PSYCHOMETRIKA
2014
DOI: 10.1007/s11336-014-9401-5

PSYCHOMETRICS BEHIND COMPUTERIZED ADAPTIVE TESTING

HUA-HUA CHANG

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

The paper provides a survey of 18 years' progress that my colleagues, students (both former and current) and I made in a prominent research area in Psychometrics—Computerized Adaptive Testing (CAT). We start with a historical review of the establishment of a large sample foundation for CAT. It is worth noting that the asymptotic results were derived under the framework of Martingale Theory, a very theoretical perspective of Probability Theory, which may seem unrelated to educational and psychological testing. In addition, we address a number of issues that emerged from large scale implementation and show that how theoretical works can be helpful to solve the problems. Finally, we propose that CAT technology can be very useful to support individualized instruction on a mass scale. We show that even paper and pencil based tests can be made adaptive to support classroom teaching.

Key words: computerized adaptive testing, multidimensional CAT, sequential design, martingale theory, a-stratified item selection, response time, constraint management, CD-CAT.

1. Introduction

Over the past 20 years, Computerized Adaptive Testing (CAT) has become an increasingly important testing mode in large scale educational assessment. An adaptive test differs greatly from a linear test. In the former, items are selected sequentially, according to the current performance of an examinee. In the latter, examinees are tested with a preassembled set of items. A CAT is typically engineered to tailor the test to each examinee's trait level, thus matching the difficulties of the items to the examinee being measured. So, the examinees are always challenged during the entire course of the testing. The major advantage of CAT is that it provides more efficient latent trait estimates (θ) with fewer items than would be required in conventional tests (e.g., Weiss, 1982).

A growing body of evidence shows that CAT can help with classroom assessment and also facilitate individualized learning. CAT is revolutionary in this context because it is changing the way we address challenges in assessment and learning. CAT has already had a substantial influence on the functioning of society by affecting how people are selected, classified, and diagnosed.

This paper is a summary from the presidential address I made at the 78th Annual Meeting of the Psychometric Society, July 23–26, 2013, Arnhem, the Netherlands. As the president of the society, it is my great honor to share my view concerning some current and future developments in Psychometrics. In this regard, I would like to address a number of issues that were inspired by my work with my colleagues and students on Computerized Adaptive Testing. This paper presents (i) a review of item selection algorithms from Robbins–Monro to Fred Lord; (ii) the establishment of a large sample foundation for Fred Lord's maximum information based item selection methods; (iii) some issues that emerged from large scale implementation and the remedies to solve the problems; (iv) further developments of CAT, including multidimensional CAT, response time in CAT, constraint control in CAT, cognitive diagnostic CAT, multistage testing, and CAT in adaptive learning; and (v) a brief discussion.

This article is based on the Presidential Address Hua-Hua Chang gave on June 25, 2013 at the 78th Annual Meeting of the Psychometric Society held in Arnhem, the Netherlands.

Requests for reprints should be sent to Hua-Hua Chang, University of Illinois at Urbana-Champaign, 430 Psychology Building, 630 E. Daniel Street, M/C 716, Champaign, IL 61820, USA. E-mail: hhchang@illinois.edu

Published online: 06 February 2014

2. Establish a Large Sample Foundation for Computerized Adaptive Testing

2.1. CAT—From Robbins–Monro to Fred Lord

The most important element in CAT is the item selection procedure that is used to select items during the course of the test. According to Fred Lord (1970), an examinee is measured most effectively when test items are neither too difficult nor too easy. Heuristically, if the examinee answers an item correctly, the next item selected should be more difficult; if the answer is incorrect, the next item should be easier. To carry out the branching rule, we need to know the difficulty levels for all items in the pool. Let $b_1, b_2, \ldots, b_n$ be a sequence of the difficulty parameters in the Rasch model for the n items already administered to an examinee. Intuitively, the next item should be selected such that b_{n+1} be close to a point of interest b_0, where b_0 represents the difficulty level of an item that the examinee has about a 50 % chance to answer correctly. If b_n convergences to b_0 as $n \to \infty$, b_0 is a reasonable guess for the true θ, and thus we can characterize $\hat{\theta} = b_0$. This process is called a Robbins and Monro (1951) process, and Fred Lord is the first one who introduced the Robbins–Monro process to application in adaptive testing.

Among several procedures Lord (1970) proposed as applications of the Robbins–Monro process, the sequential design of updating the b-parameters works the most promising: $b_{n+1} = b_n + d_n(x_n - m)$, where $x_n = 1$ if the answer is correct, $x_n = 0$ otherwise; $d_1, d_2, \ldots$ is a decreasing sequence of positive numbers chosen before the testing; and m is a point of interest (denoting an examinee's unknown ability), say, $m = 1/2$. Assume the item pool is so rich that we can select any b-value from the range of $(-\infty, +\infty)$. If d_1 is not too small, according to Hodges and Lehmann (1956), the sequence of d can be chosen as $d_i = d_1/i$, $i = 1, 2, 3, \ldots$. Interestingly, the convergence of b_n to $b_0 = m$ does not require strong assumptions such as local independence or the exact shapes of the item characteristic curves to be known.

To make adaptive testing operational, the size of the item pool must be large enough to cover all possible values of b. Moreover, it may need many items for $\hat{\theta}_n$ to be close to θ. Another problem frequently encountered in CAT design is efficiency in addition to consistency. We would like to know whether our estimate has the smallest variance among other consistent estimates, and hence it is needed to include more information in the adaptive design, such as the exact shapes of the item response functions and Fisher information functions as well.

For conventional paper-and-pencil tests, it is well known that, under suitable regularity conditions, $\hat{\theta}_n$ is consistent and asymptotically normal, centered at the true θ and with variance approximated by $I^{-1}(\hat{\theta}_n)$, where $I(\theta)$ is the Fisher test information function. Under the local independence condition, an important feature of $I(\theta)$ is that the contribution of each item to the total information is additive: $I(\theta) = \sum_{j=1}^{n} I_j(\theta)$, where $I_j(\theta)$ is Fisher item information for item j. Thus, under the local independence assumption, the total amount of information for a test can be readily determined. This feature is highly desirable in CATs because it enables test developers to separately calculate the information for each item and combine them to form updated test information at each stage. To make the sample variance of $\hat{\theta}_n$ small, we can sequentially select n items so that their information at $\hat{\theta}_j$, $j = 1, 2, \ldots, n$, are as large as possible.

To make $\hat{\theta}_n$ the most efficient, Lord (1980) proposed a standard approach for item selection in CAT, which is to select the item with the maximum Fisher item information as the next item. Such efficiency can be achieved by recursively estimating θ with current available data and assigning further items adaptively. Note that in Item Response Theory (IRT), the large sample properties of $\hat{\theta}_n$ were established under the local independence assumption. In adaptive design, the selection of the next item is dependent on the basis of the examinee's responses to the items previously administered (Mislevy & Chang, 2000). Since the maximum information method has been the most popular item selection method for more than three decades, it is essential to establish a large sample foundation for the approach.

2.2. *Martingale and CAT*

Establishing the limiting distribution of $\hat{\theta}$ is important in advancing CAT research. When analyzing independent random variables, we are so lucky to have plentiful theoretical tools for getting large sample results, such as the central limit theorem, law of large numbers, Chebyshev's inequality, etc. However, in CAT, the large sample property of $\hat{\theta}$ should be derived from a sequence of dependent random variables because the kth item is chosen according to the preceding $k-1$ responses $X_1, \ldots, X_{k-1}$, and hence $X_1, \ldots, X_k$ are generally dependent. Yet, such dependence is formed in a probabilistically interesting way. Note that $\{X_1, \ldots, X_k\}$, which represents the realized sequence at a particular time k, contains the previous sequence $\{X_1, \ldots, X_{k-1}\}$.

