ITEM SELECTION METHODS IN POLYTOMOUS COMPUTERIZED ADAPTIVE TESTING

BY

USAMA SAYED AHMED ALI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Carolyn J. Anderson, Chair
Professor Hua-Hua Chang, Director of Research
Professor Jeffery Douglas
Professor Jinming Zhang

Abstract

Given the rapid advancement of computer technology, the importance of administering adaptive tests with polytomous items is in great need. With regard to the applicability of adaptive testing using polytomous IRT models, adaptive testing can use polytomous items of either rating scales, or in some testing situations of multiple choice. Additionally, the availability of computerized polytomous scoring of open-ended items enhances such applicability. This need promotes the research in polytomous adaptive testing (PAT). This dissertation is an effort to focus on item selection methods, as a major component, in polytomous computerized adaptive testing. So, it consists of five chapters that cover the following:

Chapter 1 focuses on a thorough introduction to the item response theory (IRT) models and adaptive testing related to polytomous items. Such an important overview and introduction to basic concepts in test theory and mathematical models for polytomous items is needed for the flow of consequent chapters. Chapter 2 is devoted to the development of a central location index (LI) to uniquely represent the polytomous item with a scale value parameter using most commonly used polytomous models. The motivation and rationale to search for a central or an overall location parameter is twofold: a) the confusion of multiple and different parameterizations for a polytomous item even for the same model, and b) the unavailability of such single location parameter block the usage of certain item selection methods in adaptive testing. Two approaches are used to derive the proposed LIs, one is based on the item category response functions (ICRFs) and the other is based on the polytomous item response function (IRF). As a result, four LIs are proposed. Chapter 3 is particularly assigned to development of an item selection method based on the developed location index and primarily assess its performance in the PAT context relative to existing methods. This method belongs to the non-information based item selection methods and we referred it as Matching-LI method. The results support that this proposed method is promising and is capable to produce accurate ability estimates and successfully manage the item pool usage. Chapter 4 introduces new item selection methods taking in consideration the previous chapter's results. The new methods are the hybrid, stage-based information, polytomous *a*-stratification methods.

The first two methods try to merge more than one criterion for selecting items of each PAT (e.g., the hybrid method merges both the Matching-LI and maximum information (MI) methods). The last method uses Matching-LI method within each stratum. Chapter 5 provides discussion, conclusions, and limitations and future research directions with respect to important components of an adaptive testing program (i.e., item selection methods, item response models, item banks, and trait versus attribute estimation).

To my family: Yasmine, Mahmoud, Arwa, and Yusuf.

Acknowledgments

First and foremost, all my gratitude is due to God Almighty for guiding me through, and aiding me in completing this work. Then, I would like to give thanks to my family, especially my parents for their support, encouragement, and advice. I would also like to give my sincere thanks to my advisor Professor Hua–Hua Chang who provided me with technical experience in educational measurement research and life skills. I really appreciate his persistence in advising to uplift my performance and to hasten to the top.

I would like to give my gratitude to Professor Carolyn Anderson who passionately helped me to her best in different aspects. I really appreciate her great efforts toward furthering my career.

Much gratitude to Professor Jeffery Douglas and Professor Jinming Zhang for serving as committee members, and constantly being there for me. I would like to thank two of my colleagues Chia-Yi Chiu and Aaron Oaks for their endless technical support.

Also I would like to thank the Egyptian Government, the Ministry of Higher Education, and the Egyptian Cultural and Educational Bureau in Washington D.C. that supported me for four years to fulfill my mission here in the States.

Finally, I do appreciated the great role of my wife who stood besides me to attain such position. I ask Allah to help and reward her, and all who have a right on me, the best and guide us to please Him.

Table of Contents

List of Tables

List of Figures

Chapter 1

Polytomous Item Response Models and Adaptive Testing

The current chapter provides an important overview and introduction to basic concepts in test theory and mathematical models for polytomous items. This is required to understand the proposed methods for item selection in computerized adaptive testing for polytomous items that are described and studied in this dissertation.

Introduction

Item response theory (IRT) is critical to large-scale assessment, and computerized adaptive testing (CAT) is considered one of the major modern developments of IRT. Until recently, most of the research and applications of CAT focused on dichotomous items, and only a few studies have investigated CAT with polytomous items and more specifically polytomous item selection methods (Choi & Swartz, 2009; Dodd, De Ayala, & Koch, 1995; van Rijn, Eggen, Hemker, & Sanders, 2002). The polytomous items are important for various testing purposes such as education, personality, attitudes and more (Embretson & Reise, 2000).

Both dichotomous and polytomous items are used in many standardized tests such as state assessment measures. In addition, tests consisting of polytomous items are preferable for one or more of the following reasons: (a) fewer polytomous items can attain the same reliability compared to the dichotomous items, (b) the easiness of assessing some traits using rating scales, and (c) the suitability of expressing item responses on an ordinal scale (van der Ark, 2001).

Different models are available for modeling polytomous item responses. Examples of such models are: the graded response model (GRM; Samejima, 1969), the nominal response model (NRM; Bock, 1972), the partial credit model (PCM; Masters, 1982), and the generalized PCM (GPCM; Muraki, 1992). The current study focuses on the work related to the polytomous adaptive (PAT) system. Basically as in dichotomous CAT, PAT consists mainly of four components (e.g., Dodd et al., 1995; Lima Passos, Berger, & Tan, 2007):

1. Item pool (or item bank).

2. Item selection method.
3. Ability estimation procedure.
4. Stopping rule.

The first component in CAT is the pool of items that is used as a resource to deliver adaptive tests. In dichotomous CAT, the item pool needs to contain enough number of items to satisfy the purpose of testing (Lord & Novick, 1968). On the other hand, an item bank composed of 30 polytomous items was considered sufficient to successfully estimate the ability or trait level in PAT. This result did not take into consideration the security of item banks in terms of item exposure rate, or the issue of content balancing. In addition, the distribution of item bank information is another problematic issue, such as skewed or bimodal distributions (Dodd et al., 1995). Therefore, a larger item bank is required to deal with these issues.

The second critical component in adaptive testing to efficiently utilize the item pool to sequentially select item after item depending on the examinee's current ability estimate (i.e., $\hat{\theta}$) until the end of test. This dissertation focuses on new item selection rules for polytomous items.

The third issue in PAT is ability (or latent trait) estimation approaches. In the literature, they can be classified into frequentist and Bayesian methods. Maximum likelihood estimation (MLE) is a popular estimation method belonging to the first category. Bayes modal (BM), maximum a posteriori (MAP), and expected a posteriori (EAP) are Bayesian estimators used in CAT. Here we need to mention that ability estimation procedures are under continuous development. Tao, Shi and Chang (2009) proposed a new version of MLE to include scoring weights, called item-weighted likelihood estimator (IWLE), into the process of estimation. That IWLE method is different from Warm's (1989) weighted likelihood estimation method.

The final component of a CAT algorithm is a stopping rule that terminates the test. Different termination criteria to end a test exist. One approach is to administer a fixed number of items. This approach is called fixed-length adaptive testing and it is preferable in situations where the number of delivered items is a concern. Another approach that is used as a stopping rule is whether a predetermined measurement precision has been achieved. The administration of items will continue until the standard error of measurement is below an acceptable value (Embretson & Reise, 2000). The latter approach is called variable-length adaptive testing. A third approach is to terminate the test after a predetermined time is elapsed. A mixture of these approaches can be practically used as well such as using both target precision and maximum number items a stopping rule.

Toward the enhancement of PAT, the development and enhancement of item selection criteria is our starting point.

The dissertation consists of five chapters: (a) the first chapter focuses on an thorough introduction to the IRT models and adaptive testing related polytomous items, (b) the second chapter is devoted to the development of a central location index to uniquely represent the polytomous item with a scale value parameter using most commonly used polytomous models, (c) the third chapter is particularly assigned to development of an item selection method based on the developed location index and primarily assess its performance in the PAT context relative to existing methods, (d) the fourth chapter introduces new item selection methods taking in consideration the previous chapter's results, and (e) the fifth chapter provides discussion, conclusions, and limitations and future research directions.

The reminder of this chapter covers different mathematical IRT models of polytomous items, the connection among them, and the parameters describing item properties, followed by explaining two item information measures, other modifications applied to the original measures, and their relation to item selection criteria. Then, the chapter concludes with defining the key terms used through out the research.

Polytomous IRT Models

Test items that have more than two response categories are called polytomously-scored items. Both dichotomously and polytomously-scored items are ubiquitous in educational and psychological testing. Due to the different scoring scheme, polytomous IRT models should be used in item parameter calibration and ability estimation.

In the dichotomous item case, the test administrator classifies the observed responses into one of two categories, correct or incorrect. This dichotomization treats all incorrect answers as equivalent to one another. As such all the mathematical operations required to answer an item on a mathematics test are considered *una voce* (i.e., as a "single operation"). If an examinee correctly performs all the operations and their response is categorized as a correct response, they receive a value of 1. Otherwise, the response is categorized in the incorrect category with assigned value of 0. The value assigned to responses reflects only whether the examinee has correctly performed (heuristically) one "operation" (De Ayala, 2009).

In contrast, consider the following mathematical test item as an illustration:

$$(6/3)\ +\ 2 = ?.$$

A scoring rubric for this item might be based on the operations or subtasks needed to answer this item correctly. Therefore, for this item i there exist three possible integer scores x_i for this item, $x_i = 0, 1, \text{or}\, 2$. These scores are called category scores that might indicate the number of successfully completed operations. In general, polytomous test items are scored in a way that reflects a particular score category out of $m + 1$ scores (i.e., $x = 0, 1, ..., m$) that an examinee has achieved, been classified into, or endorsed. Examples of this includes the case where a person gets a score of 2 on an item that is scored from 0 to 4 or a person selects a "disagree" option on a 5-point Likert-scale item (De Ayala, 2009; Ostini & Nering, 2010).

With regard to the applicability of adaptive testing using polytomous IRT models, CAT can use polytomous items of either rating scales, or in some testing situations of multiple choice (Wang & Wang, 2001). Additionally, the availability of computerized polytomous scoring of open-ended items enhances such applicability as suggested by Bennet, Steffen, Singley, Morely, & Jacquemin's (1997) research.

For the purpose of modeling polytomous data, polytomous IRT models were proposed. One classification of the polytomous IRT models is in terms of whether the category responses are treated as nominal and ordinal variables. In another classification scheme, polytomous IRT models fall in one of two main categories based on the mathematical form of the model (Thissen & Steinberg, 1986). One category consists of *difference* models such as Samejima's graded response model (GRM; Samejima, 1969), and the other consists of *divided-by-total* models such as the nominal response model (NRM; Bock, 1972). For ordinal polytomous models, Mellenbergh (1995) defined different classification scheme in terms of the models' different order-preserving mechanisms in forming the dichotomies of response categories. He classified them into three classes:

1. the cumulative probability (or graded-response) models.
2. continuation ratios (or sequential) models.
3. the adjacent-category (or partial-credit) models.

In the following subsections, a brief description of each model of the most commonly used models, its parameterization, the meaning of the models' parameters, and relationships among these models are provided.

Nominal response model. For nominal item responses, Bock (1972) proposed that the probability of response x to an item i by an examinee with ability θ, denoted by $P_{ix}(\theta)$, equals

$$P_{ix}(\theta) = \frac{\exp(c_{ix} + a_{ix}\theta)}{\sum_{x=0}^{m} \exp(c_{ix} + a_{ix}\theta)}, \qquad \text{x= 0, 1, \ldots, m}, \tag{1}$$

where a_{ix} is a slope (analog to discrimination) parameter of category x, c_{ix} is a location (analog to difficulty) parameter of category x, and $m + 1$ is the number of response categories to item i.

For model identification, either sums are set to zero (i.e., $\sum a_{ix} = \sum c_{ix} = 0$), or one response category is set to to zero (e.g., the parameters of the lowest response category, $a_{i0} = c_{i0} = 0$). For simplicity, there is no examinee subscript on the ability parameter θ in the current and subsequent models. For each category of an item, there exists an item category response function (ICRF), and the set of ICRFs per item is uniquely determined given a set of identification constraints. Figure 1 provides the item category characteristic curves (ICCCs) obtained from the ICRFs of a 4-category item; one ICCC corresponds to an ICRF. At each level on the ability continuum, the sum of ICRFs equals 1 (i.e., $\sum_{x=0}^{m} P_{ix}(\theta = \theta_0) = 1$).

Bock's NRM is a general polytomous model that is used for mutually exclusive response categories. Item response categories are not necessarily ordered and the NRM is the only model for nominal categories. Mellenbergh (1995) has showed that the NRM can be reformulated in terms of m log odds for a nominal variable of $(m+1)$-response categories. He treated them as m dichotomous response variables that correspond to choices of m categories to a reference category. The log odds of choosing a score category x over the first category is as follows:

$$\ln\left({}^{P_{ix}(\theta)}/_{P_{i0}(\theta)}\right) = \ln\left(\frac{\exp(c_{ix} + a_{ix}\theta)}{\exp(c_{i0} + a_{i0}\theta)}\right) = (c_{ix} - c_{i0}) + (a_{ix} - a_{i0})\theta = (c_{ix} + a_{ix}\theta), \quad x = 0, 1, \ldots, m. \tag{2}$$

If $m = 1$ (i.e., dichotomous item), then the NRM reduces to the two-parameter logistic model, where the probability of answering the item correctly, $P_i(\theta) = P_{i1}(\theta)$ and the probability of answering the item incorrectly, $Q_i(\theta) = 1 - P_{i1}(\theta) = P_{i0}(\theta)$. This can be described by

$$\ln\left({}^{P_i(\theta)}/_{1-P_i(\theta)}\right) = \ln\left({}^{P_{i1}(\theta)}/_{P_{i0}(\theta)}\right) = (c_{i1} - c_{i0}) + (a_{i1} - a_{i0})\theta = (c_{i1} + a_{i1}\theta), \tag{3}$$

and the parameters of the reduced model are $a_i = a_{i1}$, $b_i = -{}^{c_{i1}}/_{a_{i1}}$.

Graded response model. Samejima (1969) proposed a model for items with ordinal response categories (e.g., Likert-scale items). Samejima's graded response model (GRM) is one of the difference models (Thissen & Steinberg, 1986); even Samejima is against such categorization and her model can be expressed as a divide-by-total model (Samejima, 2010). The GRM expresses the cumulative probability of getting at least a score x on item i

$$P_i^*(X \geq x) = P_{ix}^* = \frac{\exp(a_i(\theta - b_{ix}))}{1 + \exp(a_i(\theta - b_{ix}))}, \tag{4}$$

where $x = 1, 2, \ldots, m$, and therefore the probability of responding in a specific category score is $P_{ix} = P^*_{ix} - P^*_{i,x+1}$. Note that it is assumed in GRM that $P^*_{i0} = 1$ and $P^*_{i,m+1} = 0$. Also, the b_{iv}s are ordered such as $b_{i1} < b_{i2} < \ldots < b_{im}$.

Partial credit model. Masters (1982) developed a polytomous model, the partial credit model (PCM), that is different in terms of its parameterization and conceptualization from the GRM. This model is an extension of the dichotomous Rasch model. The PCM belongs to the adjacent-category models in Mellenbergh's classification of IRT models, and to the divide-by-total models in Thissen and Steinberg's classification. Items are viewed as requiring multiple steps to obtain a correct answer. Incorrect response options reflect partial knowledge; therefore, partial credit is given to each step completed. For an examinee with ability θ, the probability of selecting category or response option x of item i, denoted by $P_{ix}(\theta)$, equals

$$P_{ix}(\theta) = \frac{\exp \sum_{v=1}^{x} (\theta - b_{iv})}{1 + \sum_{c=1}^{m} \exp \sum_{v=1}^{c} (\theta - b_{iv})}, \tag{5}$$

where b_{iv} is an item-category location parameter. For notational convenience, $\sum_{v=0}^{0} (\theta - b_{iv}) = 0$. The location parameter b_{iv} can be broken down into an item location parameter (i.e., b_i), and a category threshold (i.e., d_{iv}) such that $b_{iv} = b_i - d_{iv}$. The difference in ability levels between two adjacent categories, x and $(x + 1)$, is called the step difficulty or threshold, d_{ix}. The log odds based on the PCM equals

$$\ln \left({}^{P_{ix}(\theta)} / {}_{P_{i,x-1}(\theta)} \right) = (\theta - b_{ix}), \quad x = 0, 1, \ldots, m. \tag{6}$$

In the PCM, the thresholds need not to be ordered; harder steps may be preceded by easier steps and vice versa.

Generalized partial credit model. As noticed above the PCM is based on the Rasch model and considers items to be equally discriminating. Relaxing this assumption and allowing items to be differentially discriminating led to the invention of the generalized partial credit model (GPCM; Muraki, 1992, 1993). The ICRFs of GPCM can be written as

$$P_{ix}(\theta) = \frac{\exp \sum_{v=1}^{x} Z_{iv}(\theta)}{1 + \sum_{c=1}^{m} \exp \sum_{v=1}^{c} Z_{iv}(\theta)}, \tag{7}$$

and

$$Z_{iv}(\theta) = Da_i(\theta - b_{iv}) = Da_i(\theta - b_i + d_v), \tag{8}$$

where D is a scaling constant that puts the trait scale in the same metric as the normal ogive model ($D = 1.7$) or stays on the metric of the logistic model ($D = 1$), a_i is a slope parameter for item i, b_{iv} is an item-category parameter, b_i is an item location parameter, and d_v is a category parameter. If $m = 1$, the model reduces to Birnbaums (1968) two-parameter logistic model. If all a_i are equal, the GPCM reduces to the PCM. These two restrictions combined yield a polytomous Rasch model.

The log odds based on the GPCM is

$$\ln\left({}^{P_{ix}(\theta)}/_{P_{i,x-1}(\theta)}\right) = a_i(\theta - b_{ix}), \quad x = 0, 1, \ldots, m. \tag{9}$$

Without assuming the order of response categories the log-odds in Equation 2, following the NRM log odds, can be expressed as

$$\ln\left({}^{P_{ix}(\theta)}/_{P_{i,x-1}(\theta)}\right) = \ln\left(\frac{\exp(c_{ix} + a_{ix}\theta)}{\exp(c_{i,x-1} + a_{i,x-1}\theta)}\right) = (c_{ix} - c_{i,x-1}) + (a_{ix} - a_{i,x-1})\theta,$$
$$x = 0, 1, \ldots, m. \tag{10}$$

Constraints can be replaced on parameters of the NRM such that it is a GPCM, specifically, $(c_{ix} - c_{i0}) + (a_{ix} - a_{i0})\theta$ must equal $a_i(\theta - b_{ix})$. Therefore, the relationship between parameters in the NRM and GPCM can be expressed as

$$a_i = (a_{ix} - a_{i,x-1}) \text{ and } b_{ix} = -\frac{(c_{ix} - c_{i,x-1})}{(a_{ix} - a_{i,x-1})} \tag{11}$$

Both the PCM and GPCM contain a similar term $Z_{ix}^{+}(\theta)$ (i.e., the sum of $Z_{iv}(\theta)$). In terms of the GPCM, this term can be written as

$$Z_{ix}^{+}(\theta) = \sum_{v=1}^{x} Z_{iv}(\theta) = \sum_{v=1}^{x} Da_i(\theta - b_i + d_v) = Da_i \left[x(\theta - b_i) + \sum_{v=1}^{x} d_v \right], \tag{12}$$

and can be rewritten as

$$Z_{ix}^{+}(\theta) = Da_i \left[T_x(\theta - b_i) + X_x \right]. \tag{13}$$

Andrich (1978) called T_x the scoring function and X_x the category coefficient.

Information Indices and Related Item Selection Rules

In this section, information functions for polytomously scored items are presented. These information indices are very important; because item selection methods were built using them. Examples of such information-based item selection criteria are: maximum information criterion and maximum Kullback-Leibler information criterion. Subsequently more sophisticated versions of item selection rules based on Fisher and Kullback-Leibler information indices are described.

Fisher and observed information. The observed information is given by,

$$J(\theta) = -\left(\frac{\partial^2 \log L}{\partial \theta^2} \right), \tag{14}$$

where L is the likelihood of a sample of n independent item response observations (van der Linden & Pashley, 2000). A well known index in statistics is Fisher information that gives the amount of information provided by data on an unknown parameter, θ. It is given by,

$$I(\theta) = -E\left(\frac{\partial^2 \log L}{\partial \theta^2} \right) = E(J(\theta)), \tag{15}$$

where E is the expectation to be found with respect to the responses. These two information measures are equivalent under dichotomous IRT models (van der Linden, 1998). Also, they are the same under the GPCM (Donoghue, 1994; Muraki, 1993). However, this equality is not satisfied under other models such as GRM with items of three or more response categories. This was the reason for Choi and Swartz (2009) to investigate the performance of some Bayesian item selection methods based on these two measures under GRM.

