Parameter Drift Detection in Multidimensional Computerized Adaptive Testing Based on Informational Distance/Divergence Measures

Applied Psychological Measurement
2016, Vol. 40(7) 534–550
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621616663676
apm.sagepub.com

Hyeon-Ah Kang¹ and Hua-Hua Chang¹

Abstract

An informational distance/divergence-based approach is proposed to detect the presence of parameter drift in multidimensional computerized adaptive testing (MCAT). The study presents significance testing procedures for identifying changes in multidimensional item response functions (MIRFs) over time based on informational distance/divergence measures that capture the discrepancy between two probability functions. To approximate the MIRFs from the observed response data, the *k*-nearest neighbors algorithm is used with the random search method. A simulation study suggests that the distance/divergence-based drift measures perform effectively in identifying the instances of parameter drift in MCAT. They showed moderate power with small samples of 500 examinees and excellent power when the sample size was as large as 1,000. The proposed drift measures also adequately controlled for Type I error at the nominal level under the null hypothesis.

Keywords

item parameter drift, multidimensional item response function, multidimensional computerized adaptive testing

Multidimensional computerized adaptive testing (MCAT; Bloxom & Vale, 1987; Segall, 1996; Tam, 1992) has recently gained much popularity as a diagnostic measurement tool in educational, psychological, and clinical settings. It allows more comprehensive description of examinees' test performances than traditional unidimensional item response theory (IRT) and achieves much efficiency in test administration by taking advantage of adaptive item selection technique (Chang, 2015). One of the critical issues in administering MCAT is to ensure item parameters remain stable over time. Because scoring of examinees' latent traits is predicated on the item parameter invariance, any systematic instability in item parameters can adversely affect the scoring process. If test items are used to grant professional competence such as in

¹University of Illinois at Urbana–Champaign, USA

Corresponding Author:

Hyeon-Ah Kang, University of Illinois at Urbana–Champaign, 1310 S. Sixth Street, #210 Educational Psychology, Champaign, IL 61820-6990, USA.
Email: hkang31@illinois.edu

certification or licensure assessments, shifts in item parameters can lead to false decisions when classifying examinees. Hence, it is important to identify instances of parameter drift as soon as possible to minimize any repercussions on the validity and comparability of the scaled scores.

There are a number of factors causing the item parameter drift in practice. The occurrence of parameter drift (Goldstein, 1983) can be attributed to changes in instruction or curriculum (Bock, Muraki, & Pfeifferberger, 1988; Cook, Eignor, & Taft, 1988), changes in constructs (Chan, Drasgow, & Sawin, 1999), or changes made to items between test administrations (Sykes & Ito, 1993). The parameter drift can also appear as a consequence of item disclosure by previous test takers, test fraud, item overexposure, or inappropriate testwise training (Li, 2008).

A common practice for identifying parameter drift in unidimensional items is to recalibrate items and evaluate changes in the estimated parameter values over testing occasions. The instance of parameter drift can be determined by comparing the item parameter estimates directly (Bock et al., 1988) or comparing model fit statistics obtained from the different testing events (Glas, 2000). Alternatively, one can compare test characteristic curves across time points (Guo, Zheng, & Chang, 2015; Wollack, Cohen, & Wells, 2003). Statistical measures that are originally developed for analysis of differential item functioning (DIF) also serve as drift detection measures (e.g., Kim & Cohen, 1991; Lord, 1980; Raju, 1988). Concerned with the computerized adaptive testing (CAT), the cumulative sum chart based on standardized differences between estimated item parameter values can be used to evaluate the drift (Veerkamp & Glas, 2000). For a discussion of causes and methods for detecting parameter drift in unidimensional tests, see Clark (2014).

Although much has been studied about the drift detection methods in unidimensional tests, there has been relatively little work developing or comparing drift procedures in MCAT. A complication of using the traditional methods for the MCAT data is that outcomes of the drift analysis are likely to be confounded with high calibration error or linking error. Because the traditional approaches use item parameter estimates from separate calibrations as a basis for drift analysis, the efficacy of the procedures is critically dependent upon the quality of item calibration (Donoghue & Isham, 1998). If the parameter estimates contain high uncertainty due to large numbers of parameters or missing responses, inferences about the existence of parameter drift may be hardly validated. Besides, calibrating items in MCAT separately time after time is an onerous task, especially when the items have varying frequencies in terms of exposure distribution. In instances where operational items need to be routinely monitored during the MCAT administrations, the computational burden of separate calibrations becomes even more serious.

The focus of this article is on developing viable options for detecting parameter drift without the need to calibrate items in the MCAT data. The new procedures do not involve item parameter estimation, and hence, they can be freed from calibration error or linking error. The procedures are based on significance testing of the difference between an observed multidimensional item response function (MIRF) and an anticipated MIRF. In a typical MCAT program, a precalibrated item pool is available such that items can be selected according to their known statistical properties. Therefore, the presence of parameter drift can be determined by comparing the probability function observed from the operational testing against the probability function derived from the initial item parameter values. The present article introduces four informational distance/divergence measures that serve this role.

The present article is organized as follows. First, the theoretical framework for the analysis of parameter drift in multidimensional items is briefly outlined. The article proceeds to introduce the informational distance/divergence measures that can be used for identifying the existence of the drift. The proposed methods are then evaluated in simulations. Finally, some

discussion on the use of informational distance/divergence measures is provided. Throughout the article, it is assumed that items used for MCAT have adequate fit for the psychometric model of concern, and they are precalibrated with enough precision before the operational use. The initial parameter values of the items can thus be treated as known or reference values in the subsequent drift analysis.