When analyzing a sequence of recursive estimates with certain dependency, martingale theory can often be used. It is remarkable that most of the original developments of martingale theory were established by an Illinois professor Joseph Leo Doob after he came to Champaign in 1935. Being a graduate student at Illinois from 1985 to 1991, I remember seeing Dr. Doob so often sitting in the back very attentively at the weekly statistical seminars. Like all the students and faculty in the department of statistics I admire Doob so much for his eminent contribution to the field of probability theory; in particular, martingale theory. However, back then martingale theory was too remote to me. I studied it because it was required by the qualifying examination, and I never thought that one day we would use martingale to solve problems in psychometrics. Thanks to late Professor Walter Philipp for his excellent teaching of Probability and Measure, a two semester series I took from him back in 1988–1989 at Illinois. I should also thank Xuming He, currently a distinguished statistics professor at University of Michigan, and Zhiliang Ying for their generous help whenever I needed clarifications for difficult contents. Now martingale theory has become known to practitioners for its great contribution to the solution of the CAT problem that emerged from educational testing which people never thought they are connected to each other.

To show a connection between CAT and martingale, first, let $X_1, \ldots, X_n$ be a sequence of dependent random variables with joint density $f(x_1, \ldots, x_n)$. Let $f(x_k|x_1, \ldots, x_{k-1})$ be the conditional density of X_k given $X_1, \ldots, X_{k-1}$. Then

$$f(x_1, \ldots, x_n) = \prod_{k=1}^{n} f(x_k|x_1, \ldots, x_{k-1}).$$

Let's consider the null hypothesis $H_0 : f = f_0$ versus the alternative $H_1 : f = f_1$. Let L_n be the likelihood ratio

$$L_n = \frac{f_1(X_1, \ldots, X_n)}{f_0(X_1, \ldots, X_n)}.$$

Then, it can be verified that under H_0, L_n is a martingale with respect to σ-filtration $\mathcal{F}_n$,

$$E\{L_n|X_1, \ldots, X_{n-1}\} = L_{n-1} \tag{1}$$

where

$$\mathcal{F}_n = \sigma\{X_1, \ldots, X_n\}. \tag{2}$$

It is interesting to mention that in most probability textbooks martingale is usually introduced by a gambling example describing a fair game. In Equation (1), if L_{n-1} represents a gambler's current fortune and $\{X_1, \ldots, X_{n-1}\}$ is a σ-field containing all the information about the game after the $(n-1)$th play, the equation states that the expected fortune after the next play is the same as the present fortune, which implies that the fortunes of the gambler and the house are equally

weighted. Martingales exclude the possibility of winning strategies based on game history, and thus they are a model of fair games.

Now we can use the theory to establish the consistency of MLE in CAT. The reason for Equation (1) to hold is that $f(x_1,\ldots,x_n) = f(x_1,\ldots,x_{n-1})f(x_n|x_1,\ldots,x_{n-1})$ implies

$$L_n = L_{n-1}\frac{f_1(X_n|X_1,\ldots,X_{n-1})}{f_0(X_n|X_1,\ldots,X_{n-1})}, \tag{3}$$

and

$$E\left\{\frac{f_1(X_n|X_1,\ldots,X_{n-1})}{f_0(X_n|X_1,\ldots,X_{n-1})}\middle| X_1,\ldots,X_{n-1}\right\} = 1. \tag{4}$$

These results are fundamental for the establishment of a CAT asymptotic theory. For a given martingale, if it has an upper or a lower bound, then the martingale must converge (a.s.). Since the likelihood is always nonnegative, 0 is a lower bound. Therefore, L_n converges to a limit. It can be shown under very mild conditions that this limit must be 0. Note that this is under H_0, i.e., f_0 is the true density function. An important circumstance about the implication of $L_n \to 0$ should also be explained. Since L_n is the ratio of f_1 over f_0, it means that f_0 must eventually become much larger than f_1. In other words, if the maximum likelihood is used to select f_0 versus f_1, f_0 will be selected, which represents the truth. We can extend this result to a finite number of possible densities: $f_0, f_1, \ldots, f_m$. Then the maximum likelihood estimate (MLE) always consistently picks the true density f_0.

In a CAT $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$ are recursively estimated along with a sequential design. Define the σ-filtration as that in (2). Now clearly, $\mathcal{F}_k \subset \mathcal{F}_{k+1}$. Zhiliang Ying and I (e.g., see Chang & Ying, 2009) demonstrated that the consistency and asymptotically normality of CAT-MLE can be established under mild regularity conditions. Taking the two-parameter logistic (2PL) model as an example, we showed that $\{X_n - P_n(\theta)\}$ is a martingale difference sequence and

$$S_n = \sum_{j=1}^{n} a_j\left\{X_j - \frac{e^{a_j(\theta-b_j)}}{1+e^{a_j(\theta-b_j)}}\right\} \tag{5}$$

is a martingale for the given σ-filtration defined in (2). Note that $S_n = 0$ is the likelihood equation for solving the nth MLE $\hat{\theta}_n$. Under very general and non-restrictive assumptions, we proved (see Chang & Ying, 2009)

$$\sqrt{\sum_{i=1}^{n} a_i^2}(\hat{\theta}_n - \theta) \xrightarrow{L} N(0,1) \tag{6}$$

as $n \to \infty$, where θ is the true ability and the normalized factor $\sqrt{\sum_{i=1}^{n} a_i^2}$ may be replaced by $\sqrt{I^{(n)}(\hat{\theta}_n)}$ or $\sqrt{I^{(n)}(\theta)}$, where $I^{(n)}$ is the observed Fisher information. Also, the results can be generalized to other models.

3. Developing New Item Selection Algorithms

Before building the asymptotic theory for Fred Lord's maximum information approach, we should ask ourselves how such a theoretical result can add information beyond the fact that many practitioners had already demonstrated that $\hat{\theta}_n$ converges to true θ by conducting Monte Carlo experimentations. Our ultimate goal to study the asymptotic problems is to bring rigorous statistical methods into applied research and develop new algorithms which are more effective and

FIGURE 1.
Highly discriminating items are like a tightly focused spotlight that shines intensely but casts little light outside a narrow bean. However, they can be more effectively used in the later stages of the test after the examinee has been roughly located.

user-friendly to practitioners. While deriving the proof of (6), we noticed that the range of the a-parameters must be bounded and also, $\hat{\theta}_n$ might be divergent if the sequence of $\{a_1, a_2, \ldots, a_n\}$ were not appropriately chosen. The maximum Fisher information method tries to find an item whose difficulty is close to the examinee's estimated proficiency and has a steep item characteristic curve. We noticed two limitations with this approach. First, it meant that sharply discriminating items were always chosen first, and indeed, left many items of appropriate difficulty but lesser discriminating ability only rarely, if ever, used. Since item development costs average more than $1,000 an item (e.g., see Downing, 2006) this was a waste and made it easier to effectively cheat on an exam (because the effective size of the item pool was severely shrunken). And second, the algorithm performed poorly at the beginning of the exam, when the examinee's proficiency was badly estimated.

The a-stratified method (Chang & Ying, 1999) was proposed which uses less discriminating items early in the test, when estimation is the least precise, and saves highly discriminating items until later stages, when finer gradations of estimation are required. One of the advantages of using the a-stratified method is that it attempts to equalize the item exposure rates for all the items in the pool. The a-stratified design has received positive remarks from many researchers. For example, Davey and Nering (2002, p. 181) wrote the following:

> "Highly discriminating items are like a tightly focused spotlight that shines intensely but casts little light outside a narrow bean. Less discriminating items are more like floodlights that illuminate a wide area but not too brightly. The idea is to use the floodlights early on to search out and roughly locate the examinee, then switch to spotlights to inspect things more closely."