Bock (1972) introduced the information due to the responses in category x as a partition of the item information, and Dodd et al. (1995) provided the general formula of the item information function (IIF) derived from a model for a polytomous item,

$$I_i(\theta) = \sum_{x=0}^{m} \frac{(P'_{ix}(\theta))^2}{P_{ix}(\theta)} = \sum_{x=0}^{m} I_{ix}(\theta), \tag{16}$$

where $P'_{ix}(\theta)$ is the first derivative of the probability of getting a score x on the i^{th} item, $P_{ix}(\theta)$, with respect to θ. $I_{ix}(\theta)$ is the item category information function and it is clear that additivity feature is applied to item information over the item categories.

Under the IRT assumption of local independence, the test information function has its additive property across a group of n items as well, $I_i(\theta) = \sum_{x=0}^{m} I_{ix}(\theta)$, (Lord & Novick, 1968). The factors that affect the item information are complex in polytomous items. For that reason, Akkermans and Muraki (1997) investigated the peaks of the IIF for trinary items (i.e., three-category items).

For a more specific formula of the polytomous item information functions, the above equation can be applied to any model. For the NRM, the IIF, $I_i(\theta)$, is given by,

$$I_i(\theta) = \sum_{x=1}^{m} a_{ix}^2 \, P_{ix}(\theta) - \left(\sum_{x=1}^{m} a_{ix} \, P_{ix}(\theta) \right)^2, \tag{17}$$

where a_{ix} is the item category discrimination parameter (Lima Passos et al., 2007). For the GRM, the IIF, $I_i(\theta)$, is given by (Ostini & Nering, 2010), (Refer to Lima Passos, Berger, and Tan (2008) for the IIF for the GRM item.)

$$I_i(\theta) = \sum_{x=1}^{m} A_{ix}. \tag{18}$$

In Equation 19, A_{ix} is described as the basic function that is given by

$$A_{ix} = D^2 \, a_i^2 \, \frac{\Big[P^*_{ix}(\theta)(1 - P^*_{ix}(\theta)) - P^*_{i,x+1}(\theta)(1 - P^*_{i,x+1}(\theta)) \Big]}{P_{ix}(\theta)}. \tag{19}$$

Under the GPCM, the IIF, $I_i(\theta)$, is given by,

$$I_i(\theta) = a_i^2 \left[\sum_{x=1}^{m} x^2 \, P_{ix}(\theta) - \left(\sum_{x=1}^{m} x \, P_{ix}(\theta) \right)^2 \right] = a_i^2 \, \mathrm{var}(X_i), \tag{20}$$

where the $\left[\sum_{x=1}^{m} x^2 \, P_{ix}(\theta) - (\sum_{x=1}^{m} x \, P_{ix}(\theta))^2\right]$ is the variance of score on the item given a specific level of ability (Donoghue, 1994). When PCM is considered, the discrimination parameter, a_i, is dropped from the above Equation. The Fisher information function is considered a local index that conveys the information around a single θ value (Chang & Ying, 1996; Chen, Ankenmann, & Chang, 2000). De Ayala (1992) found that using the category information instead of item information for item selection reduced the test length on average by one item for NRM CAT simulations.

Kullback-Leibler information. In the context of educational testing, Chang and Ying (1996) presented the Kullback-Leibler (KL) information function as a global information criterion for item selection in dichotomous CAT. The KL information function is defined by

$$KL_i \left(\theta \parallel \theta_0\right) = P_i(\theta_0) \log \left[\frac{P_i(\theta_0)}{P_i(\theta)} \right] + (1 - P_i(\theta_0)) \log \left[\frac{1 - P_i(\theta_0)}{1 - P_i(\theta)} \right], \tag{21}$$

where θ_0 represents a true ability level and $P_i(.)$ is probability of answering item i correctly. Chang and Ying (1996) mentioned several important features of the KL function:

1. It is not symmetric; that is, $KL_i(\theta \parallel \theta_0) \neq KL_i(\theta_0 \parallel \theta)$.
2. $KL_i(\theta \parallel \theta_0) \geq 0$ and $KL_i(\theta_0 \parallel \theta_0) = 0$.
3. Similar to the additive property of Fisher information, test information is the sum (over n items) of the item information; that is,

$$KL^{(n)} \left(\theta \parallel \theta_0\right) = \sum_{i=1}^{n} KL_i \left(\theta \parallel \theta_0\right). \tag{22}$$

Note that $KL_i(\theta \parallel \theta_0)$ is a function of two variables, θ and θ_0. Geometrically, the KL information function is a surface in three-dimensional space, as opposed to a function in two-dimensional space represented by $I_i(\theta)$.

As a function of two θ levels, $KL_i(\theta \parallel \theta_0)$ represents the power of an item to discriminate between these two levels. When $\theta = \theta_0$, the value of $KL_i(\theta \parallel \theta_0)$ is 0 (i.e., the item cannot distinguish between examinees at the same level of θ). When θ and θ_0 are very different, the value of $KL_i(\theta \parallel \theta_0)$ is large (i.e., the item can easily distinguish between examinees with different θs). In other words, the KL is considered as a warning signal that gives a loud alarm when there is a difference between the two θ levels; otherwise the alarm will be lower. By contrast, $I_i(\theta)$ represents the discrimination power of an item at a single θ.

The KL information index with respect to a polytomous item of $(m+1)$ categories can be generalized to the following

$$KL_i(\theta \parallel \theta_0) = \sum_{x=0}^{m} P_{ix}(\theta_0) \log \left[\frac{P_{ix}(\theta_0)}{P_{ix}(\theta)} \right], \tag{23}$$

As in the context of Fisher information, we can get the form of category KL (CKL) information function as followed,

$$KL_{ix}(\theta \parallel \theta_0) = P_{ix}(\theta_0) \log \left[\frac{P_{ix}(\theta_0)}{P_{ix}(\theta)} \right]. \tag{24}$$

From the graphical representation as shown in Figures 1 and 2, we can see some of the properties of KL and CKL under a given polytomous model, GPCM (note that these curves are part of the KL or CKL surface at a specific level of θ_0, the real ability level),

1. These curves are not symmetric; that is, $KL_{ix}(\theta \parallel \theta_0) \neq KL_{ix}(\theta_0 \parallel \theta)$ (e.g., in the graphed item, $KL_{ix}(\theta = -1 \parallel \theta_0 = 0) = .2756 \neq KL_{ix}(\theta_0 = 0 \parallel \theta = -1) = -.1206$.
2. $KL_i(\theta \parallel \theta_0) \geq 0$ and $KL_i(\theta_0 \parallel \theta_0) = 0$ but $KL_{ix}(\theta \parallel \theta_0)$ is a real-valued that can be negative as shown in Figures 1 and 2.
3. All CKL curves and the KL curve intersect at the same point. This point corresponds to the real ability value; the amount of information of each are equal at that point. For the 3-category item, it is the minimum value of middle-category KL (i.e., the CKL with respect to score 1). In addition, the CKL for first (last) categories is a monotone increasing (decreasing) function, respectively.

Modified information functions. As a step in modifying and refining information functions, Veerkamp and Berger (1997) introduced an interval information

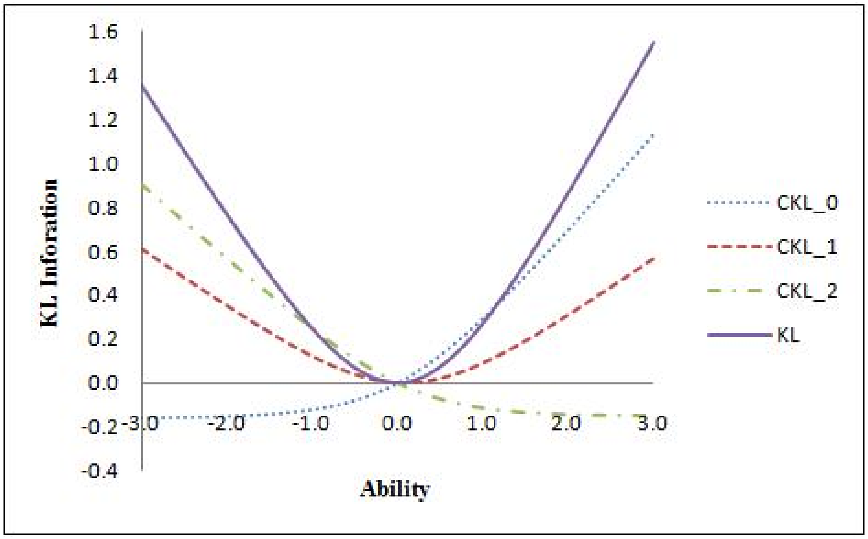


Figure 1: Item and category KL information curves ($\theta_0 = 0$).

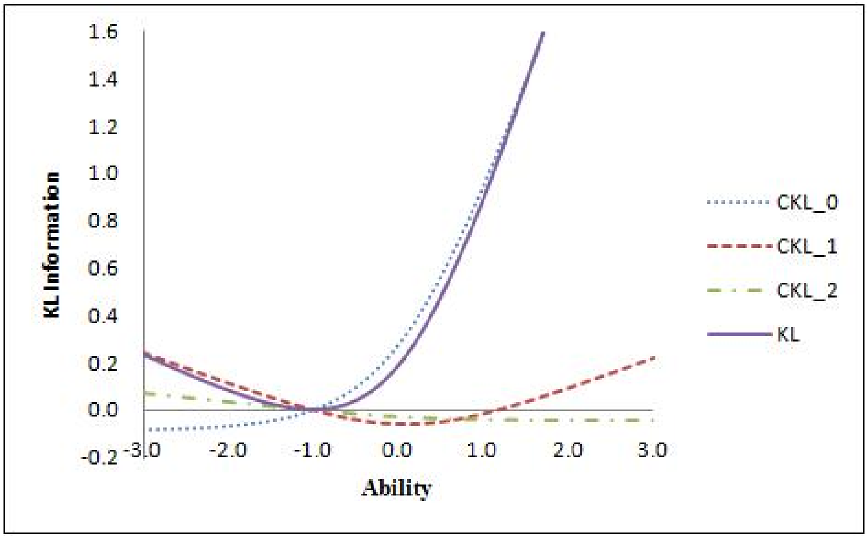


Figure 2: Item and category KL information curves ($\theta_0 = -1$).

criterion for dichotomous CAT to overcome the problems of Fisher information. Instead of maximizing Fisher information function at an ability estimate, they proposed to integrate the function over a small interval around the estimate to compensate for the uncertainty in it. In PAT, there is another reason to integrate Fisher information function over an interval. Fisher information function might be multi-modal when items are analyzed with the GPCM (Muraki, 1993). Van Rijn et al. (2002) demonstrated that a multi-peaked item might contain more information for a small interval around the ability estimate than the item that contains maximum Fisher information at the ability estimate. They proposed to select the next item with a maximum interval information criterion:

$$i = \arg\max_i \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} I_i(\theta)d\theta, \tag{25}$$

where i is a potential item to be administered and δ is a small constant defining the width of the interval.

Other variations of Fisher item information have been proposed and used as alternatives to the original index. The A-optimality or sum criterion, $\Phi_A = \sum_\theta I_i(\theta)$ and the D-optimality or product criterion is given by $\Phi_D = \prod_\theta I_i(\theta)$ where the distribution of θ is equally weighted within an interval. In general, the A-optimality criterion corresponds to the arithmetic sum and the D-optimality criterion corresponds to the geometric mean (Berger & Veerkamp, 1996). Additional indices such as maximum expected information that depends on the observed information or Fisher information exist.

The relationship between item information and IRF in PAT. The theoretical relationship between the Fisher information and KL information in the context of polytomous items is still the same as that for dichotomous items. That is, the second derivative of KL information is considered Fisher information at the true ability level.

With respect to the relationship between Fisher information function (i.e., $I_i(\theta)$), and IRF or the expected score (i.e., $E[X]$). Figure 3 graphically shows the information curve and the polytomous item response function for an item with three categories over the ability continuum.

Muraki (1993) pointed out a relationship between the IRF and item information

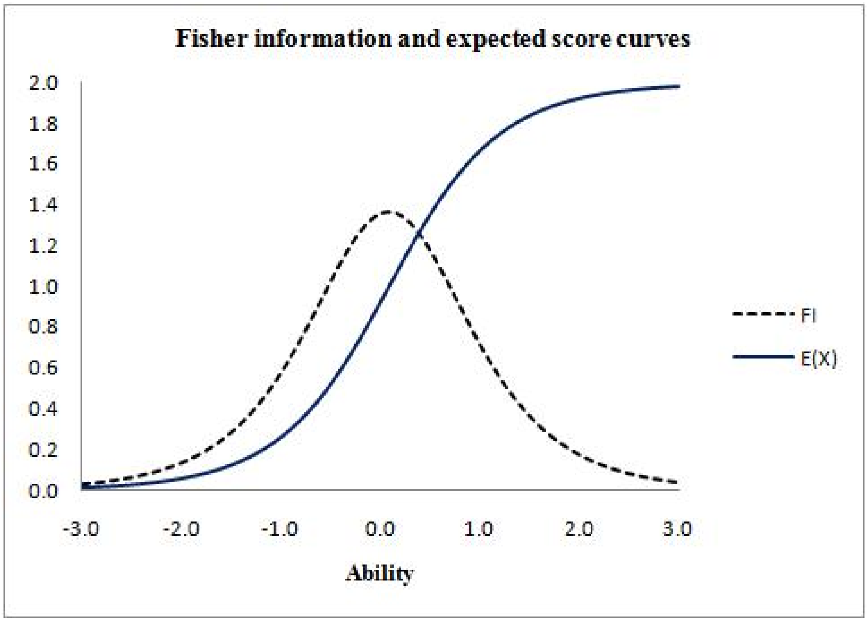


Figure 3: Item information and expected score curves for a 3-category item.

function, that can also be related to the gradient function $G_i(\theta)$ described in Equation 26 as follows,

$$G_i(\theta) = \frac{\partial \bar{T}_i(\theta)}{\partial \theta} = a_i^2 \left[\sum_{x=1}^{m} x^2\, P_{ix}(\theta) - \left(\sum_{x=1}^{m} x\, P_{ix}(\theta) \right)^2 \right] = \frac{I_i(\theta)}{a_i}, \tag{26}$$

Therefore, the curve of Fisher item information function intersects with the curve of gradient function at the point of maximum information. This is a promising result that could be beneficial and could provide a non-information based item selection algorithm in PAT depends primarily on the polytomous IRF instead of the fisher information as a criterion.

More information in polytomous items. In general, a polytomous item has more information than a dichotomous one. In reality, each pair of the adjacent categories in a polytomous item could be considered as a single dichotomous item (Dodd et al., 1995). For a fixed amount of information, this property makes the required size of the item bank in PAT smaller than its size in dichotomous CAT. In other words, given a dichotomous and polytomous item banks with same number of items the polytomous one contains more information. Also, this may affect the item selection if we have a mixture of item types, dichotomous and polytomous items. Also, there are some studies that discussed the information in polytomous items under different polytomous IRT models (e.g., Akkermans & Muraki, 1997; Donoghue, 1994).

Definition of Terminology

Below is a summary of terms used frequently through out the following chapters.

Polytomous item response model. Item response model for items with more than two response categories (e.g. multiple-choice item that allows partial credits for each of the response categories, or constructed-response item with multiple steps).

Ability estimate. The estimate of the level of a latent trait of an examinee demonstrated by their observed response pattern to a test.

Item response categories. The possible ways as assigned by the item writer that an examinee could respond to an item. In the context of multiple-choice items, item response categories are the options provided for the examinee to choose; in constructed-response items, they are the steps or parts of the solution to the item that allow different amounts of partial credit to be awarded upon their completion.

Item response function (IRF). The mathematical equation that relates the probability of answering an item correctly as a function of the ability of the examinee attempting the item and the item parameters.

Item category response function (ICRF). The mathematical equation that describes the probability of an item category being chosen as a function of the ability of the examinee and item category parameters.

Item characteristic curve (ICC). The curve that demonstrates the relationship between the ability of an examinee and the probability of the examinee answering the item correctly. Sometimes it is referred as a trace line. It is the graph of IRF plotting against the ability parameters.

Item category characteristic curve (ICCC). The curve represents the relationship between the probability of an examinee choosing an item category and the ability of the examinee. ICCCs of all the categories within an item are usually plotted on the same graph.

Chapter 2

Development of Location Indices for Polytomous Items and Their IRT Applications

The current chapter addresses the need to represent a polytomously-scored item by one index for the use in adaptive testing. Therefore, this chapter introduces the development of location indices for a polytomous item. Possible applications of such indices in CAT are reported as well.

Objective of Study

Items are the building blocks of testing. Based on the scoring of response options, items can be classified into two types: dichotomous and polytomous items. In terms of the parameters that characterize items, they depend on the IRT model that fits such items. For dichotomous items, the two-parameter logistic model are characterized by two properties: an item's discrimination power (measured by parameter a_i for item i), and difficulty (measured by parameter b_i). In the polytomous case, many polytomous models use a single discrimination parameter for the item regardless of response option, such as the GRM, PCM, and GPCM. Other IRT models have multiple parameters; one parameter per item category. With regard to item difficulty for polytomous items, models use several parameters or thresholds that depend on item categories; therefore, at least two thresholds will be used to provide an idea of the difficulty for a 3-category item (i.e., scored 0, 1, or 2).

The motivation and rationale to search for a central or an overall location parameter is twofold: a) it may be due the confusion of multiple and different parameterizations for a polytomous item even for the same model, and b) the unavailability of such a single location index blocks the usage of certain item selection methods in adaptive testing.

Based on the PAT literature, the item selection methods are sometimes natural extensions of those used with dichotomous items, such as information indices. The information-based item selection procedures may consider the item as a whole or at the score category level. Dodd et al. (1995) commented that only the information-based item selection algorithms have been investigated for the GRM, NRM, and PCM because of the unavailability of a single location index (or scale value parameter).

The potential item selection method to be developed in this dissertation intends to use the properties of item characteristic curve (ICC) and item-category characteristic curves (ICCCs) to define new polytomous item indices.

Item Location Indices

Two general approaches are used to develop item location for polytomous item. The first approach is to study the category response functions and the second one focuses on the item response function. Basically, the proposed indices are based on the ICRFs and IRF of a polytomous item.

The development of such indices given here applies to polytomous models that have the same discrimination across the different item categories where the mathematical derivations introduced here apply for the other models in such category.

Considering the GPCM, the ICRF for score x on the ith item is

$$P_{ix}(\theta) = \frac{\exp \sum_{v=1}^{m} a_i(\theta - b_{iv})}{1 + \sum_{c=1}^{m} \exp \sum_{v=1}^{c} a_i(\theta - b_{iv})}. \tag{27}$$

The definition of model parameters was introduced in Equations 7 and 8.

Studying the item category response functions (ICRFs). Assume a polytomous item with three categories under the GPCM where each category has its own ICRF. The parameters for item i are: a_i, b_{i1}, and b_{i2}. Therefore, we have three ICRFs with the possible scores (0, 1, or 2) on the item, $P_{i0}(\theta)$, $P_{i1}(\theta)$, and $P_{i2}(\theta)$. A graphical representation of such a 3-category item is given in Figure 4. This graph provides the rationale behind the following derivations. As seen in Figure 4, the peak of the partial-credit score curve occurs at the same point that the zero and perfect score curves intersect. It can be shown mathematically that this will always occur for the GPCM. Starting with the probability of attaining the partial credit (middle) score, $P_{i1}(\theta)$, and we locate the peak of this ICRF. From Equation 1,

$$P_{i1}(\theta) = \frac{\exp\left[a_i(\theta - b_{i1})\right]}{1 + \exp\left[a_i(\theta - b_{i1})\right] + \exp\left[a_i(2\theta - b_{i1} - b_{i2})\right]}, \tag{28}$$

and the first derivative of Equation 28 with respect to θ is

$$\begin{aligned}\frac{\partial P_{i1}(\theta)}{\partial\theta} &= a_i P_{i1}(\theta) - a_i P_{i1}(\theta)\sum_{c=1}^{2} cP_{ic}(\theta) \\ &= a_i P_{i1}(\theta)\left[1 - \sum_{c=1}^{2} cP_{ic}(\theta)\right]. \qquad (29)\end{aligned}$$

Setting Equation 29 equal to zero, we find that the maximum of Equation 29 occurs when any of the following three conditions hold: $a_i = 0$, $P_{i1}(\theta) = 0$, or $\sum_{c=1}^{2} cP_{ic}(\theta) = 1$ (or $E(X_i) = 1$). The fist condition, $a_i = 0$, indicates that the item has no discrimination power, hence, it would not be in an operational item pool. The second condition, $P_{i1}(\theta) = 0$, is not achievable. All response options have non-zero probability. The third condition, $\sum_{c=1}^{2} cP_{ic}(\theta) = 1$ or $E(X_i) = 1$, can be attained as seen by noting that:

$$\begin{aligned}\sum_{c=1}^{2} cP_{ic}(\theta) &= 1 \\ P_{i1}(\theta) + 2P_{i2}(\theta) &= 1 \\ \exp\left[a_i(\theta - b_{i1})\right] + 2\exp\left[a_i(2\theta - b_{i1} - b_{i2})\right] &= 1 + \exp\left[a_i(\theta - b_{i1})\right] + \exp\left[a_i(2\theta - b_{i1} - b_{i2})\right] \\ \exp\left[a_i(2\theta - b_{i1} - b_{i2})\right] &= 1 \\ a_i(2\theta - b_{i1} - b_{i2}) &= 0 \\ \theta &= \tfrac{1}{2}(b_{i1} + b_{i2}). \qquad (30)\end{aligned}$$

The point on ability continuum corresponding to the intersection between these two ICCCs of scores 0 and 2 satisfies the following condition

$$P_{i0}(\theta) = P_{i2}(\theta) \qquad (31)$$

$$\frac{1}{1 + \exp\left[a_i(\theta - b_{i1})\right] + \exp\left[a_i(2\theta - b_{i1} - b_{i2})\right]} = \frac{\exp\left[a_i(2\theta - b_{i1} - b_{i2})\right]}{1 + \exp\left[a_i(\theta - b_{i1})\right] + \exp\left[a_i(2\theta - b_{i1} - b_{i2})\right]}.$$

The equivalence in Equation 31 implies that

$$\exp\left[a_i(2\theta - b_{i1} - b_{i2})\right] = 1, \tag{32}$$

and is the same conclusion as given in Equation 30 (i.e., $\theta = \frac{1}{2}(b_{i1} + b_{i2})$).