Analysis of Parameter Drift in Multidimensional Items

Suppose a test purports to measure a set of latent proficiencies, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p) \in \boldsymbol{\Theta}$, and denote a vector of item parameters for a studied item as $\boldsymbol{\beta}$. The function relating the $\boldsymbol{\theta}$ to the probability of a correct item response is called the MIRF and denoted as $P(\boldsymbol{\theta}; \boldsymbol{\beta})$. Let $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ represent the item parameter vectors corresponding to two testing occasions. Because the item parameters uniquely dictate the shapes of item response functions (Kim, Cohen, & Park, 1995), item parameter invariance can be characterized by the equivalence of the two MIRFs:

$$P(\boldsymbol{\theta}; \boldsymbol{\beta}_0) = P(\boldsymbol{\theta}; \boldsymbol{\beta}_1), \quad (1)$$

for all levels of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. A response score for a dichotomous item, X , has the Bernoulli distribution, $f(X=x|\boldsymbol{\theta}; \boldsymbol{\beta}) = P(\boldsymbol{\theta})^x(1 - P(\boldsymbol{\theta}))^{1-x}$, for which Equation 1 can be equivalently expressed as

$$f_0 = f(X=x|\boldsymbol{\theta}; \boldsymbol{\beta}_0) = f(X=x|\boldsymbol{\theta}; \boldsymbol{\beta}_1) = f_1, \quad (2)$$

for all levels of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The probability functions, f_0 and f_1 , represent two hypotheses about the functional form of the studied item. When applied to MCAT, f_0 denotes the probability function computed from the preestimated item parameters, whereas f_1 represents the probability function observed from the response data.

Consider a dissimilarity measure, D that captures the discrepancy between the probability functions. Equation 2 indicates that the degree of heterogeneity between f_0 and f_1 is a function of $\boldsymbol{\theta}$. To obtain an overall amount of parameter drift for the studied item, an average is taken over all possible levels of $\boldsymbol{\theta}$:

$$\mathcal{D} = \int_{\boldsymbol{\Theta}} D(f_0||f_1; \boldsymbol{\theta}) \omega(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (3)$$

where $D(f_0||f_1; \boldsymbol{\theta})$ evaluates the distance or divergence of f_1 from f_0 at $\boldsymbol{\theta}$, and $\omega(\boldsymbol{\theta})$ is the density function of $\boldsymbol{\theta}$. The quantity $\mathcal{D}$ assesses the dissimilarity between f_0 and f_1 without regard to whether or not the magnitude of drift is constant with the varying proficiency levels. The hypotheses for testing the presence of parameter drift are stated as $H_0 : \mathcal{D}=0$ and $H_1 : \mathcal{D}>0$. The null hypothesis of no drift is rejected with level of significance α if the value of $\mathcal{D}$ exceeds the 100(1 - α)th percentile.

Computation of the drift statistic in Equation 3 requires a matching criterion so that the two probability functions can be evaluated at the same proficiency level. In the unidimensional tests, examinees' number-correct scores or unidimensional proficiency estimates are commonly used as a matching variable. When a test is intended to measure multiple traits, matching on the unidimensional criterion can result in Type I error inflation (Ackerman, 1992). Therefore, to meet the validity standards and to avoid high Type I error rate, the present study employs latent proficiency vectors estimated from the operational MCAT as a matching criterion. The proficiency estimates take all pertinent latent dimensions into account and are readily obtainable from the administrations of the MCAT. Note that in multidimensional tests, it is almost impossible to match examinees on the exact values of the proficiency estimates (i.e., thin

matching). A certain strategy must be employed to pool examinees based on their estimated proficiency vectors (i.e., thick matching). This goal is achieved by employing the k -nearest neighbors (k -NNs) algorithm.

A basic procedure follows from this. Given a collection of query vectors, $\mathbf{q}_l$ ($l = 1, \dots, Q$), the algorithm searches for k examinees whose estimated proficiency vectors are close to $\mathbf{q}_l$ according to a certain distance metric. Item responses for the k -nearest examinees are then collected to approximate the MIRF. The probability of a correct response for $\mathbf{q}_l$ is calculated by the sample proportion of the correct responses among the nearest neighbors as follows:

$$\tilde{P}(\mathbf{q}_l) = \frac{1}{k} \sum_{i \in N_k(\mathbf{q}_l)}^k x_{ij}, \quad (4)$$

where $N_k(\mathbf{q}_l)$ represents the k -NNs to a query vector $\mathbf{q}_l$, and x_{ij} is an observed response score for an examinee i to the studied item j . It is often necessary to use a weighted average of the k -NNs so that the nearer contributes more to the average outcome than the more distant ones. Let $d(\hat{\boldsymbol{\theta}}_i, \mathbf{q}_l)$ denote the distance between the query vector $\mathbf{q}_l$ and the i th proficiency estimate $\hat{\boldsymbol{\theta}}_i$ in the neighborhood of $\mathbf{q}_l$. Then, a weighted approximation of the MIRF is obtained as

$$\tilde{P}(\mathbf{q}_l) = \sum_{i \in N_k(\mathbf{q}_l)}^k x_{ij} \omega(\hat{\boldsymbol{\theta}}_i, \mathbf{q}_l), \quad (5)$$

where

$$\omega(\hat{\boldsymbol{\theta}}_i, \mathbf{q}_l) = \frac{\exp[-d(\hat{\boldsymbol{\theta}}_i, \mathbf{q}_l)]}{\sum_{i \in N_k(\mathbf{q}_l)} \exp[-d(\hat{\boldsymbol{\theta}}_i, \mathbf{q}_l)]}. \quad (6)$$

The weights defined in this manner satisfy $\sum_{i=1}^k \omega(\hat{\boldsymbol{\theta}}_i, \mathbf{q}_l) = 1$. The type of the distance metric $d(\cdot, \cdot)$ can be determined based on the properties of the latent parameters. A common practice for a vector of a continuous variable is to use the Euclidean distance.

The main advantage of using the k -NNs technique in the MCAT data is that the algorithm always finds a proper set of matching variables for comparing the MIRFs. Despite the fact that the size or the distribution of examinee samples can change over different testing events, the k -NNs algorithm can approximate the MIRFs with comparable precision over time by taking the k -NNs close to the query vectors into consideration.

Informational Distance/Divergence Measures

To structure the heterogeneity between the MIRFs, a statistical dissimilarity index that separates pairs of probability functions is needed. This article proposes four informational distance/divergence measures that serve such purpose. The informational distance/divergence measures have been widely employed in many areas of statistics such as binary hypothesis testing, classification applications, anomaly detection in high-dimensional data, and so forth. The use of the distance/divergence measures is especially advantageous in multidimensional settings in that they can summarize the degree of heterogeneity between probability distributions into a single numeric without regard to the size of dimension or the number of parameters.