Their vivid description regarding when to use high discriminating items and when to use low discriminating items could by be imaginatively illustrated by Figures 1 and 2 (work by Wen Zeng, a current graduate student specialized in educational measurement at the University of Wisconsin at Milwaukee.)

FIGURE 2.
Less discriminating items are more like floodlights that illuminate a wide area but not too brightly. However, one can turn on the floodlight early on to approximately locate the examinee, and then switch to spotlights to inspect things more closely (See Figure 1.)

Clearly, the a-stratified method in Chang and Ying (1999) is a simplified version which did not address such issues as what the best set of stratification properties might be for a specific item pool and population distribution. Further research has taken place and yielded numerous refinements. Chang, Qian, and Ying (2001) have developed the a-stratified design with b-blocking to overcome the problem in a situation where a-parameters and b-parameters are positively correlated. The basic idea is to force each stratum to have a balanced distribution of b values to ensure a good match of θ for different examinees. Chang and van der Linden (2003) and van der Linden and Chang (2003) proposed using 0–1 mathematical programming models, together with the a-stratified method, to balance contents and improve accuracy. Yi and Chang (2003) and Leung, Chang, and Hau (2003) proposed solutions to incorporating the ability to handle item content.

The a-stratified item selection method exemplifies a very major conceptual advancement in CAT, which demonstrated that the usual practice in CAT of choosing items with higher discrimination power at earlier stages of testing was inappropriate. This is because at earlier stages of testing we have only a vague idea of the examinee's proficiency; hence, instead of utilizing items with high discrimination, we should be using items with low discrimination. This work had immediate pay-off—not only by rescuing items that previously went unused, but also by improving the efficiency of proficiency estimation.

Hau and Chang (2001) asked whether an a-stratified item selection method should select items according to ascending a-values or descending a-values. In this regard, it is interesting to notice their conclusion that the item pool stratification strategy based on ascending a-values yields clear benefits comparing with that based on descending a-values. Further, Zhiliang Ying and I (Chang & Ying, 2008) analytically demonstrated that items with high a-parameters tend to cause "big jumps" for ability estimators at the very early stage of the test. Our simulation study revealed that the ascending-a methodology is essential to overcoming the underestimation problem among GRE and GMAT takers widely reported in early 2000 (e.g., see Carlson, 2000).

For more issues in early days' CAT implementations, see Chang (2004) and Chang and Ying (2007).

3.1. KL Information

I have known Zhiliang Ying since 1987 when I was a graduate student at UIUC. At that time, Ying had just joined the Department of Statistics as an assistant professor and the department asked me to introduce him to the area. Later, I became his driving coach and one of his best friends. As such, I felt comfortable asking Ying for his help in learning asymptotic theory. Our academic relationship soon proved to be mutually beneficial, as I helped to reinforce Ying's interest in psychometric research. Later the collaboration with Zhiliang Ying has generated numerous manuscripts. One of the best is about using Kullback–Leibler (KL) information in CAT (Chang & Ying, 1996), which is actually related to my thesis. In 1990, I was persuaded by my thesis advisor William Stout, who later served as President of the Psychometric Society in 2002, to work on a research question posed by Paul Holland (1990) about the establishment of the asymptotic posterior normality of given examinee response pattern under any of a very large class of IRT models and unrestrictive regularity assumptions (see Chang & Stout, 1993). Stout also encouraged me to invite Ying as a member on my preliminary examination committee, and I am glad I did.

While working on the asymptotic posterior normality problem, I found that it was essential to prove the weak and strong consistency of MLE in IRT. Under the guidance of Stout and Ying, I soon realized that a sufficient condition for MLE convergence needs to be brought in:

$$\limsup_{n\to\infty} n^{-1} \sum_{j=1}^{n} E_{\theta_0} Z_j(\theta) \le -c(\theta), \tag{7}$$

where $E_{\theta_0} Z_j(\theta) = -K_j(\theta \| \theta_0)$, and $c(\theta)$ is a positive finite number that may depend on θ and j. It is worth noting that K_j is the KL information at an item level. Condition (7) implies that, if the n items are selected in a way such that the expected value of the log-likelihood ratio

$$K^{(n)}(\theta \| \theta_0) = E_{\theta_0} \log\left\{ \frac{P(X_1, \ldots, X_n|\theta_0)}{P(X_1, \ldots, X_n|\theta)} \right\}$$

generates enough discrimination power for any $\theta \neq \theta_0$, then θ_0 can be identified, i.e., the MLE converges to θ_0. Of particular importance, $K^{(n)}$ defined above is the KL information at a test level. That was my first experience to work with KL information. Several years later when I started CAT research at Educational Testing Service (ETS), it became clear to me that we should find a way to use KL information to build a new item selection algorithm.

KL information (distance or divergence), introduced in 1951 by Solomon Kullback and Richard Leibler, is viewed as the second most important concept in the information theory that is of statistical nature. Their motivation for this concept partly arises from "the statistical problem of discrimination", as they put it. They believed that statistically "two populations differ more or less according as to how difficult it is to discriminate between them with the best test". Indeed, the KL distance gives precisely the asymptotic power for the likelihood ratio test, which is the best (most powerful) according to the Neyman–Pearson theory. In a CAT, our statistical problem is to test the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta = \theta_1$, where θ_1 can be viewed an estimate of θ_0. From the Neyman–Pearson theory, it is well known that

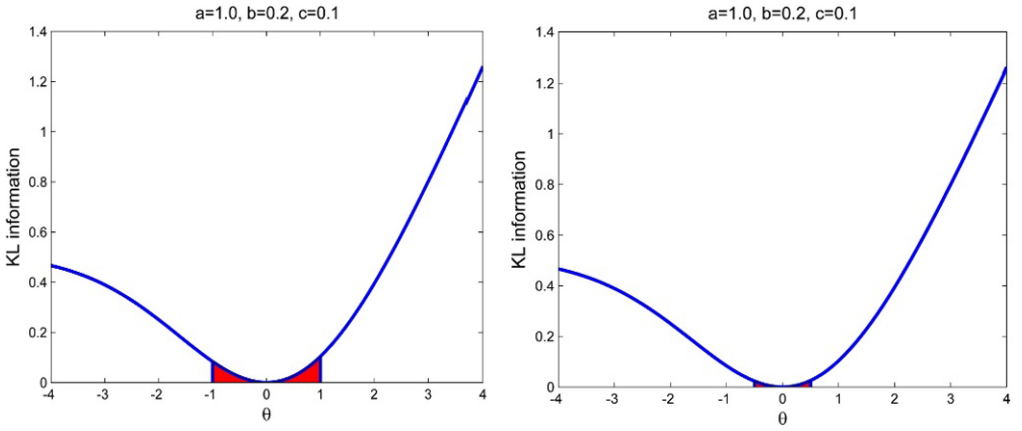

FIGURE 3.
Integration of information along a theta region.

the optimal test is the likelihood ratio based test: rejecting H_0 in favor of H_1 if and only if the likelihood ratio statistic

$$LR = \frac{P\{X_1, \ldots, X_n|\theta_1\}}{P\{X_1, \ldots, X_n|\theta_0\}}$$

exceeds a certain threshold.