It is verified that the two ICCCs for the lowest and highest scores on the item that are intersecting and we can see it is corresponding to the same point on the ability scale as the peak for the ICCC of the partial-credit score, see Figure 4 for an example of 3-category item.

For a more generalized version of a polytomous item with $m + 1$ categories where every two ICCCs of scores x and $m - x$ are intersecting in some point, (i.e., $P_{i0}(\theta) = P_{im}(\theta)$, $P_{i1}(\theta) = P_{i,m-1}(\theta)$, ..., $P_{ix}(\theta) = P_{i,m-x}(\theta)$, $\ldots$, $P_{i,\frac{m+1}{2}}(\theta) = P_{i,\frac{m+3}{2}}(\theta)$). See Figure 5 for an example of 5-category item.

$$\begin{aligned}
P_{ix}(\theta) &= P_{i,m-x}(\theta) \\
x\theta - \sum_{c=1}^{x} b_{ic} &= (m-x)\theta - \sum_{c=1}^{m-x} b_{ic} \\
[(m-x) - x]\theta &= \sum_{c=1}^{m-x} b_{ic} - \sum_{c=1}^{x} b_{ic} \\
(m-2x)\theta &= \sum_{c=x+1}^{m-x} b_{ic} \\
\theta &= \tfrac{1}{m-2x} \sum_{c=x+1}^{m-x} b_{ic}, \quad x = 0, 1, ..., \tfrac{m+1}{2}
\end{aligned} \tag{33}$$

At the two middle ICCCs, $\theta = b_{i,\frac{m-1}{2}}$. When m is an even integer, as represented by the 5-category item example in Figure 5, such that there is one middle ICCC representing the score of $\frac{m}{2}$, we need to get a point on θ scale that corresponds to the maximum of this ICCC.

To conclude, Table 1 summarizes the category characteristic curves of a polytomous item with ordered response options scored 0 to m and the formula of the corresponding intersection points on the ability scale with reference to the definition of such scale values. This overall summary of the relations among ICRFs suggests the

Table 1

Studied ICRFs and Corresponding Intersection Points

ICCCs	Intersection Point	Notes
C_0, C_1	b_1	Model definition
C_1, C_2	b_2	Model definition
C_0, C_2	$^1/_2(b_1+b_2)$	The same as the peak of C_1 for a 3-category item
C_x, C_{m-x}	$\left(^1/_{m-2x}\right)\sum_{c=x+1}^{m-x} b_{ic}$	a form of intersecting point of x & $m-x$ curves
C_x, C_y	$\left(^1/_{m-x-y}\right)\sum_{c=x+1}^{m-y} b_{ic}$	a form of intersecting point of any two curves
C_0, C_m	$\left(^1/_m\right)\sum_{c=1}^{m} b_{ic}$	a form of intersecting point of 0 & m score curves

Note.C_v=item category characteristic curve for score v.

following proposed location indices.

Conclusion 1:

Based on the mathematical derivations mentioned above, we can propose alternative forms of location index (LI) for a polytomous item. The first form of LI is the average item category difficulties that takes all ICCCs into account, ($\mathrm{LI}_{\mathrm{mean}}$), by substituting $x = 0$ into Equation 33,

$$\mathrm{LI}_{\mathrm{mean}} = \tfrac{1}{m}\sum_{c=1}^{m} b_{ic}. \tag{34}$$

The second form of LI is the truncated (trimmed or Windsor) mean; that is, the average of item category difficulties that takes all ICCCs into account except the zero- and perfect-score curves, ($\mathrm{LI}_{\mathrm{trimmed\,mean}}$), by substituting $x = 1$ in Equation (33),

$$\mathrm{LI}_{\mathrm{trimmed\ \ mean}} = \tfrac{1}{m-2}\sum_{c=2}^{m-1} b_{ic}. \tag{35}$$

The third form of LI is the median of item category difficulties, ($\mathrm{LI}_{\mathrm{median}}$), which is a

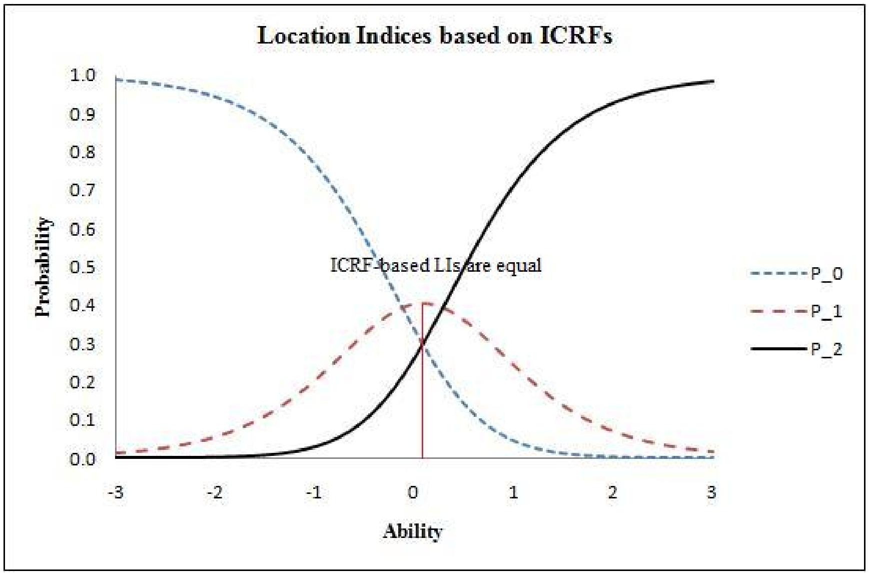


Figure 4: Item category characteristic curves (ICCCs) for a 3-category item.

possible choice in statistics,

$$\mathrm{LI}_{\mathrm{median}} = \mathrm{median}(b_{ix}s) = \begin{cases} b_{ix}^{(k)}, & \text{if } m \text{ is even} \\ 0.5(b_{ix}^{(k)} + b_{ix^*}^{(k+1)}), & \text{if } m \text{ is odd} \end{cases} \tag{36}$$

where $b_{ix}^{(k)}$ is the threshold parameter that have the k^{th} rank among the thresholds of the i^{th} item and has score x, $b_{ix^*}^{(k+1)}$ is the threshold parameter that have the $(k+1)^{th}$, and $b_{ix}^{(k)} \leq b_{ix^*}^{(k+1)}$.

Studying the item response function (IRF). The ease of calculating a polytomous IRF follows from the fact that an item response function (IRF) can be thought of as describing the rate of change of expected value of an item response as a function of the change in θ relative to an item's location b_i (Ostini & Nering, 2006). More succinctly, this can thought as a regression of the item score onto the trait ability (Lord, 1980; Chang & Mazzeo, 1994).

The previous LIs are based on the ICRFs, hence they are considered as local indices by the nature of information gained from curves of specific score categories. On the other hand, this is not the case in polytomous models. Chang and Mazzeo (1994) showed that the IRF for a polytomously scored item is defined as a weighted sum of the ICRFs

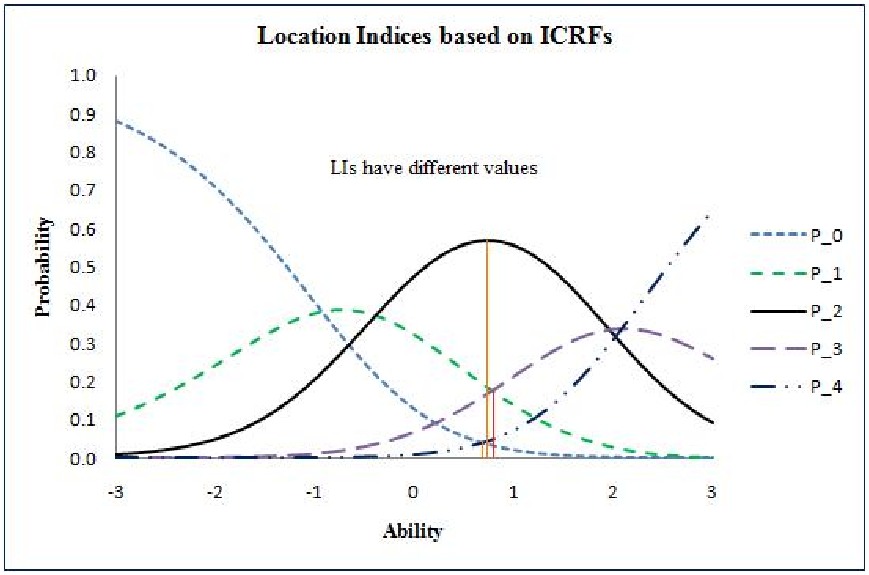


Figure 5: Item category characteristic curves (ICCCs) for a 5-category item.

(the probability of getting a particular score for a randomly sampled examinee of ability),

$$E[X_i] = \sum_x x \, P_{ix}(\theta). \tag{37}$$

The IRF, as defined in Equation 37, ranges from 0 to m (i.e., the maximum possible score category of an item). Chang and Mazzeo (1994) established the correspondence between an IRF and a unique set of ICRFs for two of the most commonly used polytomous IRT models (i.e., GRM, PCM, and GPCM), where they considered the GPCM and the PCM as one model. Specifically, they provided a proof for these models as follows: "If two items have the same IRF, then they must have the same number of categories; moreover, they must consist of the same ICRFs." The condition of the proof is that each item has its discrimination parameter that does not depend on the category on the response scale. The GRM, PCM, and GPCM satisfy this condition but the NRM does not because the latter potentially has different discrimination parameters for each the response categories.

Along the same lines, Akkermans and Muraki (1997) introduced an item response function (IRF) defined as a normalized expected score (i.e., weighted sum of ICRFs divided by the number of item categories) that ranges from 0 to 1. Akkermans and

Muraki's IRF differs in terms of the range from that introduced by Chang and Mazzeo (1994). Akkermans and Muraki introduced the gradient (i.e., first derivative) of IRF as an item discrimination function, $G(\theta)$,

$$G_i(\theta) = \frac{\partial \bar{T}_i(\theta)}{\partial \theta} = a_i^2 \left[\sum_{x=1}^{m} x^2 \, P_{ix}(\theta) - \left(\sum_{x=1}^{m} x \, P_{ix}(\theta) \right)^2 \right] = \frac{I_i(\theta)}{a_i}. \tag{38}$$

The polytomous IRF has various merits. First, the IRF carries the full information of the item and encompasses the partial amount of information included in ICRFs. Second, the expected score is valid to be applied to the most commonly used ordinal response models, (i.e., GRM, PCM, and GPCM). Third, it is well connected to Fisher information, see Equation 38. Due to these three properties of the IRF or expected score of a polytomous item, it is a worthy candidate as a central location parameter.

Conclusion 2:

The fourth form of LI is derived from the polytomous IRF. Dichotomous IRT models such as one-, two-, or three-parameter logistic models have an important feature that the conditional mean of item score (i.e., expectation) is the probability of answering the item correctly. Note that the dichotomous IRF uses the value of 0.5 (if there is no guessing) as a threshold to determine the item location where the highest score is one. Using same analogy, the index of a polytomous IRF corresponds to an expected score equals $0.5m$, where m is the highest possible score of $(m+1)$-response category item. Since this value has a global nature in that it considers the IRF, we call it $\mathrm{LI}_{\mathrm{IRF}}$.

$$\mathrm{LI}_{\mathrm{IRF}} = \theta : E[X_i] = \frac{m}{2}. \tag{39}$$

For example, the θ point that corresponds to the 0.5 m under the GPCM can be obtained through the following equation whose closed-form solution is complicated to produce,

$$\sum_{x=1}^{m} \left[(2x - m) \exp \left(a_i \left(x\theta - \sum_{c=1}^{x} b_{ic} \right) \right) \right] = m. \tag{40}$$

Therefore, an iterative algorithm is used to obtain the $\mathrm{LI}_{\mathrm{IRF}}$ for each polytomous item.

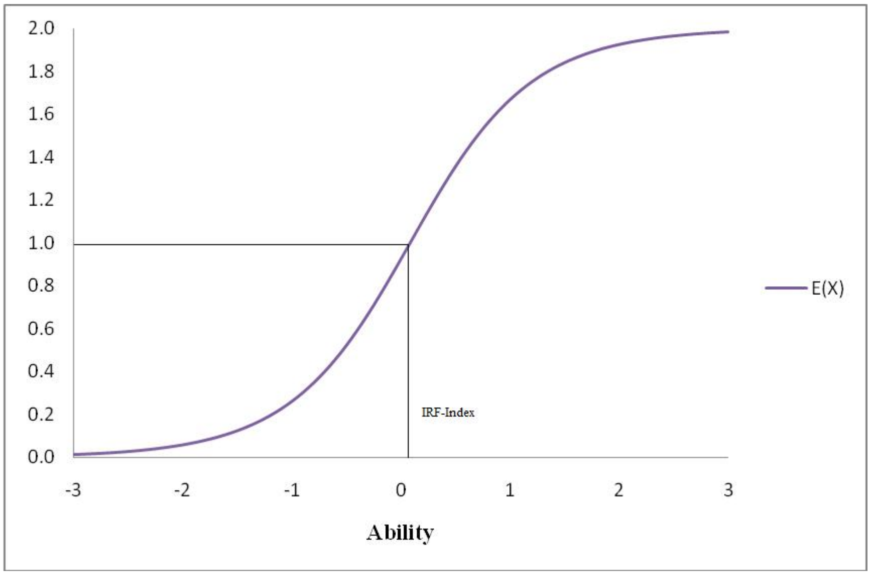


Figure 6: Item characteristic curve (ICC) for a 3-category item.

Appendix A presents the details of the Newton-Raphson method to get the approximate value of $\mathrm{LI}_{\mathrm{IRF}}$ for both the partial credit models (PCM and GPCM) and graded response model (GRM). Using the first derivative of P_{ix} to get the step of each iteration, the approximate value of $\mathrm{LI}_{\mathrm{IRF}}$ using the partial credit models is

$$\theta_{t+1} = \theta_t - \frac{f(\theta_t)}{f'(\theta_t)} \tag{41}$$

$$= \theta_t - \frac{\sum_{x=1}^{m} x\, P_{ix}(\theta_t) - \frac{m}{2}}{\sum_{x=1}^{m} x\, a_i\, P_{ix}(\theta_t)\, [x - \sum_{c=1}^{m} cP_{ic}(\theta_t)]}. \tag{42}$$

For a given polytomous item with 3-response categories, there is a correspondence between the $\mathrm{LI}_{\mathrm{IRF}}$ and the ICRFs-based LIs, see Equation 30 and Figures 4 and 6. For items with more than three response categories, the values of these indices are different, see Figures 5 and 7.

Analogy of location index in dichotomous and polytomous items. Some of location indices proposed for polytomous items reduce to indices used for dichotomous items. In other words, the location index (parameter) or difficulty parameter for a dichotomous item is based on the same techniques used to study the category response functions and item response function of a polytomous item. Using a two-parameter logistic model (Lord & Novick, 1968) to model the dichotomous item, we can have the

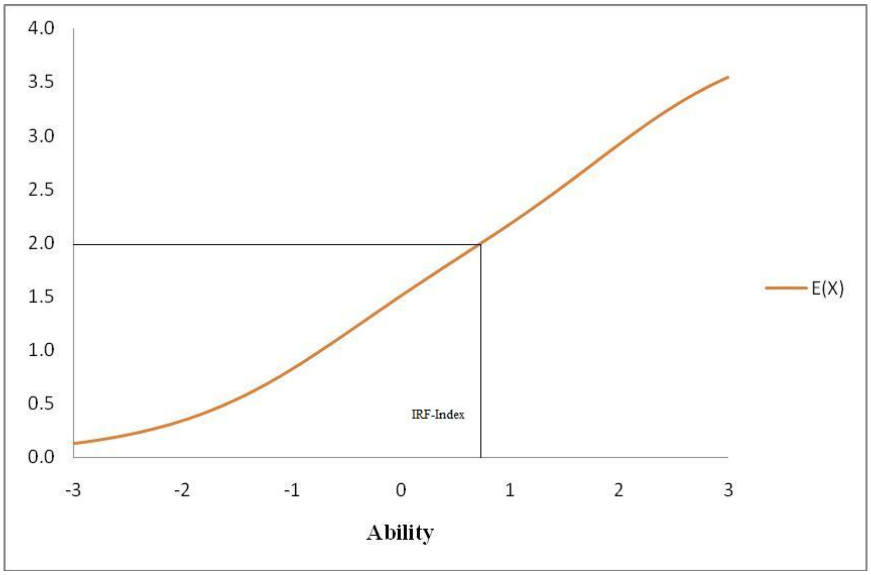


Figure 7: Item characteristic curve (ICC) for a 5-category item.

intersection point of the two characteristic curves of such an item (i.e., the curves that correspond to correct and incorrect answers) points to the difficulty (location) parameter, b_i. Figure 8 shows an example of the ICCs of correct and incorrect answers of a dichotomous item intersecting in a point corresponding to $\theta = 1.0$ (=item difficulty parameter). This was based on the two category characteristic curves, first approach used in the polytomous case.

Since there are more than two curves in the polytomous case, the median of category thresholds can act as a location parameter and it corresponds to the peak of the characteristic curve of category ${}^{m}/_{2}$ for m even, and the mean of the intersection of middle two curves of the intermediate scores, ${}^{m-1}/_{2}$ and ${}^{m+1}/_{2}$ for m odd.

An alternative is to use the truncated mean, this is a version of LI that considers only the category characteristic curves of partial scores and excludes the extreme response categories (i.e., zero and perfect scores). This form of an LI does not have a counterpart for dichotomous items because they have only two response options (correct/incorrect or perfect/zero scores).

While based on the point of view of expected scores or item response functions, the point on the ability scale corresponding to an expected score of 0.5 for dichotomous scoring represents an index of item difficulty. Figure 9 shows an ICC of the same item

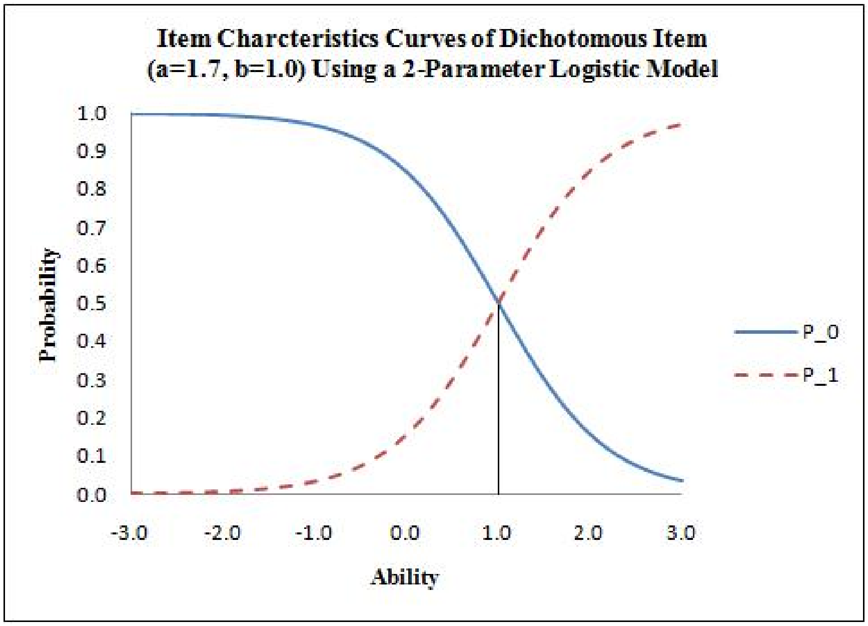


Figure 8: Item category characteristic curves (ICCCs) for a 2-category item.

whose difficulty parameter $b_i = 1$. This curve also represents the expected score conditional on ability level, and it is obvious that $\theta = 1.0$ where the expectation equals a half. This provides the basis for the second approach.