The following introduces four informational distance/divergence measures for identifying the presence of parameter drift in MCAT and presents corresponding estimation methods of the drift statistics. The drift statistics below are defined for the dichotomous items for convenience; however, their definitions stated in terms of the MIRFs are not necessarily limited to the dichotomous cases. When multidimensional polytomous items are of concern, the distance/divergence measures can be adjusted to assess the discrepancy between two multidimensional item response category functions.

In line with the notations above, the vector of reference item parameters (i.e., initial item parameter estimates) is denoted as β_0 , and the parameter vector that characterizes the item in the operational assessment is denoted as β_1 . Analogously, the MIRFs corresponding to each item parameter vector are denoted as $P_0(\theta) = P(\theta; \beta_0)$ and $P_1(\theta) = P(\theta; \beta_1)$, respectively.

Euclidean Distance

One of the simplest ways to quantify the degree of dissimilarity between two probability functions is to use the Euclidean distance. For given θ , the Euclidean distance between $f_0 = f(x|\theta; \beta_0)$ and $f_1 = f(x|\theta; \beta_1)$ is calculated as

$$D_E(f_0 || f_1; \theta) = \sqrt{\sum_{x=0}^1 [f(x|\theta; \beta_0) - f(x|\theta; \beta_1)]^2}. \quad (7)$$

The quantity defined in Equation 7 is nonnegative and equals 0 if and only if $f_0 = f_1$ for all $\theta \in \Theta$. As a special case of the Minkowski distance (also known as $\mathcal{L}_p$ norm) of order $p=2$, the Euclidean distance is always finite between the probability distributions.

The drift statistic based on the Euclidean distance is obtained by plugging Equation 7 into Equation 3:

$$\mathcal{D}_E = \int_{\Theta} D_E(f_0 || f_1; \theta) \omega(\theta) d\theta = \sqrt{2} \int_{\Theta} |P_0(\theta) - P_1(\theta)| \omega(\theta) d\theta. \quad (8)$$

In Equation 8, $P_0(\theta)$ is computed from the reference item parameters, and hence, it only needs to find $P_1(\theta)$. Applying the k -NNs technique described above, the drift statistic for the Euclidean distance is estimated as

$$\hat{\mathcal{D}}_E = \sqrt{2} \sum_{l=1}^Q |P_0(\mathbf{q}_l) - \tilde{P}_1(\mathbf{q}_l)| \omega(\mathbf{q}_l) \Delta, \quad (9)$$

where $P_0(\mathbf{q}_l)$ is the MIRF evaluated at $\mathbf{q}_l$; $\tilde{P}_1(\mathbf{q}_l)$ is the MIRF corresponding to $N_k(\mathbf{q}_l)$; $\omega(\mathbf{q}_l)$ is the distributional weight of $\mathbf{q}_l$; and Δ is an increment between $\mathbf{q}_{l-1}$ and $\mathbf{q}_l$. A common choice of Δ in the continuous space is to use the Euclidean distance.

Hellinger Distance

The Euclidean distance of the square root function is called Hellinger distance (Bhattacharyya, 1943). In the present context, the Hellinger distance between f_0 and f_1 conditioned on θ is defined as

$$D_H(f_0 || f_1; \theta) = \sqrt{\frac{1}{2} \sum_{x=0}^1 \left\{ \sqrt{f(x|\theta; \beta_0)} - \sqrt{f(x|\theta; \beta_1)} \right\}^2}. \quad (10)$$

The Hellinger distance, D_H satisfies the metric properties—nonnegativity, symmetry, and triangle inequality—and is bounded between 0 and 1 as a result of Cauchy–Schwarz inequality. Deriving the drift statistic for the Hellinger distance parallels that for the Euclidean distance. Substituting the dissimilarity measure with the Hellinger distance in Equation 3 gives the following:

$$\mathcal{D}_H = \int_{\Theta} D_H(f_0 || f_1; \boldsymbol{\theta}) \omega(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (11)$$

Based on the observed responses, this quantity can be estimated as follows:

$$\hat{\mathcal{D}}_H = \frac{1}{\sqrt{2}} \sum_{l=1}^Q \left\{ \left(\sqrt{P_0(\mathbf{q}_l)} - \sqrt{\tilde{P}_1(\mathbf{q}_l)} \right)^2 + \left(\sqrt{Q_0(\mathbf{q}_l)} - \sqrt{\tilde{Q}_1(\mathbf{q}_l)} \right)^2 \right\}^{1/2} \omega(\mathbf{q}_l) \Delta, \quad (12)$$

where $Q_0(\mathbf{q}_l) = 1 - P_0(\mathbf{q}_l)$ and $\tilde{Q}_1(\mathbf{q}_l) = 1 - \tilde{P}_1(\mathbf{q}_l)$.

Kullback–Leibler (KL) Divergence

The KL (Kullback & Leibler, 1951) divergence (also known as relative entropy or crossing entropy) has been frequently used in information theory as a measure of discriminating two probability functions. Chang and Ying (1996) introduced the KL divergence to the IRT framework as a global item information measure that discriminates examinees' true proficiencies and provisional estimates in CAT. In the context of multidimensional IRT, the KL divergence of f_1 from f_0 at $\boldsymbol{\theta}$ is defined as follows:

$$D_{KL}(f_0 || f_1; \boldsymbol{\theta}) = \sum_{x=0}^1 f(x | \boldsymbol{\theta}; \boldsymbol{\beta}_0) \log \frac{f(x | \boldsymbol{\theta}; \boldsymbol{\beta}_0)}{f(x | \boldsymbol{\theta}; \boldsymbol{\beta}_1)}. \quad (13)$$