In the 1996 paper, we discussed the statistical imperfection behind the maximum Fisher information approach. We argued that the Fisher information should not be used at the beginning of the test when $\hat{\theta}_n$ is likely not close to its destination θ_0 and suggested that the KL information be used instead. As the test moves along, when there are enough items in the test to ensure that $\hat{\theta}_n$ is close enough to θ_0, the Fisher information should then be used. We proposed a KL information index (called KL index or KI hereafter), defined as

$$KI_i(\hat{\theta}_n) = \int_{\hat{\theta}_n - \delta_n}^{\hat{\theta}_n + \delta_n} KL_i(\hat{\theta}_n, \theta) d\theta, \tag{8}$$

where $\delta_n \to 0$ as $n \to \infty$. Since the second derivative of KL_j evaluated at $\hat{\theta}_n$ equals the Fisher item information at $\hat{\theta}_n$, $I_j(\hat{\theta}_n)$ reflects the curvature of KL_j at $\hat{\theta}_n$. Geometrically, the KL index is simply the area below the KL information curve, bounded by the integration interval, as shown in Figure 3. It is obvious that when δ_n is small, the area defined on right hand side of (8) is determined by the curvature of $K_j(\theta\|\hat{\theta}_n)$ at $\hat{\theta}_n$, and thus, it is type of local information. For δ_n large, the area is also determined by the tails, and thus, it is type of global information.

However, for δ_n with an intermediate value, the area is determined by both global and local information. In a CAT, items with the maximum KL index value should be selected, which summarizes the discrimination power of an item in differentiating $\hat{\theta}$ from all its neighborhood levels. In real adaptive testing, examinee's true θ_0 is unknown and $\hat{\theta}$ is used as a point estimate; the integration interval is large at the beginning of the test when n is small so as to cover θ_0 that might be far away from $\hat{\theta}$; toward the end of the test when n is large, the interval shrinks and when n approximates infinity, maximizing the KL index is the same as maximizing the item Fisher information at $\hat{\theta}$. Simulation studies have shown that the KL index method provides more robust item selection especially when test length is short. For instance, Chang and Ying (1996, 2008) have demonstrated that by selecting globally discriminative items (i.e., items with maximum KL index) at the beginning of the test, examinees' latent traits could be more accurately

recovered, even after the first few items were answered incorrectly. In contrast, maximum Fisher information item selection would yield severely biased θ estimates.

4. Further Developments

4.1. Multidimensional CAT

Multidimensional Item Response Theory (MIRT) deals with the situations in which multiple skills or attributes are being tested simultaneously. It has gained much recent attention, for instance, many certification and admission boards are interested in combining regular tests with diagnostic services to allow candidates obtain more informative diagnostic profiles of their abilities (Mulder & van der Linden, 2009). A number of MIRT models have been proposed. Depending upon how different dimensions interact to produce a correct response, a distinction was made between compensatory and non-compensatory models. The discussion in this section will be based on compensatory MIRT models, with the item response function taking the form of (Reckase, 2009)

$$P_i(\boldsymbol{\theta}) \equiv \text{Prob}(u_i = 1|\boldsymbol{\theta}) = c_i + \frac{1 - c_i}{1 + \exp[-(\boldsymbol{a}_i^T \boldsymbol{\theta} - b_i)]},$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ 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.

Multidimensional computerized adaptive testing (MCAT) combines CAT and multidimensional trait diagnosis and so offers the advantage of both. But it also introduces more complexity in item selection, interim scoring, and item pool management. Therefore, it is highly desirable to develop more efficient item selection methods that can be immediately applied in large scale implementation. Segall (1996, 2001) generalized the maximum Fisher information method to multidimensional models and proposed to select items that maximized the determinant of Fisher test information matrix. His method was later coined as "D-optimality" by Mulder and van der Linden (2009). The D-optimality method is shown to produce smallest generalized variance of $\hat{\boldsymbol{\theta}}$ because the determinant of the Fisher information matrix is inversely related to the volume of the confidence ellipsoid around $\hat{\boldsymbol{\theta}}$. However, the D-optimality method also exhibits the same limitations as the maximum Fisher information method in one-dimensional case because it is based on local information. As a remedy, Veldkamp and van der Linden (2002) proposed to use KL information in MCAT (named as KL index) as a weighted multiple integration of the KL information over a multivariate space (as seen in Equation (9)), but it may not be easy to apply in practice due to the computational complexity:

$$KI\left(\hat{\boldsymbol{\theta}}^{k-1}\right) = \int_{\hat{\boldsymbol{\theta}}_1 - \frac{d}{n^{1/2}}}^{\hat{\boldsymbol{\theta}}_1 + \frac{d}{n^{1/2}}} \cdots \int_{\hat{\boldsymbol{\theta}}_1 - \frac{d}{n^{1/2}}}^{\hat{\boldsymbol{\theta}}_p + \frac{d}{n^{1/2}}} KL_i\left(\boldsymbol{\theta} \| \hat{\boldsymbol{\theta}}^{k-1}\right) P(\boldsymbol{\theta}|\boldsymbol{X})\, \partial\boldsymbol{\theta} \tag{9}$$

where $\boldsymbol{X}$ denotes the vector of responses for a particular examinee, $\hat{\boldsymbol{\theta}}^{k-1}$ is the intermediate ability estimate for that examinee after $(k-1)$ items, and d is a constant that is usually selected to be 3. Figure 4 illustrates the KL index in a two-dimensional case, where it represents the volume under the KL information surface enclosed by a square region.

Chun Wang and I derived two theorems to quantify the relationships between the KL index and item parameters. The first theorem relates the magnitude of KL index with the item discrimination parameters (see Wang, Chang, & Boughton, 2011a, 2011b).

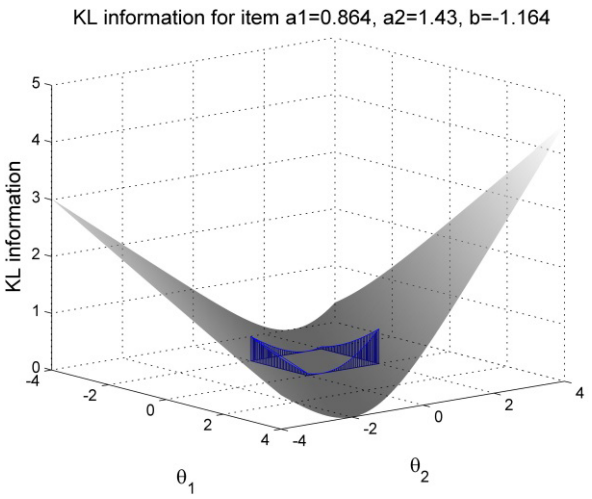

FIGURE 4.
Illustration of KL information in two-dimensional case for a single item.

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 $KI(\boldsymbol{\theta}_0) = \int\int_D KL_j(\boldsymbol{\theta}\|\boldsymbol{\theta}_0)\,\partial\boldsymbol{\theta}$, 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, $f(\underset{\sim}{a}) = a_1^2 + a_2^2$ when D is a square or a circle, and $f(\underset{\sim}{a}) = (a_1 r_1)^2 + (a_2 r_2)^2$ when D is a rectangle or an ellipse.*

We show that when the two dimensions are given equal weights, 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 $a_1^2 + a_2^2$ (also known as the square of the multidimensional discrimination, Reckase & McKinley, 1991) are more likely to be chosen. This conclusion was further generalized to a p-dimensional case ($p \geq 2$) where we had (Wang & Chang, 2011)

$$KI \approx \frac{2^{p-1}}{3}\left(\frac{r}{\sqrt{n}}\right)^{p+2} \frac{Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i]^2}{P_i(\boldsymbol{\theta})(1 - c_i)} \sum_i a_i^2.$$

Wang further proved that items with high discrimination power on all dimensions are preferred. This is equivalent to that maximizing the KL index in a limiting case is the same as maximizing the trace of the Fisher information matrix. The second theorem relates the size of the KL index with the item difficulty parameter as follows.

