KULLBACK-LEIBLER INFORMATION AND ITS APPLICATIONS IN MULTI-DIMENSIONAL ADAPTIVE TESTING

BY

CHUN WANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Master's Committee:

Professor Hua-Hua Chang, Chair
Professor Jeff Douglas

ABSTRACT

This paper first discusses the relationship between Kullback-Leibler information (KL) and Fisher information in the context of multidimensional item response theory and is further interpreted for the two-dimensional case, from a geometric perspective. This explication should allow for a better understanding of the various item selection methods in multi-dimensional adaptive tests (MAT) which are based on these two information measures. The KL information index (KI) method is then discussed and two theorems are derived to quantify the relationship between KI and item parameters. Due to the fact that most of the existing item selection algorithms for MAT bear severe computational complexity, which substantially lowers the applicability of MAT, two versions of simplified KL index (SKI), built from the analytical results, are proposed to mimic the behavior of KI, while reducing the overall computational intensity.

Dedicated to loved ones in my heart, my parents, husband, and professors.

ACKNOWLEDGEMENTS

I am heartily thankful to my adviser, Hua-Hua Chang, whose encouragement, guidance and support from the initial to the final. I would also like to thank Dr. van der Linden and Dr. Richard Patz from CTB/McGraw-Hill, who gave me generous financial support. Lastly, I offer my regards and appreciation to all of those who supported me in any respect during the completion of the project, including Li Wang, Erkao Bao, Jibo He, and Lihua Yao. This dissertation is reproduced with kind permission of Springer Science + Business Media. The original publication is **Wang, C**., Chang, H-H., & Boughton, K. (2011). Kullback-Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika*, 76(1), 13-39.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER1 INTRODUCTION

Multi-dimensional item response theory (Reckase, 1985, 1997) is gaining more attention recently due to the increased interest in testing for diagnosis. Many certification and admission boards are trying to combine regular tests with diagnostic services to allow candidates to obtain more informative diagnostic profiles of their abilities (Mulder & van der Linden, 2009). The diagnostic feature of MIRT is reflected by viewing the underlying latent ability as a multi-dimensional vector, typically denoted as $\theta_i = (\theta_{i1}, ...\theta_{ip})^T$, where p is the number of dimensions or subscales analogous to the number of attributes in cognitive diagnosis. In addition to getting one overall summative score, this approach will provide a finer break down of the domain score for each dimension. Moreover, we can get a continuous estimate of each subscale as an alternative to the dichotomous master/non-master results provided by many cognitive diagnosis models, thereby gaining more information on each subscale for every examinee.

Building adaptive tests based upon MIRT, called multi-dimensional adaptive testing (MAT), offers at least two advantages over unidimensional adaptive testing (UAT): (a) MAT includes more information than UAT since the multiple subscales being measured are often correlated, and (b) MAT can balance content coverage automatically without fully resorting to content balancing techniques (e.g., Segall, 1996). Just like UAT, the most important component in MAT is the item selection algorithm, which selects items during the course of the test. To date, several methods for item selection have been proposed. For example, Bloxom and Vale

(1987) put forward a suggestion to generalize Owen's (1969; 1975) Bayesian procedure from UIRT to MIRT. Segall (1996) proposed an item selection criterion to maximizing the determinant of Fisher information matrix (Mulder & van der Linden, 2009), which is further extended to include prior information. Luecht (1996) subsequently implemented this criterion in the context of licensure testing with various non-statistical constraints. Van der Linden (1999) developed a novel approach for estimating the weighted sum of ability elements via the trace of the asymptotic covariance matrix of the ability estimates (known as *A*-optimality). It is important to note that all these criterions were based on Fisher information (FI). However, Veldkamp and van der Linden (2002) did later introduce a multi-dimensional Kullback-Leibler information (KL) based criterion and according to Chang and Ying (1996), FI is local information and KL is global information. Global information should be used when *n* (i.e., test length) is small, and local information when *n* is large. An important aspect of the index proposed by Veldkamp and van der Linden (2002) is that it combines both local and global information and thus makes appropriate usage of information throughout the entire test.

Although many item selection methods have been proposed for MAT, it remains a matter of debate as to which method is the most appropriate in certain applications. To this end, the distinctive feature of each method needs to be explored. Specifically, the properties of FI and KL established in unidimensional IRT need to be validated in the multi-dimensional context. For example, in UIRT the FI at a specific value of ability θ_0 is the second derivative of the KL with respect to θ

evaluated at θ_0. According to Chang and Ying (1996), this feature indicates that the value of the FI is the curvature of the KL curve at θ_0. Thus, maximizing the area under KL is equivalent to maximizing FI when the test length is long. Therefore, this research will assess whether this relationship can be generalized to the multi-dimensional space. The second goal of this research is to characterize the selective mechanism of the KL index (KI), such as the parameter patterns favored by KI. This investigation will facilitate item pool development and maintenance by diagnosing which items are more likely to be under- or over-exposed.

In this paper the results will be presented in three-dimensional space in order to provide a more intuitive geometric visualization, and thus the latent variable space is two-dimensional. In fact, due to technical complexities in both formulation and computation, most current MAT applications are essentially based on two or three dimensional IRT models (Allen, Ni, & Haley, 2008; Haley, Ni, Ludlow, & Fragala-Pinkham, 2006; Li & Schafer, 2005; Mulder & van der Linden, 2009; van der Linden, 1999; Veldkamp & van der Linden, 2002). However, and more importantly, these theoretical results can be generalized to any number of dimensions. The generalization is also discussed in this paper.

The rest of the paper is arranged by first discussing the key concepts and utilization of information in MIRT, followed by a set of analytical results. The next section then proposes a new item selection index, called the simplified KL information index (SKI), followed by two supporting simulation studies. The final section discusses directions for future work in this area.

CHAPTER2 TYPES OF INFORMATION MEASURES IN MAT ITEM SELECTION

2.1 Multi-dimensional item response model

MIRT models have been developed to capture the complexity of modern assessments (Adams, Wilson, & Wang, 1997), with the multi-dimensional three-parameter model (M3PL) taking the form of (Reckase, 2009)

$$p_i(\boldsymbol{\theta}) \equiv \text{Prob}(u_i = 1 | \boldsymbol{\theta}) = c_i + \frac{1 - c_i}{1 + \exp[-(a_i^T \boldsymbol{\theta} - b_i)]} \quad , \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, ...\theta_p)^T$ is the ability vector for an examinee and p is the number of dimensions or subscales. u_i is a binary random variable containing the response to item i, c_i is the pseudo-guessing parameter, b_i is the intercept term playing the role of item difficulty, and a_i^T is a $1 \times p$ vector of discrimination parameters for item i. The form of item response function in (1) is a direct generalization of the three-parameter logistical model (Birnbaum, 1968) to the multi-dimensional case. If the guessing parameter c is equal to 0, then the model reduces to the multi-dimensional 2-parameter model (M2PL). Note that the multiple subscales follow a compensatory rule in this model and each item has only one difficulty parameter, as separate difficulty parameters for each dimension would render the model unidentifiable (Reckase, 1985). The discrimination parameter vector indicates the relative importance of each ability to answer item i correctly. Notice that due to the rotational indeterminacy of the $\boldsymbol{\theta}$-space, the elements of $\boldsymbol{\theta}$ may not automatically represent the desired abilities. However, while important, this is a scaling issue and

is beyond the current scope of this paper, so we assume the item pool is pre-calibrated with the correct rotation of the ability space determined.

2.2 Fisher information

Item information is typically defined as FI, which is a function of true θ and therefore differs from examinee to examinee in the population. FI measures the amount of information that an observable random variable, for example, the item response *X*, carries about an unknown parameter θ. It can be formularized as

$$I(\theta) = E\left\{[\frac{\partial}{\partial\theta}\ln f(X;\theta)]\middle|\theta\right\}, \tag{2}$$

where $f(X;\theta)$ is the likelihood function computed from item response functions (IRF), which usually takes the form of

$$f(X;\theta) = L(\theta_0; X_1, X_2, ...X_n) = \prod_{i=1}^{n}[P_i^{X_i}(\theta_0)Q_i^{1-X_i}(\theta_0)], \tag{3}$$

and θ is the latent ability (Lord, 1980). IRFs can come from one-, two-, or three-parameter models. Built upon (2), one item selection method, namely, the maximum Fisher information method (MFI; Thissen & Mislevy, 2000) is proposed in unidimensional computer-adaptive testing (CAT). This criterion tries to maximize the FI at the current ability estimate, $\hat{\theta}^{(t)}$, after *t* items have been administered. FI is additive, meaning that for a test consisting of items, *i=1,2,….n*, the test information is simply the sum of the individual item information, expressed as $I^{(n)}(\theta) = \sum_{i=1}^{n} I_i(\theta)$. Test FI is inversely related to the variance of the maximum likelihood estimator (MLE) following an asymptotic theory, which says that $\hat{\theta}^{mle} \sim N(\theta, I^{-1}(\hat{\theta}^{mle}))$, and

this is the foundation of MFI method in CAT. Other things being equal, the larger the FI, the more precise the $\hat{\theta}^{mle}$ will be. By selecting items that maximize FI at the interim ability estimate, the MFI method can force $\hat{\theta}^{mle}$ to converge to the true θ as quickly as possible.

In the multi-dimensional case, the FI extends to a matrix instead of a scalar. For item *i*, the matrix is defined as

$$I_i(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\log f(u_i|\boldsymbol{\theta}))|\boldsymbol{\theta}\right], \tag{4}$$

and if the MIRT model in (1) is taken, the information matrix becomes

$$I_i(\boldsymbol{\theta}) = \frac{Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta})-c_i]^2}{P_i(\boldsymbol{\theta})(1-c_i)}\begin{bmatrix} a_{i1}^2 & a_{i1}a_{i2} & \ldots & a_{i1}a_{ip} \\ a_{i1}a_{i2} & a_{i2}^2 & \ldots & a_{i2}a_{ip} \\ : & & & \\ a_{i1}a_{ip} & a_{i2}a_{ip} & \ldots & a_{ip}^2 \end{bmatrix}, \tag{5}$$

where $Q_i(\boldsymbol{\theta})$=1- $P_i(\boldsymbol{\theta})$. The item information matrix will add up to form the test information matrix, maintaining the additive property. In this case, the asymptotic property of MLE and its relationship to the Fisher information also holds. Specifically, assuming $\boldsymbol{\theta}$ is a *p*-dimensional vector, the MLE of $\hat{\boldsymbol{\theta}}^{mle}$ is distributed asymptotically as $\hat{\boldsymbol{\theta}}^{mle} \sim N(\boldsymbol{\theta}, I_s^{-1}(\hat{\boldsymbol{\theta}}^{mle}))$, where $I_s^{-1}(\hat{\boldsymbol{\theta}}^{mle})$ is the inverse of the information matrix evaluated at $\hat{\boldsymbol{\theta}}^{mle}$, with each element representing either the variance of one ability dimension or the covariance between two ability dimensions. For a more general case, please refer to Lehmann (1999) or Mulder and van der Linden (2009).

Note from (2) and (3) that FI is only a function of a single point, say, $\hat{\theta}^{(t)}$;

indicating that FI represents the item discrimination power only around $\hat{\theta}^{(t)}$ (Hambleton & Swaminathan, 1985). Thus, it is not a good indicator of item discrimination power when $\hat{\theta}^{(t)}$ is far from true θ, which is often the case in the early stage of a CAT. Due to this limiting feature, FI is termed as local information (Chang & Ying, 1996).