The KL divergence is nonnegative because of Gibbs inequality and equals 0 when the two distributions coincide almost everywhere. The drift statistic for the KL divergence is obtained as

$$\mathcal{D}_{KL} = \int_{\Theta} D_{KL}(f_0 || f_1; \boldsymbol{\theta}) \omega(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (14)$$

and estimated as

$$\hat{\mathcal{D}}_{KL} = \sum_{l=1}^Q \left[P_0(\mathbf{q}_l) \log \frac{P_0(\mathbf{q}_l)}{\tilde{P}_1(\mathbf{q}_l)} + Q_0(\mathbf{q}_l) \log \frac{Q_0(\mathbf{q}_l)}{\tilde{Q}_1(\mathbf{q}_l)} \right] \omega(\mathbf{q}_l) \Delta. \quad (15)$$

While the KL divergence has been considered as one of the most powerful discriminating measures of probability functions in information theory, it has a number of drawbacks that may detract from its usefulness as a pairwise drift measure in applied settings. By definition, the KL divergence is not symmetric in f_0 and f_1 , and hence, the values of the KL divergence can differ depending on the order of the arguments. In addition, although the KL divergence is lower bounded by the Hellinger distance such that $2D_H^2 \leq D_{KL}$, the values of the KL divergence can potentially equal infinity. Because of this unboundedness, the distribution of test statistics from the KL divergence can be severely right-tailed, resulting in unstable critical values when significance testing is implemented using the empirical sampling distributions. On this account, this study proposes an alternative dissimilarity measure that is bounded and invariant with respect

to permutations of the arguments but still preserves most of the desirable properties of the KL divergence.

Jensen–Shannon (JS) Divergence

Let λ be a probability for X being drawn from a probability distribution f_0 , and $1 - \lambda$ be a probability for X having a counterpart probability distribution f_1 . A symmetric version of the KL divergence can be obtained by a weighted average as follows:

$$D_\lambda(f_0 \| f_1; \boldsymbol{\theta}) = \lambda D_{KL}(f_0 \| \lambda f_0 + (1 - \lambda)f_1; \boldsymbol{\theta}) + (1 - \lambda) D_{KL}(f_1 \| \lambda f_0 + (1 - \lambda)f_1; \boldsymbol{\theta}). \quad (16)$$

The quantity D_λ evaluates the expected information gain about X from discovering which probability function X is sampled from. By imposing the same degree of uncertainty on the two probability functions, the JS (Rao, 1982) divergence is obtained as follows:

$$D_{JS}(f_0 \| f_1; \boldsymbol{\theta}) = \frac{1}{2} [D_{KL}(f_0 \| f_m; \boldsymbol{\theta}) + D_{KL}(f_1 \| f_m; \boldsymbol{\theta})], \quad (17)$$

where $f_m = (f_0 + f_1) / 2$. The JS divergence satisfies the metric properties and is bounded between 0 and $\log 2$. For evaluating the parameter drift in the multidimensional IRT framework, the JS divergence-based drift measure is defined as follows:

$$\mathcal{D}_{JS} = \frac{1}{2} \int_{\boldsymbol{\theta}} \sum_{x=0}^1 \left[f(x | \boldsymbol{\theta}; \boldsymbol{\beta}_0) \log \frac{f(x | \boldsymbol{\theta}; \boldsymbol{\beta}_0)}{f_m(x | \boldsymbol{\theta})} + f(x | \boldsymbol{\theta}; \boldsymbol{\beta}_1) \log \frac{f(x | \boldsymbol{\theta}; \boldsymbol{\beta}_1)}{f_m(x | \boldsymbol{\theta})} \right] \omega(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (18)$$

where $f_m(x | \boldsymbol{\theta}) = [f(x | \boldsymbol{\theta}; \boldsymbol{\beta}_0) + f(x | \boldsymbol{\theta}; \boldsymbol{\beta}_1)] / 2$. Based on the k -NNs, the quantity is approximated as follows:

$$\begin{aligned} \hat{\mathcal{D}}_{JS} = \sum_{l=1}^Q \frac{1}{2} \left[P_0(\mathbf{q}_l) \log \frac{P_0(\mathbf{q}_l)}{P_m(\mathbf{q}_l)} + Q_0(\mathbf{q}_l) \log \frac{Q_0(\mathbf{q}_l)}{Q_m(\mathbf{q}_l)} + \dots \right. \\ \left. + \tilde{P}_1(\mathbf{q}_l) \log \frac{\tilde{P}_1(\mathbf{q}_l)}{P_m(\mathbf{q}_l)} + \tilde{Q}_1(\mathbf{q}_l) \log \frac{\tilde{Q}_1(\mathbf{q}_l)}{Q_m(\mathbf{q}_l)} \right] \omega(\mathbf{q}_l) \Delta, \end{aligned} \quad (19)$$

where $P_m(\mathbf{q}_l) = [P_0(\mathbf{q}_l) + \tilde{P}_1(\mathbf{q}_l)] / 2$ and $Q_m(\mathbf{q}_l) = [Q_0(\mathbf{q}_l) + \tilde{Q}_1(\mathbf{q}_l)] / 2$.

Simulation Study

A simulation study is used to examine the performance of the distance/divergence-based drift measures. For the drift analysis to be fully sequential, significance testing must be carried out after each included observation; however, such procedure becomes inefficient in terms of a central process unit time. The present study conducts drift analysis at three predefined time points instead. The drift statistics were computed at each time an item was administered to $n = 500$, 1,000, and 1,500 examinees, and significance testing was carried out on the difference between the original and observed probability functions.

Determination of k

Because the performance of the drift measures depends on the k -NNs algorithm, it is important to carefully choose the size of k . An overly large value of k would fail to evaluate the

heterogeneity in the item response functions properly. If k is too small, the approximation of item response functions could be unstable because only a bit of responses are utilized for a given query vector. (In the extreme case of $k = 1$, P_0 and $\tilde{P}_1$ are computed based on only one response.) In both cases, test statistics would become powerless. To ensure stable estimation of the drift statistics and adequate statistical power, the present study examined the performance of the drift measures using sets of possible k values. The values of k were determined as n/Q such that no information was lost from the response data. When Q was not a factor of n , k was determined as the nearest integer greater than n / Q so that a rehash of the response data could be minimized in the computation of the drift statistics. For controlling the quality of the drift statistics, the k -NNs algorithm was required to have a minimum of six and a maximum of 15 query vectors.