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 jth item is denoted as $KL_j(\boldsymbol{\theta}\|\boldsymbol{\theta}_0)$. If we define the item KL information index as $KI(\boldsymbol{\theta}_0) = \int\int_D KL_j(\boldsymbol{\theta}\|\boldsymbol{\theta}_0)\,\partial\boldsymbol{\theta}$, 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 $c = 0$.*

Figure 5 illustrates the size of the KL index as a function of θ_1 and θ_2 for a given item. It is clearly shown that the KL index is maximized along a line formed by this linear equation of $\boldsymbol{a}'\boldsymbol{\theta}_0 - \boldsymbol{b} = 0$. Such a conclusion is beneficial in adaptive testing because when the interim

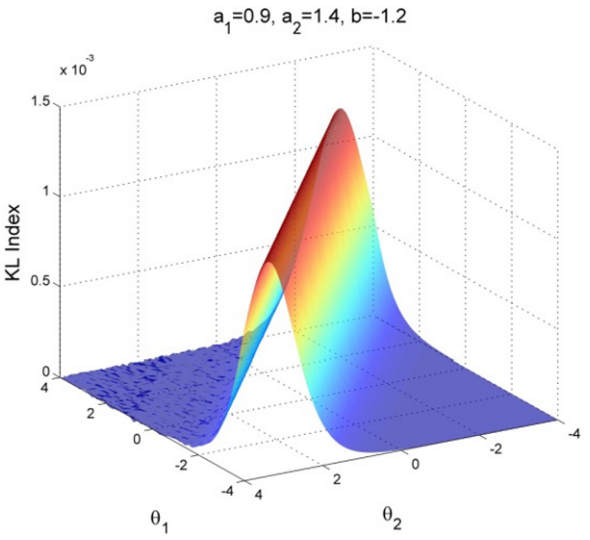

FIGURE 5.
Illustration of KL information index in two-dimensional case for a single item.

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 one-dimensional CAT (Chang & Ying, 1999).

Wang's work may also be helpful to identifying over-used and under-used items, and thus it may provide guidelines for item pool management and replenishment. Inspired by Wang's finding, we propose a simplified KL index (SKI), which is computationally more efficient without sacrificing the accuracy (Wang et al., 2011a, 2011b). An example for a two-dimensional case is given in the following:

$$SKI = \left(a_1^2 + a_2^2\right) \frac{1}{|\boldsymbol{a}'\boldsymbol{\theta} - b|} \exp\left\{-\left[\frac{\text{rank}(a_1^2 + a_2^2)}{N/L} - 1\right]^2\right\}.$$

The Wang and Chang paper then proposed another new method called KL information with Bayesian update (KLB), which calculates the KL distance between two subsequent posterior distributions of ability, indicating the amount of useful information carried by the candidate item. KLB offers a new way of utilizing the KL information and outperforms the original KL index regarding both estimation accuracy and item pool usage via a series simulation studies (Wang & Chang, 2011). Most recently Wang, Chang, and Boughton (2013b) extended the fixed-length MCAT to a variable-length version, a generalization that guarantees examinees of different ability levels are estimated with approximately equal precision and therefore enhances test fairness.

4.2. Modeling Response Time in Computerized Testing

Response time (RT) has been a popular dependent variable in cognitive psychology since the mid-1950s because the amount of time that it takes a person to respond to a stimulus is believed to indicate some aspects about the underlying cognitive process. In educational measurement, RT also provides valuable information. For instance, analysis of the time spent on the items in a basic mathematics test may provide insights into test-takers' learning processes, solution strategies, the quality of the test-takers' knowledge, and the cognitive demands of the items. One advantage of computerized testing in educational assessment is that it makes possible the collection of an examinee's response time.

Analyzing RT along with other dependent variables, such as response accuracy, requires sophisticated statistical models. The most commonly used models are parametric models, such as the exponential model, the gamma model, the Weibull model, and the lognormal model. These models differ in their assumptions about the response time distribution and the relation between response time and ability, and in the type of test item; therefore, selection of an appropriate model for a real data set can be problematical. We proposed two types of hierarchical semi-parametric models, one derived from the Cox proportional hazard model (Wang, Fan, Chang, & Douglas, 2013c) and the other building on the linear transformation model (Wang, Chang, & Douglas, 2013a).

To be more specific, Klein Entink, van der Linden, and Fox (2009) summarized three different approaches that have been taken in the past to model RTs. The first approach models response times exclusively, so that it is mainly applicable to speed tests which have strict time limits. Such models include the exponential model (Scheiblechner, 1979), the Weibull model (Rouder, Sun, Speckman, Lu, & Zhou, 2003), and the gamma model (Maris, 1993), among others. The second approach focuses on separate analysis of RTs and response accuracy. This approach assumes that RTs and responses vary independently, which might not be true in educational measurement. The third approach advocates joint modeling of both RTs and responses, and such models include those of Thissen (1983), van der Linden (1999), Roskam (1997), Wang and Hanson (2005), among others.

Our lab's contribution is built upon van der Linden (2007)'s hierarchical model, in which RT and responses are modeled separately at the measurement model level; and at a higher level, a population model for the person parameters (speed and ability) is constructed to account for the correlation between them. For instance, at the first level, responses are assumed to follow a three-parameter logistic (3PL) model, whereas for response times, a lognormal model with separate person and item parameters was adopted (van der Linden, 2006),

$$T_{ij} \sim f(t_{ij};\tau_j,\alpha_i,\beta_i) \equiv \frac{\alpha_i}{t_{nj}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i\left(\ln t_{ij} - (\beta_i - \tau_j)\right)\right]^2\right\} \tag{10}$$

where τ_j is the latent speed of examinee j and β_i is the time intensity of item i. α_i serves as the discrimination parameter for item i. At the second level, examinees' latent traits (θ_j, τ_j) is assumed to be randomly drawn from a bivariate normal distribution, with a certain mean vector and a covariance matrix.

Although lognormal distribution seems to show adequate fit to most item response time distributions collected from the testing data, it is still possible that item RT distributions will differ dramatically from one item to another, which calls for the need of a flexible model that relaxes such distributional assumptions (Ranger & Kuhn, 2011). Most recently, our lab proposed to replace the lognormal model in (10) with either a Cox proportional hazard model taking the following form

$$h_{ij}(t|\tau_i) = h_{0j}(t)\exp(\beta_j\tau_i), \tag{11}$$

or a semi-parametric linear transformation model with the following regression form

$$H_i(t_{ij}) = \beta_i\tau_j + \varepsilon_{ij}. \tag{12}$$

In both (11) and (12), τ_j is the latent speed of examinee j and β_i is the item-level slope parameter. In (11), h_{ij} is the hazard function[1] for item i and person j, and $h_{0j}(t)$ is the non-parametric

[1]The hazard function is the instantaneous rate at which events occur. In psychological terms, the hazard rate is the conditional probability of finishing the task in the next moment, which is therefore, also viewed as the processing capacity of an individual.