An example. The following is a numerical example of calculating the LIs for a polytomous item. Table 1 provides GPCM parameters of five items with four or five score categories and their corresponding LIs. Since, the four proposed LIs (i.e., $\mathrm{LI}_{\mathrm{mean}}$, $\mathrm{LI}_{\mathrm{trimmed\,mean}}$, $\mathrm{LI}_{\mathrm{median}}$, and $\mathrm{LI}_{\mathrm{IRF}}$) are identical for the case of 3-category items, therefore, they are not included in the table. The Table shows that the LIs differentiate in values from item to item. For example, they have similar values for items 3 and 5 but have different values for items 1, 2 and 4. For items more than five response options, the $\mathrm{LI}_{\mathrm{trimmed\,mean}}$ and $\mathrm{LI}_{\mathrm{median}}$ start to differentiate. From the table it is obvious these two LIs are the same.

Applications of Item Indices in Assessment

The availability of an index that represents the item location parameter (equivalent to difficulty or location parameter in dichotomous items) provides a summarized parameter of multiple category thresholds. Therefore, a polytomous item can also have two parameters to ease the usage of it in some situations, in addition to the

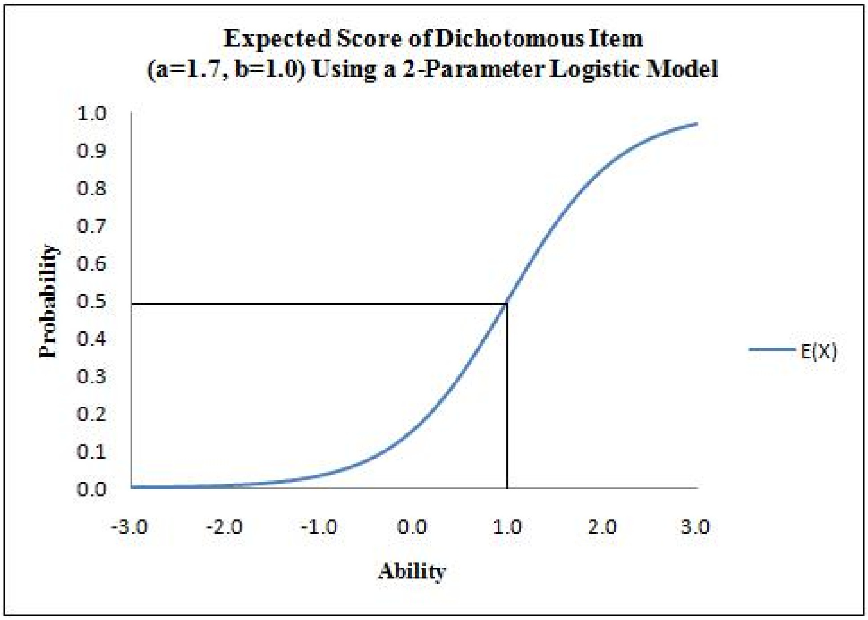


Figure 9: Expected score curve for a 2-category item.

Table 2

Item Parameters and the Corresponding Location Indices (LIs)

	GPCM Parameters					Location Indices			
Id	a	b_1	b_2	b_3	b_4	LI_{mean}	$\text{LI}_{\text{trim mean}}$	$\text{LI}_{\text{median}}$	LI_{IRF}
1	1.578	-2.718	0.183	2.725		0.063	0.183	0.183	0.170
2	0.894	-2.435	-1.215	0.734		-0.972	-1.215	-1.215	-1.050
3	1.688	-2.758	-1.352	-1.050	0.676	-1.121	-1.201	-1.201	-1.190
4	0.459	-1.102	-0.596	2.114	2.208	0.656	0.759	0.759	0.690
5	1.072	-2.343	-1.842	-1.328	-0.925	-1.610	-1.585	-1.585	-1.600

category-related information in each item. The b_i parameter of a polytomous item that is included in model parameterization such GPCM is meaningful now and has a theoretical background and beneficial usage.

In the context adaptive testing, non-information based item selection approaches can be presented such that an individual's estimated ability level is matched to a polytomous item's location index (LI). In particular, four proposed item selection methods in PAT are built based on the alternative forms of polytomous item LI. The choice of the next item to be administered is based on each form of the proposed index that matches the current ability estimate. For example, considering the $\mathrm{LI}_{\mathrm{IRF}}$, a global item index, computed for an item under a polytomous response model, the next item for administration is chosen based on matching $\mathrm{LI}_{\mathrm{IRF}}$ to the current estimate of examinee's ability.

Lima Passos et al.'s (2008) paper presented some findings regarding the item's ${}^{1}/{}_{2}(b_{i1} + b_{i2})$. This index, based on our analytical results, corresponds to the mean of item category thresholds, $\mathrm{LI}_{\mathrm{mean}}$, and the other LIs such as $\mathrm{LI}_{\mathrm{IRF}}$ in the case of 3-category items are equal as well, Equation 30. They found that the smaller the difference given by ${}^{1}/{}_{2}(b_{i1} + b_{i2}) - \theta$, the better (i.e., the more accurate) the tailoring between a selected item i and the underlying trait θ. This is the core idea of the Matching-LI procedure in polytomous adaptive testing and one of the main applications of polytomous item LIs.

Chapter 3

Matching Location Index as a Non-information Item Selection Approach in Polytomous Adaptive Testing

Introduction

One area of computerized adaptive testing (CAT) that received substantial attention in the measurement literature concerns the method of determining which item in an item pool should be administered at a given time of the adaptive-testing session. Currently, the available item selection approach is information-based (i.e., uses information functions as criteria for choosing items). Traditionally, the most widely used item selection approach is the maximum information (MI) criterion, whereby the item in the pool that has the highest information function value at the interim value of the estimated ability ($\hat{\theta}$) is selected for next administration (van der Linden & Pashley, 2000). The drawbacks of the MI approach is that the selected item maximizes the information at the current value of $\hat{\theta}$, not the true value of ability level (θ). As a result, for a test taker having an ability level θ, the MI produces a group of selected items that is optimal for the ability level $\hat{\theta}$ rather than θ). The extent to which $\hat{\theta}$ is apart from θ will affect how optimal the set of selected items (van der Linden, 1998).

Another drawback of the MI approach is the skewed distribution of item bank usage. This approach heavily selects items with high-a parameters. As a consequence, these high discriminating items are overexposed and this affects the security of the test and consequently its validity. Additionally, there are more items of low and medium discrimination in the item bank that are underexposed or are not used at all. It is known that the item pool provides a collection of pretested and qualified items. Therefore, each item passed a long process of investigation and hence it waste of time and money not to use all.

A non-information item selection approach is proposed to help enhance and solve the problems raised by using information-based methods (e.g., MI). The development of an overall location index (LI) for polytomous items provides the basis for that alternative approach in item selection. The Matching-LI item selection method is an attractive approach that uses the distance between the value $\hat{\theta}$ and the item LI as a criterion for item selection.

One main goal of the current study is to evaluate the performance of proposed method in terms of the estimation efficiency and item pool usage. The interest in exploring the properties of the Matching-LI approach of polytomous item selection stems from two observations. First, the use of polytomous items in adaptive testing is increasing (e.g., Dodd et al., 1995), and the development of testlets is paving the road for greater potential for polytomous item response models in adaptive testing environment. Second, the information functions of polytomous items can be irregular and even multimodal (Muraki, 1993), unlike the the information functions of dichotomous items that are always unimodal and symmetric.

The following sections are organized to provide a description to the non-information item selection approach through a Matching-LI method, followed by a presentation to the design of simulation study. Some results that compare the performance of Matching-LI and MI methods are presented followed by a discussion of the results and their implications.

Method

This section introduces the methods used to study adaptive selection of polytomous items for computerized tests. First, the traditional item selection method in PAT is reviewed because it provides a benchmark for the comparisons with the LI methods. Subsequently, a description of each form of the Matching-LI methods is provided.

Maximum information method. Maximum information is the standard method in item selection. The criterion to maximize is Fisher information measure at the interim ability estimate. Therefore, the next item after administering t items is to search for an item that provides the maximum information at $\hat{\theta}^{(t)}$,

$$i_{t+1} = \arg\max_{h} \left\{ I_h\left(\hat{\theta}^{(t)}\right) : h \in R_t \right\}, \tag{43}$$

where $I_h\left(\hat{\theta}^(t)\right)$ is the information function of item h at a specific trait estimate.

The information function depends on the polytomous item response model considered and fit to the data. In the current study, it is of interest to investigate the properties of the studied item selection approaches for an item bank consisting of ordered response items, such items that are fit by the GPCM (Muraki, 1992).

Matching-LI methods. In the last Chapter, four polytomous item location indices (LIs) were proposed to represent a central or overall difficulty-like parameter for each item. The first formula for a location index, LI_{mean}, is the mean of step difficulty parameters: $b_{i1}, ..., b_{im}$. The second formula for a location index, $\text{LI}_{\text{trimmed mean}}$, is the truncated mean of step difficulty parameters that $b_{i2}, ..., b_{i,m-1}$. The third proposal of a location index, $\text{LI}_{\text{median}}$, is the median of step difficulty parameters: $b_{i1}, ..., b_{im}$. The fourth form of a location index is related to the polytomous item response function, LI_{IRF}.

Polytomous item selection that is based on its unique LI is presented. Therefore, four proposed item selection methods in PAT are built based on the alternative forms of item LI. The choice of the next item to be administered is based on each form of the proposed index, LI_h, that matches the current ability estimate $\hat{\theta}^{(t)}$; that is, this method searches for an item of minimum distance as follows

$$i_{t+1} = \arg\min_{h} \left\{ \left| \hat{\theta}^{(t)} - \text{LI}_h \right| : h \in R_t \right\}, \tag{44}$$

where LI_h is a location index of a polytomous item h from the remaining items in the item pool not administered yet, R_t.

Simulation Study Design

To assess the performance of Matching-LI Method as an item selection rule in PAT, a simulation study was conducted. The study used Monte carlo simulation to achieve its objectives of assessing the performance of the proposed item selection methods and comparing its effectiveness to other existing methods. This section describes the simulation design.

Data generation. The input used to simulate data consists of: (a) an item pool and item parameter distributions, (b) distribution of target population, and (b) test length and specification in the following details are described:

Item pool and distributions of item parameters. The item pool consists of 300 polytomous items. The pool has different items with different numbers of score categories, 3, 4, and 5. The item pool contains 200 3-category items (refer it as Type III), 60 4-category items (Type IV), and 40 5-category items (Type V). This satisfies the rule of thumb that the item pool size is at least 12 times the test length (Stocking, 1994).

The item discrimination parameters (i.e., a_i) were generated from uniform distribution $U[.25, 3.0]$ and category parameters were generated from uniform distribution $U[-3, 3]$. The use of a uniform distribution ensured that an adequate number of items existed at low, medium, and high levels of difficulty (e.g., Penfield, 2006).

The distance between the first and last category parameters (i.e., thresholds) affected the shape of the information function of an item. The items with the same set of thresholds yielded the same total amount of information across the entire ability continuum, but different ordering of the thresholds affected the peakedness of the item information curve. The peakedness of the curve increased as the degree of deviation from the sequential order of the thresholds increased (Dodd & Koch, 1987).

The item parameters considered in the item bank have no reversals (i.e., the category parameters are in ascending order). Also, we have considered reversals in 20% of the item pool by altering the order of the same set of thresholds to form different items and a new item pool.

Ability distributions. One thousand values of θ were drawn from a standard normal distribution (i.e., $N(0,\ 1)$). Every simulated examinee receives a number of PATs, each is directed by a specific item selection method. Note that the first item was always selected randomly from the Type III items (3-category items).

Test length and specification. Test length is also important to a PAT. It affects the accuracy of the final ability estimation. A fixed-length test of 15 items is delivered to each examinee. The specification of test should consist of 10 Type-III items, 3 Type-IV items, and 2 Type-V items. The different item types have the same scoring weights so the test total score will be 52 (i.e., $10 \times 3 + 3 \times 4 + 2 \times 5$).

PAT setting. A FORTRAN program was written to provide PATs for each examinee. It has a main program and several sub-routines that required tasks such as calculating the LIs for each item, the probabilities of GPCM, ability score estimates (e.g., EAP estimates), and item information indices.

Item selection algorithms. Two item selection algorithms were considered in this investigation. The first method selects the next item by matching the current ability estimate to an item location index (LI) as described previously (Matching-LI). There were four different forms based on these different proposed indices. The first version uses the mean item category location index for matching, (LI_1), the second version uses the truncated mean of item category parameters, (LI_2), the third version uses the median of item category parameters, (LI_3), and the fourth version uses the polytomous IRF whose location index is corresponding to a expected score of $.5m$, ($\mathrm{LI}_{\mathrm{IRF}}$).

The second method is maximum item information (MI), the traditional method where the next item is selected based on maximizing information index at the present ability estimate.

Ability estimation. Expected a Posteriori (EAP) with a prior distribution $N(0, 1)$ was used for scoring examinees. With regard to the initial ability estimate, it was set equal to zero, the mean of the ability distribution for the population.

Evaluation criteria. The study used two main evaluation criteria for comparison among the different item selection methods: (a) measurement precision and (b) item pool usage as measured by item exposure rates to show how effectively the item pool was used.

Measurement precision. The ability parameters recovery was used to evaluate the proposed Matching-LI method compared to a standard method. Three statistics were used to determine how close the estimated ability estimates were to the original ability estimates. Evaluation criteria used were: (a) Bias, (b) Mean Square Error (MSE), and (c) Pearson correlation coefficient. These indices capture the measurement precision.

Average bias (Bias) was estimated using Equation 45 below. In Equation 45, let θ_j , $j = 1, , N$ be the original ability of N examinees and $\hat{\theta}_j$ be the respective estimator from the PAT using different item selection methods. Then the estimated bias was computed as

$$\mathrm{Bias} = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_j - \theta_j\right). \tag{45}$$

MSE was calculated using

$$MSE = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_j - \theta_j\right)^2. \tag{46}$$

The smaller the bias and the MSE, the better item selection method. Also, the conditional bias and MSE are also considered given a small range over the ability continuum. Therefore, the ability continuum is divided into five homogenous groups: the lowest 20%, 20-40%, 40-60%, 60-80%, and the highest 20%.

The third statistic considered for the measurement precision was the Pearson product moment correlation between the estimated and original ability, $\rho_{\theta,\hat{\theta}}$; that is,

$$\rho_{\theta,\hat{\theta}} = \frac{\sum_{j=1}^{N}(\theta_j - \bar{\theta}_j)(\hat{\theta}_j - \bar{\hat{\theta}}_j)}{S_\theta S_{\hat{\theta}}}. \tag{47}$$

Item pool usage. Using the item exposure rate provide a measure of which items are selected by different algorithms. Item exposure rate is defined as the ratio of the number of simulees who receive an item and the total number of examinees. Useful information can be obtained through exposure rates such as ratios of over-, under-, and never-exposed items in the item pool. The χ^2 statistic, a descriptive measure to indicate the skewness of item exposure rate distribution (Chang & Ying, 1999), was computed by

$$\chi^2 = \frac{\sum_{i=1}^{M}(r_i - L/M)^2}{L/M}, \tag{48}$$

where r_i is the exposure rate of item i, L is the test length, and M is the item pool size. It quantifies the discrepancy between the observed and the ideal, uniform distribution and is considered a good indicator of the efficiency of item pool usage. The smaller the χ^2 statistic the better the exposure control.

Results

This section presents the relationship between the a-parameter and the one corresponding LI for each item in the item pool, followed by the comparison of the results for both measurement precision and item pool usage indices. (Note that all results of the method based on truncated-mean, $\text{LI}_{\text{trimmed mean}}$, is same as the method based on median,

Table 3

Descriptive Statistics of Simulated Item Pool (No Reversals)

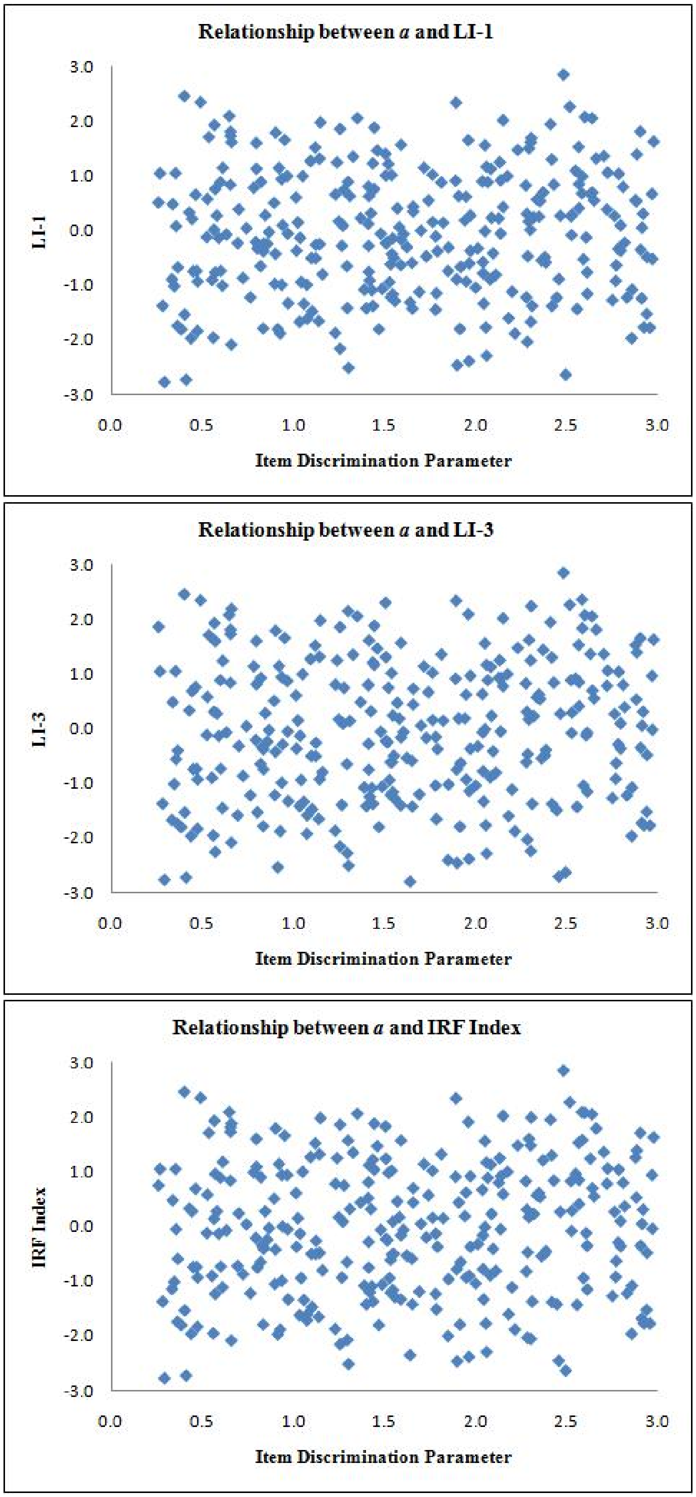
Statistic	a	b_1	b_2	b_3	b_4	$b_{im} - b_{i1}$
No. Items	300	300	300	100	40	300
Mean	1.621	−1.279	0.522	1.362	1.849	2.461
SD	0.775	1.313	1.522	1.209	0.954	1.497
Minimum	0.255	−2.992	−2.799	−2.265	−0.925	0.015
Maximum	2.975	2.772	2.993	2.992	2.965	5.931

$\text{LI}_{\text{median}}$; because there are no reversals within any item of the item pool.)

Distribution of *a*-parameter and corresponding LIs in the item pool. Descriptive statistics (i.e., mean, standard deviation, minimum, and maximum) of the GPCM parameters of the item pool are provided in Table 3. In addition to that the distances between the first and last threshold parameters of items in the pool are provided. It ranges from almost very small range, 0.015, to large range, 5.931, with mean = 2.461 and SD = 1.497.

