Educational and Psychological Measurement

http://epm.sagepub.com/

Item Selection for the Development of Parallel Forms From an IRT-Based Seed Test Using a Sampling and Classification Approach

Pei-Hua Chen, Hua-Hua Chang and Haiyan Wu

Educational and Psychological Measurement 2012 72: 933 originally published online 16 May 2012

DOI: 10.1177/0013164412443688

The online version of this article can be found at:
http://epm.sagepub.com/content/72/6/933

Published by:
SAGE
http://www.sagepublications.com

Additional services and information for *Educational and Psychological Measurement* can be found at:

Email Alerts: http://epm.sagepub.com/cgi/alerts

Subscriptions: http://epm.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Oct 16, 2012

OnlineFirst Version of Record - May 16, 2012

What is This?

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

Educational and Psychological
Measurement
72(6) 933–953
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164412443688
http://epm.sagepub.com

Item Selection for the Development of Parallel Forms From an IRT-Based Seed Test Using a Sampling and Classification Approach

Pei-Hua Chen[1], Hua-Hua Chang[2], and Haiyan Wu[3]

Abstract

Two sampling-and-classification–based procedures were developed for automated test assembly: the Cell Only and the Cell and Cube methods. A simulation study based on a 540-item bank was conducted to compare the performance of the procedures with the performance of a mixed-integer programming (MIP) method for assembling multiple parallel test forms. The study investigated the statistical equivalence of the forms generated by the three test assembly methods (Cell Only, Cell and Cube, and MIP) in terms of test information functions, test characteristic curves, mean square deviations, and practical constraints, such as content balancing and nonoverlap among forms. The results indicated that the 13-point MIP method outperformed the other two methods in terms of the "closeness" test information functions between the reference form and the assembled parallel tests. Regarding test characteristic curves, the Cell Only and Cell and Cube methods yielded more similar test characteristic curves than the MIP method. Constraining test information functions apparently does not guarantee that the assembled forms will yield similar test characteristic curves. Overall, the Cell Only and Cell and Cube methods have the potential to provide results similar to the optimization approach.

[1]National Chiao Tung University, HsinChu, Taiwan
[2]University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA
[3]Florida State University, Tallahassee, FL, USA

Corresponding author:
Pei-Hua Chen, National Chiao Tung University, Department of Management Science, 1001 University Road, HsinChu 30010, Taiwan
Email: paulachen@g2.nctu.edu.tw

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

Keywords

automated test assembly, randomization, seed test

Large-scale assessments are usually required to administer tests at multiple locations and sessions. Testing agencies and organizations, therefore, typically desire to generate multiple parallel test forms. In recent years, advanced innovations in computer technology and psychometric theory have led to changes in test assembly practices. Most researchers (Breithaupt & Hare, 2007; Finkelman, Kim, & Roussos, 2009; Luecht, 1998; Sanders & Verschoor, 1998; Swanson & Stocking, 1993; van der Linden, 2005; Veldkamp, 2002) solve test assembly problems by using the *constrained combinatorial optimization* approach, which seeks the values of the decision variables that optimize the objective function, subject to specified constraints (Hillier & Lieberman, 2001). In such a context, a test assembly problem can be treated as trying to match a certain target test information function (objective function) subject to test length and content coverage (constraints).

Optimization approaches to automated test assembly (ATA) identify optimal solution under the specific constraints of statistical and content-related test specifications within a given item bank. Therefore, having successively constructed forms means that items selected for previous forms are removed from the item pool, and new forms are constructed with the remaining items. Thus, forms built later are not nearly as parallel as forms built earlier. A heuristic correction or replacement phase (Swanson & Stocking, 1993; van der Linden, 2005) is usually required to avoid such a problem. Although multiple-form assembly problems can be simultaneously modeled and constructed, the increase in computation time is commensurate with the expansion in the size of the problem. Furthermore, optimization approaches find the optimal test forms based mainly on the objective function; therefore, test forms are not built by means of uniform sampling, meaning that there is an equal chance for each feasible test to be selected (Belov & Armstrong, 2005).

