelor of Arts in Psychology in 1995. During her undergraduate studies, Aline became interested in Neuroscience and joined the graduate program at the University of Pittsburgh. She received a Master's degree in Neuroscience 1999 and Ph.D in the same field in 2003. During the following years, she worked in research laboratories at Rutgers University and at the Children's Hospital in Dallas. In August, 2007, she entered The Graduate School at the University of Texas at Austin to pursue graduate level studies in Educational Psychology.

Permanent Address: 11100 Sierra Blanca,
Austin, Tx 78726.

This report was typed by the author.