Also, the bivariate distribution of a-parameter and LIs for all items in the item pool is depicted in Figure 10. It shows that the a-parameter is uniformly distributed and the different LIs are uniformly distributed as well, which is expected as the category parameters are generated originally from a uniform distribution. It is noticeable that there is no positive relationship between a-parameter and the proposed LIs in the item pool.

Measurement precision. Table 4 presents the overall measurement precision indices. It is expected that the maximum information method will be the most preferable with respect to measurement precision. It is considered as a baseline here for high precision. Also, it is obvious that the four matching-LI methods using different item indices result in a slight loss in measurement precision compared with the maximum information method. Among these four forms of matching methods, the IRF-index based method is slightly more precise than the other three matching methods; the IRF-index method yields slightly smaller bias, MSE, and larger $\rho_{\theta,\hat{\theta}}$. It is more adequate to claim


Relationship between *a* and LI-1

Relationship between *a* and LI-3

Relationship between *a* and IRF Index

Figure 10: Distribution of a-parameter and LIs for all items in the item pool.

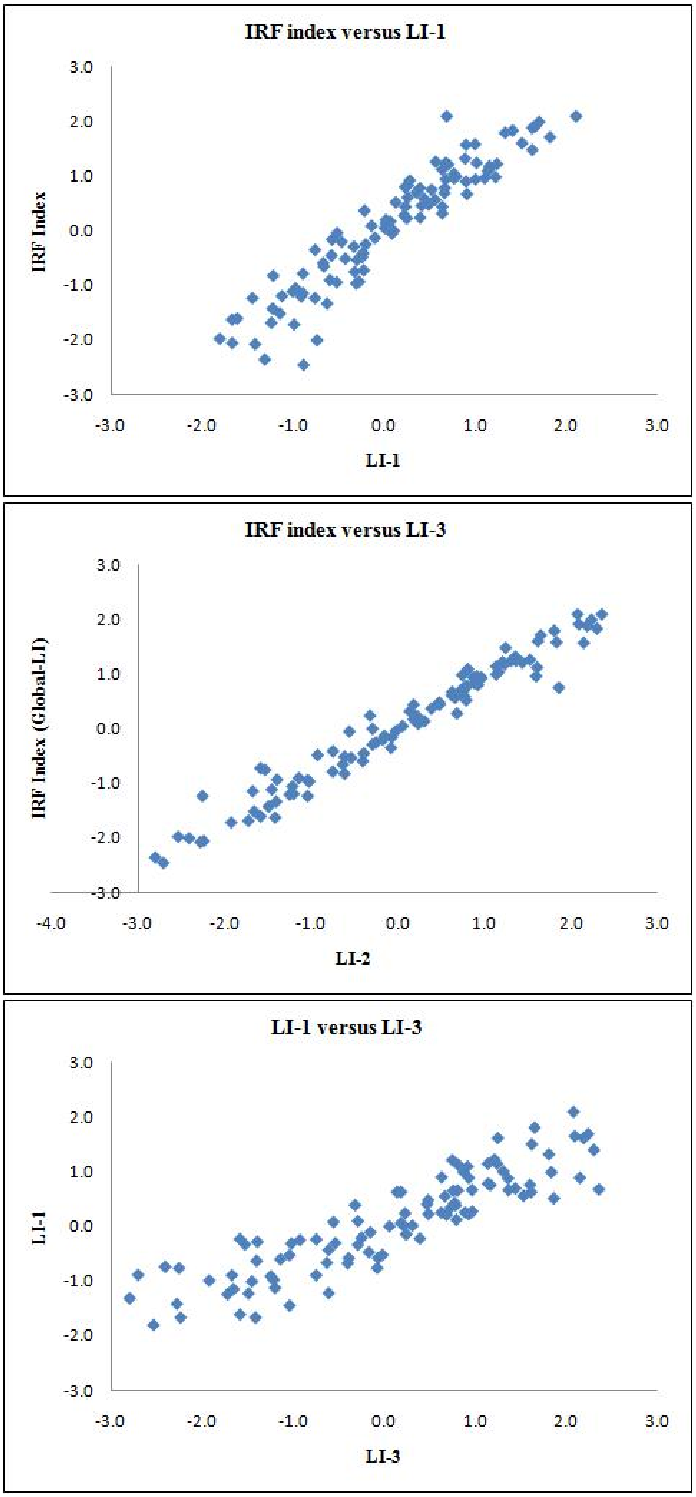


Figure 11: Distribution of LIs for all items in the item pool.

Table 4

Overall Measurement Precision Indices Under Different Item Selection Methods (N=1000, M= 300, L=15)

Methods	Bias	MSE	$\rho_{\theta,\hat{\theta}}$
Matching-LI$_{\text{mean}}$	0.004	0.086	.953
Matching-LI$_{\text{trimmed mean}}$	0.008	0.078	.957
Matching-LI$_{\text{median}}$	0.008	0.078	.957
Matching-LI$_{\text{IRF}}$	0.002	0.076	.958
MI	0.002	0.029	.984

Note. MSE=mean squared error; LI=location index; MI=maximum information.

that these four matching methods are comparable in terms of measurement precision.

Conditional Bias and MSE are reported in Table 5. It is clear that the MI method surpasses all the other methods across different ability levels. Also, the overall conditional bias and MSE at different levels are of very low level for all considered item selection methods; conditional bias reaches a maximum of 0.12 in absolute value and conditional MSE reaches a maximum of 0.10. In general, at all levels of θ, the loss in measurement precision for the four LI methods is very small compared to the MI method. The overall measurement precision indices supports the claim that the matching-LI methods are comparable, and this also applied to conditional bias and MSE. All selection methods work slightly better at the three intermediate groups than at the lowest- and highest-20% groups of examinees.

Table 5

Conditional Bias and MSE Under Different Item Selection Methods (N=1000, M= 300, L=15)

		Matching-LI				
Statistic	Group	LI_{mean}	$\text{LI}_{\text{trimmed mean}}$	$\text{LI}_{\text{median}}$	LI_{IRF}	MI
Conditional Bias	0-20%	-0.121	-0.114	-0.114	-0.097	-0.006
	20-40%	-0.027	-0.026	-0.026	-0.041	0.011
	40-60%	0.015	0.025	0.025	-0.016	0.006
	60-80%	0.053	0.044	0.044	0.044	0.000
	>80%	0.099	0.113	0.113	0.122	0.001
Conditional MSE	0-20%	0.085	0.089	0.089	0.085	0.029
	20-40%	0.085	0.067	0.067	0.079	0.032
	40-60%	0.099	0.081	0.081	0.075	0.027
	60-80%	0.079	0.069	0.069	0.065	0.024
	>80%	0.083	0.082	0.082	0.076	0.034

Note. MSE=mean-squared error of estimation; LI=location index; MI=maximum information.

Item pool usage. Table 6 shows the item pool usage indices. The matching method based on LI_{mean} had no items that were overexposed, and the other three matching-LI methods have only one overexposed item (i.e., less than 1% of the item pool). Also, the mean method manages to control item exposure rate in that; it leads to the lowest maximum exposure rate (.158) and the other three methods are slightly higher than the standard exposure rate, .20, while the MI method have a fairly large maximum exposure rate, .743. All the four proposed methods have no items that have been never exposed. Although, these four methods have a percentage of approximately 4–7% of underexposed items; underexpose rates are very low. The MI is much worse compared to the four matching LI methods. The MI results in 10% overexposed item, about 22% never-exposed items, and more than 50% underexposed items. Regarding the skewness of item exposure rates, the matching methods are the best (have lower statistics), and the

Table 6

Overall Item Pool Usage Indices Under Different Item Selection Methods (N=1000, M= 300, L=15)

	Matching-LI				
Index	LI_{mean}	$\text{LI}_{\text{trimmed mean}}$	$\text{LI}_{\text{median}}$	LI_{IRF}	MI
Max. Exposure Rate	.158	.232	.232	.201	.743
Overexposed (%)	0	0.33	0.33	0.33	10.00
Underexposed (%)	3.67	7.00	7.00	6.67	51.33
Never-exposed (%)	0	0	0	0	21.67
χ^2	6.86	9.13	9.13	8.46	81.39

Note. LI=location index; MI=maximum information.

MI method is the worst.

Figure 12 provides a visual representation of the item exposure rate distribution for each item selection method (the method based on truncated-mean, $\text{LI}_{\text{trimmed mean}}$, is not presented because its exposure-rate distribution is same as the method based on median). The items were ranked based on the value of discrimination parameter, a. It shows clearly the uniform usage of items based on proposed item selection methods. On the other hand, as expected the MI method depends largely on the a parameter.
The Figure 12 provides the actual values of discrimination parameters to give an indication of which values have been chosen.

Table 7 presents the percentages of the overexposed items from the different number of item categories of the mixed item pool. The over-exposed items were mostly from the 4-category items compared to the other items of 3 and 5 categories.

Summary

The implementation of large scale assessment via PAT has generated great challenges to practitioners. The matching methods based on the proposed location indices are clearly superior to the MI method in terms of balancing item pool usage. The Matching-LI methods used over 99% of the item pool thus use all available items and do not waste any of them. It is well known that item writing is expensive and time

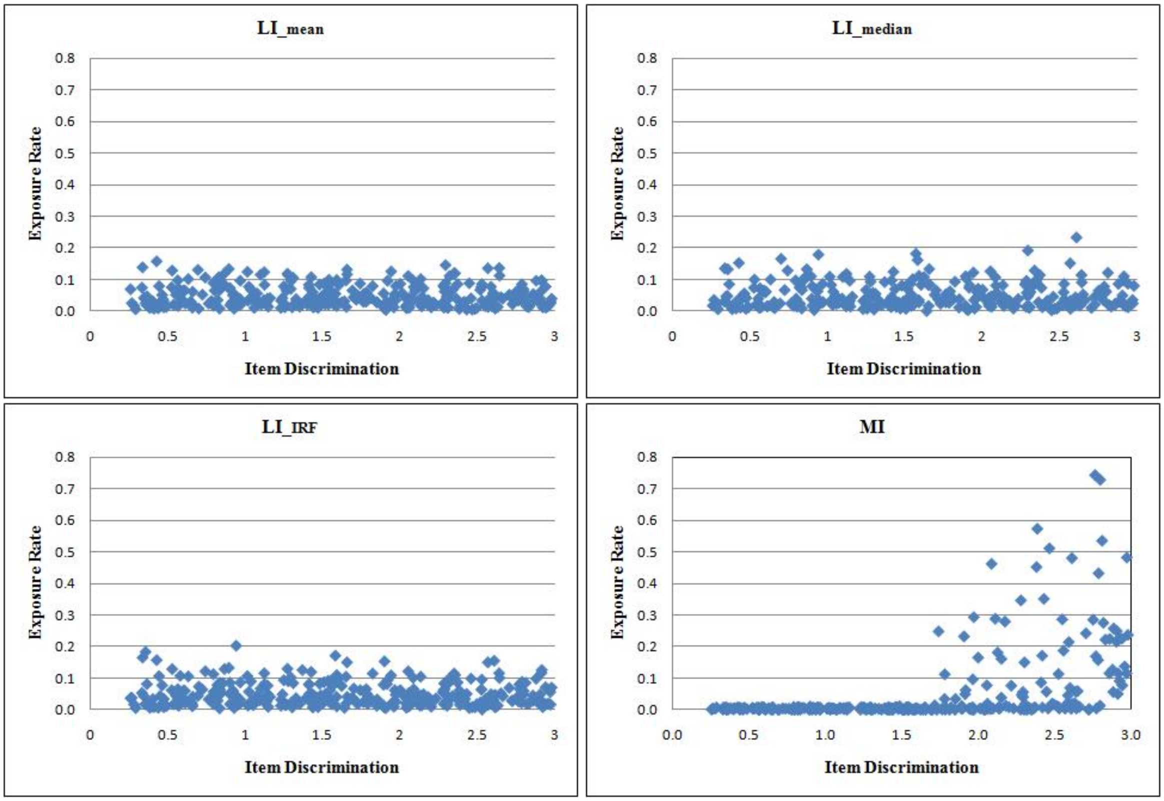


Figure 12: Distribution of item exposure rates of the item selection methods.

Table 7

Percentages of Exposure Rates for Over-exposed Items (N=1000, M= 300, L=15)

	Matching-LI			
Number of Categories	LI$_{\text{mean}}$	LI$_{\text{median}}$	LI$_{\text{IRF}}$	MI
3	0	0	0.5	9.00
4	3.33	3.33	0	8.33
5	0	0	0	7.50

Note. LI=location index; MI=maximum information.

consuming. These items must pass a long review process to be included in an operational item pool. Therefore, these proposed LI methods have a very desirable feature of effectively utilizing the item pool. The decrease in performance of the proposed LI methods in terms of precision of ability estimation is negligible. An added feature of the LI methods is that they are easy and fast to implement because they are based on simple statistical algorithms. The findings of current study support the findings of Lima Passos et al. (2008) showing that a criterion that allows for a quick convergence to the most suitable location parameters produces steady RMSE and bias curves.

The Matching-LI_{IRF} (or matching-expected-score) method is slightly better than the other three LI methods in terms of having a high level of measurement precision and also has a successful usage of item pool and exposure control. All the proposed methods perform as if each has an exposure control inherited that prevent items from being overexposed or never being exposed, and keeps most of the items away from being underexposed. Recall the test length used here is a short test (i.e., 15 items). In conclusion, the matching-LI method is very promising in PAT. The LI methods could be modified to deal with more constraints, such as content balancing, or different item pool structure. These modifications and their performance with skewed ability distributions require further investigation.

In terms of future research, the non-information item selection approach needs to be investigated more in other contexts other than fixed-length tests, such as in delivering an adaptive test of variable-length. A new version of polytomous a-stratification strategy can be developed and investigated. In other words, the Matching-LI within stratum approach is applied. Other factors and developments to study including the item pool structure and the effect of a correlation between the item discrimination parameters and the corresponding LI, that would necessitate blocking LI within each stratum. The item pool used here provided no relationship between a-parameter and the LI so it is sufficient to stratify the item pool using a-parameter only without considering blocking the distribution of LI.

Based on the results displayed in the current Chapter with regard to the primary assessment of the the Matching-LI method's characteristics. In the following Chapter, a follow-up study of the Matching-LI method was conducted in a real setting. Additionally, more item selection procedures were developed and investigated.

Chapter 4

Improving the Performance of Matching-Location-Index Method With Information Indices

Introduction

The mechanism of item selection is just one crucial component along with other components in adaptive testing that needs to be improved and be flexible to different purposes of testing. Ranking students in terms of their abilities and classifying these students into performance levels are considered two major purposes in traditional adaptive testing. Each item selection method used in adaptive testing has its strengths and weaknesses. Therefore, it is better to know these characteristics that may enable us to incorporate them in a procedure of combined benefits. The idea of combining more than one procedure in selecting test items is to acquire as much as possible of these procedures' pros together.

The item selection criteria, such as Fisher information, Kullback-Leibler (KL) global information, or others, have no general recommendation to be used in polytomous adaptive testing (Veldkamp, 2003). During the early stage of adaptive testing, item selection criteria based on Fisher's information often produce less stable latent trait estimates than the KL global information criterion (Lima Passos, Berger & Tan, 2008). This may cause problems; when the ability estimate is not close to the true value of the ability parameter, Fisher information will produce inefficient item selection.

Cheng, Chang, Douglas and Guo (2009) reported that simulation studies for dichotomous items revealed that the global information method outperformed the maximum information early in the sequence (Chang & Ying, 1996), indicating it would be a better choice for short adaptive tests. Also, because it would not always select the item with the highest discrimination parameter at every difficulty level, it takes some steps toward addressing item exposure. However, neither the maximum information nor the global information method adequately address balancing item exposure and have no features for satisfying test constraints.

Under the GPCM, Pastor, Dodd and Chang (2002) used the *a*-stratification design to control item exposure. Due to the presence of multiple category difficulty parameters per item, the item selection was based on the information function. They did not use the a-stratification design to its full capacity in PAT. So, we suggest using one index rather than several step difficulties per item. One possibility is the summation of the step difficulties for an item as an index for item pool stratification. To choose the next item, we may find some sort of matching this proposed index with the current estimated ability is used.

Another approach to enhance the item selection algorithm is the use of suitable item information measures based on the stage of test (i.e., Kullback-Leibler (KL) information in the early stages and maximum information (MI) that uses Fisher information in the later stage(s) of CAT course). The usage of multiple item selection criteria was previously used in ability estimation; expected a posterior (EAP) used early in a test and later maximum likelihood estimator (MLE).

Certainly, the development of item selection criteria in polytomous adaptive testing is possible as some new ideas can be achieved and can be promising solutions to problems raised by multi-response items. The performance of an item selection method is affected by the other components in PAT, so we consider them as well in the current study for better assessment of the proposed LI.

To conclude, two main points were raised here, first, dealing with the inaccurate trait estimates at the early test stage, second, heavily usage of high-a items on the expense of low-a items (i.e., the unbalanced utilization of item pool). To circumvent such problems alternative item selection criteria are introduced. Therefore, the purpose is to refine the performance of Matching-LI method that were introduced previously and measure the behavior of the new methods and evaluate their strengths and weaknesses for efficient measurement.

In the following section, three proposed item selection methods are introduced: (a) hybrid method, (b) polytomous a-stratification strategy, and (c) stage-based information method.

Item Selection Methods

In the current section, we introduce the new procedures for selecting polytomous items for adaptive testing. Based on the preliminary results of Ali and Chang (2010) and the previous chapter that assess the performance of the Matching-LI method compared to the maximum information (MI) method, it was clear the advantage of using such method for maintaining item pool and controlling exposure rates without applying strategies of item exposure control. In the current study, a modified version of Matching-LI method that incorporates an information measure into the item selection criterion is presented. Also, two other procedures are presented: polytomous *a*-stratification strategy and alternating information index (or stage-based Information) method.

Hybrid matching-LI and information method.

This version of item selection method combines two procedures in one to form a new criterion in item selection for polytomous adaptive testing. It combined matching-LI and Fisher information criteria together.

Originally the non-information item selection procedure depends on matching item's LI (e.g., $\mathrm{LI}_{\mathrm{mean}}$ or $\mathrm{LI}_{\mathrm{IRF}}$) to the interim ability estimate, $\hat{\theta}$, during the test course. The criterion is to minimize the absolute difference between $\hat{\theta}$ and LI_h. So the i_{t+1} item to be selected has the minimum distance between the current $\hat{\theta}^{(t)}$ based on t items and its location parameter, as follows

$$i_{t+1} = \arg\min_{h} \left\{ \left| \hat{\theta}^{(t)} - \mathrm{LI}_h \right| : h \in R_t \right\},$$

where LI_h is a location index of a polytomous item h from the remaining items in the item pool not administered yet, R_t. Minimizing the distance between the estimated ability level and LI is equivalent to maximizing the reciprocal of this distance.

As indicated by previous results, the Matching LI procedure is performing very well in terms of the item pool usage, but it is slightly less efficient in ability estimation compared to the criterion that maximizes Fisher information function. To help enhance the precision of test taker's ability estimation in addition to the efficiency of item pool usage, a hybrid method that combines both the information function and Matching LI approach in one item selection is proposed where the item is selected by maximizing the following criterion

$$i_{t+1} = \arg\max_{h} \left\{ \frac{1}{\left|\hat{\theta}^{(t)} - \mathrm{LI}_h\right|} I\left(\hat{\theta}^{(t)}\right) : h \in R_t \right\}. \tag{49}$$

This new criterion is maximizing the product of the information function $I\left(\hat{\theta}^{(t)}\right)$ and the reciprocal of the distance between item's location parameter and the interim trait estimate, as appeared in Equation 1. The techniques plays the role of balancing the powers of both procedures and hopefully perform better. In other words, the hybrid method tries to take the estimation efficiency from the first term, MI criterion, and to get the balance pool utilization from the second term, the Matching-LI criterion. The hybrid comes from the nature of each term, information and non-information bases, respectively.

Alternating information index (or stage-based information) method. The MI criterion's strong dependence on the item's discrimination parameter, for instance, gives rise to a twofold nuisance. Because the most discriminative items are also the most informative, their selection leads not only to the unwelcomed side effect of overexposure of valuable items from the outset of the test. It also underlies a problem known as the attenuation paradox (Lord & Novick, 1968; van der Linden & Pashley, 2000). The attenuation paradox can be a serious hindrance in the early stages of an adaptive test, where the bias of the estimator $\hat{\theta}$ can be relatively high (i.e., $\hat{\theta}$ might strongly deviate from the true trait value θ_0). Selecting a highly discriminative item to match an uncertain $\hat{\theta}$ might provide little information on the true trait. Consequently, a mismatch between the items, selected to fit the newly updated estimator, and the true value can arise, leading to further inefficiency of the succeeding $\hat{\theta}$. A considerable delay of $\hat{\theta}$ sequential convergence to the true value can follow.