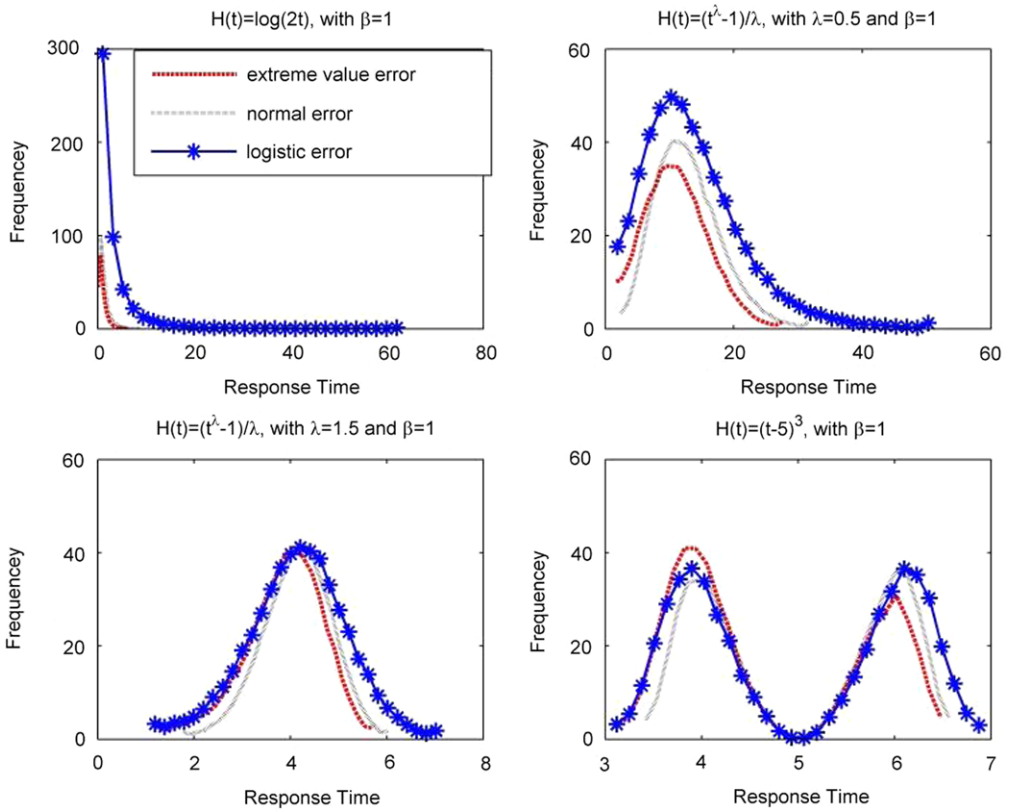

FIGURE 6.
The possible RT distributions from different combinations of error term distribution and transformations.

baseline hazard for item j. In (12), H_i represents a monotone but non-parametric transformation for item i. The error term, ε_{ij}, is independent and identically distributed with a certain known distribution F, the latter of which does not necessarily have to follow a normal distribution. When the error term follows extreme value distribution, model (12) reduces to a re-parameterized version of (11). In both models, it is the non-parametric term (either the baseline hazard in (11) or monotone transformation in (12) that makes the model more flexible and versatile. Figure 6 illustrates the possible RTs distributions given different combinations of the error term distributions and forms of the continuous transformations. Clearly, lognormal distribution assumption in van der Linden (2007) might sometimes be too restrictive to account for a variety of possible RT distributions, such as panel (4) in Figure 6. In this example, the RTs distribution follow a bimodal distribution, indicating that there are possibly two solution strategies for this given item, with one solution taking a shorter time and the other one taking a longer time.

We demonstrated that almost all existing parametric models can be viewed as special cases of the semi-parametric models by fixing either the baseline hazard or the error-term distributions. The inclusiveness of the new models adds flexibility to the model-fitting of real data sets. The new models also relax distributional assumptions, useful when response-time distributions differ dramatically across items within a test, as is often the case (Ranger & Kuhn, 2011). We developed two-stage model-estimation methods for both new models and provided indices for model fit checking.

In addition to the psychometric models for response times, my colleagues and I also worked on developing efficient item selection algorithms utilizing response time as a source of informa-

tion. Note that the traditional CAT item selection methods would only depend on item information without taking into account the amount of time required to answer each item. As a result, some examinees may receive a set of items that take too long to finish, making item selection less efficiently. Fan, Wang, Chang, and Douglas (2012) proposed two item-selection criteria—the first modifies the maximum information criterion to maximize information per time unit, and the second is an inverse time-weighted version of α-stratification that achieves more balanced item exposure than the information-based techniques. Simulation studies have shown that the two new methods were able to generate accurate ability estimation within shorter time limit, thereby yielding a more efficient test.

4.3. Constraint Control in CAT

In a CAT, items can be selected by optimizing the statistical property of the latent trait estimate, so that the CAT can provide more efficient trait estimates with fewer items than traditional non-adaptive tests. Test development, however, is a complicated process where statistical properties are not the only concern. For instance, for a general math test involving both algebra and trigonometry questions, we need to ensure that both content areas are adequately covered. An item selection algorithm taking into account item difficulty alone may result in tests that do not have content validity.

In real testing programs, dozens, even hundreds of such constraints exist. This makes item selection a complex constrained optimization problem. Well-established methods such as 0–1 integer programming may be unwieldy in CAT due to the fact that the items are sequentially and adaptively chosen for each examinee. Cheng, Chang and Yi (2007) specifically deals with content balancing in CAT; Cheng and Chang (2009) proposed the maximum priority index (MPI) method that can handle multiple constraints simultaneously. The example shown in this paper included 36 constraints, and the MPI method was able to meet them all in CAT. Cheng, Chang, Douglas, and Guo (2009) further combined the MPI method with the stratification design for exposure control purpose. Notably, the concept of content balancing has also been applied to multidimensional CAT, e.g., a CAT built on the bi-factor model (Zheng, Chang, & Chang, 2013) in the context of patient-reported outcome research.

4.4. CD-CAT

When a CAT is built on latent class models instead of latent trait models, the algorithms need to be adapted. Cognitive diagnostic models are constrained latent class models that attempt to identify examinees' latent cognitive profiles, or skill mastery patterns. Each pattern is a latent class. The traditional item selection algorithms developed for IRT models do not apply to latent class models. Xu, Chang, and Douglas (2003) proposed two algorithms for cognitive diagnostic computerized adaptive testing or CD-CAT, one based on the Kullback–Leibler information and the other based on the expected predictive Shannon entropy, both well-defined on latent classes. McGlohen and Chang (2008) further incorporated constraints into item selection in CD-CAT by using the shadow-test approach. Cheng (2009) further improved the algorithms by Xu et al. (2003) and proposed the posterior-weighted Kullback–Leibler information method and the hybrid Kullback–Leibler information method. To ensure adequate coverage of all attributes/skills, Cheng (2010) imposed constraints on the number of times each attribute is represented in the CD-CAT. Wang, Chang and Douglas (2012) then discussed algorithms that serve dual purposes, i.e., estimating both the latent trait and knowledge profile (latent classes) efficiently and accurately. Wang (2013) further proposed the mutual information method for item selection in a CD-CAT and notably a computationally easier formula to facilitate real-time application of this method.

Figure 7 shows a picture taken from a large scale CD-CAT assessment in China in 2012. Liu, You, Wang, Ding, and Chang (2013) found that CD-CAT can be effectively utilized to

FIGURE 7.
Students in Zhengzhou, China, actively participate in class discussion with the help a CD-CAT system, and such learning is more enjoyable than the regular classroom instruction.

help teachers in classroom teaching. A survey conducted in Zhengzhou found that CD-CAT encourages critical thinking, making students more independent in problem solving, and offers easy to follow individualized remedy, making learning more interesting. Thanks to the cutting-edge Browser/Server (B/S) architecture, today schools could implement their CAT systems with little to no additional cost using their current computer labs and networks (e.g., see Chang, 2012). Clearly, more classroom based CAT methods should be developed in the future.