Determination of Query Vectors

Among a number of heuristic techniques for determining the query vectors, grid search and random search are the most widely used methods for hyperparameter optimization of the k -NNs algorithm. The grid search method uses grid points manually drawn from a subset of a hyperparameter space (i.e., Θ) as query vectors, whereas the random search method uses query vectors that are randomly sampled from the Θ . Because the grid search operates on a grid that increases by a specific value on a finite interval, it can suffer from the curse of dimensionality. Compared with the grid search, the random search is more efficient and computationally less intensive (Bergstra & Bengio, 2012). For this reason, the current study employed the random search method for determining the query vectors in the computation of the drift statistics. To prevent attenuation of drift impact from occurring, the upper and lower limits of Θ were set at 2 and -2 , respectively, and query vectors were randomly sampled from this subset.

Determination of Critical Values

The asymptotic distributions of the drift statistics proposed are unknown. Thus, critical values were obtained empirically through bootstrap resampling under the null hypothesis. A brief description of finding the critical values at significance level α is given as follows.

Step 1: Compute drift statistics at each time an item is assigned to n examinees in the null case of MCAT administration.

Step 2: For each n condition, treat the original set of the drift statistics as a population of scores and perform resampling with replacement.

Step 3: Repeat Step 2 to generate m resamples of test statistics.

Step 4: Calculate the $100(1 - \alpha)$ quantile in each m resample to obtain critical values.

Step 5: Determine an empirical critical value as an average of the critical values over the m resamples.

The number of resamples, m , should be large enough to get a good estimation of the quantile. The present study used $m = 1,000$ resamples and determined the critical values at significance level $\alpha = 0.05$ in the upper tail of the sampling distributions. The critical values resulting from the above steps are size-corrected that do not depend on unknown population parameters but are likely to yield a test with little power. Therefore, to secure adequate power of the drift statistics, $n = 500$ was considered as the minimum sample size for drift analysis.

Data Generation

MCAT was administered based on the multidimensional three-parameter logistic model (Reckase, 1997):

$$P(\boldsymbol{\theta}; \mathbf{a}, b, c) = c + \frac{1 - c}{1 + \exp[-1.7\mathbf{a}^T(\boldsymbol{\theta} - b)]}, \quad (20)$$

where $\mathbf{a} = (a_1, \dots, a_p)$ is the item discrimination vector, b is the intercept parameter, and c is the lower asymptote. The dimension of $\mathbf{a}$ depends on the number of latent traits being measured by tests. The present study considered two scenarios, two-dimensional and three-dimensional CAT, to examine the performance of the drift measures under the different levels of dimensionality. Examinees' latent proficiency levels were simulated from the multivariate normal distribution with zero means, unit variances, and correlations of 0.3. MCAT was administered to 5,000 examinees based on the Bayesian D-optimality item selection criterion (Segall, 1996). For proficiency estimation, the maximum a posteriori was used. Test length was fixed at 35. Item exposure rate was controlled by setting the maximum item exposure rate at 0.3 such that no more than 30% of examinees in the sample received the same item. Each item pool had 300 test items with approximate simple structure (Zhang & Stout, 1999). Item parameters were randomly sampled from the following distributions: $a \sim U(0.5, 1.2)$ for the primary dimension(s), $a \sim U(0, 0.4)$ for the secondary dimension(s), $b \sim N(0, 1)$, and $c \sim \text{Beta}(100, 400)$.

Parameter Drift Simulation

Within each MCAT scenario, performances of the drift measures were evaluated with or without parameter drift occurring. The study design included three proportions of item pool-wide parameter drift: 0%, 5%, and 10%. Motivated by the tendency in practice in which overexposed items tend to have high likelihood of being flagged for parameter drift (Zhang, 2014), drift items were randomly selected from a set of frequently used items in the null case of MCAT. All items chosen for parameter drift were made to experience the same type of parameter drift to examine the impact of drift type systematically. Two types of parameter drift were created. The parallel shifts in the MIRFs were made by changing the b -parameters of the drift items by 0.3 or 0.5 units from the original values. The nonparallel transformations of the MIRFs were simulated by changing the a -parameters by 0.5 units and the b -parameters by 0.3 or 0.5 units. The direction of the parameter drift was determined such that the drift items became less discriminating or less difficult over the testing events (e.g., DeMars, 2004; Veerkamp & Glas, 2000). Because the proposed drift measures do not place any constraints on the direction, similar inferences can be made about the performance of the measures in the opposite direction.

Crossing the conditioned factors resulted in 18 different MCAT scenarios. Each scenario was executed with 100 replications to regulate the sampling error, and the results were averaged over the replications.

Results

Determination of Q

The preparatory study for determining the number of query vectors suggested that Q values between eight and 10 have minor differences in terms of power and Type I error rates. (The performances of the drift measures in relation to varying Q values are presented in the appendix.) For the sake of simplicity in presentation, results provided below are summarized using the

Figure 1. Empirical sampling distributions in the two-dimensional computerized adaptive testing.

fixed values of Q . Specifically, the Euclidean distance and the KL divergence used eight query vectors. Results for the Hellinger distance and the JS divergence were reported based on 10 and nine query vectors, respectively. These choices were motivated by the consideration that the resulting tests had the highest power against the alternative hypothesis across all simulation conditions. On the whole, as the Q values were further from these optimal values, both the power and Type I error tended to decrease.

Empirical Sampling Distributions

To serve as test statistics for significance testing, drift measures must have distinct limiting distributions under the null hypothesis. As their asymptotic distributions are unknown, the current study obtained empirical sampling distributions via simulation and evaluated their characteristics in the null case of no drift. Figure 1 presents empirical sampling distributions of the drift statistics obtained from the two-dimensional CAT. All drift measures demonstrated explicit forms of sampling distributions. The distance-based statistics had the symmetric sampling distributions. The divergence-based measures appeared positively skewed distributions. The impact of increasing n on the drift statistics was manifested by shifts of the sampling distributions toward the left. In general, the empirical sampling distributions became less dispersed, less skewed, and less peaked as n increased.