In addition to the aforementioned reasons, the idea behind proposing the stage-based information method comes from the stages that a test goes through. It has been distinguished between three main stages of ability estimation; namely, early, interim, and final stages as suggested by van der Linden and Pashley (2000). The logic behind alternating the information criterion used for item selection was used before in the context of trait estimation; where the different stages affect the choice of the ability estimators. For example, in dichotomous CAT more than one method is used during the course of test starting with expected a posteriori (EAP) at the initial stage where a higher possibility to have zero or perfect score pattern that makes maximum likelihood estimator (MLE) undefined. Later when both zeros and ones are present at response pattern, switching to the MLE would be reasonable. Using such alternation would benefit the trait estimation in adaptive testing.

By the same token in the context of item selection criteria, it was known that KL global information criterion is better than the MI in the early stages of the test. Therefore, the proposed method suggests to use a global information index for the early phase where much uncertainty about the trait estimate is available that enable us using the MI. On the other hand, it has been argued that the global information seems more appropriate to counteract the attenuation paradox (Lima Passos et al., 2007). Therefore, it is beneficial to globally search for the polytomous items until a point where enough items are administered and consequently high expectation of more accurate trait estimate. At that time, the information index used is shifted to another index suitable to such stage. It is needed to know if the estimation accuracy will be affected if the transition point differs. In other words, when to alternate the information index, e.g., do we use the KL information at the early and interim stages of a test or only for the first stage.

This alternating information index method is another way of combining two criteria to get the utmost benefits out of them by getting their unique features and avoiding their weak points.

Polytomous *a*-stratification strategy. Dichotomous *a*-stratification method was proposed for selection as a solution to problems raised by the traditional method based on Fisher information (see for more details Chang & Ying, 1999; Qi, Chang, & Ying, 2001). In the dichotomous stratified design, the test is partitioned into a number of stages and the pool is partitioned into strata, where items of each stage are selected from a particular stratum. The selection is done by matching the difficulty parameter of an item with the examinee's trait estimate. In the polytomous version of *a*-stratification design, Pastor et al. (2002) provided an alternative item selection procedure where maximum information is used within each stratum to choose the next item for adminstration. This is because there are multiple difficulty steps per polytomous item and there is no single item parameter available to be used for selection.

After the development of such single parameter (i.e., LI, that represents the difficulty of a polytomous item), matching such index within each stratum becomes possible. LI can be one form of $\mathrm{LI}_{\mathrm{mean}}$, $\mathrm{LI}_{\mathrm{trimmed\ mean}}$, $\mathrm{LI}_{\mathrm{median}}$, or $\mathrm{LI}_{\mathrm{IRF}}$. The proposed method is Matching-LI method with stratification to the item pool. Therefore, polytomous *a*-stratification strategy (PASS) uses one of these proposed polytomous item indices in replacement of the *b*-parameter in the dichotomous *a*-stratification strategy. Blocking the LI within each stratum may not be needed. In other words, we may not need to have a uniform distribution of that index through the strata if there is no positive significant relationship between *a*-parameter and the LI for the assigned item pool. At that time, it is sufficient to stratify the item pool using *a*-parameter only without considering the distribution of LI in each stratum. The selection within each stage will be based on matching the estimated trait as well. The main difference of this method compared to that used by Pastor et al. (2002) is using a non-information approach, e.g., Matching-LI, instead of using the information-based item selection method at each stage of the test.

Also, it may consider alternative criterion other than the discrimination parameter to stratify the item pool. One possibility is to use the width of item information function. Dodd and Koch (1987) reported that shape of item information is not solely related to the distance between step values for such an item but also it is related to the sequence of these step values (i.e., the existence of reversal).

Data and Study Design

The purpose of the current study is to evaluate the performance of proposed item selection methods in addition to using the Matching-LI methods using two indices, $\mathrm{LI}_{\mathrm{mean}}$ and $\mathrm{LI}_{\mathrm{IRF}}$ and the MI and KL as reference information criteria. This results eight item selection procedures: Hybrid, Stage, PASS, Matching-$\mathrm{LI}_{\mathrm{mean}}$, Matching-$\mathrm{LI}_{\mathrm{IRF}}$, MI, and KL.

The study design used item parameters from real items to get a more realistic results. The following section provides more information about the study design and the implementation of PATs.

Data.

Item parameters. An item bank has been formed from ninety-three (93) real items collected from a real application. The item parameters were calibrated using a large sample of students who took the National Assessment of Educational Progress (NAEP) reading main assessments from 2000-2007. Each item has four category response options. The estimates of item parameters were obtained by fitting the GPCM to the data and using a NAEP BILOG/PARSCALE program. The item parameters, the discrimination and three step difficulty for each item, are reported in addition to the LIs corresponding to each item in Appendix A.

In this simulation, a short-length test of nine items was used. With regard to the early stages, we can see the performance of such methods in the first few items for example five items or so to get the difference in ability estimation and item pool management. Also, it shows the type of items that are recommended by each criterion.

The item parameters considered in the item bank have a number of items with reversals (i.e., the category parameters are not in ascending order from score zero to score m). The items of reversed item category thresholds had only one reversal and there are 19 such items (i.e., 20% of the actual item pool). Table B1 in Appendix B provides the item parameters of item pool.

Response generation. One thousand simulees were generated from a standard normal distribution. Each simulated examinee received eight PATs, each of which was directed by a specific item selection method.

The response of each simulee was generated in the following manner: The probability of answering each of the four score categories (i.e., 0 to 3) conditional on the known item parameters and simulees true trait value was calculated. Thereafter, three cut scores for the cumulative probabilities were determined, t_1, t_2, and t_3, with $t_1 = P_{i0}$, $t_2 = P_{i0} + P_{i1}$ and $t_3 = P_{i0} + P_{i1} + P_{i2}$ where P_{i0}, P_{i1} and P_{i2} are the probability of getting a score category 0, 1, or 2, respectively, calculated via Equation 7 in Chapter 1. Based on t_1, t_2 and t_3, four category intervals were obtained, namely, $(0, t_1)$, (t_1, t_2), (t_2, t_3), and $(t_3, 1)$. A random number was drawn from the uniform distribution $U(0, 1)$.

The category (i.e., the interval containing the random number) was assigned an item-score of 1, and 0 otherwise.

Procedure.

PAT initialization. The polytomous adaptive tests start with a randomly chosen item from the item pool by the item selection methods and the initial trait value was assumed to be zero, the mean of trait distribution.

Stoping rule. The stopping rule used here is administering a fixed number of items to all simulees. Test length is an important factor to a PAT; it affects the accuracy of the final trait estimation. Therefore, each delivered adaptive test consisted of nine items.

Item selection procedures. We compared the proposed item selection methods to the existing methods. The first method was non-information based that selects the next item by matching the current trait estimate to an item index as described before (Matching-LI); hence, it has two versions based on these different indices. The first version uses the average item category location index that considers all ICCCs into account, ($\mathrm{LI}_{\mathrm{Mean}}$). The second version considered the polytomous IRF as an item index ($\mathrm{LI}_{\mathrm{IRF}}$).

The hybrid method combined maximizing Fisher information and matching $\mathrm{LI}_{\mathrm{IRF}}$. The alternating information index (Stage) Method was applied where MI was used toward the end of test. Therefore, first at the selection was based on KL information criterion until the administration of 6 items then we switched to MI.

To apply the PASS, the item bank was divided into a number of strata. Based on the polytomous item' a_i parameter, where $i = 1, 2, \ldots, 93$, and their relationship between items' LIs, the stratification was conducted. At the current, we used only the discrimination parameter to sort the items in an ascending order. For this case, there were three strata, each stratum equally consisted of 31 items and 3 items were selected from each stratum. The items from 1-31, 32-62, and 63-93 formed the three strata from low to high discrimination. For item selection within each stratum, two method were of interest. Matching-LI method was applied and the $\mathrm{LI}_{\mathrm{IRF}}$ was the chosen index used to represent the item's location parameter, (PASS with Matching-$\mathrm{LI}_{\mathrm{IRF}}$). Also, we applied the PASS with MI where MI was the criterion for selecting items at each test phase.

Standard methods were used as MI and KL. With regard to the KL method, as we do not know the true trait value, an interval is used. The trait confidence interval of the nth step in the implementation of PATs, is defined as $[\hat{\theta}_{n-1} - \delta_{n-1}, \hat{\theta}_{n-1} + \delta_{n-1}]$ where $\delta_{n-1} = \frac{c}{\sqrt{n}}$, where c is a constant selected according to a coverage probability (Chang & Ying, 1996), herein $c = 3$.

Scoring method. The expected a posteriori (EAP) with a prior distribution $N(0,\ 1)$ is considered for scoring simulees.

Evaluation measures. The study used two main evaluation criteria for comparison among the different item selection methods. Measurement precision was the first criterion for comparison. The other one was pool utilization as expressed by item exposure rates, where they showed how effectively the item pool was used.

Accuracy of measurement. The ability parameters recovery was used to evaluate the proposed item selection methods. Two statistics are used to determine how close the estimated ability estimates were to the original ability estimates. Evaluation criteria fit here were: (a) Bias, (b) Root Mean Square Error (RMSE), and (b) Relative Efficiency. These indices were to capture the measurement precision.

Average bias (*bias*) was estimated. Let θ_j , $j = 1, , N$ be the original trait value of N examinees and $\hat{\theta}_j$ be the respective estimator from the PAT using different item selection methods.

The RMSE was calculated using Equation 50 as follows

$$RMSE = \sqrt{\frac{1}{N} \sum_{j=1}^{N} \left(\hat{\theta}_j - \theta_j\right)^2}. \tag{50}$$

The smaller bias and RMSE the better item selection method would be. Relative efficiency also is a good indicator for the degree of procedure enhancement relative to a reference procedure. As a standard we used the MI as reference for our the relative efficiency. The close the values to one, the more efficient the procedure is.

Item utilization. To determine the item exposure of each item selection method, the probability of administering each item was computed by dividing the number of times the item was administered by the total number of simulees. The minimum and

Table 8

Descriptive Statistics of Real Item Pool (M=93)

Statistic	a	b_1	b_2	b_3	Distance
Mean	0.59	-0.72	0.47	1.88	2.72
SD	0.17	1.17	0.93	1.18	1.22
Minimum	0.28	-4.01	-1.82	-1.36	0.26
Maximum	1.19	1.56	2.27	4.47	8.15

maximum of item exposure rates in addition to the three quartiles were calculated as well. The closer these values are, the more uniformly the distribution of item exposure rates will be. Also, the percentage of pool of over- and under-exposed items provided useful information. The χ^2 statistic, a descriptive statistic that indicates the skewness of item exposure rate distribution (Chang & Ying, 1999). It quantifies the discrepancy between the observed and the ideal, uniform distribution and is considered a good indicator of the efficiency of pool utilization.

For the current study, the adaptive tests was considered as medium-stakes test. Therefore, we tolerate the target exposure rate to be .30. Also, three quartiles in addition to maximum and minimum exposure rates were considered. Also, the type of items heavily selected by each method was studied.

Results

The current section presents a descriptive analysis to the item pool, followed by the comparison of the results for both measurement precision and item pool usage indices.

Description of the NAEP item pool. Descriptive statistics (i.e., mean, standard deviation, minimum, and maximum) of the NAEP reading item parameters are provided in Table 8. In addition, the distances between the largest and smallest threshold parameters of items are provided. It ranges from almost very narrow width, 0.26, to a larger width of overage, 8.15, with mean =2.72 and SD = 1.22.

Also, the distribution of a-parameter and LIs for all items in the item pool is depicted in Figure 13. Note that the discrimination parameters were limited in range,

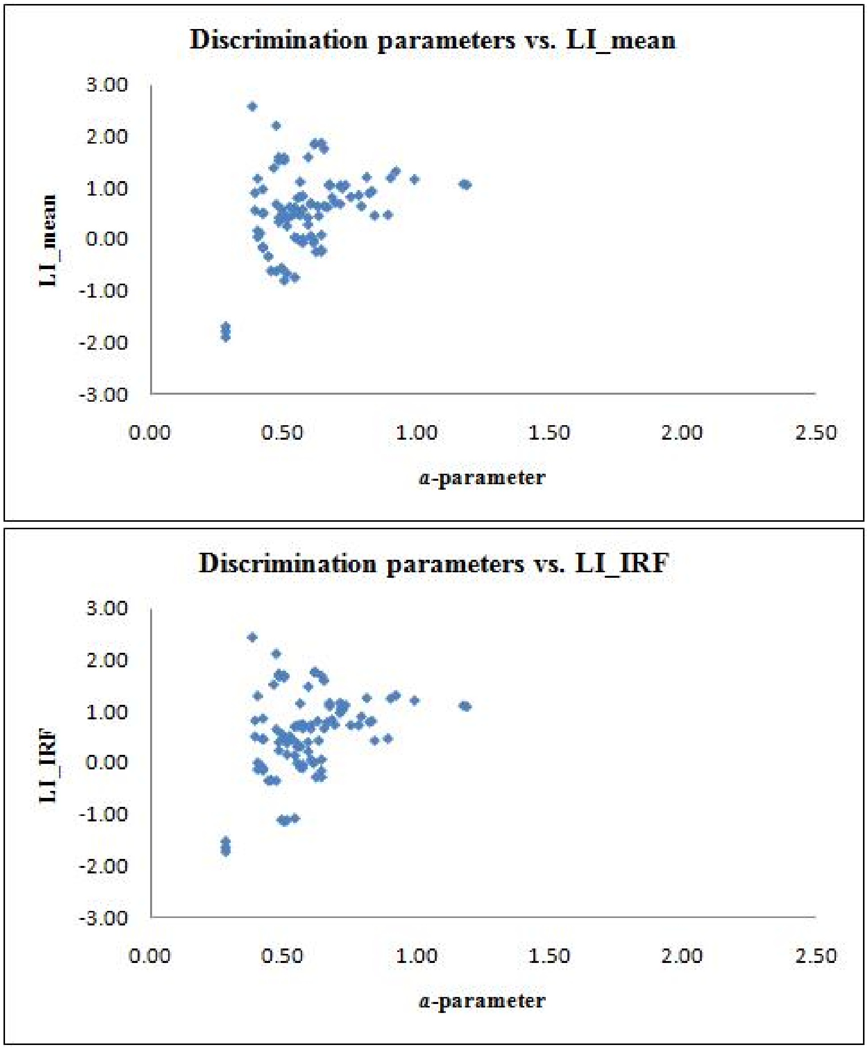


Figure 13: Relationship of LI_{mean} and LI_{IRF} with discrimination parameters.

[0.28, 1.19]; therefore, the item bank is not rich enough especially for high-ability students. Idea: we can add items for the item pool to fill the gap found in the items available. For example, high-discriminating items need to be added to the item bank. With regard to the item category thresholds, the item bank has items that threshold mean covers the range of [-1.91, 2.58]. Also, there is a lack of items that cover the right area (at the right of the vertical line of a=1.20).

Total item information and the standard error of measurement are displayed in Figure 14. Commenting of the match between the distribution of item bank information and the trait distribution in the target population, it is obvious that the total information given by the item bank is not enough to secure accurate trait estimates along the scale

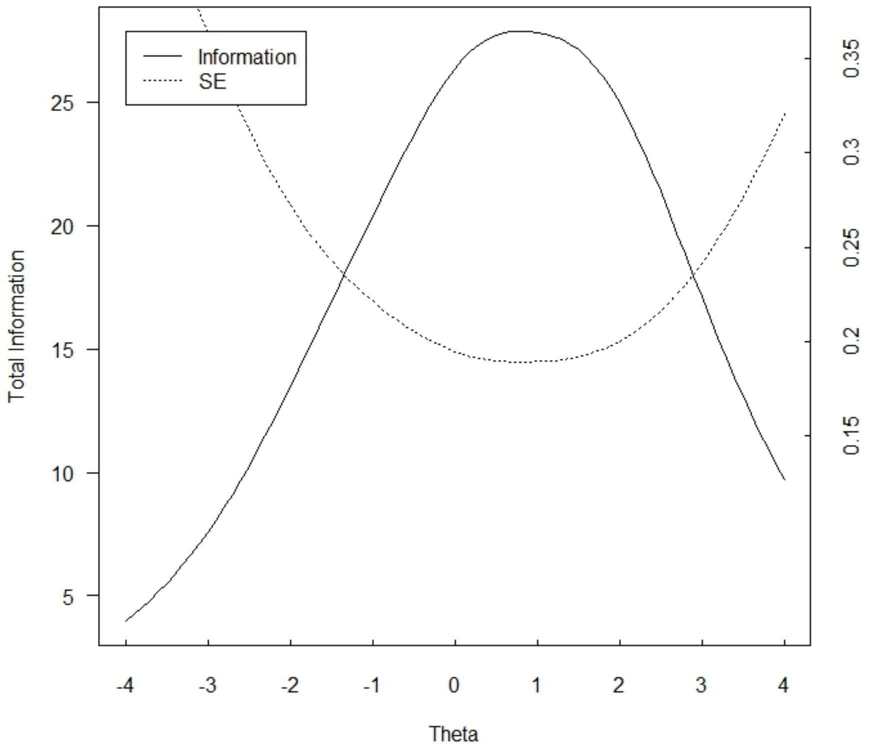


Figure 14: Total item bank information.

range especially at the lower and higher ends. The curve of total information is peaked at the interval of (0.5, 1) up to 30 logits. Also, the minimum value possible for standard error of an estimate is about 0.20 and getting much larger at the two tails. It may be due to the way of collecting these items; they are not designed to be grouped in one pool.

Accuracy of measurement. Table 9 presents the bias, RMSE, and relative efficiency for different item selection methods. As expected, the MI is the most efficient in trait estimation, the PASS is the least, and all the remaining methods are intermediate. Measurement using the Stage-based information method is very precise and very similar to the precision of MI. Relative to the MI, the Hybrid and KL methods are very efficient and are superior to the two methods of non-information approach, Matching-LI$_{\text{mean}}$ and Matching-LI$_{\text{IRF}}$. The PASS methods had a largely biased estimates and thus larger RMSEs. All the remaining methods had very small bias and RMSE.

Pool utilization. Table 10 summarizes the distribution of item exposure rate (r)

Table 9

Measurement Accuracy Indices for Different Item Selection Methods (N=1000, M= 93, L=9)

Methods	Bias	RMSE	Relative Efficiency
MI	0.017	0.462	1.00
KL	0.008	0.515	0.90
Matching-LI$_{\text{mean}}$	-0.003	0.565	0.82
Matching-LI$_{\text{IRF}}$	0.025	0.604	0.76
Hybrid	-0.004	0.531	0.87
Stage	0.022	0.480	0.96
PASS with MI	1.663	1.877	1.00*
PASS with Matching-LI$_{\text{IRF}}$	1.835	2.037	0.92*

Note. RMSE=root mean squared error; LI=location index; MI=maximum information; PASS=polytomous a-stratification strategy.
* The relative efficiency is with regard to PASS with MI.

across the pool under different item selection rules. As we know that the smaller the χ^2 statistic the better the exposure control. It is clear the Hybrid and Matching-LI methods have the best pool utilization indices, where the fives values of minimum, quartiles, and maximum exposure rate are closer to each other than those of the other methods.

Table 11 shows the statistics of over- and under-exposed items and it also provides the preferred items selected by different item selection criteria. The item that has been mostly selected by all item selection methods is item 62: $a_{62} = 0.62$,
$\mathbf{b}_{62} = (-0.39, -0.59, 0.23)$ with LI$^{62}_{\text{mean}}$= -0.25 and LI$^{62}_{\text{IRF}}$= -0.29.
Then there two other examples of mostly selected items, items 66 and 67. Their properties are as follows,
item 66: $a_{62} = 0.64$, $\mathbf{b}_{62} = (-0.17, -0.20, 0.61)$ with LI$^{66}_{\text{mean}}$= 0.05 and LI$^{66}_{\text{IRF}}$= 0.08.
item 67: $a_{67} = 0.64$, $\mathbf{b}_{67} = (-0.39, -0.60, 0.25)$ with LI$^{67}_{\text{mean}}$= -0.25 and LI$^{67}_{\text{IRF}}$= -0.29.