2.3 Kullback-Leibler information

Generally, KL measures the divergence (i.e., non-symmetric distance) between two probabilities over the same parameter space (Cover & Thomas, 1991; Lehmann & Casella, 1998), and it is usually defined as

$$KL[g \| f] = E_f \left[\log \frac{f(X)}{g(X)} \right]. \quad (6)$$

Here, $f(X)$ and $g(X)$ are two probability distributions. The expectation here is taken over $f(X)$, which usually represents the "true" distribution of the observed data. $g(X)$ often represents an approximation of $f(X)$. Following Renyi (1970, 1961), KL is sometimes called the *information gain* by *X* if f can be used instead of g. It is also called the *relative entropy* for using g instead of f. $KL[g \| f]$ measures how easy it is to tell apart the two probability distributions (Henson & Douglas, 2005). Statistically, KL is derived from the well-known likelihood ratio test. Assume $f(X)$ is the likelihood function when $\theta = \theta_0$ and $g(X)$ is the likelihood function when $\theta = \theta_1$. Using the Neyman-Pearson theory (Lehmann, 1986), the likelihood ratio test is the best test of $\theta = \theta_0$ versus $\theta = \theta_1$. In this

regard, the expected value of log-likelihood ratio, also the definition of KL (see Equation 6), quantifies how powerful the statistical test is, and therefore measures the discrimination power of an item for distinguishing θ_1 from θ_0. Note that the application of KL in the context of CAT was first introduced by Chang and Ying (1996). For item *i*, KL is expressed explicitly as

$$KL_i(\theta_1 \| \theta_0) = p(\theta_0)\log(\frac{p(\theta_0)}{p(\theta_1)}) + (1-p(\theta_0))\log(\frac{1-p(\theta_0)}{1-p(\theta_1)}). \quad (7)$$

One important feature of KL is that it is a function of two ability levels, θ_0 and θ_1, and it does not require θ_1 to be close to θ_0, which makes it suitable for use in the early stages of item selection in CAT (Chang & Ying, 1996). For an *n*-length test, the test KL is the summation of the item KL. In the multi-dimensional case, the only changes to the KL are that θ becomes a vector instead of a scalar in (7), and the item response function follows the MIRT model in (1).

CHAPTER 3 ANALYTICAL RESULTS OF KL

3.1 Connections between Fisher Information and KL Information

The asymptotic theory introduced in the previous section explicitly relates the performance of MLE $\hat{\theta}^{mle}$ to the item/test FI. The expectation of the likelihood ratio, which is just the KL information, serves as one assumption in the course of the proof (Chang, 1996; Chang & Stout, 1993). In this sense, KL only has an indirect effect on the property of $\hat{\theta}^{mle}$, but understanding the relations between FI an KL will help identify the roles of KL in the adaptive item selection process.

In UIRT, Chang and Ying (1996) showed that FI at θ_0 equals the second derivative of KL evaluated at the same true value θ_0, which is expressed as

$$\frac{\partial^2}{\partial\theta^2} KL(\theta \| \theta_0)\Big|_{\theta=\theta_0} = I(\theta_0). \tag{8}$$

For any given θ, KL represents the ease or difficulty of distinguishing θ from θ_0. In particular, for θ varying around θ_0, KL reduces to FI. Geometrically, if KL is viewed as a curve on the plane, FI becomes the curvature of the curve at $\theta = \theta_0$. Extending this relationship to the multi-dimensional case, it is explicitly verified that the FI matrix is equal to the Hessian (i.e., second partial derivative) matrix of the KL, mathematically expressed as

$$I_{ij}(\boldsymbol{\theta}_0) = \frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j} KL(\boldsymbol{\theta} \| \boldsymbol{\theta}_0). \tag{9}$$

In the two-dimensional case, the above relationship can be displayed geometrically. First, KL is in fact a function of four random variables

(θ_{10}, θ_{20}, θ_1, θ_2) where $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20})$ is the examinee's true ability level. However, in CAT, we always assume the interim point estimate $\hat{\boldsymbol{\theta}}^{k-1}$ as the "true" value in calculating KL, that is, $\theta_{10} = \hat{\theta}_1^{k-1}, \theta_{20} = \hat{\theta}_2^{k-1}$, so KL is reduced to a function of only two random variables (θ_1, θ_2). In a graphical sense, imagine a three-dimensional space ($\lambda, \upsilon, \kappa$), with λ corresponding to θ_1, υ to θ_2 and κ to KL. Figure 1 displays the KL surface (KLS) for two different items with discrimination parameters (a_1, a_2) and a difficulty parameter (*b*), assuming the true ability point is $(\theta_{10} = -1, \theta_{20} = 1)$. KLS intersects with the undersurface (λ, υ) exactly through this true ability point (-1, 1), meaning that KL is 0 at this point. However, the KLS intersects with the undersurface not only through this single point, but through the line $a_1(\theta_1 - \theta_{10}) + a_2(\theta_2 - \theta_{20}) = 0$, denoted here as the "zero-KL information line". The zero-KL information line will be elaborated upon later in this section. Figure 1 displays the KLS for two items, the upper panel is a general item with non-zero a_1 and a_2, whereas the lower panel is the item with a_2 equal to zero.

Let us focus on the first item (i.e., upper panel in Figure 1), which is the more general case. If we cut the KLS by a plane $\theta_1 = \theta_{10}$ parallel to the vertical plane $\upsilon = \theta_2$, the resulting curve is just the KL information curve of θ_2, as shown in Figure 2. In general, the curvature of a curve $y = f(x)$ at $x = x_0$ is

$$\kappa = \left. \frac{\left| y'' \right|}{(1 + y'^2)^{3/2}} \right|_{x = x_0} . \qquad (10)$$

In the case provided here, the function of the curve can be expressed as

$KL = y = f(\theta_2)$, and $y' = \frac{d}{d\theta_2} f(\theta_2)$ and $y'' = \frac{d^2}{d\theta_2^2} f(\theta_2)$. Since this curve intersects the under surface at $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20})$ and $f'(\theta_2)\big|_{\theta_2=\theta_{20}} = 0$, the curvature reduces to $\kappa = \left|y''\right\|_{\theta_1=\theta_{10},\theta_2=\theta_{20}}$. Therefore, it is fairly straightforward to verify that the curva curvature of the intersected curve at $\theta_2 = \theta_{20} = 1$ is $\kappa = P(\underset{\sim}{\theta}_0)(1-P(\underset{\sim}{\theta}_0))a_2^{\ 2} = I_{22}$, the second diagonal element in the FI matrix. This result is in fact consistent with Chang and Ying's (1996) conclusion. Similarly, the first diagonal element corresponds to the curvature of the resulting curve by cutting the KLS with $\theta_2 = \theta_{20}$. However, if the vertical cutting plane is not parallel to either $\upsilon = \theta_2$ or $\lambda = \theta_1$, for example, the intersection of the vertical plane and undersurface is a line drilling through $(\theta_{10}, \theta_{20})$ with angle α from θ_1-axis (as shown in Figure 3), then the curvature of the resulting curve at $(\theta_{10}, \theta_{20})$ becomes slightly more complicated. For ease of interpretation and derivation, we transform the original rectangular coordinates to a cylindrical coordinate system, i.e.,

$$\begin{cases} \theta_1 = \theta_{10} + r\cos\alpha \\ \theta_2 = \theta_{20} + r\sin\alpha \\ y = y \end{cases}, \tag{11}$$

with *y* referring to the KL information. Now the curvature reduces to $\kappa = \left|y''\right\|_{r=0}$ for a fixed α. Expanding this second derivative yields,

$$\boldsymbol{\kappa} = P(\boldsymbol{\theta}_0)(1-P(\boldsymbol{\theta}_0))(a_1\cos\boldsymbol{\alpha} + a_2\sin\boldsymbol{\alpha})^2. \tag{12}$$

Given this conclusion, it is interesting to find that the curvature is zero when

$a_1 \cos\alpha + a_2 \sin\alpha = 0$, where $\alpha = \alpha_0 = \arctan(-\frac{a_1}{a_2})$ and α_0 is exactly the angle of the zero-KL information line from the θ_1-axis. In fact, this result is quite obvious if viewed from Figure 3. That is, if the vertical cutting plane intersects with the undersurface through the zero-KL line, the resulting curve is nothing but a flat line. On the other hand, the curvature in (12) is maximized when $\alpha = \frac{\pi}{2} - \alpha_0$, which means that the curvature becomes the largest along the line that is perpendicular to the zero-KL line. This finding is important for accurately estimating the composite ability $\theta = a_1\theta_1 + a_2\theta_2$. For an item with discrimination parameters $(\lambda a_1, \lambda a_2)$, where λ is a constant, it is most informative (in terms of KL information) with respect to the composite ability $\theta = a_1\theta_1 + a_2\theta_2$. In other words, this item can best distinguish θ from $\theta_0 = a_1\theta_{10} + a_2\theta_{20}$, because θ moves along the direction of the largest curvature (or one can imagine this as the largest "gradient" of KLS) toward /away from θ_0. This conclusion illustrates that if estimating a linear composite ability $\theta = a_1\theta_1 + a_2\theta_2$ is desirable, then items with discrimination parameters $(\lambda a_1, \lambda a_2)$ are favored.

Overall, although the off-diagonal elements of the FI matrix are not explicitly displayed from the geometric representation, we can still conclude that the whole FI matrix can be fully recovered from KL by taking derivatives as shown in (9). In other words, whenever KL is available, the FI matrix can be determined, but KL cannot be recovered from FI.

3.2 KL Information in Adaptive Tests

The primary source for the application of KL information in CAT is Chang and Ying's (1996) pioneering approach. In this section we will first introduce this approach in unidimensional CAT and then its extension to MAT. Secondly, a thorough discussion of the specific item parameter patterns favored by KL is given. Finally, a discussion of global information and local information in the context of item selection in MAT is presented.

3.2.1 KL information index. Motivated by the findings that KL should be used when *n* (i.e., number of items administered during the test) is small and FI when *n* is large, Chang and Ying (1996) constructed a single index, called the KL information index (KI),

$$KI(\hat{\theta}_n) = \int_{\hat{\theta}_n-\delta_n}^{\hat{\theta}_n+\delta_n} K(\hat{\theta}_n \| \theta) d\theta . \tag{13}$$

Here, δ_n determines the size of the interval over which the average is computed. Following the general asymptotic theory for ML estimators that $\hat{\theta}_n$ is asymptotically normal with mean θ_0 and variance $[I^{(n)}(\theta_0)]^{-1}$, δ_n is reasonably chosen as $\delta_n = \frac{d}{n^{1/2}}$ (Chang & Ying, 1996), because $I^{(n)}(\theta_0)$ is of order *n*; *d* is a user-defined constant. This selection of δ_n also reflects the smooth transition from KL to FI in the KI. Initially when *n* is small, this index summarizes the information of the item with respect to a wide spectrum of θ levels, which is extremely useful at the beginning of the test when $\hat{\theta}$ is far away from θ. As the test proceeds with large *n*, the magnitude of KI is essentially determined by the curvature of KI at $\hat{\theta}_n$.

Viewing KI as the area under the KL information curve, it follows that the maximum area is equivalent to the maximum curvature and therefore the maximum FI.