4.5. Multistage Testing and CAT

For the past three decades the marketplace for item selection algorithms has been dominated by Fred Lord's maximum information method. However, several high stakes testing programs that built their CAT essentially based on the maximum information method have run into difficulties during their large scale operations and decided to give up CAT and replace it with Multi-Stage Testing (MST). While MST has many advantages, it does not mean that the problems encountered with CAT operations have no solutions. In my opinion, most of these issues can be resolved by redesigning the item selection algorithms appropriately. For instance, as it was discussed early in the paper that high discriminating items should be avoided at the early stages of the testing. Indeed, after studying a widely reported problem that GRE CAT and GMAT CAT did not produce reliable scores for thousands examinees in early 2000s (e.g., see Carlson, 2000 and Merritt, 2003), we provided both analytical and empirical evidence that those incidents could be avoided by fine-tuning the existing algorithms (Chang and Ying, 2008). Zhiliang Ying and I feel deeply honored for receiving the 2008 Annual Award of National Council on Measurement in Education for the research.

MST is a special case of CAT, in which an examinee receives a set of items (instead of a single item in CAT) that are matched to his or her provisional ability estimates. In fact, such design can be viewed as a *group sequential* design in clinical trials, whereas a CAT can be viewed a *fully sequential* design. In clinical trials, a fully sequential design is preferred if the next experiment can be based on the observations from all preceding experiments. The advantage of using a fully sequential design is that it usually reduces sample size and increases power

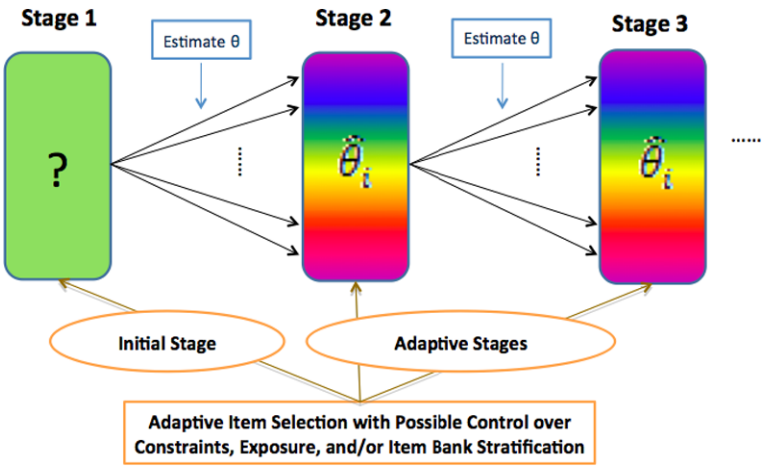

FIGURE 8.
Prototype of on-the-fly multistage adaptive testing (OMST).

(accuracy). However, "fully sequential" may not be practical in applications when the waiting time to get outcome from an experiment is too long. With its introduction in the 1970s for clinical trials (e.g., see Armitage, 2002; Pocock, 2002; O'Brien & Fleming, 1979, and Lan & DeMets, 1983), the group sequential method became fashionable to perform updating (interim analyses) for only very few times during the study period. Because of this, utilizing sequential methods in clinical trials has become routine nowadays. However, such "grouping" may not be necessary in an educational test since the response to each item can be immediately observed.

Currently, there is an enormous interest in developing methods for MST, such as automatically assembling parallel subtest forms, optimal routing strategies, constraints management, and test security. Publications concerning the topic are pouring in journals, conference sessions, technical reports, etc. However, many manuscripts were simply making inconsequential extensions over the very original paper of Luecht and Nungester (1998). In my opinion, much remains to be done at least on two frontlines: one is to develop innovative MST test-form assembly strategies and the other is to explore the relationship with CAT so that the future research of MST and CAT can be merged within a big framework of sequential design. Most recently, Zheng and Chang (2011) proposed an adaptive testing design called "on-the-fly multistage adaptive testing" (OMST, Figure 8) that merges CAT and MST into one big flexible framework. The basic idea is to utilize the multitude of item selection methods in CAT to assemble multistage test stages on the fly. With the flexibility of on-the-fly assembly, a test can have a variable number of stages and within each stage can have a variable number of items with varying degree of global and local information. One example of such "hybrid" designs is to gradually shrink the stage length as the test proceeds. More specifically, at the beginning of the test, when little information about examinee ability is available, longer stage lengths can provide a more accurate estimate before the first adaptation occurs (Chang & Ying, 1996); in later stages, when the estimate is closer to its true value, shorter stage lengths can provide more opportunities to adapt. This hybrid design is in fact an illustration of the smooth transition from MST to CAT under the big framework of OMST.

OMST could combine the merits of CAT and MST and offset their limitations. On the one hand, like MST, examinees in OMST are allowed to skip and go back to earlier items within a stage; on the other hand, like CAT, OMST can adapt more precisely to individual examinees than MST, especially for those at the polar ends of the ability scale. Also note that a well-liked feature of MST is that test developers can review the assembled test forms before administration.

When there are a great number of test forms, however, human review of all forms could become overwhelming. In OMST, the quality control can be partially automated and efficiently assisted by computers.

4.6. CAT and Adaptive Learning

Increasingly, CAT technology is being used worldwide for implementing large scale admissions and licensure testing. Still, the traditional paper-and-pencil (P&P) based test is the most common method of assessment in the classroom. To be truly useful for teachers in instructional planning, however, assessments should be tailored to each individual student. This would provide more precise and reliable diagnostics regarding each student's understanding and thought processes, thereby enabling teachers to better pinpoint areas in which students require further instruction. Indeed, a growing body of evidence shows that CAT has enormous potential to revolutionize classroom assessment and greatly facilitate individualized learning.

Recently, my colleagues in China, Yong-Qin Wang, Honyun Liu and Xiaofeng You (see Y.-Q. Wang, Liu, & You, 2013d), have made a marvelous breakthrough on making paper-and-pencil tests adaptive. Their design includes a PC server and a smart printer-scanner, such as a high-end model of Ricoh. Every time the students finish an in-class exercise, their booklets are scanned into the system by the printer-scanner. Then, the system automatically scores the exercise and generates individualized diagnostic reports. Based on the processed information, the system can also generate a stapled booklet for each student that contains a tailor-made homework assignment with clear instructions. I believe Wang and his colleagues have set an extraordinary example demonstrating how CAT technology can be used to support individualized instruction on a mass scale.

In a one-to-one instructional environment, the content and pace of instruction can be completely customized to best fit the observed progress of a particular student allowing the teacher to better focus on the individual's specific needs and problems. Certainly, such personalized instruction would be ideal. However, as a recent TIME article titled "A is for Adaptive" (Webley, 2013) explains: "It's impossible to provide one-to-one teaching on a mass scale, but technology enables us to get closer than ever before. As schools increasingly invest in computers and other digital products, students have access to a wider range of study materials, and teachers and administrators have the ability to view precise analyses of how they respond to that material, adjusting as needed." I found it truly exciting and encouraging to read this article featuring the Knewton platform, an innovative computer-based instruction system that allows schools, publishers, and developers to provide adaptive learning for students. However, instead of personal instruction by human teachers, the individualized teaching is performed by the Knewton system based on the contents and strategies developed by the educators. In contrast, a significant benefit worth emphasizing of Wang's model is that it does not diminish the teachers' role in classroom teaching. On the contrary, teachers can teach more effectively in their classrooms using the CAT enabled system, which provides constant individualized feedback to students as well as regular diagnostic reports to teachers regarding their students' performance. In particular, a built-in tracking mechanism ensures that students with special needs will receive prompt attention.

Clearly, CAT methods emphasizing individualized teaching and learning are invaluable and have shown great promise. In this regard, I believe both Wang's model and the Knewton platform, despite their differences, will play important roles in the future of classroom assessment.