We can see that items 62 and 67 are very much the same in terms of their characteristics and were their behavior. Note that the items were rank ordered based on

Table 10

Descriptive Statistics of Pool Utilization by Different Item Selection Methods

Methods	Min.	1st Qtile	2nd Qtile	3rd Qtile	Max.	χ^2
MI	.004	.009	.013	.017	.909	43.29
KL	.005	.009	.012	.019	1.000	44.51
Matching-LI$_{\text{mean}}$	.006	.032	.059	.114	.466	11.71
Matching-LI$_{\text{IRF}}$	.007	.029	.059	.122	.539	12.02
Hybrid	.006	.025	.059	.130	.599	13.06
Stage	.003	.009	.012	.021	1.000	44.24
PASS with MI	0	0	0	.032	1.000	63.73
PASS with Matching-LI$_{\text{IRF}}$	0	0	.031	.069	1.000	47.44

Note. LI=location index; MI=maximum information; KL=Kullback-Leibler information; PASS=polytomous *a*-stratification strategy.

the discrimination parameter so as the item number increases, the discrimination power of item is getting larger. Based on that we can see the pattern of selected items attached to the different item selection methods.

On the other hand, in terms of the over- and under-exposed indices, the Hybrid and Matching-LI$_{\text{IRF}}$ methods provided the best performance of utilizing the item pool and the two forms of PASS were the least.

Summary

The current study presented new item selection methods in the context of adaptive testing with polytomous items. They are the Hybrid, Stage, PASS methods. These new methods were compared to two Matching-LI procedures and two information procedures, MI and KL. The notion of combining the information and non-information approaches as expressed in the Hybrid method was excellent. It provided a balanced performance on both measurement precision and item pool usage.

The other example of item selection method is stage-based method that paired between the MI and KL criteria to be used exchangeably in delivering test items throughout the test course. The method is efficient in trait estimation but did not

Table 11

Item Utilization Indices and **Mostly Selected** Items by Different Item Selection Methods

Methods	Over- exposed (%)	Under- exposed(%)	Over-exposed Items
MI	13.98	27.96	**62**, **66**, **67**, 82, 85–93
KL	10.75	27.96	12, 13, 17, 20, 59, **62**, 64–**67**
Matching-LI$_{\text{mean}}$	8.60	1.08	16, 17, 20, 26, 33, 34, 37, **62**
Matching-LI$_{\text{IRF}}$	5.38	3.23	16, 17, 20, 37, **62**
Hybrid	5.38	3.23	20, **62**, 65–**67**
Stage	11.83	29.03	12, 13, 59, **62**, 64–**67**, 88, 92, 93
PASS with MI	9.68	55.91	22, 31, 53, 58, 59, **62**, **67**, 87, 88
PASS with Matching-LI$_{\text{IRF}}$	9.68	38.71	9, 16, 23, 33, 37,**62**, 65–**67**

Note. LI=location index; MI=maximum information; KL=Kullback-Leibler information; PASS=polytomous a-stratification strategy.

enhance the usage of items available to be selected from.

Stratification of the item pool and selecting item within each stratum by any of the MI or matching the LI$_{\text{IRF}}$ do not help enhance the estimation accuracy and the estimates were biased. The real item pool used in the study was not rich enough for better stratification. This supports the influence of item pool quality and also the item pool information/size on the performance of item selection methods in adaptive testing (Lima Passos, et al., 2007; Pastor et al., 2002). The study of Lima Passos et al. used two item pools with 300 and 600 items. We found that Pastor et al. manipulated the item pool by modifying their item parameters.

In terms of the items that have been labeled as over-exposed items, the item selection methods that apply the information measures as criteria (e.g., MI and Stage procedures) tend to be attached to the items of relatively high discrimination. (Consider that the largest a-parameter is 1.19). On the contrary, the item selection methods of non-information approach (e.g. Matching-LI) tend to select more the item of high-low to relatively-medium discrimination. The highest a-parameter for an item from this group is 0.64 and the lowest was 0.47. The Hybrid method tends to search for items similar to the Matching-LI methods. Generally, most of the reported items are items with reversal. This is related to Dodd and Koch's (1987) claim that the items with the same set of thresholds yielded the same total amount of information across the entire ability continuum, but different ordering of the thresholds affected the peakedness of the item information curve. The peakedness of the curve increased as the degree of deviation from the sequential order of the thresholds increased.

Chapter 5

Discussion and Conclusion

Discussion

The technology of adaptive testing with polytomous items, especially item selection procedures, was the main focus of this dissertation. The polytomous items were introduced starting with their properties using an IRT framework. Different model parameterizations were introduced to show its suitability to different types of polytomous items. The most commonly used polytomous models were critically analyzed in terms of the parameters that describe an item that has $m + 1$ ordered response categories scored from 0 to m. Consequently, the analysis, in addition the literature review of PAT, concluded that there was a need to have an overall location index or parameter for polytomous items.

The importance of such an index is critical for a PAT environment. The need to have such an index that allow the researcher to have more and more item selection criteria other than information-based criteria is urgent . The unavailability of an index to represent the difficulty or location of a polytomous item prevents the application of PAT non-information item selection procedures. Therefore, the primary focus here was to develop a novel polytomous item location index.

The mathematical derivation allows for four possibility variants of the index. Three indices, LI_{mean}, $\text{LI}_{\text{trimmed mean}}$, and $\text{LI}_{\text{median}}$, were developed based on studying the interrelations among the item category response functions (ICRFs). One more index was developed on the basis of polytomous item response function (IRF), LI_{IRF}. Studying the polytomous IRF has the property that it can be applied to the most commonly used item response models, GRM, PCM, and GPCM. The proposed polytomous LIs are related to item difficulty parameters and have as special cases those used for dichotomous responses.

The success of developing a single location parameter for a polytomous item opens the door for more and more innovative procedures, especially the presence of non-information item selection approach opened the door for the hybrid approach in item selection that combined both information and non-information criteria. The results of the simulation study provided that the Matching-LI methods balanced item pool utilization without wasting any items. This is compared to MI that wasted more than 50% of the available items, which is undesirable property considering that item writing and review process are tedious, time-consuming, and expensive. On the other hand, the non-information approach does not lose much in measurement precision. Another added benefits of the new method is that it is easy and fast.

All the Matching-LI methods perform as if each method has its exposure control mechanism inherited that effectively manages the item pool (i.e., prevent items from being overexposed or never being exposed, and keeps most of the items away from being underexposed). The simulation study findings suggested the modification of these non-information methods and this was the motivation for the second study.

Other item selection procedures were proposed. These methods are hybrid (information/non-information), stage-based information, and polytomous a-stratification strategy. A second study was conducted to evaluate their behavior. The second study was conducted based on NAEP data. The hybrid method was proposed as a response to the first study; it helps enhance the estimation accuracy by adding Fisher information to the selection criterion. The study findings successfully supports the logic of building the hybrid method on both information and non-information criteria by merging them in one single criterion.

With regards to the stage-based information method, the two criteria MI and KL, were use sequentially to deliver an adaptive test. This method was efficient in trait estimation but did not improve the usage of item pool. The third method that applied item pool stratification was the worse in teems of both estimation accuracy and item utility.

The type of items preferred by different item selection methods were investigated. The item selection methods, such as MI and Stage-based, that apply information indices have a higher tendency used the relatively high-discriminating items in the operational item pool. The methods of non-information approach (e.g., Matching-LI) tended to use items with medium discrimination and the hybrid method has the same tendency. Generally, most of the reported items are items with reversal. Dodd and Koch (1987) stated that the amount of information available by each item depends on several factors such as ordering of the thresholds. The ordering of the thresholds affects the peakedness of the item information curve and is consequently related to the deviation from the sequential order of the thresholds.

Finally, from the results obtained here, the stratification process of polytomous item pool is not as straightforward as that of the dichotomous case. This matches the results of Pastor et al. (2002) in their conclusion to deeply think more about a polytomous stratification design (PSD). Such PSD can use an index that considered not only the discrimination parameter in stratifying the item bank but also considered the single location parameter of a polytomous item and other item properties for a better PSD such as width of information coverage and sequentiality of step parameters.

Conclusion

Given the the rapid advancement of computer technology, the importance of administering adaptive tests with polytomous items is in great need. This need promotes the research in polytomous adaptive testing. From the current research that was conducted here, it is obvious that the Matching-LI methods are very promising item selection procedures in polytomous adaptive testing environment. These methods are considered under a new category of polytomous item selection methods called non-information approach. It was very convenient to build such methods based on the derived LIs. Also, the hybrid method is an added-value procedure to the literature in such field.

These conclusions can generalized to fixed-length adaptive tests fitted by the GPCM, selected form an item bank, and scored by the EAP. The research designs were as close to the real application as possible. Two test lengths were considered 9 and 15 items which are considered short to medium length. The size of item pool ranges from 93 to 300 items.

Limitations and Future Research

We will address the limitations and our recommendations for future research in terms of the different aspects of adaptive testing.

Item selection methods. The research in testing is very active. As we are interesting in developing procedures for selecting the best items in polytomous adaptive testing, it is also important to match practitioner's needs regarding the test structure. The structure of a test is predesigned by content experts that provides the user with adequate representation of content areas. According to Boyd, Dodd and Choi (2010) there is no research conducted in severely constrained PATs; that is, tests that incorporate many non-statistical constraints. Therefore, a natural extension of the current research is to add a mechanism to the item selection procedures to satisfy these constraints without affecting the precision of measurement and maintaining optimal item pool usage.

The severely constrained adaptive tests need specific selection methods for polytomous items. Cheng and Chang (2009) introduced a maximum priority index as a mechanism used for delivering of severely constrained adaptive tests with dichotomous items. This index performs very well in that context compared to existing methods such as the Swanson and Stocking (1993) method. This priority index can be incorporated with the Matching-LI methods.

In terms of information measures used, Fisher and KL information indices are two information measures considered the building blocks of the selection criteria applied in adaptive testing with both dichotomous and polytomous items. Other information measures can be useful to be merged to enhance the efficiency of CAT system. In the case of stage-based method presented here, these two measures have been alternated based on their superiority attached to each stage. Other available measures can be used such as mutual information (Cover & Thomas, 1991).

The item selection methods developed and studied here can be refined and further investigated. One possibility is the continuous stage-based information method where a linear combination of global and local information measures is introduced. The linear combination assigns specific weights to each information measure that are monotonically increasing or decreasing as the number of items administered increases.

Item banks. In terms of structure of the item bank: peaked versus flat. Both peaked and flat item banks were used in the polytomous adaptive system. The peaked has more information at small range; whereas, the flat one covers a wider range. An item bank that has uniform information coverage along the trait range offers the best grounds for a PAT setting, irrespective of the criteria. In practice, however, item banks tend to have a bell-shaped information curves, with relatively poor information at extreme values of the traits. In this case, the issue of choosing an item selection criterion becomes more of clear decisive relevance (Lima Passo et al., 2007). Also, the effect that an item bank information/size has on quality of the trait estimation has been detected and reported in previous studies on dichotomous and polytomous CAT (Dodd et al., 1995; Roberts, Lin, & Laughlin, 2001). Therefore, the item bank, in relation to the item characteristics and newly developed LIs, needs to be investigated more.

Polytomous IRT models. The item analysis using simulated- and real-data were performed using traditional polytomous IRT models; that is, dominance models. The most commonly used dominance models are the GRM, NRM, PCM, and GPCM. Other models can fit specific polytomous item data (Tay, Ali, Drasgow, & Williams, in press) such as the ideal point models including the generalized graded unfolding models (GGUM; Roberts, Lin, & Laughlin, 2001).

Trait versus attribute estimation. In terms of overall trait estimation versus student's profile, the field of applying polytomous cognitive diagnostic models is in its infancy and there is no research that provides adaptive tests and consider the cognitive diagnosis. Ali, Shuliang and Chen (in preparation) tries to use the partial-credit DINA model (de la Torre, 2010) in polytomous adaptive testing with cognitive diagnosis (CD-PAT).

References

Ali, U. S. (2010, April). *New location indices for polytomous items.* Paper presented at the 1st Annual Graduate Student Conference, Champaign, IL.

Ali, U. S., & Chang, H.-H. (2010, July). *Developing item selection methods for polytomous CAT (PCAT).* Paper presented at the 75th annual meeting of Psychometric Society, Athens, GA.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Akkermans, W., & Muraki, E. (1997). Item information and discrimination functions for trinary PCM items. *Psychometrika, 62,* 569–578.

Bennet, R. E., Steffen, M., Singley, M. K., Morely, M., & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computer-adaptive tests. *Journal of Educational Measurement, 34,* 162–176.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (1968). *Statistical theories of mental test scores* (Chaps. 17 to 18). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 27,* 29–51.

Boyd, A., Dodd, B., & Choi, S. (2010). Polytomous models in computerized adaptive testing. In M. L. Nering, & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 229–255). New York, NY: Routledge.

Chang, H.-H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59,* 391–404.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20,* 213–229.

Chen, S.-Y., Ankenmann, R. D., & Chang, H.-H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24,* 241–255.

Cheng, Y., Chang, H.-H., Douglas, J., & Guo, F. (2009). Constraint-weighted *a*-stratification for computerized adaptive testing with nonstatistical constraints: Balancing measurement efficiency and exposure control. *Educational and Psychological Measurement, 69,* 35–49.

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria of polytomous items. *Applied Psychological Measurement, 33,* 419–440.

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement, 16,* 327–243.

De Ayala, R. J. (1993). An introduction to polytomous item response theory models. *Measurement and Evaluation in Counseling and Development, 25,* 172–189.

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

de la Torre, J. (2010, July). The partial-credit DINA model. Paper presented at the 75th Meeting of Psychometric Society, July 7–9, Athens: GA.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19,* 5–22.

Deng, H., Ansley, T., & Chang, H.-H. (2010). Stratified and maximum item selection procedures in computerized adaptive testing. *Journal of Educational Measurement, 47,* 202–226.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement, 41,* 295–311.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kingsbury, G.G., & Zara, A.R.(1989). Procedures for selecting items for computerized adaptive tests . *Applied Measurement in Education, 2,* 359–375.

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

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

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

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19,* 91–100.

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

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17,* 351–363

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19,* 73–90.

Nering, M. L., & Ostini, R. (2006). *Polytomous item response theory models.* Thousands Oaks, CA: SAGE.

Nering, M. L., & Ostini, R. (2010). New perspectives and applications. In M. L. Nering, & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 3–20). New York, NY: Routledge.

Lima Passos, V., Berger, M. P. F., & Tan, F. E. (2007). Test design optimization in CAT early stage with the nominal response model. *Applied Psychological Measurement, 31,* 213–232.

Lima Passos, V., Berger, M. P. F., & Tan, F. E. (2008). The D-optimality item selection criterion in the early stage of CAT: A study with the graded response model. *Journal of Educational and Behavioral Statistics, 33,* 88–110.

Pastor, D. A., Dodd, B. G., & Chang, H.-H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement, 26,* 147–163.

Penfield, R. D. (2006). Applying bayesian item selection approaches to adaptive tests with polytomous items. *Applied Measurement in Eduction, 19,* 1–20.

Roberts, J. S., Lin, Y., & Laughlin, J. E. (2001). Computerized adaptive testing with the generalized graded unfolding Model. *Applied Psychological Measurement, 25,* 177–196.

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

Samejima, F. (2010). The general graded response model. In M. L. Nering, & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 77–108). New York, NY: Routledge.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problem. *Applied Psychological Measurement, 17,* 151–166.

Tao, J., Shi, N.-Z., & Chang, H.-H. (2009). Item-weighted WLE methods for ability estimation in mixed tests composed of both dichotomously and polytomously scored items. *Applied Psychological Measurement.*

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. (in press). Fitting IRT models to dichotomous and polytomous data: Assessing the relative model-data fit of ideal point and dominance models. *Applied Psychological Measurement.*

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567–577.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43,* 39–55.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2010). *National Assessment of Educational Progress (NAEP), 2000-2007 Reading Assessments.* Retrieved from http://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_irt_read.asp

van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25,* 273–282.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63,* 201–216.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1–25). Boston, MA: Kluwer Academic Publishers.

van Rijn, P.W., Eggen, T. J. H. M., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26,* 393–411.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22,* 203–226.

Veldkamp, B. P. (2003). Item selection on polytomous CAT. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), New developments in psychometrics (pp. 207–214). Tokyo: Springer-Verlag.

Veldkamp, B. P., & van der Linden, W. J. (in press). Multidimensional adaptive testing with constraints on test content. *Psychometrika.*

Wang, S., & Wang, T. (2001). Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25,* 317–331.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427–450.

Appendix A

Approximation of $\mathrm{LI}_{\mathrm{IRF}}$ Using Newton-Raphson Method

The Newton-Raphson method is a numerical method to solve nonlinear equations of the form of $f(x) = 0$. The approximate solution to the equation is

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}, \tag{A-1}$$

where x_{t+1} is the updated approximation based on the previous estimate, x_t, and $f'(x_t)$ is the fist derivative of $f(x_t)$ with respect to x.

In the following section we introduced the approximation of $\mathrm{LI}_{\mathrm{IRF}}$ using the three different polytomous IRT models: (a) the partial credit models (Masters, 1982; Muraki, 1992) and (b) the graded response model (Samejima, 1969).

$\mathrm{LI}_{\mathrm{IRF}}$ of the Partial Credit Models' (PCM and GPCM) Items

In the current case, consider a polytomous item with $m + 1$ response categories ranging from 0 to m. The formula of getting a category x on item i using a general form to the partial credit models is given by

$$P_{ix}(\theta) = \frac{\exp \sum_{v=1}^{x} a_i(\theta - b_{iv})}{1 + \sum_{c=1}^{m} \exp \sum_{v=1}^{c} a_i(\theta - b_{iv})}, \tag{A-2}$$

where the discrimination parameters a_i for all items are equal for PCM and different for GPCM. The expected score given a specific ability value θ is given by

$$E(X) = \sum_{x=1}^{m} x\, P_{ix}(\theta), \tag{A-3}$$

Assuming that $\frac{m}{2}$ is our critical point to get the corresponding θ value that satisfies such criterion, so the following is to satisfy

$$\sum_{x=1}^{m} x P_{ix}(\theta) = \frac{m}{2}, \tag{A-4}$$

therefore, the function that needs to be solved is

$$f(\theta) = \sum_{x=1}^{m} x\, P_{ix}(\theta) - \frac{m}{2} = 0. \tag{A-5}$$

Given that

$$\frac{\partial P_{ix}(\theta)}{\partial \theta} = a_i P_{ix}(\theta) \left[x - \sum_{c=1}^{m} c P_{ic}(\theta) \right], \tag{A-6}$$

so the the first derivative of $f(\theta)$ with respect to θ, $f'(\theta)$, is given by

$$f'(\theta) = \sum_{x=1}^{m} x\, a_i P_{ix}(\theta) \left[x - \sum_{c=1}^{m} c P_{ic}(\theta) \right], \tag{A-7}$$

The approximate value of $\mathrm{LI}_{\mathrm{IRF}}$ using the partial credit models is

$$\begin{aligned} \theta_{t+1} &= \theta_t - \frac{f(\theta_t)}{f'(\theta_t)} && \text{(A-8)} \\ &= \theta_t - \frac{\left[\sum_{x=1}^{m} x\, P_{ix}(\theta_t)\right] - \frac{m}{2}}{\sum_{x=1}^{m} x\, a_i\, P_{ix}(\theta_t)\left[x - \sum_{c=1}^{m} c P_{ic}(\theta_t)\right]}. && \text{(A-9)} \end{aligned}$$

$\mathbf{LI_{IRF}}$ of the Graded Response Model's (GRM) Items

The formula of getting a category x on item i using a general form to the graded response model is given by

$$P_{ix}(\theta) = P^*_{ix}(\theta) - P^*_{i,x+1}(\theta), \tag{A-10}$$

where $P^*_{ix}(\theta)$ is given by the formula of two-parameter logistic model as follows

$$P^*_{ix}(\theta) = \frac{1}{1 + exp[-a_i(\theta - b_{ix})]}. \tag{A-11}$$

therefore, the function that needs to be solved is

$$f(\theta) = \sum_{x=1}^{m} x\, (P^*_{ix}(\theta) - P^*_{i,x+1}(\theta)) - \frac{m}{2} = 0. \tag{A-12}$$

Given that

$$\frac{\partial P^*_{ix}(\theta)}{\partial \theta} = a_i P^*_{ix}(\theta) \left[1 - P^*_{ix}(\theta)\right], \tag{A-13}$$

so the the first derivative of $f(\theta)$ with respect to θ, $f'(\theta)$, is given by

$$f'(\theta) = \sum_{x=1}^{m} x\, a_i \left[P_{ix}^{*}(\theta)\left(1 - P_{ix}^{*}(\theta)\right) - P_{i,x+1}^{*}(\theta)\left(1 - P_{i,x+1}^{*}(\theta)\right)\right], \tag{A-14}$$