Based upon Chang and Ying's (1996) index, Veldkamp and van der Linden (2002) then proposed a Bayesian version of the KL information index for MAT which is expressed as

$$KI(\hat{\boldsymbol{\theta}}^{k-1}) = \int_{\boldsymbol{\theta}} ..... \int KL_i(\boldsymbol{\theta} \| \hat{\boldsymbol{\theta}}^{k-1}) P(\boldsymbol{\theta} | X_1, X_2, ... X_n) \partial \boldsymbol{\theta} \,, \tag{14}$$

and it is equivalent to

$$KI(\hat{\boldsymbol{\theta}}^{k-1}) = \int_{\theta_{10}-\frac{d}{n^{1/2}}}^{\theta_{10}+\frac{d}{n^{1/2}}} ..... \int_{\theta_{p0}-\frac{d}{n^{1/2}}}^{\theta_{p0}+\frac{d}{n^{1/2}}} KL_i(\boldsymbol{\theta} \| \hat{\boldsymbol{\theta}}^{k-1}) P(\boldsymbol{\theta} | X_1, X_2, ... X_n) \partial \boldsymbol{\theta} \quad . \tag{15}$$

Here, $X_1, X_2, ..., X_n$ are the responses, and the posterior probability $P(\boldsymbol{\theta} | X_1, X_2, ... X_n)$ serves as a "weight" in constructing the index. To show the property of this index and to graphically represent it, we will assume a flat posterior in which each θ is given equal weight. Following Veldkamp and van der Linden's (2002) logic, in the two-dimensional case, they assume the integration domain is a square centered at $(\theta_{10}, \theta_{20})$ with side length of $\frac{6}{n^{1/2}}$, and the two dimensions are of equal priority in the item selection. However, this integration domain can be adjusted according to the specific test requirements, which reflects the potential flexibility of this method. Denote the integration domain as *D*, which is central symmetric with center $(\theta_{10}, \theta_{20})$, and we can consider several cases:

1. Square domain $D = [\theta_{10} - r, \theta_{10} + r] \times [\theta_{20} - r, \theta_{20} + r]$.

2. Rectangle domain $D=[\theta_{10}-r_1,\theta_{10}+r_1]\times[\theta_{20}-r_2,\theta_{20}+r_2]$.

3. Circular domain $D=\{(\theta_1,\theta_2)\left|\theta_1^2+\theta_2^2\le r^2\right.\}$.

4. Elliptic domain $D=\{(\theta_1,\theta_2)\left|\frac{\theta_1^2}{r_1^2}+\frac{\theta_2^2}{r_2^2}\le 1\right.\}$.

The first and third cases presume that both dimensions are equally important in a test, while the second and fourth cases assume the two dimensions are weighted differently. In terms of a graphical interpretation, the KI is actually the volume of the three-dimensional region between the KLS and the (λ,ν)-plane, bounded laterally by the circular cylinder (or other cylinders depending on the integration domain) as shown in Figure 4. In adaptive testing where true values $(\theta_{10},\theta_{20})$ are unknown, the interim point estimates $(\hat{\theta}_1,\hat{\theta}_2)$ are used instead as the centroid.

3.2.2 KI and item discriminations.

Theorem 1. Let $\boldsymbol{\theta}_0$ be the true ability vector of the examinee and $\boldsymbol{a}$ be the vector of item discrimination parameters. For any given $\boldsymbol{\theta}$, let $KL_j(\boldsymbol{\theta}\|\boldsymbol{\theta}_0)$ be the KL item information. Define the item KL information Index as $KI(\boldsymbol{\theta}_0)=\iint_D KL_j(\boldsymbol{\theta}\|\boldsymbol{\theta}_0)\partial\boldsymbol{\theta}$, where D is the central symmetric domain centered around $\boldsymbol{\theta}_0$. For the two-dimensional case, $KI(\boldsymbol{\theta}_0)\propto f(a)$ as $D\to 0$. In particular, $\boldsymbol{f}(\boldsymbol{a})=a_1^2+a_2^2$ when D is a square or circle, and $\boldsymbol{f}(\boldsymbol{a})=(a_1r_1)^2+(a_2r_2)^2$ when D is a rectangle (defined by $D=[\theta_{10}-r_1,\theta_{10}+r_1]\times[\theta_{20}-r_2,\theta_{20}+r_2]$) or ellipse (defined by $D=\{(\theta_1,\theta_2)\left|\frac{\theta_1^2}{r_1^2}+\frac{\theta_2^2}{r_2^2}\le 1\right.\}$).

Outline of Proof. The proof focuses on the case when *D* is either a circle or an ellipse by use of the cylindrical coordinates; the square and rectangular case, which can be obtained through the original Cartesian coordinates, is omitted here. This

theorem focuses on the two-dimensional case; however, the conclusion can be generalized to more than two dimensions by algebraic derivation, which will be the subject of future research.

Presumably if the integration domain is a circle with diameter *2r*, with the random variables (θ_1, θ_2) taking the form in (11), the KL is expressed as

$$KL(\theta_{10}, \theta_{20}, r, \alpha) = P(\theta_{10}, \theta_{20}) \log(\frac{P(\theta_{10}, \theta_{20})}{P(r, \alpha)}) + Q(\theta_{10}, \theta_{20}) \log(\frac{Q(\theta_{10}, \theta_{20})}{Q(r, \alpha)}), \quad (16)$$

and the KL information index is reformulated as

$$KI(\theta_{10}, \theta_{20}) = \iint_D KL(\theta_{10}, \theta_{20}, r, \alpha) r dr d\alpha . \quad (17)$$

The Taylor expansion of $KL(\theta_{10}, \theta_{20}, r, \alpha)$ at $r = 0$ is written as

$$KL(\theta_1, \theta_2, r, \alpha) = KL(\theta_1, \theta_2, 0, \alpha) + rKL'(\theta_1, \theta_2, 0, \alpha) + \frac{1}{2} r^2 KL''(\theta_1, \theta_2, 0, \alpha) + o(r^2), \quad (18)$$

where $o(r^2)$ is the error term that can be ignored. The derivative is taken with respect to *r*. It is fairly straightforward to confirm that

$$KL(\theta_1, \theta_2, 0, \alpha) = 0, \quad (19)$$

$$KL'(\theta_1, \theta_2, 0, \alpha) = 0, \text{ and} \quad (20)$$

$$KL''(\theta_1, \theta_2, 0, \alpha) = \frac{(p')^2}{p(1-p)}\bigg|_{r=0} = \frac{(1-p_0)(p_0-c)^2}{p_0(1-c)} (a_1 \cos\alpha + a_2 \sin\alpha)^2, \quad (21)$$

where *c* is the guessing parameter, *p* is a short form of $P(r, \alpha)$, p_0 is the short form of $P(\theta_{10}, \theta_{20})$, and both *p and* p_0 are item response functions following a two-dimensional IRT model in (1). Substituting (18)~(21) into (17) and ignoring the error term yields

$$KI(\theta_{10},\theta_{20}) \approx \frac{\pi}{8} r^4 \frac{(1-p_0)(p_0-c)^2}{p_0(1-c)}(a_1^2+a_2^2). \quad (22)$$

The magnitude of KI is proportional to $\boldsymbol{f}(\boldsymbol{a}) = a_1^2 + a_2^2$ as long as r goes to zero. In addition, comparing (22) to (5), we can conclude that the size of KI is in effect proportional to the trace of the FI matrix.

Now let us consider the situation in which two abilities are weighted differently and take the elliptical domain in case 4 as an example. We first rewrite the cylindrical coordinates as

$$\begin{cases} \theta_1 = \theta_{10} + rt\cos\alpha \\ \theta_2 = \theta_{20} + rs\sin\alpha \\ y = y \end{cases}, \quad (23)$$

where *t* and *s* are determined by the shape of the ellipse, and they are given in advance to reflect the relative importance of the two abilities in the test. We can then verify that

$$KI(\theta_{10},\theta_{20}) \approx \frac{\pi}{8} r^4 st \frac{(1-p_0)(p_0-c)^2}{p_0(1-c)}(a_1^2t^2+a_2^2s^2), \quad (24)$$

and consequently, $KI(\theta_{10},\theta_{20}) \propto (a_1^2t^2 + a_2^2s^2)$. In summary, the theorem indicates that when the area of the integration domain approximates to zero, the magnitude of KI is proportional to the function of the two item discrimination parameters. The form of the function depends on the shape of the integration domain.

Specifically, if the domain is a circle, then the magnitude of KI, which closely resembles its unidimensional counterparts, is proportional to $a_1^2 + a_2^2$ (Wang & Chang, 2009). Notice that this relationship holds only under large sample assumption that "the domain area approximates to zero", which is satisfied if we

choose $r = 3/\sqrt{n}$ (r is chosen in this way to form a 99% confidence interval around $\hat{\theta}$; Chang & Ying, 1996; Veldkamp & van der Linden, 2002) and the test length is long enough. As an outcome, it is expected that when the two dimensions are given equal weight, item selection may capitalize on large values of $a_1^2 + a_2^2$ at a later stage in the test, and therefore items with large $a_1^2 + a_2^2$ are more likely to be chosen. In fact, $\sqrt{a_1^2 + a_2^2}$ is the so-called multi-dimensional discrimination (MDISC; Reckase & McKinley, 1991). As an analogue to unidimensional IRT, in which the discrimination parameter is related to the slope of the item response curve at the point where the slope is steepest, the MDISC is defined as the steepest slope on the item response surface (IRS) and is expressed as MDISC= $(\sum_{k=1}^{K} a_{ik}^2)^{1/2}$ if the test measures *K* dimensions. The MDISC is an overall measure of the capability of an item to distinguish between individual examinees that are in different locations in the ability space.

One important question is if two items have equal MDISC, which one is preferred? The answer is the item with larger discrimination difference, i.e., $|a_1 - a_2|$ should have higher priority in the two-dimensional case. Denote the FI matrix as $\begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$, it can be verified that the variance of $\hat{\theta}_1$ is $\text{var}(\hat{\theta}_1) = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}$ and the variance of $\hat{\theta}_2$ is $\text{var}(\hat{\theta}_2) = (I_{22} - I_{12}I_{11}^{-1}I_{21})^{-1}$. To minimize $\text{var}(\hat{\theta}_1)$, we need to simultaneously maximize I_{11} and minimize I_{22}^{-1}, thus items with maximum a_1 and minimum a_2 are desired, and vise versa for

minimizing $\text{var}(\hat{\theta}_2)$. Therefore, items with larger $|a_1 - a_2|$ are preferred in this sense.

3.2.3 KI and item difficulty. The objective of CAT is to select items that are tailored to an examinee's ability, and thus items with difficulty values close to an examinee's ability level are more likely to be selected. This trend is reflected in the item selection rules such as "match-*b* criterion" for the *a*-stratified method in UAT (Chang & Ying, 2008). In the two-dimensional case, the question would be whether there is an analytical form that the item difficulty *b*-parameter takes such that KI reaches maximum. The analysis of such a question may in fact illuminate the *b*-parameter pattern favored by KI. In terms of FI, we already know that *b* should follow a certain linear combination of θs (Mulder & van der Linden, 2009). Thus, intuitively, we would expect the difficulty-parameter to follow a similar function of the abilities in order to maximize the item KL. Figure 5 displays a snapshot of KI as a function of $(\theta_{10}, \theta_{20})$. Obviously, it is not at a single point where the KI reaches maximum, but along a line which is a linear combination of θ_{10} and θ_{20}.

The ideal situation is to find the function along which the peak line follows. Let us return to the previous derivation in (22), when the integration domain is small enough, the size of KI is approximately proportional to $\frac{(1-p_0)(p_0-c)^2}{p_0(1-c)}(a_1^2 + a_2^2)$.