Figure 2. Empirical sampling distributions in the three-dimensional computerized adaptive testing.

Similar patterns were observed for the three-dimensional case. Figure 2 provides empirical sampling distributions obtained from the three-dimensional CAT. All drift test statistics showed distinct sampling distributions under the null. Compared with Figure 1, the sampling distributions in Figure 2 had smaller averages and smaller standard deviations (SDs), suggesting that drift statistics from the higher dimension were less distinguishable. The trends associated with n were consistent with those for the two-dimensional CAT. When the Q was held constant for each drift measure, the sampling distributions tended to shift toward the left and became less dispersed, less skewed, and less peaked as n increased.

The figures presented above suggest that the distance/divergence-based drift statistics have explicit sampling distributions under the null hypothesis. Based on these distributions, empirical critical values were obtained through bootstrap resampling at the significance level of 0.05. Table 1 reports the average and SD values of the critical values used in the study. In line with the prior results, the empirical critical values tended to decrease as n or p increased. Along with these changes, the critical values became less dispersed, suggesting that the critical values became more consistent.

Type I Error Study

In Table 2, Type I error rates of the drift measures are summarized for the null and drift cases. The Type I error rate was defined as the percentage of nondrift items that were erroneously identified as drift in each MCAT administration. The values reported for the drift cases were

Table 1. Averages and SDs of Empirical Critical Values.

<i>n</i>	<i>p</i> = 2				<i>p</i> = 3			
	EU	HE	KL	JS	EU	HE	KL	JS
500 (<i>SD</i>)	.107 (.009)	.268 (.019)	.030 (.004)	.009 (.001)	.036 (.003)	.154 (.011)	.010 (.001)	.003 (.000)
1,000 (<i>SD</i>)	.080 (.008)	.205 (.016)	.018 (.003)	.005 (.001)	.027 (.003)	.119 (.009)	.006 (.001)	.002 (.000)
1,500 (<i>SD</i>)	.066 (.007)	.170 (.015)	.012 (.002)	.004 (.001)	.022 (.002)	.095 (.007)	.004 (.001)	.001 (.000)

Note. *n* = sample size; *p* = number of latent dimensions; EU = Euclidean distance; HE = Hellinger distance; KL = Kullback–Leibler divergence; JS = Jensen–Shannon divergence.

Table 2. Type I Error Rates in the Null and Drift Cases.

Drift	<i>n</i>	<i>p</i> = 2				<i>p</i> = 3			
		EU	HE	KL	JS	EU	HE	KL	JS
0%	500	.043	.042	.042	.043	.046	.046	.046	.047
	1,000	.037	.039	.037	.037	.044	.044	.044	.043
	1,500	.037	.036	.037	.038	.042	.045	.045	.044
5%	500	.054	.067	.068	.064	.057	.061	.065	.060
	1,000	.068	.076	.072	.076	.075	.072	.076	.072
	1,500	.080	.090	.088	.091	.111	.112	.114	.112
10%	500	.076	.107	.112	.100	.074	.083	.093	.081
	1,000	.113	.146	.146	.149	.114	.123	.133	.123
	1,500	.153	.191	.192	.197	.188	.209	.213	.206

Note. *n* = sample size; *p* = number of latent dimensions; Drift = proportion of drift items in the item pools; EU = Euclidean distance; HE = Hellinger distance; KL = Kullback–Leibler divergence; JS = Jensen–Shannon divergence.

obtained by averaging over the drift levels and drift types because no systematic pattern was found across these factors. Overall, the drift measures maintained good adherence to the nominal level of significance under the null hypothesis. They constantly kept Type I error rates below the nominal level without regard to *n* or *p*. *SDs* of the Type I errors remained small and slightly increased as *n* increased. The average *SDs* were .006, .008, and .009 for each *n* condition.

The drift measures began to display Type I error inflation as the item pools were contaminated with drift items. The magnitude of the inflation was somewhat minor or moderate when 5% of the items in the pool were flagged for parameter drift. The Type I error inflation problem appeared substantial as 10% of the item pools were flagged for drift. Among the distance/divergence measures, the Euclidean distance was found most conservative in terms of Type I error inflation. Increasing *n* in the drift analyses resulted in increased Type I error rates for all drift measures possibly due to the greater sensitivity with the larger *n*. Consistent with the null case, *SDs* of the Type I error rates were influenced by the level of *n*, that is, the larger the *n*, the more the variability in the Type I error rates. In the 5% drift case, *SDs* of the Type I error rates were .031, .044, and .054 for increasing *n*. In the 10% drift case, *SDs* of the Type I error rates increased to .046, .073, and .094. Overall, the impact of the dimensionality on the averages and *SDs* of the Type I error rates was not immediately obvious.

Table 3. Power Rates When 5% of Item Pools Were Flagged for Parameter Drift.

$\Delta(a, b)$	n	$p = 2$				$p = 3$			
		EU	HE	KL	JS	EU	HE	KL	JS
(0, .3)	500	.754	.772	.769	.758	.753	.763	.784	.775
	1,000	.836	.839	.861	.850	.818	.847	.853	.836
	1,500	.886	.896	.901	.895	.868	.889	.886	.883
(0, .5)	500	.751	.769	.777	.781	.751	.777	.771	.787
	1,000	.838	.843	.853	.851	.827	.838	.851	.845
	1,500	.882	.895	.904	.908	.876	.887	.893	.893
(.5, .3)	500	.755	.756	.767	.764	.746	.759	.770	.771
	1,000	.835	.833	.852	.861	.821	.831	.847	.843
	1,500	.885	.880	.906	.898	.874	.875	.895	.889
(.5, .5)	500	.738	.760	.771	.765	.758	.775	.775	.770
	1,000	.839	.848	.845	.851	.825	.841	.846	.837
	1,500	.889	.896	.904	.907	.872	.891	.897	.884

Note. $\Delta(a, b)$ = changes in a - and b -parameters; n = sample size; p = number of latent dimensions; EU = Euclidean distance; HE = Hellinger distance; KL = Kullback–Leibler divergence; JS = Jensen–Shannon divergence.