A random search approach has been recently introduced into the area of ATA (Belov & Armstrong, 2005; Chen, 2005; Chen & Chang, 2005, 2006). Unlike the optimization approaches that pick the best possible solutions first, the random search approach has the demonstrable advantage of producing uniform tests (Belov, 2008). One test assembly method, the Monte Carlo random search with tabu search elements (Glover, Taillard, & de Werra, 1993) proposed by Belov and Armstrong (2005), has the major advantage of uniform test assembly. Based on the random search characteristics, each form has an equal chance of selection, thereby avoiding the sequential degradation problem without any need for a second step modification. The present study also proposes a similar randomization-based approach to assemble multiple parallel test forms.

The goal of the present study was to develop two random sampling and classification approaches—the Cell Only and the Cell and Cube methods—to build multiple forms based on a reference form (i.e., a target test or a seed test). The performance of

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

the proposed methods was measured against the baseline of one of the commonly used approaches in test assembly: the mixed-integer programming (MIP) method. This article is organized as follows. We briefly review optimization approaches and follow with discussion of some of the issues encountered in multiple-form ATA. We then provide the rationale for our sampling and classification methods. Finally, we compare test assembly methods using several criteria, such as mean square deviations between the reference form and the assembled form, target test information functions, and test characteristic curves.

A Brief Review of Optimization Approaches in ATA

Previous studies based on optimization approaches have used the MIP (Breithaupt & Hare, 2007; van der Linden, 2005; Veldkamp, 2002) and the enumerative heuristic methods (Finkelman et al., 2009; Luecht, 1998; Luecht & Hirsch, 1992; Sanders & Verschoor, 1998; Swanson & Stocking, 1993) to tackle various ATA problems. The MIP method has been popular in test assembly practice mainly because of its flexibility to deal with very complicated test specifications—such as content areas, enemy items, item type, and word counts—as well as its superiority in assembling numerous forms with respect to large item banks. For example, the MIP method can specify a test assembly problem by using a weighted target information function in the objective function with appropriate constraints to construct forms that can yield uniform test information functions along the ability scale.

An effective solution to a complicated and large MIP problem usually requires sophisticated algorithms. Commercial optimization software packages (usually called solvers), such as the IBM ILOG CPLEX Optimization Studio 12 (IBM, 2009), have readily built-in algorithms, and they require only very little customized programming. Large or mid-sized testing organizations can benefit from commercially available solvers to handle computationally intensive real assembly problems. Furthermore, MIP results provide a global optimal solution, whereas other heuristic approaches generally produce a suboptimal solution.

Although the MIP method offers a powerful solution to test assembly problems, the cost of a commercial license for specialized solvers ranges from hundreds to thousands of U.S. dollars, which makes it difficult for small or not-for-profit testing agencies to afford (Davey, 2009). Although free-version solvers are available, such as the standard solver built in to Microsoft Excel 2007 (Cor, Alves, & Gierl, 2008, 2009), limits on decision variables and constraints make them unsuitable for handling most real test assembly problems. Moreover, while researchers can write their own MIP code in any programming language, such an undertaking is usually very complicated and time consuming. Therefore, practitioners are greatly interested in employing alternative approaches, such as heuristic methods, which can use existing programming software to assemble test forms, particularly for some testing programs with relatively simple test specifications.

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

Enumerative heuristics like the Weighted Deviation Model (Swanson & Stocking, 1993) or the Normalized Weighted Absolute Deviation Heuristic (NWADH) proposed by Luecht (1998), on the other hand, do not rely on commercial optimization software packages, and they often provide a relatively quick solution. Enumerative heuristics, however, often require computation of the differences between the target values and the current values at each iteration, and therefore, the computation time is proportional to item pool size and test length. Building tests using enumerative heuristics typically requires customized programming, and therefore, it usually involves intensive computation for tests with complicated specifications (Luecht, 1998; Swanson & Stocking, 1993).

Issues in Constructing Parallel Test Forms

Meaning of Parallel Test Forms

Parallel test forms are required to be equivalent in terms of their statistical and content-related properties. When the test information functions of alternative test forms are identical, the statistical equivalence of these forms is called weakly parallel in an item response theory (IRT) framework (Samejima, 1977). Most test assembly studies restrict the test information function within a certain range of the target by matching information functions pointwise along the ability scale (Luecht, 1998; van der Linden, 2005). For a new testing program, generating test forms that target a given shape of the test information function is generally useful. There is no need to equate the forms to the reference form.