Given the fixed a_1, a_2, and c, one only needs to maximize

$$g(b) = \frac{(1-p_0)(p_0-c)^2}{p_0(1-c)}. \tag{25}$$

Since $g(b)$ is actually the scalar part of the FI matrix (refer to Equation 5), it is

easy to verify that *b* should satisfy the following equation to maximize KI (Mulder & van der Linden, 2009),

$$b_{\max} = \begin{cases} a_1\theta_1 + a_2\theta_2 & \text{for c=0} \\ a_1\theta_1 + a_2\theta_2 - \log(\dfrac{-1+\sqrt{1+8c}}{4c}) & \text{for c>0} \end{cases} . \quad (26)$$

However, (26) is restricted to the assumption that the integration domain approximates zero, and thus it will be more interesting to speculate on the more general case. Unfortunately, there is no closed analytical form for *b* when *c* is greater than zero, but we are able to obtain an analytical form for *b* when *c* is assumed to be zero.

Theorem 2. Let $\boldsymbol{\theta}_0$ be the true ability vector of the examinee, $\boldsymbol{a}$ be the vector of item discrimination parameters, and b be the item difficulty parameter. For any $\boldsymbol{\theta}$, the KL information for the j^{th} item is denoted as $KL_j(\boldsymbol{\theta}\|\boldsymbol{\theta}_0)$. If we define the item KL information index as $KI(\boldsymbol{\theta}_0) = \iint_D KL_j(\boldsymbol{\theta}\|\boldsymbol{\theta}_0)\partial\boldsymbol{\theta}$, where D is the central symmetric domain centered around $\boldsymbol{\theta}_0$, then for the two-dimensional case, when $\boldsymbol{a}$ and $\boldsymbol{\theta}_0$ are fixed, the $KI(\boldsymbol{\theta}_0)$ is maximized when $\boldsymbol{a}'\boldsymbol{\theta}_0 - \boldsymbol{b} = 0$ for c=0.

Outline of Proof. First simplify the integrand in KI as

$$KL_j(\boldsymbol{\theta}\|\boldsymbol{\theta}_0) = P_0 \log\frac{P_0}{P} + Q_0 \log\frac{Q_0}{Q} = P_0(t - t_0) + \log\frac{1+\exp(t)}{1+\exp(t_0)}, \quad (27)$$

where $$P = P(\theta_1, \theta_2) = \frac{\exp(t)}{1+\exp(t)}, \quad Q = 1 - P \quad ;$$

$$P_0 = P(\theta_{10}, \theta_{20}) = \frac{\exp(t_0)}{1+\exp(t_0)}, \quad Q_0 = 1 - P_0 \qquad t = a_1\theta_1 + a_2\theta_2 - b; t_0 = a_1\theta_{10} + a_2\theta_{20} - b \quad .$$

Regard $KL_j(\boldsymbol{\theta}\|\boldsymbol{\theta}_0)$ as a function of *b*, denoted as $f(b)$. For an item with fixed discrimination parameters, integrating (27) causes the first part in (27) to drop out because it is an odd function and the integration domain is symmetric, so that the integral of $f(b)$ further simplifies to a function *g(b)* as given by,

$$g(b) = \iint_D \log \frac{1+\exp(t)}{1+\exp(t_0)} d\theta_1 d\theta_2 \quad . \tag{28}$$

Choose two new coordinates $s_1 = \theta_1 - \theta_{10}$, $s_2 = \theta_2 - \theta_{20}$ for convenience and let $A = a_1 s_1 + a_2 s_2$; since D is central symmetric, $(-s_1, -s_2)$ is inside D whenever (s_1, s_2) is also inside, and thus

$$\begin{aligned} 2g(b) &= \iint_D (\log \frac{1+\exp(t_0 + a_1 s_1 + a_2 s_2)}{1+\exp(t_0)} + \log \frac{1+\exp(t_0 - a_1 s_1 - a_2 s_2)}{1+\exp(t_0)}) ds_1 ds_2 \\ &= \iint_D (\log(1+\exp(2t_0) + \exp(t_0 + A) + \exp(t_0 - A)) - \log(1+\exp(2t_0) + 2\exp(t_0))) ds_1 ds_2 \end{aligned} \quad . \tag{29}$$

Therefore, the integrand in (28) is rewritten as,

$$\varphi(b) = \log(1 + \frac{\exp(t_0)}{(1+\exp(t_0))^2}(e^A + e^{-A} - 2)) \quad . \tag{30}$$

Because $e^A + e^{-A} - 2 \geq 0$, it is sufficient to maximize $\frac{e^{t_0}}{(1+e^{t_0})^2}$. Taking the derivative with respect to *b*, one obtains

$$\frac{d\varphi(b)}{db} = \frac{\exp(t_0)(1-\exp(t_0))}{(1+\exp(t_0))^3} \quad . \tag{31}$$

When $t_0 > 0$, $\varphi(b)$ is decreasing, when $t_0 < 0$, $\varphi(b)$ is increasing, thus $t_0 = 0$ is the unique value to maximize $\varphi(b)$, which also maximizes $f(b)$ when a_1 and a_2 are fixed. Thus, $a_1\theta_{10} + a_2\theta_{20} - b = 0$ is the function that maximizes the KI.

In all, the above discovery points to a way to select items based upon item

difficulty. In adaptive testing, when interim estimate $\hat{\boldsymbol{\theta}}$ is updated, we can choose the item with a *b*-parameter as close as possible to $\boldsymbol{a}'\hat{\boldsymbol{\theta}}$, which is similar to the "match-*b*" criterion in unidimensional CAT (Chang & Ying, 1999). In fact, this finding will help in constructing a simplified version of KI (SKI), which will be explicated in the following section. Although the theorem only holds when the M2PL is employed, the result is still useful, as M2PL is extensively discussed in the literature (see van der Linden, 1996; Veldkamp & van der Linden, 2002; Reckase & McKinley, 1991). It is important to note that even if the M3PL is employed, the results in (26) could still be used, albeit for the slightly more limited situation due to the requirement of a longer test length.

Note that both theorems can be generalized to the higher-dimensional case (for any $p>2$). As to Theorem 2, the conclusion can be extended in a fairly straightforward manner. In fact, the proof from equations (27) to (31) is applicable to any given number of integrations as long as the integration domain is central symmetric, and therefore we can conclude that KI is maximized when $\boldsymbol{a}'\boldsymbol{\theta}_0 - \boldsymbol{b} = 0$ for c=0, where $\boldsymbol{\theta}_0$ is a *p*-dimensional vector for any given number of *p*. As to Theorem 1, although the conclusion still holds for $p>2$, the strategy adopted to prove for $p=2$ is different than that for $p>2$. The former employs cylindrical coordinates, and the latter has to utilize rectangular coordinates. The application of cylindrical coordinates yields an interesting geometric explanation for the relationship between KL and Fisher information. However, it is not clear how to show such a relationship for $p>2$ geometrically. Therefore, to be consistent with the geometric explanations discussed in the paper, we will only present the proof for

p=2. As to p>2, please refer to Wang and Chang (2010) for the proof.

3.3 Global Information vs. Local Information

KL is termed as "global information" because it quantifies the discrimination power between two ability levels, θ_0 and θ_1, whether they are close together or not; whereas FI only measures the item discrimination close to θ_0 and in this sense it is called "local information". When a local information criterion is used, item selection procedures may favor items with optimal properties that are far from the examinee's actual ability level, a phenomenon called the "attenuation paradox" in test theory (Lord & Novick, 1968). A general approach to deal with the attenuation paradox is to replace a maximum-point information criterion with an interval-based criterion so that items that provide information over a larger range of trait values are preferred (van Rijn, Eggen, Hemker, & Sanders, 2002; Veerkamp & Berger, 1997). Such criterions include, for example, maximum interval information $\int_{\theta_L}^{\theta_U} I(\theta)d\theta$, where $[\theta_L, \theta_U]$ forms a latent trait interval over which the information is accumulated (Veerkamp & Berger, 1997; Passos, Berger, & Tan, 2008); maximum posterior weighted information $\int_{\theta_L}^{\theta_U} I(\theta)P(\theta|X_1, X_2, ...X_n)d\theta$ (van der Linden, 1998), and maximum expected information, such as $\sum_x \int_{\theta_L}^{\theta_U} P(X_n = x|X_1, X_2, ...X_{n-1})I(\theta|X_1, X_2, ...X_n = x)d\theta$, where $P(X_n = x|X_1, X_2, ...X_{n-1})$ is the posterior predictive distribution (van der Linden, 1998). However, KI in (14) is by definition associated with an interval which is

close in proximity to the true ability value θ_0. The main difference between KI and the above mentioned interval-based criterions is that KI is a dynamic combination of KL and FI. In UAT when *n* is small, KI relies on the item KL; and when *n* is large, KI gradually transitions to rely on FI, and thus KI behaves quite similar to the maximum FI criterion. However, in two-dimensional adaptive testing, this smooth transition is not so obvious. Although we can still relate the KI directly to the FI matrix under large-sample approximation, KI and FI based methods have distinct selective mechanisms, even when the test length is longer. Specifically, the *D*-optimality (Segall, 1996) criterion is meant to maximize the determinant of the **FI** matrix, which is algebraically equivalent to maximizing the product of the eigenvalues of the **FI** matrix. As a result of taking the determinant, items that mainly test a single ability are generally most informative. Alternatively, KI tries to maximize the trace of the **FI** matrix, which is the same as the summation of the eigenvalues. Therefore, the distinction between *D*-optimality and KI reduces to the comparison of the multiplicative and additive rules.

CHAPTER 4 SIMPLIFIED KL INDEX

A MAT algorithm can be used to simultaneously asses multiple abilities, while tailoring the test to match an examinee's set of latent abilities, thus offering greater precision. However, the actual application of MAT is greatly limited by its computational intensity. In order to be able to use MAT in a real-time application, the item selection process needs to be fast and efficient, however, the existing KL based methods, which contain multiple integrations, are extremely time-consuming. Given this critical drawback in the application of MAT, it would seem promising to construct a simple index that could provide an adequate approximation to the item KL information.

Stemming from the pervious discussion of the relationship between KI and item parameters, a simplified KL information index (SKI) is proposed. Depending upon whether the term $|a_1 - a_2|$ is considered in the item selection, two versions of SKI can be formulized as follows,

$$SKI1 = \frac{1}{|\mathbf{a}'\theta - b|}(\mathbf{a}'\mathbf{a})$$

and

$$SKI2 = |a_1 - a_2| \frac{1}{|\mathbf{a}'\theta - b|}(\mathbf{a}'\mathbf{a}),$$

where $\boldsymbol{a}$ is the column vector of item discrimination parameters, and $\mathbf{a}' = (a_1, a_2)$ is a special case in the two-dimensional space. The parameter b represents item difficulty. Note that $SKI1$ is exactly constructed from the two theorems provided

earlier in the paper, whereas *SKI*2 adds an additional term with the understanding that $|a_1 - a_2|$ is another criterion KI uses to select items. In fact, Theorem 1 has shown that the value of KI is proportional to MDISC. Thus, if KI is the primary criterion for item selection, then MDISC plays a pivotal role, and therefore $|a_1 - a_2|$ should be treated as a secondary criterion. Note that this belief will be further supported by the simulation results presented in the next section and its understanding is important for future development of an exposure control method. For example, the *a*-stratification method in unidimensional adaptive tests (Chang, Qian, & Ying, 2001; Chang & Ying, 1999; Yi & Chang, 2003) is one promising exposure control method. The motivation of the *a*-stratification method is that the magnitude of item Fisher information largely depends on the item discrimination. Thus, if we want to apply the same idea to multi-dimensional adaptive tests, we can stratify the item pool according to MDISC.