The approximate value of $\mathrm{LI}_{\mathrm{IRF}}$ using the GRM is

$$\begin{aligned} \theta_{t+1} &= \theta_t - \frac{f(\theta_t)}{f'(\theta_t)} & \text{(A-15)} \\ &= \theta_t - \frac{\left[\sum_{x=1}^{m}\left(x\left(P_{ix}^{*}(\theta_t) - P_{i,x+1}^{*}(\theta_t)\right)\right)\right] - \frac{m}{2}}{\sum_{x=1}^{m} x\, a_i \left[P_{ix}^{*}(\theta_t)\left(1 - P_{ix}^{*}(\theta_t)\right) - P_{i,x+1}^{*}(\theta_t)\left(1 - P_{i,x+1}^{*}(\theta_t)\right)\right]}. & \text{(A-16)} \end{aligned}$$

Appendix B

IRT Parameters and Information Curves of 93 NAEP Reading Assessment Items

Table B1

Item Parameters of 93 NAEP Reading Assessment Items

1st Stratum					2nd Stratum					3rd Stratum				
i	a_i	b_{i1}	b_{i2}	b_{i3}	i	a_i	b_{i1}	b_{i2}	b_{i3}	i	a_i	b_{i1}	b_{i2}	b_{i3}
1	0.28	-3.33	-0.71	-1.36	32	0.51	0.30	0.29	0.55	63	0.63	-1.59	1.25	2.24
2	0.28	-3.33	-0.46	-1.32	33	0.51	-3.90	-1.74	3.63	64	0.63	0.17	0.25	0.92
3	0.28	-3.78	-0.69	-1.26	34	0.51	-1.36	-0.22	2.33	65	0.64	-1.57	-0.02	0.96
4	0.38	1.56	1.70	4.47	35	0.52	-0.97	0.11	2.72	66	0.64	-0.17	-0.20	0.61
5	0.39	0.59	0.20	1.90	36	0.53	0.41	0.33	0.64	67	0.64	-0.39	-0.60	0.25
6	0.39	-0.31	0.18	1.80	37	0.54	-3.54	-1.63	2.92	68	0.64	0.75	1.05	3.79
7	0.40	-1.02	1.79	2.75	38	0.54	-1.21	0.64	0.66	69	0.65	0.40	1.08	3.78
8	0.40	0.08	-1.39	1.43	39	0.54	-1.49	0.95	2.37	70	0.65	-0.51	0.75	1.68
9	0.40	0.12	-1.17	1.53	40	0.55	-1.40	-0.10	3.12	71	0.66	-1.51	1.21	2.15
10	0.41	0.11	-1.25	1.48	41	0.55	-1.27	0.47	3.20	72	0.67	-0.77	1.51	2.35
11	0.42	-0.36	0.15	1.72	42	0.55	-0.93	-0.33	2.73	73	0.67	-0.08	1.21	2.06
12	0.42	-0.23	0.03	-0.33	43	0.55	-1.08	-0.14	1.24	74	0.68	-0.30	0.90	1.83
13	0.42	-0.20	0.12	-0.37	44	0.56	-0.95	-0.26	2.62	75	0.69	-0.37	0.80	1.70
14	0.42	0.97	-0.04	1.98	45	0.56	-0.12	1.27	2.18	76	0.71	-0.68	1.48	2.32
15	0.42	-0.39	0.17	1.73	46	0.56	-1.24	0.45	3.25	77	0.71	-1.74	1.59	2.19
16	0.44	-1.60	-0.45	1.03	47	0.56	-1.68	-0.30	1.86	78	0.71	-0.04	1.17	1.90
17	0.45	-2.64	1.10	-0.32	48	0.57	-1.18	0.49	3.19	79	0.72	-0.04	1.15	1.88
18	0.46	-0.79	2.03	2.90	49	0.57	-1.86	0.03	1.62	80	0.73	-0.20	1.35	1.99
19	0.47	1.29	1.63	3.69	50	0.57	-2.20	-0.36	2.56	81	0.75	0.00	0.34	2.11
20	0.47	-2.70	0.99	-0.18	51	0.57	-1.62	0.88	2.47	82	0.78	0.17	0.21	2.17
21	0.47	0.54	0.34	1.15	52	0.59	0.11	0.22	0.92	83	0.79	-1.63	1.40	2.14
22	0.48	0.11	0.21	0.94	53	0.59	0.06	-0.25	1.03	84	0.81	0.14	1.49	1.97
23	0.48	-1.18	-0.16	2.33	54	0.59	-0.03	1.11	3.69	85	0.82	0.05	0.36	2.26
24	0.48	-0.35	2.25	2.87	55	0.60	-0.99	0.01	1.16	86	0.83	0.12	0.33	2.34
25	0.48	-0.48	2.27	2.77	56	0.60	-0.20	0.48	1.76	87	0.84	-1.11	0.35	2.12
26	0.49	-4.01	-1.82	4.14	57	0.60	-0.49	0.81	1.77	88	0.89	-1.13	0.43	2.11
27	0.49	-0.72	0.41	2.05	58	0.61	-1.01	-0.04	1.02	89	0.90	0.13	1.43	2.00
28	0.50	-3.64	-1.75	2.96	59	0.61	-1.15	0.38	0.55	90	0.92	1.08	1.18	1.69
29	0.50	-0.35	2.11	2.98	60	0.61	0.35	1.45	3.72	91	0.99	0.16	1.34	1.98
30	0.50	-0.25	2.19	2.64	61	0.62	0.37	1.47	3.73	92	1.17	0.36	1.21	1.64
31	0.50	0.15	0.32	0.97	62	0.62	-0.39	-0.59	0.23	93	1.19	0.37	1.20	1.59

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000-2007 Reading Assessments (National Center for Education Statistics [NCES], 2010).

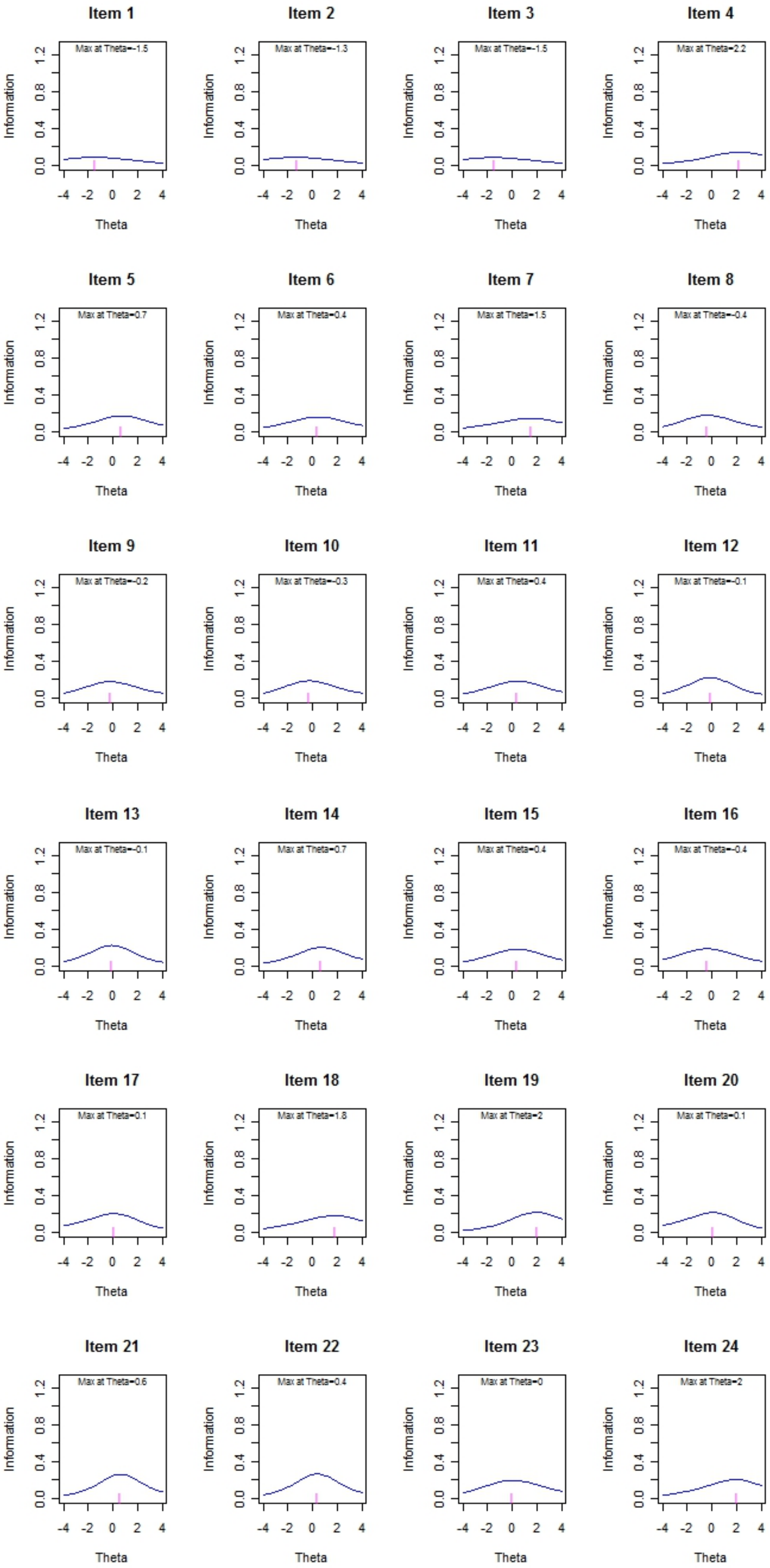


Figure B1: Information functions for items 1-24

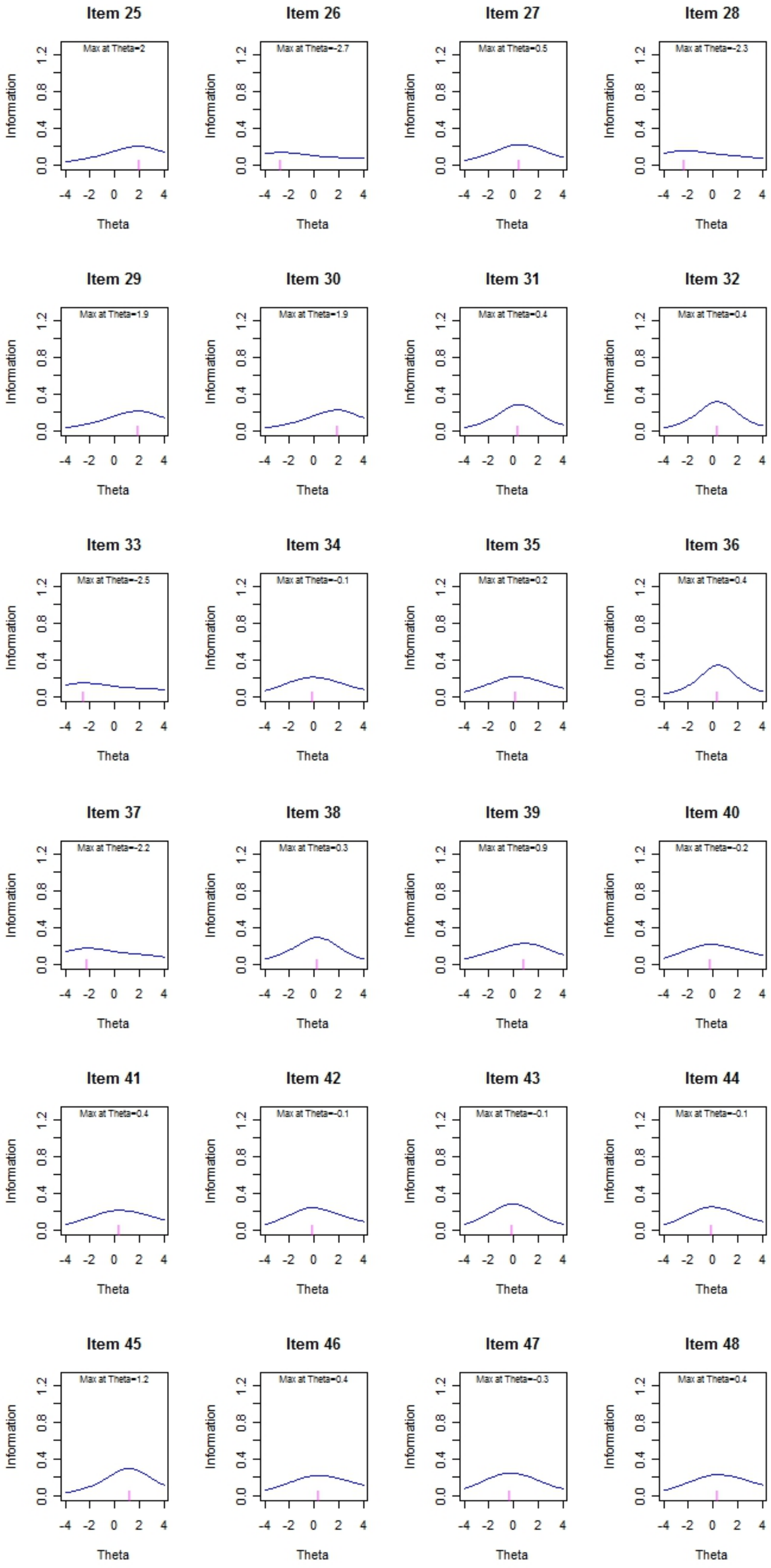


Figure B2: Information functions for items 25-48

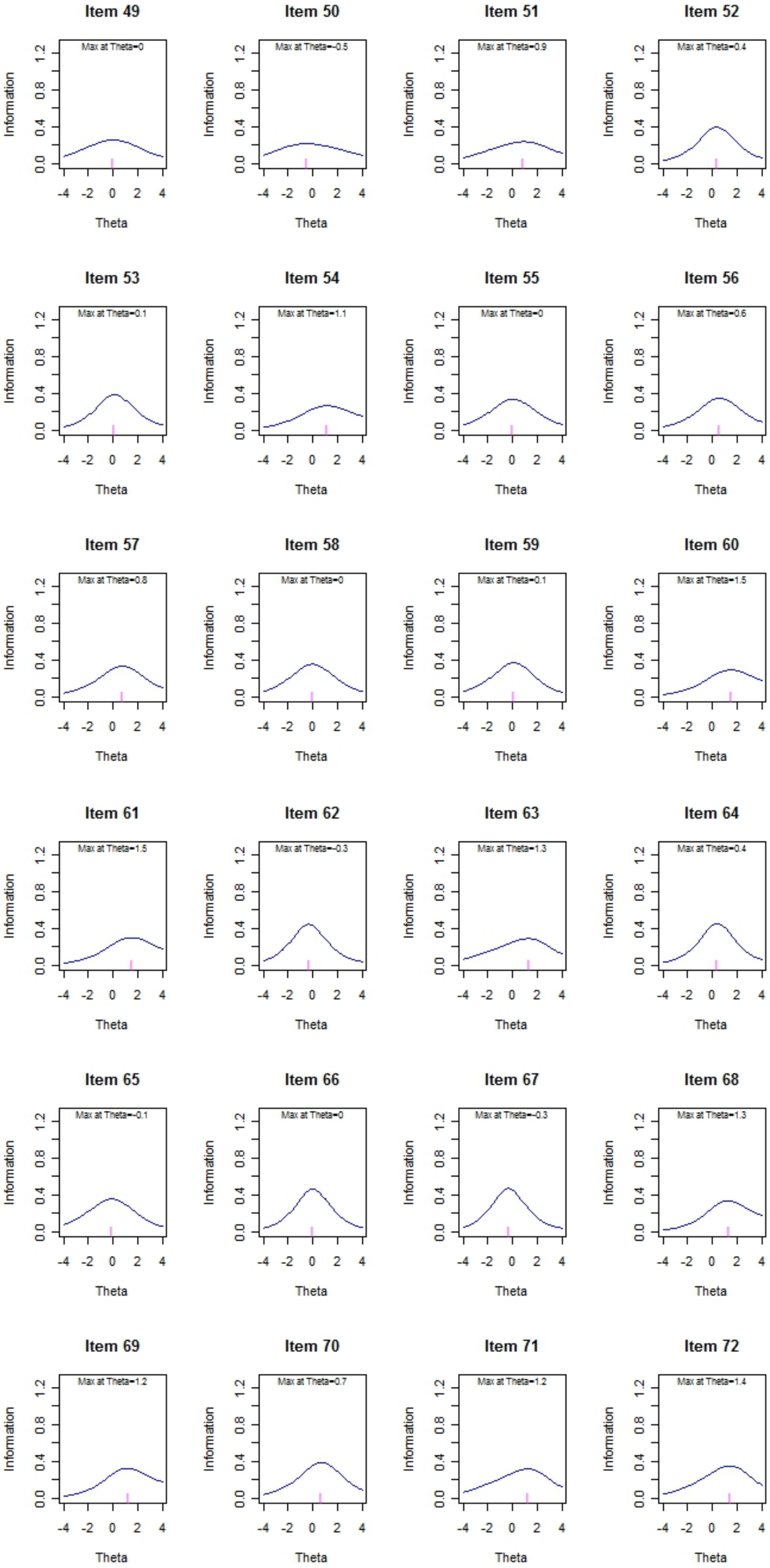


Figure B3: Information functions for items 49-72

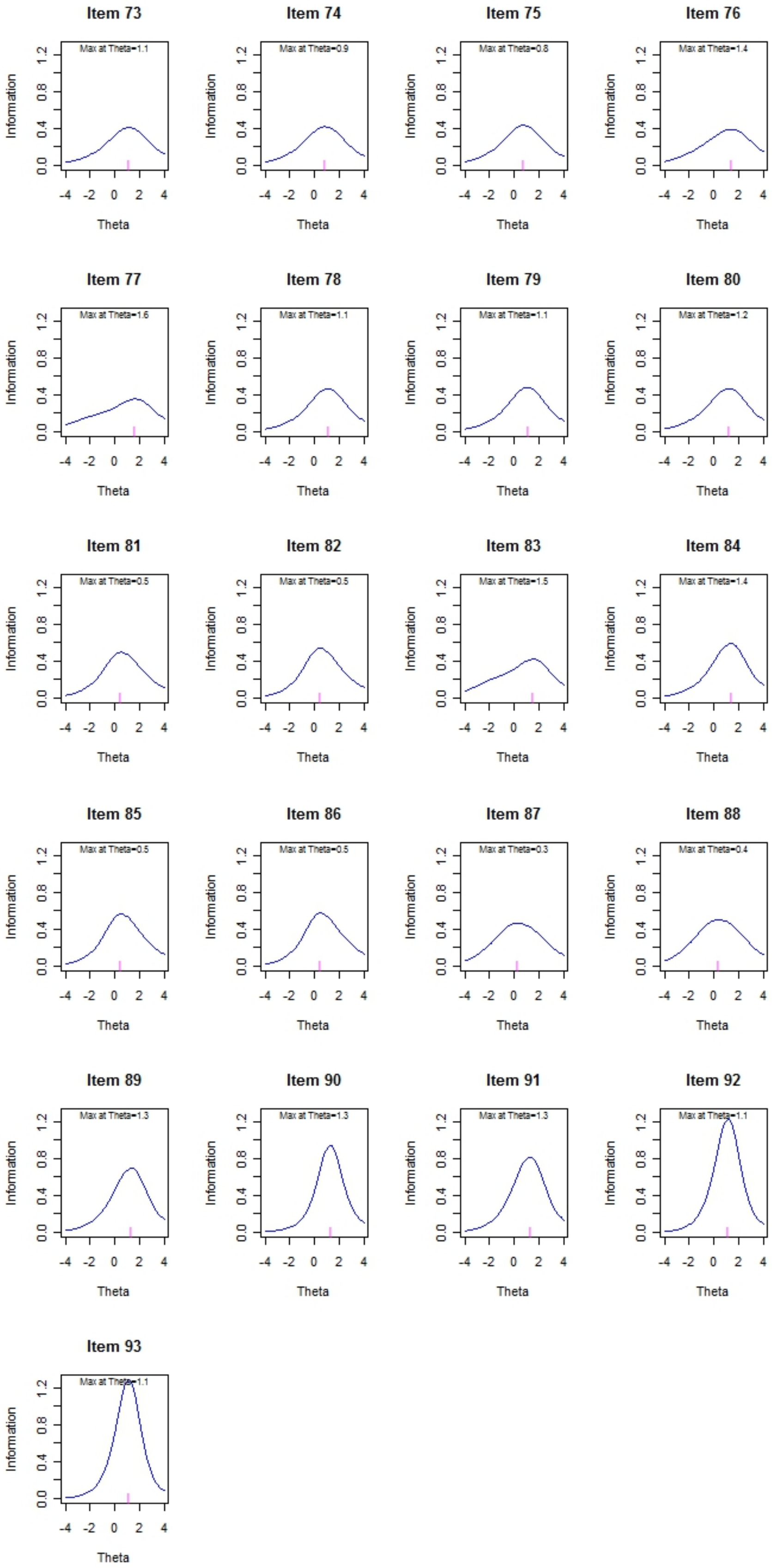


Figure B4: Information functions for items 73-93