Although computerized test assembly typically uses test information functions when evaluating parallel test construction, traditional paper-and-pencil forms are assembled according to score distributions. Matching forms using test information functions does not guarantee matched score distributions (Wightman, 1998). However, for an existing assessment, reference forms are available, and it makes sense to build new forms based on the reference form for at least two reasons. First, matching forms item by item guarantees matched test information functions and score distributions. Second, because testing agencies have expended considerable money and effort validating reference forms, their scores are typically more readily interpretable by test professionals. For example, Armstrong, Jones, and Wu (1992) created parallel test forms to match an existing seed test by formulating the test assembly problem into an MIP model. The present study also built parallel forms by matching an existing reference form, but it did so using a different approach that employed random sampling and classification.

A Sequential Versus a Simultaneous Approach

The major challenge of ATA is to construct multiple parallel forms as opposed to a single form. Multiple test forms can be assembled sequentially or simultaneously. A sequential approach assembles forms successively. When building multiple forms

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

sequentially based on the optimization approach, the algorithm tends to pick the best set of items for the first form, the second best set of items for the second form, and so on. As the number of assembled forms increases, the later forms are not as nearly parallel as the first form. Therefore, a heuristic correction, such as the Big-Shadow-Test (BST) method (van der Linden, 2005) or a replacement phase (Swanson & Stocking, 1993), is needed to successively replace and swap previously selected items until no further improvement can be made to mitigate such a problem (Swanson & Stocking, 1993; van der Linden, 2005).

Conversely, building forms sequentially based on the randomization approach does not cause deterioration problems because of its stochastic nature. Earlier built forms are not necessarily superior to or inferior to their successors. Therefore, it is worth exploring the feasibility of applying the randomization approach in test assembly.

Another approach is to assemble forms simultaneously based on the MIP method as proposed by Boekkooi-Timminga (1990). The simultaneous approach builds multiple test form assembly problems into a single model, using new decision variables to link item-level and form-level variables. This approach builds the required forms all at the same time and is known to be computationally intensive; computation time increases exponentially as problem size grows. Solving an MIP problem requires exponential running time in general, and so the solution time for the MIP method depends highly on the computer processor and the memory space in the hardware. For example, using quad core capability processors can halve the solution time or even reduce it further. Such high-speed processors and high-capacity memories are quite common in most modern computers, so they can help reduce the solution time for large MIP problems.

Rationale for the Sampling and Classification Approaches

The rationale for the sampling and classification approaches is somewhat similar to the method of ''matched random subsets,'' which constructs parallel tests based on the classical test theory proposed by Gulliksen (1950). Under the matched random subsets method, items that are close to one another on the scatter diagram of item difficulty (p) and the point–biserial correlation with total test score (r_{pbs}) are grouped into ''subsets.'' Items selected from each subset are randomly assigned to each of the required forms. Matching items based on item statistics and assigning them to subsets ensures parallel subsets and thus parallel test forms (Gulliksen, 1950).

The IRT-based sampling and classification approach to ATA uses a similar rationale to use the joint distribution of item parameter values, which determine the individual item information functions and the collective test information function. In a three-parameter logistic (3PL) IRT model, the item information function is determined by three item parameters—for item discrimination, item difficulty, and item guessing (a_i, b_i, c_i)—and one examinee ability parameter (θ). If $\hat{\theta}$ is approximated close to b_i, then the larger the value of a_i, the larger the value of item information.

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

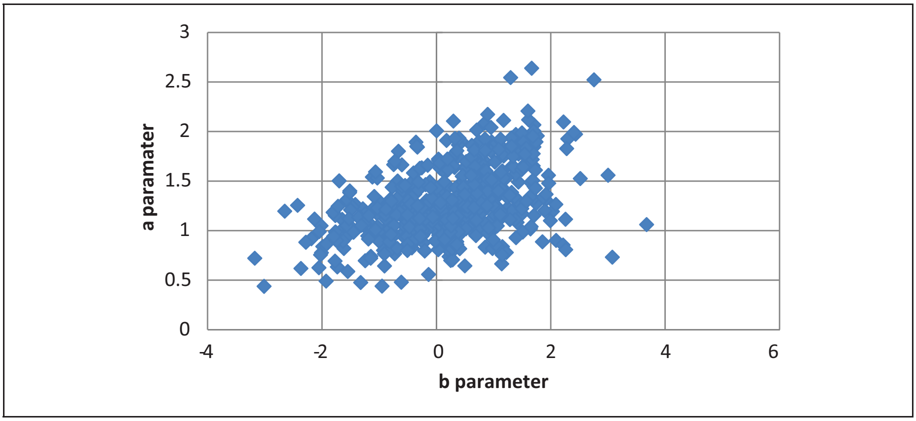

Figure 1. An example of item pool categorization for scatter plot of *a*- and *b*-parameters

Item information values, in other words, are mainly determined by the item discrimination parameter (a_i), which is conditional on ability level.