In addition, SKI involves a multiplier, $\frac{1}{|\mathbf{a}'\theta - b|}$, assuming the M2PL model is employed. If the M3PL model is used, then this multiplier would need to be modified to incorporate *c* with (26) as a reference. If some of the abilities are more important then others, then SKI would need to be modified by substituting $(\mathbf{a}'\mathbf{a})$ with $(\mathbf{a} \bullet \mathbf{w})'(\mathbf{a} \bullet \mathbf{w})$, where $\mathbf{w}$ is a column vector (same length as $\mathbf{a}$) of weights, and $\bullet$ means a one-to-one multiplication of each element in $\mathbf{a}$ and $\mathbf{w}$. As described, SKI greatly reduces the computational intensity by avoiding the multiple integrations that have burdened the application of MAT.

CHAPTER 5 SIMULATION STUDIES AND RESULTS

Simulation studies were carried out in this research for two important purposes: (a) To gather numerical evidence in support of our theoretical findings about MDISC and KI, and (b) to show that the two versions of SKI outperform the original KI. In order to show support for both of these purposes, two separate simulation studies were implemented.

5.1 Simulation 1

This simulation study was done in order to verify whether or not the items with larger MDISC values are more likely to be chosen when KI is used for item selection. In addition, this part of the study tried to obtain empirical evidence to show for items with similar MDISC, those with higher $|a_1 - a_2|$ are more preferable.

5.1.1 Item Pool Structure

In this simulation, two dimensions are considered, and we assume the item pool is sufficiently rich with items of various difficulty values, so that for every given θ value there is a corresponding item difficulty parameter *b* that follows the form in (26). For the first item pool, three types of item discrimination parameters are generated. The first type has $a_1 \sim U(1.5, 2)$, $a_2 \sim U(1.5, 2)$ (denoted as Type 1); the second type has $a_1 \sim U(1.5, 2), a_2 = 0$ (denoted as Type 2); and the last type has $a_1 = 0, a_2 \sim U(1.5, 2)$ (denoted as Type 3). Note that the items in Type 1 have the largest MDISC and smallest $|a_1 - a_2|$ values. The second item pool also has three

types of item discrimination parameters with the last two remaining the same as the first pool, with the only difference being in the first type, which changes to $a_1 \sim U(1.0,1.5)$ and $a_2 \sim U(1.0,1.5)$ (denoted as Type 4). Note that items in Type 2 to 4 have similar MDISC, but items in Type 2 and 3 have larger $|a_1 - a_2|$ values. For simplicity, the guessing *c* parameters for all items were set to zero.

5.1.2 Examinee generation

The true ability vector was generated from a multivariate normal distribution with mean of zero, and with a correlation of 0.5 between the two dimensions. The examinees were simulated in this way to represent a typical population of examinees, in which the traits are moderately correlated, with a sample size of 1000 chosen in order to produce stable results.

5.1.3 Ability Estimation

The Expected A Posterior (EAP) method was used to update $\hat{\theta}$s. Specifically, suppose (k-1) items have been administered, $\hat{\theta}^{k-1}$ is calculated as,

$$\hat{\theta}_l^{k-1} = \boldsymbol{E}(\theta_l \,|\, \boldsymbol{u}_{k-1}) \frac{\int \theta_l \int .... \int \boldsymbol{L}(\boldsymbol{\theta}; \boldsymbol{u}_{k-1}) \boldsymbol{f}(\boldsymbol{\theta}) \partial \boldsymbol{\theta}}{\int .... \int \boldsymbol{L}(\boldsymbol{\theta}; \boldsymbol{u}_{k-1}) \boldsymbol{f}(\boldsymbol{\theta}) \partial \boldsymbol{\theta}}, \tag{32}$$

where $\boldsymbol{u}_{k-1}$ is the response vector. The integration can be approximated using Gauss-Hermite quadrature (Stroud & Sechrest, 1966). The prior density $\boldsymbol{f}(\boldsymbol{\theta})$ is chosen to match the multivariate normal distribution used to generate examinees' abilities.

5.1.4 Item selection method and termination rule

In this simulation study, only the original KL index (KI) was considered. The integration domain was chosen to be a circle, although it would not make any

difference if it were a square. The test length was fixed to be 40 items, deliberately chosen in order to be able to show a clear trend.

5.1.5 Evaluation Criterion

At each stage of the test (n=1,2,….40), the frequency of each item type will be recorded.

5.1.6 Results

Figure 6 shows the frequency of each item type being exposed during the tests following the first item pool structure. Type 1 items are shown to be the most preferred items throughout the whole test and their frequencies are even higher than the sum of the Type 2 and 3 items. Admittedly, allowing items to be selected from an infinite large item pool is not very likely in reality, however, this pool was created in this way in order to more clearly track the exposure frequencies associated with each number of administered items, which allowed for a more complete examination of the selection mechanism behind KI. This result shows that MDISC is the primary underlying criterion that controls the item selection. However, it is interesting to note that although KI depends heavily on MDISC only in the later stage of the test according to the analytical derivation, the simulation results actually posit that KI still tends to pick the items with high MDISC at each stage of the test as long as the item bank is not exhausted of high MDISC items. This result also explains the unbalance in item exposure under the KI method, which further highlights the need for the development of an SKI approach that will balance the exposure by forcing the item selection to follow a certain order of MDSIC.

Figure 7 shows the longitudinal frequency results for the second item pool

structure. As can be seen, when MDISC is equal for each type of item, those with higher $|a_1 - a_2|$ are much more likely to be chosen.

5.2 Simulation 2

The second simulation study was implemented in order to provide empirical evidence to show that both versions of SKI can actually provide comparable or even higher estimation accuracy when compared to the original KI, but with less computational intensity.

5.2.1 Item bank construction

A test is multi-dimensional when it assesses more then one latent trait, however, there are two kinds of multi-dimensionality that have been categorized. The first is between-item multi-dimensionality and the second is called within-item multi-dimensionality (Adams et al., 1997). In fact, between-item multi-dimensionality can be regarded as a special case of within-item multi-dimensionality, in which each item is constrained to measure only one trait. This research utilizes the within-item multi-dimensionality, which is the more general condition, as the basis for this simulation study. The item bank is constructed following a two-dimensional two-parameter IRT model, with the item bank size set to 900. Although there is a rule of thumb that states the pool needs to have at least 12 times as many items as the test length (Stocking, 1994) due to the item pooling effect, other researchers have recommended even larger ratios (Chang & Zhang, 2002). Thus, an item pool with 900 items should be large enough for the purposes of this research. The two discrimination parameters were generated from a log-normal distribution, bounded

within 0.25 and 1.5; and the difficulty parameters were generated from a standard normal distribution.

5.2.2 Examinee generation

Two groups of examinees were generated. The *first* group of examinees were simulated to evaluate the overall estimation accuracy of each method in a general examinee population. The examinees were generated in the same way as in the first simulation study (see Section *5.1.2*). The *second* group was generated to assess the conditional estimation accuracy of each method, therefore, as in Finkelman, Nering, and Roussos (2009), examinees were simulated with true abilities on a two-dimensional grid spanning the square $\theta_{01}, \theta_{02} = (-2.0, -1.6, ...2.0)$. By crossing 11 discrete points over the two dimensions, the simulation is performed over a grid of 121 θ values. At each θ, 500 simulations are run, and the total number of simulated tests is 60,500. In both cases, the test length is set to be 40.

5.2.3 Item Selection Rules

The original KL index, both versions of the simplified KL index, and randomized item selection methods were used in the following study. In addition, we added another comparison for the first group of examinees, which was a 40-item non-adaptive test given to all the examinees. In order to select the 40 most informative items from the item bank, an index ***K*** was utilized, indicating the overall KL information carried by each item, which is defined as,

$$K = \int KI(\boldsymbol{\theta}_0) f(\boldsymbol{\theta}_0) \partial \boldsymbol{\theta}_0 = \int_{\boldsymbol{\theta}_0} \int_{\boldsymbol{\theta}_1} KL(\boldsymbol{\theta}_1 \| \boldsymbol{\theta}_0) f(\boldsymbol{\theta}_0) \partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_0$$
$$= \int_{-\infty}^{+\infty} \left[\int_{\boldsymbol{\theta}_0 - \boldsymbol{\delta}}^{\boldsymbol{\theta}_0 + \boldsymbol{\delta}} KL(\boldsymbol{\theta}_1 \| \boldsymbol{\theta}_0) \partial \boldsymbol{\theta}_1 \right] f(\boldsymbol{\theta}_0) \partial \boldsymbol{\theta}_0 \quad . \quad (33)$$

Here $KL(\boldsymbol{\theta}_1 \| \boldsymbol{\theta}_0)$ is defined in equation (7), $f(\boldsymbol{\theta}_0)$ is the prior density and the

integration is approximated by Gauss-Hermite quadrature (Stroud & Sechrest, 1966). The ***K*** index is basically the integration of the KL index over the entire ability space weighted by the prior density of each ability point, and in this way it measures the overall discrimination power of an item with respect to all possible ability points in the space. KL information is employed here because the contribution of each item to the test information is independent, which simplifies the assembly of the non-adaptive test (Veldkamp & van der Linden, 2002). The 40 items with the largest ***K*** index were chosen to form the psychometrically optimal non-adaptive test. These 40 items are given to all examinees repeatedly. Lastly, the *D*-optimality method (Segall, 1996, 2001; Mulder & van der Linden, 2009) based on the Fisher information matrix is also carried out for the first group of examinees.

5.2.4 Evaluation Criterion

The estimation accuracy of each method was measured by mean squared error (MSE),

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_{ij} - \theta_{ij})^2, \quad j = 1, 2, \tag{34}$$

with bias of each element in $\boldsymbol{\theta}$,

$$Bias = \frac{1}{N}\sum_{i-1}^{N}\left(\hat{\theta}_{ij} - \theta_{ij}\right), \ j = 1, 2. \tag{35}$$

N is the population size, *j* indicates the ability dimension, *i* denotes the examinee, $\hat{\theta}_{ij}$ is the final EAP estimate for each examinee on each dimension and, θ_{ij} is the corresponding true value. To evaluate the performance of the different methods conditioning on each ability point, the above two criterions are calculated as conditional MSE and Bias, namely, the true ability is fixed at every single value. In

addition, the approximate computational time taken by each approach was recorded to show the computational efficiency of each method.

To investigate whether the original and simplified KL indices tend to select the same items, an overlap rate was calculated between these two indices, which is defined as the average proportion of common items selected by the two methods:

$$overlap = \frac{1}{N}\sum_{i=1}^{N}\frac{\#(s_{i,KI} \cap s_{i,SKI})}{L}. \qquad (36)$$

$s_{i,KI}$ is the set of items administered to the i^{th} examinee by the original KI method, $\#(s_{i,KI} \cap s_{i,SKI})$ is the number of items in the intersection of the sets $s_{i,KI}$ and $s_{i,SKI}$. In addition, the empirical frequencies of the item exposure against the item discrimination parameters were assessed (Mulder & van der Linden, 2009).