Power Study

Table 3 reports power rates of the distance/divergence measures when 5% of the items in the pool were flagged for parameter drift. The power rate was defined as the proportion of correctly identified drift items in each MCAT administration. For evaluation, power rates above .80 were considered excellent (Cohen, 1992) and rates between 0.70 and 0.80 were considered moderate. In Table 3, the distance/divergence measures identified changes in the item parameters quite effectively despite the presence of multidimensionality. The measures showed moderate power when n was as small as 500 and showed excellent power when n was equal to or greater than 1,000. Overall, increasing n led to substantial improvement in the detecting power for all drift measures. The increase in n also resulted in decreasing SD s of the power rates. SD s of the power rates were on average .108, .094, and .082 in the two-dimensional CAT, and .110, .096, and .084 in the three-dimensional CAT at each conditioned n . These results suggest that the drift measures performed more consistently along with the larger n .

In Table 3, the divergence measures in general outperformed the distance measures. The differences between the lowest and highest power rates, however, occurred less than 0.05 in all occasions, indicating that the drift measures under evaluation performed very comparably in detection of drift items. Overall, the impact of the drift level and the drift type on the power rates seemed minor. One plausible explanation for this pattern would be the use of random query vectors in the k -NNs algorithm. As the query vectors were randomly selected, absolute values of the drift statistics could vary depending on the selected query vectors. A direct comparison of the power performances will be less meaningful in such cases despite the systematic differences in the drift level and the drift type. The grid search method, on the contrary, may show distinct patterns related to these factors. According to preliminary studies, when a set of fixed query vectors (e.g., a grid between $[-2, 2]$ at increments of 1) was used in the k -NNs algorithm, the drift measures displayed clear patterns in the power rates as the drift level and the drift type changed. The larger the drift level, the higher the power. The b -drifted items were constantly better detected compared with those with ab -parameter drift.

Table 4 provides power rates of the drift measures when 10% of the items in the item pools were flagged for parameter drift. Comparison between Tables 3 and 4 reveals that the

Table 4. Power Rates When 10% of Item Pools Were Flagged for Parameter Drift.

$\Delta(a, b)$	n	$p = 2$				$p = 3$			
		EU	HE	KL	JS	EU	HE	KL	JS
(0, .3)	500	.713	.734	.743	.740	.717	.746	.753	.753
	1,000	.805	.812	.826	.826	.802	.809	.827	.816
	1,500	.857	.867	.880	.874	.849	.870	.878	.868
(0, .5)	500	.713	.741	.746	.741	.724	.736	.750	.739
	1,000	.807	.810	.824	.827	.795	.820	.824	.818
	1,500	.858	.871	.881	.871	.851	.868	.881	.881
(.5, .3)	500	.707	.729	.736	.725	.712	.741	.734	.736
	1,000	.804	.799	.808	.816	.794	.808	.817	.811
	1,500	.847	.856	.862	.868	.845	.870	.868	.874
(.5, .5)	500	.707	.720	.732	.724	.718	.746	.741	.743
	1,000	.805	.809	.817	.806	.791	.798	.813	.803
	1,500	.853	.849	.879	.860	.845	.860	.873	.863

Note. $\Delta(a, b)$ = changes in a - and b -parameters; n = sample size; p = number of latent dimensions; EU = Euclidean distance; HE = Hellinger distance; KL = Kullback–Leibler divergence; JS = Jensen–Shannon divergence.

proportion of the drift items in the pools had a distinct impact on the power performances of the drift measures. As the item pools included more drift items, the detecting power of the drift measures decreased across all conditions. The extent of the reduction in the power was, however, rather modest; the differences in the power rates occurred less than 0.05 under all conditions controlled for in this study.

Conclusion

Preserving the quality of an item pool is essential for any continuous testing program to ensure that test scores have the same meaning over time. The traditional approach to quality control is to routinely recalibrate items and compare parameter estimates with initial values. When data from MCAT are analyzed for parameter drift, difficulties arise due to the multidimensionality and sparseness in the test data. In this article, significance testing procedures were proposed to identify the presence of parameter drift in MCAT without the need for calibrating items and, thereby, without invoking problems related to calibration or linking error. The drift indices were based on four widely used informational distance/divergence measures such that the heterogeneity in the MIRFs could be summarized into a single numeric value. The significance testing was conducted for the difference between two MIRFs, one from the initial item parameter estimates and the other from operational testing. For approximating the MIRFs from the observed data, the k -NNs algorithm was employed with the random query search method.

The simulation study demonstrated the potential of the distance/divergence measures as a drift measure in the MCAT. The drift statistics had explicit sampling distributions and adequate control over Type I error in the null MCAT administrations. The drift measures produced moderate power under the small samples ($n = 500$) and excellent power as the sample size increased to 1,000. Contamination of the item pools by drift items degraded the power and Type I error performances of the drift measures, yet in a predictable manner.

The procedures developed in this study do not assume any prior model nor require item calibration, and hence, they are inexpensive to use and much less cumbersome than the traditional calibration-based methods. Furthermore, as the procedures need only the knowledge about the response function for computing the drift statistics, they can be easily generalized to other

parametric models. The *k*-NNs technique introduced in this study also promises high potential for approximating the MIRFs in operational testing. Despite changes of samples (and therefore, in data sparseness) over time, it always finds proper sets of matching variables for comparing the MIRFs with comparable precision.

The performance of the distance/divergence measures can be further examined in future studies by systematically varying the item parameter values, the level of parameter drift, the degree of item pool contamination, and so on. In addition, since both the drift analysis and DIF analysis concern whether an item functions the same in different sets of data within the IRT framework, performances of the distance/divergence measures may well be investigated in detecting the existence of DIF. DIF studies will call for additional deliberation on factors such as sample sizes or proficiency differences between examinee groups.