The relationship of item information functions to the item parameters used in computerized adaptive testing can also be applied in ATA. If the joint distribution of *a*- and *b*-parameters for a reference form is already known, then the target test information curve across ability levels can be determined. That is, mimicking the item parameter distribution of the reference form results in achieving the target test information. Therefore, if the scatter plot of the *a*- and *b*-parameters in the reference form is provided, the task of assembling a statistically equivalent test becomes the task of creating another scatter plot, with the new test having a joint distribution of *a*- and *b*-parameters similar to the reference form.

Cell Only Method

Mimicking the scatter plot of *a*- and *b*-parameters in the reference form is not easy. Classification of the item pool and reference form, however, can play a crucial role in simplifying the process. Items in the pool and reference form can be partitioned into M groups from $m = 1, \ldots, M$ according to their *a*-parameters. Similarly, items can be divided into N groups from $n = 1, \ldots, N$ according to their *b*-parameters. Thus, the scatter plot of the *a*- and *b*-parameters for the whole item pool and the reference form can be categorized into $M \times N = MN$ cells. Figure 1 shows a representation of the item pool categorization for the Cell Only method. The values of *a* parameters in Figure 1 are between 0 and 3 and are categorized into six groups, whereas the values of *b* parameters are between -4 and 4 and are categorized into four groups. Item selection under this method involves randomly selecting the number of items within each cell in the reference form from the item pool such that the number of items from each cell in the reference form and in the assembled test is the same. This process

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

approximates the joint distribution of the *a*- and *b*-parameters and is called the Cell Only method.

Cell and Cube Method

Although the Cell Only method appears to be a reasonable approach, some items in the pool remain unselected if there is no item in the corresponding cell in the reference form. To improve item pool utilization and the item exposure rate, the present study introduced a two-stage sampling method: the Cell and Cube method.

Van der Linden (2005) proposed a BST method to solve a large simultaneous ATA problem by treating it as a sequence of smaller simultaneous problems. Current constructed forms are assembled along with a shadow test that includes the remaining test items in the pool. Items in the shadow test are returned to the pool once current forms are built. This process is repeated until all test forms are constructed.

The Cell and Cube method is similar to the BST method in the sense that both decompose a big problem into a series of smaller subproblems and the test items are partially filled up at each stage. Under the BST method, the variation in the number of subproblems depends on the number of forms one tries to build, whereas under the Cell and Cube method there are only two stages, making it a special case of the BST method.

Luecht and Hirsch (1992) proposed a heuristic search to iteratively calculate the moving average of the differences between target information and previously selected items as the basis for sequentially selecting the next item. This algorithm tends to select items with moderate information at each iteration. The proposed Cell and Cube method tries to improve the selection probability of items with relatively low or high information. This method is described below.

Suppose the test length of the reference form is n and that n_1 items were selected during the cell stage. Then, n_2 items will be selected during the cube stage, where $n = n_1 + n_2$. Test information functions are smooth and well-behaved continuous curves, and optimizing a test information function with three to five well-chosen ability points is appropriate (van der Linden, 2005). Suppose we consider T ability points at θ_t $(t = 1, \ldots, T)$. We then have T target test information values $[T(\theta_1), T(\theta_2), \ldots, T(\theta_T)]$. The value of the target test information $T(\theta_t)$ at ability point θ_t can be decomposed into two parts: the target test information for the cell stage $(T_1(\theta_t))$ and the target test information for the cube stage $(T_2(\theta_t))$, where $T(\theta_t) = T_1(\theta_t) + T_2(\theta_t)$. After selecting items during the cell stage, the required test information function for the cube stage can be determined to be $[T_2(\theta_1), T_2(\theta_2), \ldots, T_2(\theta_T)]$. If there are n_2 items to be selected during the cube stage, the average contribution of each item to the new target information function can be written as

$$\left[\frac{T_2(\theta_1)}{n_2}, \frac{T_2(\theta_2)}{n_2}, \ldots, \frac{T_2(\theta_T)}{n_2}\right] = [\bar{T}_2(\theta_1), \bar{T}_2(\theta_2), \ldots, \bar{T}_2(\theta_T)].$$

The item information at θ_t for each item i in the subpool can be calculated as $[I_i(\theta_1), I_i(\theta_2), \ldots, I_i(\theta_T)]$.