5.2.5 Results

Table 1 shows the performance of each method for the simulated examinee group 1. Both the original KI and two versions of SKI produce comparable and satisfactory estimation accuracy as shown by the small MSE and Bias compared to the Random method. In addition, each of the methods outperforms the non-adaptive counterpart. D-optimality based on Fisher information matrix is also considered here, and this method produces similar results to KI and both versions of SKI.

TABLE 1 Bias and MSE of θ_1 and θ_2 for different item selection rules.

	Mean Squared Error		Bias		Computation Time[1] (seconds)
	θ_1	θ_2	θ_1	θ_2	
Original KI	0.112	0.103	-0.025	-0.020	0.18
*SKI*1	0.115	0.102	-0.057	-0.035	0.06
*SKI*2	0.117	0.104	-0.032	-0.011	0.06

Random	0.242	0.251	-0.153	-0.147	0.01
Non-adaptive	0.205	0.157	-0.105	-0.113	N/A
D-optimality	0.120	0.103	0.007	-0.019	0.11

1. Computation time is the average CPU time needed for selecting a single item for one examinee; the program was run on a 2.2GHz processor with Compaq Visual Fortran version 6.6.

Since the results of the conditional MSE and Bias for both dimensions are quite similar, only the results for θ_1 are displayed in Figure 8. As can be seen when reviewing the plots of MSE, all three methods exhibited the same pattern of precision. In particular, the lowest precision was found when θ_1 was high and θ_2 was low, or vice versa. Note that this pattern was also reported in Finkelman et al. (2009), and they ascribe it to the prior distribution used in EAP estimation because little weight is given to those points in the prior. It should be noted that Figures 8a, 8b, and 8c are all on different scales, which may indicate that both *SKI*1 and *SKI*2 outperform KL for most of the theta points. Furthermore, *SKI*2 produced uniformly smaller MSE values at every single ability point among the three methods. However, the differences in absolute bias between the three methods were consistently smaller than the differences in MSE.

The overlap rate between KI and *SKI*1 is 0.603, whereas the overlap rate between KI and *SKI*2 is 0.365. A higher overlap rate means that the two methods tend to select the same items, and thus the results indicate that compared with *SKI*2, *SKI*1 is a closer approximation to the KL index. Figure 9 illustrates the item exposure frequency against the item discrimination parameters, in which each circle represents an item with its discrimination parameters as coordinates. The area of the circle is proportional to the item exposure rate, and in general, the largest circles

correspond to roughly 0.5 exposure rate.

As shown in Figure 9, both KL and *SKI*1 display similar patterns of item exposure, i.e., both methods prefer to select items with a high multi-dimensional discrimination. However, *SKI*2 tends to select items with high $|a_1 - a_2|$, which explains why the overlap rate between *SKI*2 and KI is relatively low. Note that the frequency of difficulty parameters are omitted here because for any item selection algorithm, the distribution of difficulty parameters is close to standard normal, which is the same distribution used to generate the *b*-parameters.

CHAPTER 6 CONCLUSIONS

This paper first discusses the relationship between KL information and the Fisher information matrix in the context of multi-dimensional IRT. From the mathematical connections between the two approaches and the fact that the FI matrix can be fully recovered from KL, the explication should allow for a better understanding of the various item selection methods that are based on these two different information measures.

This research then discusses the multi-dimensional version of the KL index, highlighting its characteristics for the two-dimensional case. In the two-dimensional case, KI is viewed as the volume under the KL information surface. Specifically, two analytical results are provided to explore the full capacity of the KL index. The first analytical result shows that the magnitude of KI, when the test length is long enough, is asymptotically equivalent to the trace of the FI matrix and, consequently, proportional to the square of the item MDISC. Although KI tends to prefer items that are sensitive to multiple abilities, KI also favors items that have larger differences between the two discrimination parameters. This is indeed similar to Fisher information based methods, such as *D*-optimality, which favor items that are highly discriminating on a single ability (Mulder & van der Linden, 2009). This comparable parametric targeting illuminates the underlying connections between the KL and FI matrix. The second analytical result shows that an item is most informative only when its difficulty matches the linear combination of the current ability estimates.

One important finding is that in the multi-dimensional case, KI can no longer approximate the FI based methods, even when the test is long enough. This is seen from our first theorem, which posits that maximizing KI in the course of the test will reduce to maximizing the trace of FI matrix, whereas none of the FI based methods (such as *D*-optimality or *A*-optimality) maximize the trace directly. This is quite different from how KI behaves in the unidimensional case. In fact, the principle initiative of constructing KI is to let KL dominate item selection early in the test and let FI direct item selection later in the test, which follows from the belief that global information should be used first. In the unidimensional case, this smooth transition from KL and FI is perfectly mirrored in KI through the changes of integration size δ_n. Despite the fact that the desired smooth transition is not directly reflected in the multi-dimensional case, the "global information first" idea is still embodied. Future research should consider defining a new KL index that can more closely mimic the KI in the unidimensional case.

According to the analytical results, the KL index can be approximated by less computationally intensive methods, and therefore two versions of SKI were proposed. The results from the simulation studies indicate that *SKI*1 approximates KL quite accurately and that *SKI*2 outperforms KL by producing more accurate ability estimation. Furthermore, both *SKI1* and *SKI2* can easily incorporate the need for weighting each dimension differently, which is often desired in practice (van der Linden, 1996; Veldkamp & van der Linden, 2002; Mulder & van der Linden, 2009). In future research, various non-statistical constraints need to be considered in the item selection process, as we believe the benefit of SKI is more

apparent in particular when adaptive tests are trying to optimize a solution based on a larger number of constraints. In addition, the new finding that KI primarily depends on MDISC may in fact facilitate further development of a stratification-based exposure control method (e.g., Chang & Ying, 1999, 2008). Overall, both *SKI*1 and *SKI*2 should be given precedence over the original KI for longer tests and tests with more constraints, due to their numerical simplicity, and thus its increased efficiency in real-time computer-based testing. However, the SKI developed in this research is only a prototype and many issues still remain to be further investigated. First, the construction of *SKI*1 is directly from the two theorems, whereas in *SKI*2, we intentionally embed a term $|a_1 - a_2|$ because the first simulation results indicated that this term also played an important role in item selection. In fact, adding this additional term does improve the accuracy of the adaptive tests, as reflected by the smaller conditional MSE. However, within *SKI*2, MDISC and $|a_1 - a_2|$ are simply combined through multiplication, which does not reflect the relative importance of each component in the item selection, and thus further modifications are needed. Another issue originates from the fact that the relationship between KI and item MDISC depends on large sample approximation, which means the relationship may not be as strong at the beginning of the test. Although the first simulation results indicated a strong relationship at all stages of the test, considering the large item pool used in the simulation, which would most likely not be found in practice, a future research study should work on building an index that only relies upon discrimination in the later stages of the test. Another line of future research is to derive mathematically how the bias or MSE relies upon the item selection rule in

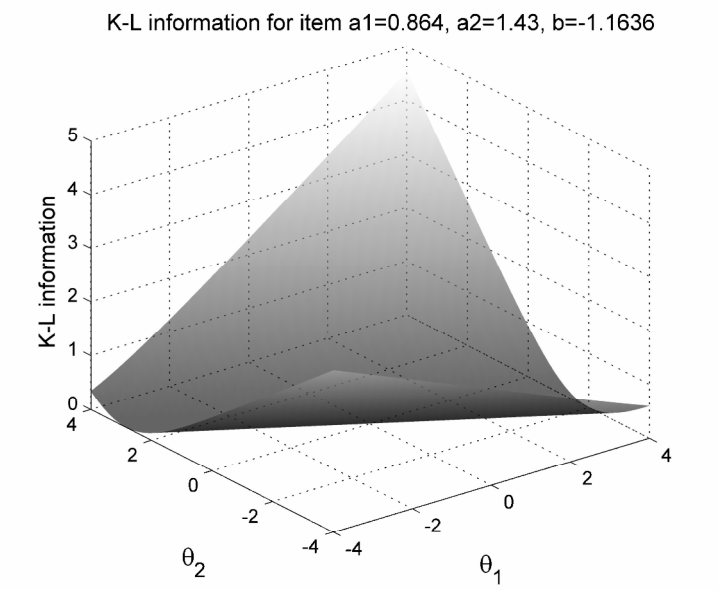
MAT, just as Chang and Ying (2008) did in the unidimensional case, which would allow one to assess how the different item selection rules affect estimation accuracy.

This paper focuses on the two-dimensional case which is readily visible by geometric representation, displaying mathematical elegance. However, the conclusion can also be extended to more than two dimensions (Wang & Chang, 2010). Note that the relationship between KI and item discrimination are derived from a geometric perspective in this paper, but for higher dimensional space, which is not observable, algebraic derivation is required. The two versions of SKI can also be generalized to more than two dimensions, with the only modification being to include the corresponding multi-dimensional discrimination parameters and a dissimilarity measure among the multiple discrimination parameters per item.

.

FIGURES

Figure 1



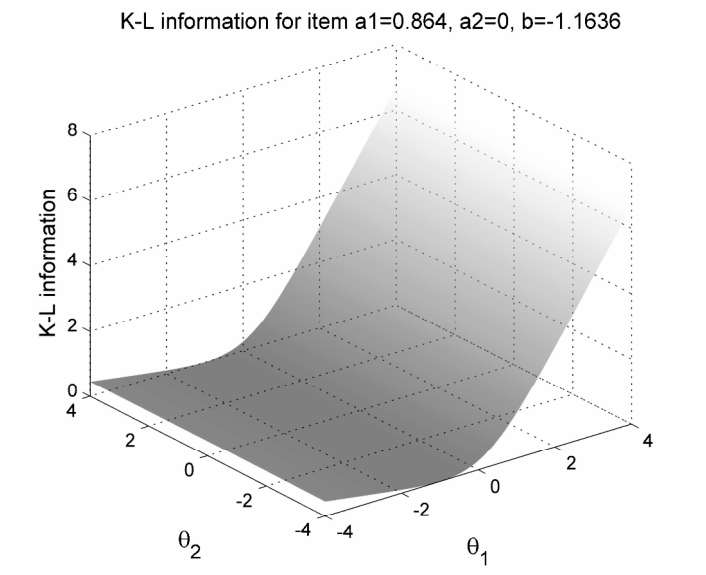
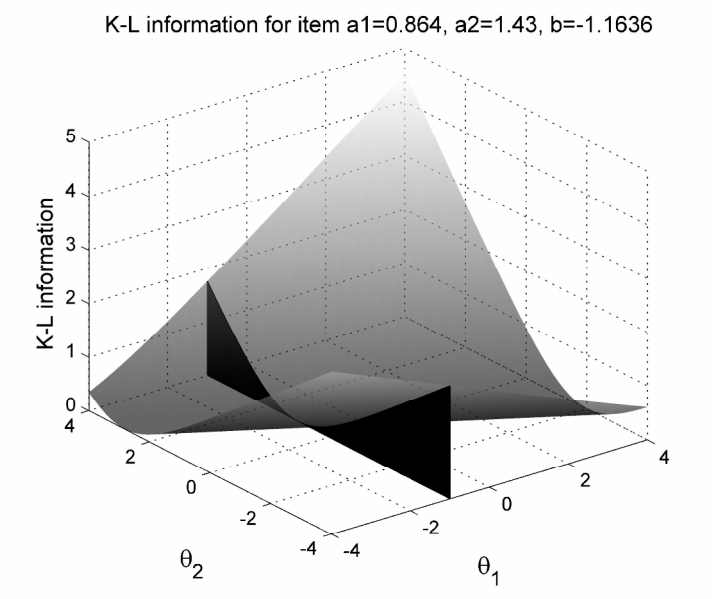