Another issue that merits a systematic study is whether the distance/divergence measures would remain desirable in the presence of MIRF misspecification. As one anonymous reviewer pointed out, items may falsely be flagged for parameter drift because of misfit of the response model in spite of the absence of drift. Previous research has shown that even small amounts of model misfit can result in serious Type I error inflation for parametric DIF detection procedures (Bolt, 2002). A future investigation might well focus on to what extent model misfit affects the performance of the drift measures and whether the distance/divergence measures might offer advantages over fully parametric calibration-based procedures. The present study relied on the assumptions that the item response model being used for MCAT is precisely specified, and the item pools from which items are selected are precalibrated with enough accuracy. Hence, robustness of the drift procedures against the violations of these assumptions will warrant future discussion. A systematic study of this issue would entail an evaluation of Type I error performances under minor violations of the response model or based on an empirical data set in which no parameter drift is expected.

Appendix

Determination of *Q*

Table A1 provides power and Type I error performances of the drift measures as a function of *Q* values under consideration. Results were averaged across all simulation factors—the drift type, drift level, *n*, and the proportion of drift items in the item pools—under the two-dimensional CAT to get a clear picture of the impact of *Q*. Table A1 suggests that the Euclidean distance generally resulted in the highest power when the *Q* equaled seven or eight. The Hellinger

Table A1. Power and Type I Error Rates in the Two-Dimensional CAT.

Criterion		Number of query vectors (<i>Q</i>)									
		6	7	8	9	10	11	12	13	14	15
EU	Power	.795	.806	.807	.800	.792	.785	.773	.759	.744	.737
	Type I	.073	.087	.091	.100	.101	.102	.098	.098	.090	.092
HE	Power	.764	.792	.809	.816	.816	.815	.811	.799	.789	.778
	Type I	.053	.069	.082	.102	.113	.124	.129	.129	.126	.124
KL	Power	.797	.816	.827	.822	.806	.787	.766	.737	.712	.688
	Type I	.064	.091	.113	.132	.139	.140	.135	.131	.123	.119
JS	Power	.775	.802	.818	.824	.820	.815	.808	.793	.782	.770
	Type I	.052	.072	.091	.113	.124	.131	.133	.134	.128	.123

Note. Power = power rate; Type I = Type I error; CAT = computerized adaptive testing; EU = Euclidean distance; HE = Hellinger distance; KL = Kullback–Leibler divergence; JS = Jensen–Shannon divergence.

Table A2. Power and Type I Error Rates in the Three-Dimensional CAT.

		Number of query vectors (Q)									
	Criterion	6	7	8	9	10	11	12	13	14	15
EU	Power	.782	.798	.801	.806	.801	.801	.795	.787	.783	.776
	Type I	.079	.096	.103	.110	.112	.114	.114	.110	.109	.111
HE	Power	.769	.791	.805	.814	.819	.823	.822	.817	.817	.812
	Type I	.055	.067	.087	.096	.110	.122	.127	.126	.132	.135
KL	Power	.802	.819	.826	.827	.824	.818	.808	.790	.778	.765
	Type I	.069	.093	.115	.128	.139	.147	.147	.142	.140	.140
JS	Power	.782	.803	.815	.822	.825	.828	.824	.816	.815	.807
	Type I	.058	.075	.095	.109	.122	.132	.135	.135	.138	.138

Note. Power = power rate; Type I = Type I error; CAT = computerized adaptive testing; EU = Euclidean distance; HE = Hellinger distance; KL = Kullback–Leibler divergence; JS = Jensen–Shannon divergence.

distance appeared most sensitive to the presence of parameter drift when the Q equaled nine, 10, or 11. The KL and JS divergence measures seemed to produce the highest power when the Q was set eight and nine, respectively. The performances of the drift measures under the three-dimensional CAT were less distinguishable than under the two-dimensional CAT. Results were, however, generally in agreement with the two-dimensional case. Table A2 indicates that the Q values between eight and 11 produced the most comparable performances for the Euclidean distance under the three-dimensional CAT. The Hellinger distance was found to perform most effectively when the Q values were between 10 and 12. The KL and JS divergences displayed the highest power when the Q values were between eight and 10 and between nine and 12, respectively. Note that Type I error inflation in Tables A1 and A2 was expected because the results were obtained from the drift cases.

Acknowledgment

The authors would like to thank Associate Editor, Dr. Daniel Bolt, and anonymous reviewers for their constructive comments and suggestions to improve the quality of the paper.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67-91.
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281-305.
- Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of Calcutta Mathematical Society*, 35, 99-109.
- 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 annual meeting of the Psychometric Society, Montreal, Quebec, Canada.

- Bock, R. D., Muraki, E., & Pfeifferberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, 25, 275-285.
- Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, 15, 113-141.
- Chan, K.-Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a Cognitive Ability Test battery. *Journal of Educational Measurement*, 84, 610-619.
- Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80, 1-20.
- Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20, 213-229.
- Clark, A. K. (2014, April). *Parameter drift methodology and operational testing application*. Poster presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.
- Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement*, 25, 31-45.
- DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, 17, 265-300.
- Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, 22, 33-51.
- Glas, C. A. W. (2000). Item calibration and parameter drift. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 183-199). Norwell, MA: Kluwer Academic.
- Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, 20, 369-377.
- Guo, R., Zheng, Y., & Chang, H.-H. (2015). A stepwise test characteristic curve method to detect item parameter drift. *Journal of Educational Measurement*, 52, 280-300.
- Kim, S.-H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement*, 15, 269-278.
- Kim, S.-H., Cohen, A. S., & Park, T. H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, 32, 261-276.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 79-86.
- Li, X. (2008). *Multidimensionality and item parameter drift: An investigation of linking items in a large-scale certification test* (Unpublished doctoral dissertation). Michigan State University, East Lansing.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495-502.
- Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21, 24-43.
- Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25-36.
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61, 331-354.
- Sykes, R. C., & Ito, K. (1993, April). *Item parameter drift in IRT-based licensure examinations*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.
- Tam, S. S. (1992). *A comparison of methods for adaptive estimation of a multidimensional trait* (Unpublished doctoral dissertation). Columbia University, New York, NY.
- Veerkamp, W. J. J., & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, 25, 373-389.
- Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement*, 40, 307-330.
- Zhang, J. (2014). A sequential procedure for detecting compromised items in the item pool of a CAT system. *Applied Psychological Measurement*, 38, 87-104.
- Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, 64, 213-249.