5. Conclusion

Over the past 30 years the CAT research has become an increasingly important in the field of psychometric research. Although it was originally inspired by problems in high stakes testing, its

findings have been beneficial to other domains such as quality of life measurement, patient report outcome, K-12 accountability assessment, survey research, media and information literacy measure, etc. It is anticipated that further research is much needed to address issues and challenges in a range of fields. Now I conclude my address enthusiastically that CAT is making a substantial influence on the functioning of society by affecting how people are selected, classified, and diagnosed; CAT research will lead to better assessment, and hence benefit society.

Acknowledgements

I wish to thank Ying Cheng, Edison Choe, Rui Guo, Hyeon-Ah Kang, Justin Kern, Ya-Hui Su, Poh Hua Tay, Chun Wang, Shiyu Wang, Wen Zeng, Changjin Zheng, and Yi Zheng for their suggestions and comments which lead to numerous improvements.

References

Armitage, P. (2002). *Statistical methods in medical research* (4th ed.). Bodmin: MPG Books.

Carlson, S. (2000). ETS finds flaws in the way online GRE rates some students. *The Chronicle of Higher Education*, *47*(8), A47.

Chang, H.-H. (2004). Understanding computerized adaptive testing—from Robbins—Monro to Lord, and beyond. In D. Kaplan (Ed.), *The Sage handbook of quantitative methods for the social sciences* (pp. 117–133). Thousand Oaks: Sage.

Chang, H.-H. (2012). Making computerized adaptive testing diagnostic tools for schools. In R.W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessments: recent history and predictions for the future* (pp. 195–226). Charlotte: Information Age Publisher.

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., & van der Linden, W.J. (2003). Optimal stratification of item pools in *a*-stratified computerized adaptive testing. *Applied Psychological Measurement*, *27*(4), 262–274.

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

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

Chang, H.-H., & Ying, Z. (2007). Computerized adaptive testing. In N. Salkind (Ed.), *The Sage encyclopedia of measurement and statistics* (pp. 170–174). Thousand Oaks, CA: Sage.

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

Chang, H.-H., & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, *37*(3), 1466–1488.

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

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*(4), 619–642.

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: the modified maximum global discrimination index method. *Educational and Psychological Measurement*, *70*, 902–913.

Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*, 369–383.

Cheng, Y., Chang, H.-H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement*, *31*(6), 467–482.

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

Davey, T., & Nering, N. (2002). Controlling item exposure and maintaining item security. In C.N. Mills, M.T. Potenza, J.J. Fremer, & W.C. Ward (Eds.), *Computer-based testing: building the foundation for future assessments* (pp. 165–191). Mahwah: Lawrence Erlbaum.

Downing, S.M. (2006). Twelve steps for effective test development. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 3–25). Mahwah: Lawrence Erlbaum Associates.

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *37*(5), 655–670.

Hau, K., & Chang, H.-H. (2001). Item selection in computerized adaptive testing: should more discriminating items be used first? *Journal of Educational Measurement*, *38*(3), 249–266.

Hodges, J.I., & Lehmann, E.L. (1956). The efficiency of some nonparametric competitors of t-test. *The Annals of Mathematical Statistics*, *27*(2), 324–335.

Holland, P.W. (1990). The Dutch identity: a new tool for the study of item response theory model. *Psychometrika*, *55*, 577–601.

Klein Entink, R.H., van der Linden, W.J., & Fox, J.-P. (2009). A Box–Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*, 621–640.

Lan, K.K.G., & DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, *70*(3), 659–663.

Leung, C., Chang, H.-H., & Hau, K. (2003). Computerized adaptive testing: a comparison of three content balancing methods. *The Journal of Technology, Learning, and Assessment*, *2*(5), 2–15.

Liu, H., You, X., Wang, W., Ding, S., & Chang, H.-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification*, *30*, 152–172.

Lord, M.F. (1970). Some test theory for tailored testing. In W.H. Holzman (Ed.), *Computer assisted instruction, testing, and guidance* (pp. 139–183). New York: Harper and Row.

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

Luecht, R.M., & Nungester, R.J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229–249.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their applications as psychometric models for response times. *Psychometrika*, *58*, 445–469.

McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*(3), 808–821.

Merritt, J. (2003). Why the folks at ETS flunked the course—a tech-savvy service will soon be giving B-school applicants their GMATs. *Business Week*, Dec. 29.

Mislevy, R., & Chang, H.-H. (2000). Does adaptive testing violate local independence? *Psychometrika*, *65*(2), 149–156.

Mulder, J., & van der Linden, W.J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*(2), 273–296.

O'Brien, P.C., & Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, *35*, 549–556.

Pocock, S.J. (2002). *Clinical trials: a practical research approach*. Padstow: TJ International.

Ranger, J., & Kuhn, J.T. (2011). A flexible latent trait model for response times in tests. *Psychometrika*, *77*, 31–47.

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.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*, 400–407.

Roskam, E.E. (1997). Models for speed and time-limit tests. In W.J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York: Springer.

Rounder, J.N., Sun, D., Speckman, P.L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, *68*, 589–606.

Scheiblechner, H. (1979). Specific objective stochastic latency mechanisms. *Journal of Mathematical Psychology*, *19*, 18–38.

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.

Thissen, D. (1983). Timed testing: an approach using item response theory. In D.J. Weiss (Ed.), *New horizons in testing* (pp. 179–203). New York: Academic Press.

van der Linden, W.J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement*, *23*, 21–29.

van der Linden, W.J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204.

van der Linden, W.J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W.J., & Chang, H.-H. (2003). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement*, *27*(2), 107–120.

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

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, *73*, 1017–1035.

Wang, C., & Chang, H.-H. (2011). Item selection in multidimensional computerized adaptive testing—gaining information different angles. *Psychometrika*, *76*(3), 363–384.

Wang, T., & Hanson, B.A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339.

Wang, C., Chang, H.-H., & Huebner, A. (2011a). Restrictive stochastic item selection methods in cognitive diagnostic CAT. *Journal of Educational Measurement*, *48*(3), 255–273.

Wang, C., Chang, H.-H., & Boughton, K. (2011b). Kullback–Leibler information and its applications in multidimensional adaptive testing. *Psychometrika*, *76*(1), 13–39.

Wang, C., Chang, H.-H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: a weighted item selection approach. *Behavior Research Methods*, *44*, 95–109.

Wang, C., Chang, H.-H., & Douglas, J. (2013a). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical & Statistical Psychology*, *66*, 144–168.

Wang, C., Chang, H., & Boughton, K. (2013b). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *37*, 99–122.

Wang, C., Fan, Z., Chang, H.-H., & Douglas, J. (2013c). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 381–417.

Wang, Y.-Q., Liu, H., & You, X. (2013d). Learning diagnosis—from concepts to system development. Paper presented at the Anual Meeting of Assessment and Evaluation, the Chinese Society of Education, Dalian, China, May.

Webley, K. (2013). A is for adaptive—personalized learning is poised to transform education. Can it enrich students and investors as the same time? *Time*, June 17, 40–45.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473–492.

Xu, X., Chang, H., & Douglas, J. (2003). A simulation study to compare CAT strategies for cognitive diagnosis. Paper presented at the annual meeting of National Council on Measurement in Education, Chicago.

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

Zheng, Y., & Chang, H.-H. (2011). Automatic on-the-fly assembly for computer adaptive multistage testing. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA, April.

Zheng, Y., Chang, C.-H., & Chang, H.-H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research*, *22*, 491–499.

Manuscript Received: 27 OCT 2013