Figure 2


K-L information for item a1=0.864, a2=1.43, b=-1.1636
K-L information
θ2
θ1

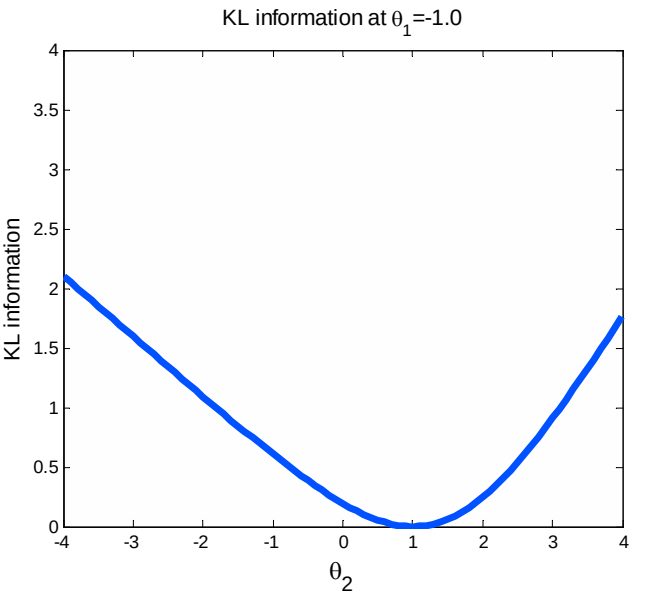
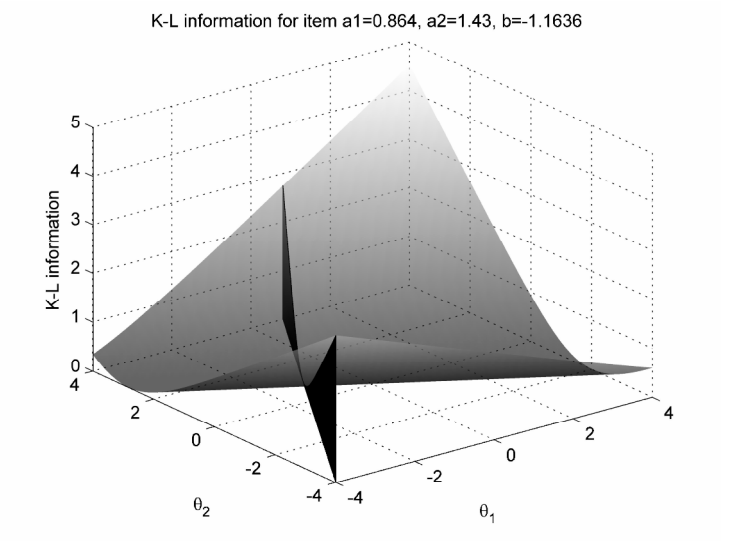

KL information at θ1=-1.0
KL information
θ2

Figure 3


K-L information for item a1=0.864, a2=1.43, b=-1.1636
K-L information
θ2
θ1

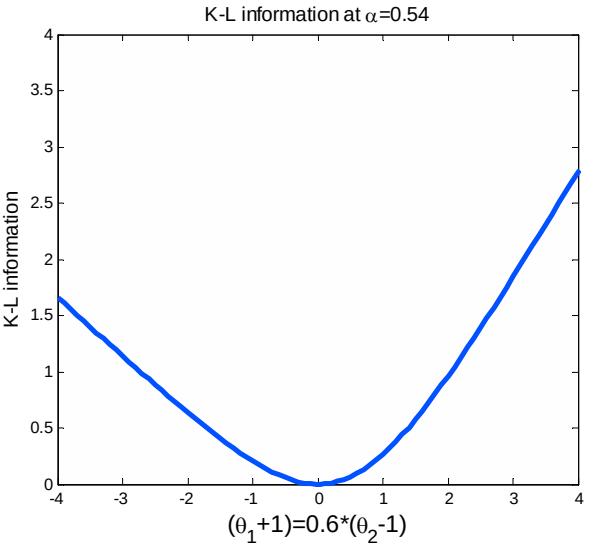

K-L information at α=0.54
K-L information
(θ1+1)=0.6*(θ2-1)

Figure 4

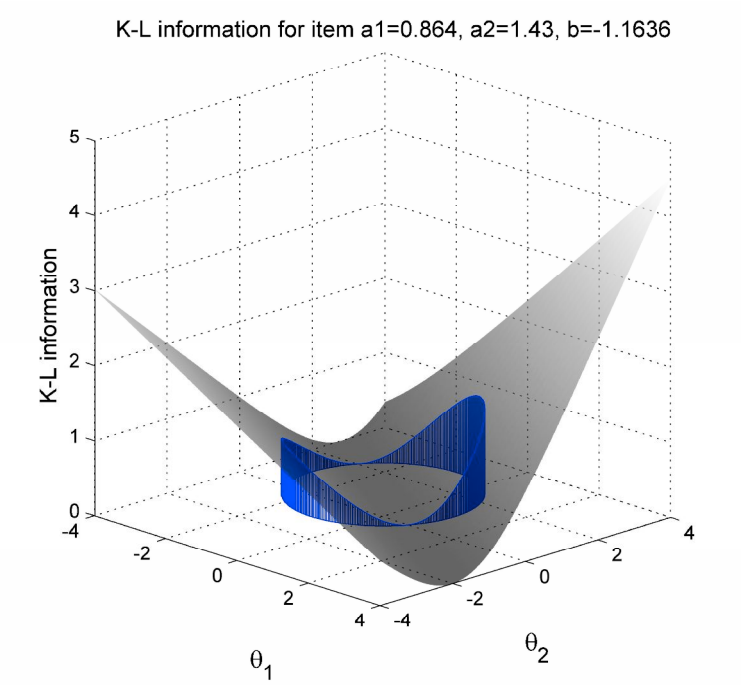

K-L information for item a1=0.864, a2=1.43, b=-1.1636
K-L information
0
1
2
3
4
5
-4
-2
0
2
4
θ1
θ2

Figure 5

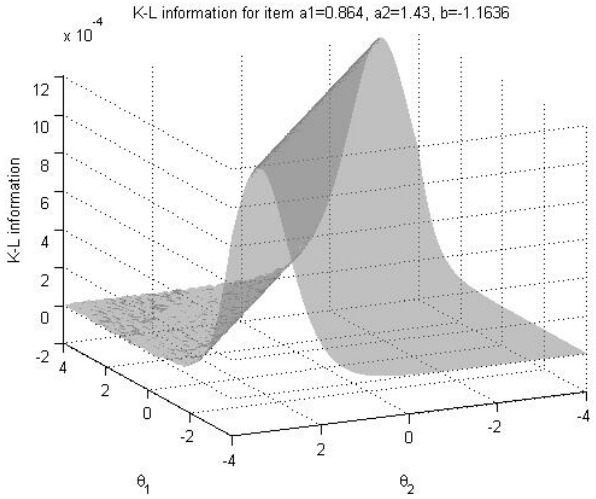

K-L information for item a1=0.864, a2=1.43, b=-1.1636
x 10^-4
K-L information
12
10
8
6
4
2
0
-2
4
2
0
-2
-4
θ1
2
0
-2
-4
θ2

Figure 6

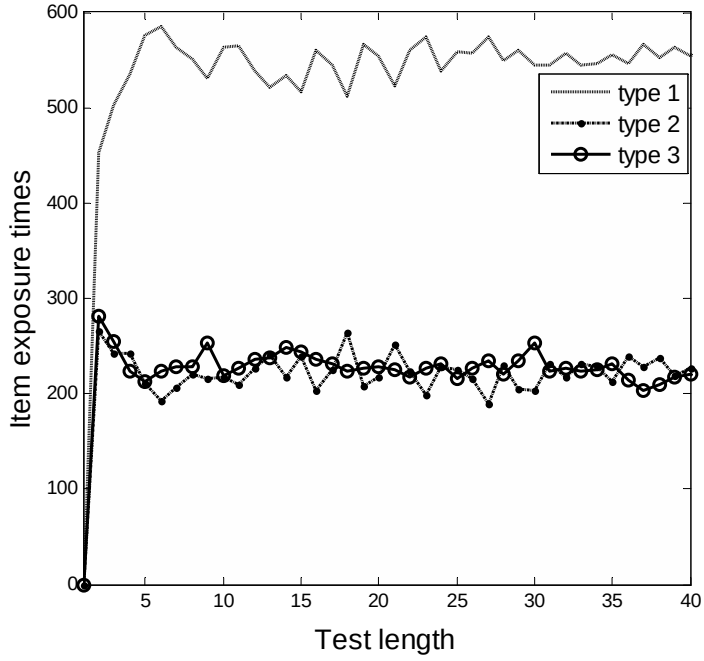

600
500
400
300
200
100
0
5
10
15
20
25
30
35
40
Item exposure times
Test length
type 1
type 2
type 3

Figure 7

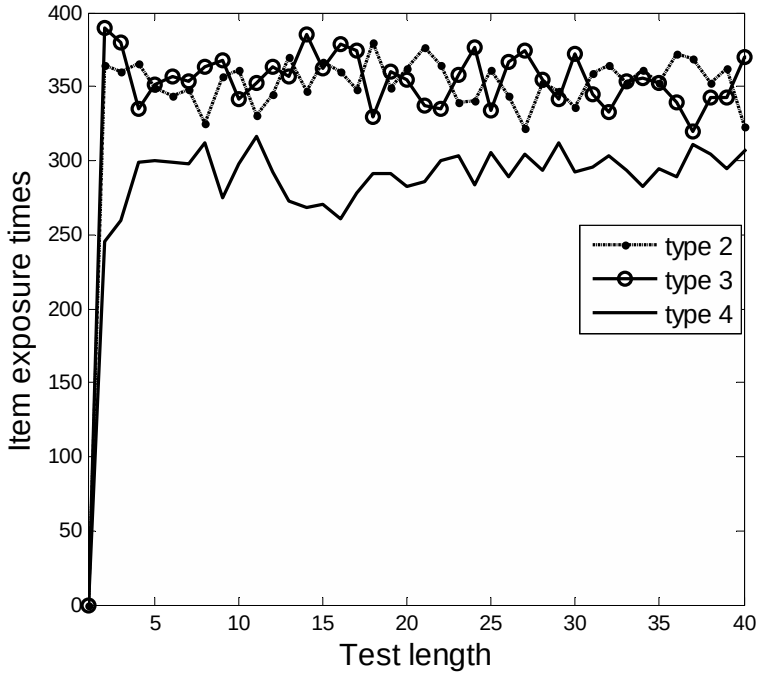
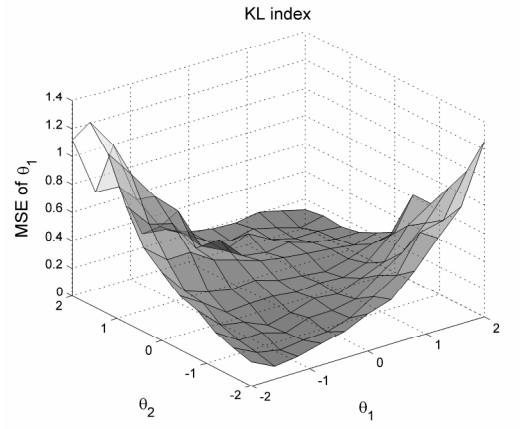
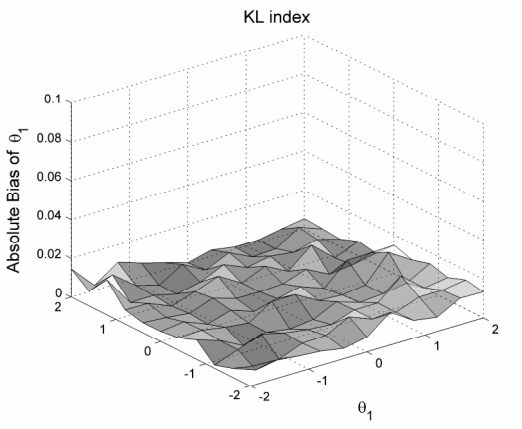

Item exposure times
Test length
type 2
type 3
type 4
0
50
100
150
200
250
300
350
400
5
10
15
20
25
30
35
40

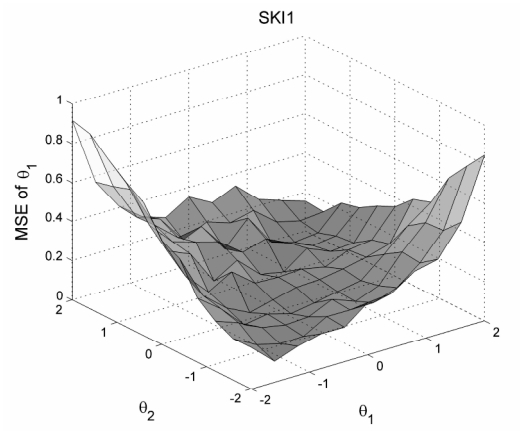
Figure 8



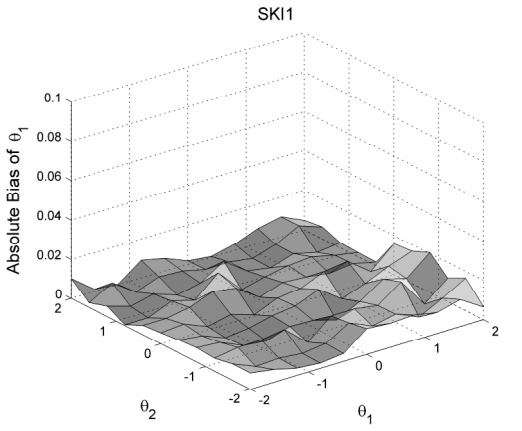
(c)



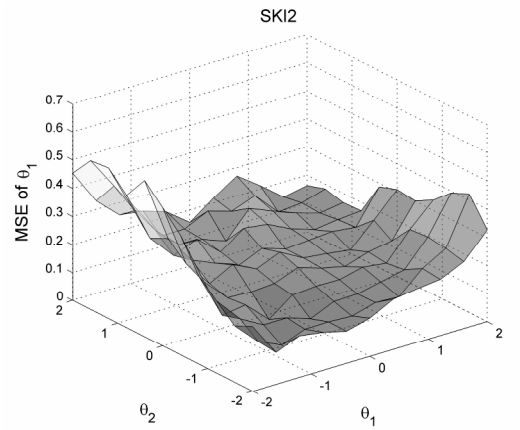
(d)



(e)



(f)



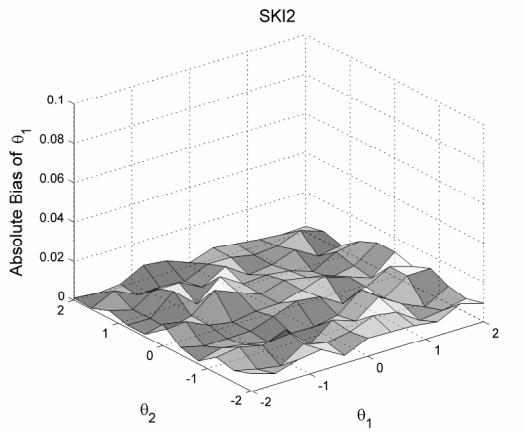
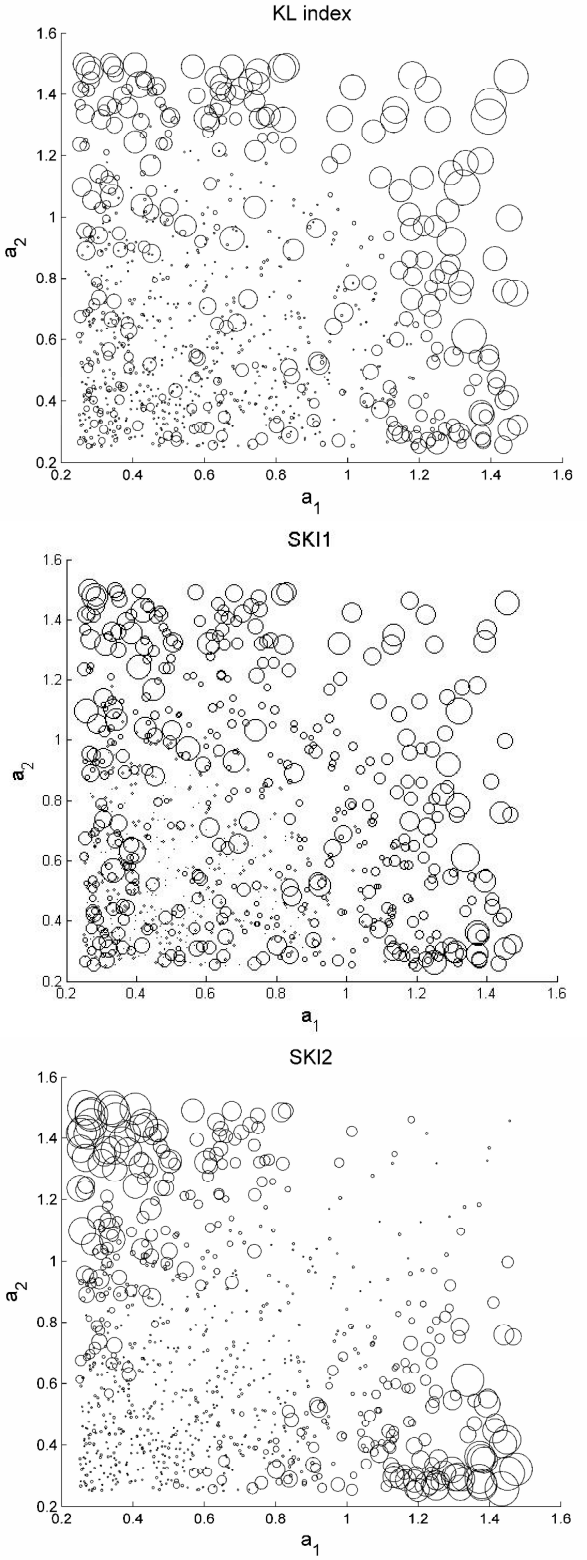


Figure 9



REFERENCES

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1-23.

Allen, D. D., Ni, P. S., & Haley, S. M. (2008). Efficiency and sensitivity of multidimensional computerized adaptive testing of pediatric physical functioning. *Disability and Rehabilitation*, 30(6), 479-484.

Birnbaum, A. (1968) Some latent trait models and their use in inferring an examinee's ability. In F.M.Lord & Novick (Eds.), *Statistical theories of mental test scores* (pp. 379-479). Reading, MA: Addison-Welsey.

Bloxom, B. M. & Vale, C. D. (1987, June). *Multidimensional adaptive testing: A procedure for sequential estimation of the posterior centroid and dispersion of theta*. Paper presented at the meeting of the Psychometric Society, Montreal, Canada.

Chang, H. H. (1996). The asymptotic posterior normality of the latent trait for polytomous IRT models. *Psychometrika*, 61(3), 445-463.

Chang, H. H., Qian, J. H., & Ying, Z. L. (2001). a-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, 25(4), 333-341.

Chang, H. H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, 58(1), 37-52.

Chang, H. H., & Ying, Z. L. (1996). A global information approach to computerized adoptive testing. *Applied Psychological Measurement,* 20(3), 213-229.

Chang, H. H., & Ying, Z. L. (1999). a-stratified multistage computerized adaptive testing. Applied Psychological Measurement, 23(3), 211-222.

Chang, H. H., & Ying, Z. L. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73*(3), 441-450.

Chang, H. H., & Zhang, J. M. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika*, 67(3), 387-398.

Cover, T. & Thomas, J. (1991). *Elements of Information Theory*. John Wiley & Sons, Inc.

Finkelman, M., Nering, M. L., & Roussos, L. A. (2009). A Conditional Exposure Control Method for Multidimensional Adaptive Testing. *Journal of Educational Measurement*, 46(1), 84-103.

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

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, 29(4), 262-277.

Haley, S. M., Ni, P. S., Ludlow, L. H., & Fragala-Pinkham, M. A. (2006). *Measurement precision and efficiency of multidimensional computer adaptive testing of physical functioning using the pediatric evaluation of disability inventory*. Archives of Physical Medicine and Rehabilitation, 87(9), 1223-1229.

Lehmann, E.L. (1986) *Testing Statistical Hypotheses (2nd edition)*. New York: Wiley.

Lehmann, E. L.& Casella, G. (1998). *Theory of Point Estimation (2nd ed.).* Springer

Lehmann, E.L. (1999). *Elements of large-sample theory*. New York: Springer.

Li, Y. H., & Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement*, 29(1), 3-25.

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

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

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20(4), 389-404.

Mulder, J., & van der Linden, W. J. (2009). Multidimensional Adaptive Testing with Optimal Design Criteria for Item Selection. *Psychometrika*, 74(2), 273-296.

Owen, R.J. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ, Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351–356.

Passos, V. L., Berger, M. P. F., & Tan, F. E. S. (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(1), 88-110.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9(4), 401-412.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21(1), 25-36.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York: Springer.

Reckase, M. D., & McKinley, R. L. (1991). The discrimination power of items that measure more than one dimension. *Applied Psychological Measurement*, 15(4), 361-373.

Renyi, A. (1961). *On measurs of entropy and information*. Proceeding Fourth Berkeley Symposium on Mathematical Statististics and Probability, 1, 547-561.

Renyi, A. (1970). *Probability theory*. North-Holland, Amsterdam.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61(2), 331-354.

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, 66(1), 79-97.

Stocking, M.L. (1994). *Three practical issues for modern adaptive testing item pools* (Research Rep. 94-5). Princeton, NJ: Educational Testing Service.

Stroud, A. H. & Secrest, D. (1966). *Gaussian Quadrature Formulas*. Englewood Cliffs, NJ: Prentice-Hall, 1966.

Thissen, D. & Mislevy, R.J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer*, (pp. 101–133). Hillsdale NJ: Erlbaum.

van der Linden, W. J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement*, 20, 373-388.

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

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, 24(4), 398-412.

van Rijn, P. W., Eggen, T., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 26(4), 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(2), 203-226.

Veldkamp, B. P., & van Der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, 67(4), 575-588.

Wang, C. and Chang, H. (2009). *Kullback-Leibler information in multidimensional adaptive testing: theory and application*. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Wang, C. & Chang, H. (2010, July). *Item Selection in Multidimensional Computerized Adaptive Testing----the New Application of Kullback-Leibler Information*. Paper presented at 2010 International Meeting of Psychometric Society, Athens, Georgia.

Yi, Q., & Chang, H. H. (2003). alpha-Stratified CAT design with content blocking. *British Journal of Mathematical & Statistical Psychology*, 56, 359-378.