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

The standard deviations of the item information functions at T ability points in the item pool can also be obtained, $(s(\theta_1), s(\theta_2), \ldots, s(\theta_t))$. Because only a few items are selected in the cell stage, using the complete item pool to calculate the standard deviation of item information at three ability points is appropriate. If the subpool is divided into K categories by T item information functions at T ability levels, the subpool will be categorized into K^T cubes. The partition points of the K categories can be determined by the required average target information function per item $\bar{T}_2(\theta_t)$ and the standard deviation of the item information $s(\theta_t)$ at tth ability point (θ_t). When three ability points ($T = 3$, $\theta_t = -1$, 0, and 1) and three item information categories ($K = 3$) are considered, the information function of an item for each ability point θ_t can be categorized as below $\bar{T}_2(\theta_t) - s(\theta_t)$, between $\bar{T}_2(\theta_t) \pm s(\theta_t)$, and above $\bar{T}_2(\theta_t) + s(\theta_t)$. The cutoff points for dividing the item subpool into 3^3 cubes are $\bar{T}_2(\theta_t) - s(\theta_t)$ and $\bar{T}_2(\theta_t) + s(\theta_t)$, where $\theta_t = -1$, 0, and 1.

Now each item can have an address d_{θ_t} $(\theta_t = -1, 0, 1)$ for the cube it belongs to (d_{-1}, d_0, d_1). Any item i in the subpool can be put into one of the three groups according to its item information $I_i(\theta_t)$ at each ability point θ_t:

$$I_i(\theta_t) \leq \bar{T}_2(\theta_t) - s(\theta_t), \quad (d_{\theta_t} = -1), \tag{1}$$

$$\bar{T}_2(\theta_t) - s(\theta_t) < I_i(\theta_t) < \bar{T}_2(\theta_t) + s(\theta_t), \quad (d_{\theta_t} = 0), \tag{2}$$

and

$$I_i(\theta_t) \geq \bar{T}_2(\theta_t) + s(\theta_t), \quad (d_{\theta_t} = 1). \tag{3}$$

Matrix 4 shows the categorization into three groups of the information function values at three ability levels ($\theta_t = -1$, 0, 1), which can be coded accordingly as in Matrix 5. Figure 2 shows a representation of the information function divided into three groups at three ability points.

$$\begin{bmatrix} \leq \bar{T}_2(-1) - s(-1), & \leq \bar{T}_2(0) - s(0), & \leq \bar{T}_2(1) - s(1) \\ \bar{T}_2(-1) - s(-1) \sim \bar{T}_2(-1) + s(-1)_, & \bar{T}_2(0) - s(0) \sim \bar{T}_2(0) + s(0)_, & \bar{T}_2(1) - s(1) \sim \bar{T}_2(1) + s(1) \\ \geq \bar{T}_2(-1) - s(-1), & \geq \bar{T}_2(0) + s(0), & \geq \bar{T}_2(1) + s(1) \end{bmatrix} \tag{4}$$

$$\begin{bmatrix} d_{-1} = -1, & d_0 = -1, & d_1 = -1 \\ d_{-1} = 0, & d_0 = 0, & d_1 = 0 \\ d_{-1} = 1, & d_0 = 1, & d_1 = 1 \end{bmatrix} \tag{5}$$

After the subpool is categorized into 27 cubes, each item can only belong to one cube. Now, we can select n_2 items in the subpool according to the following algorithm:

Step 0: Items are selected in pairs (x and y).

Step 1: Select item x randomly from the subpool and notice its address according to the cube it belongs to, denoted $[d_{x,-1} = -1,\ d_{x,0} = 0,\ d_{x,1} = -1]$.

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

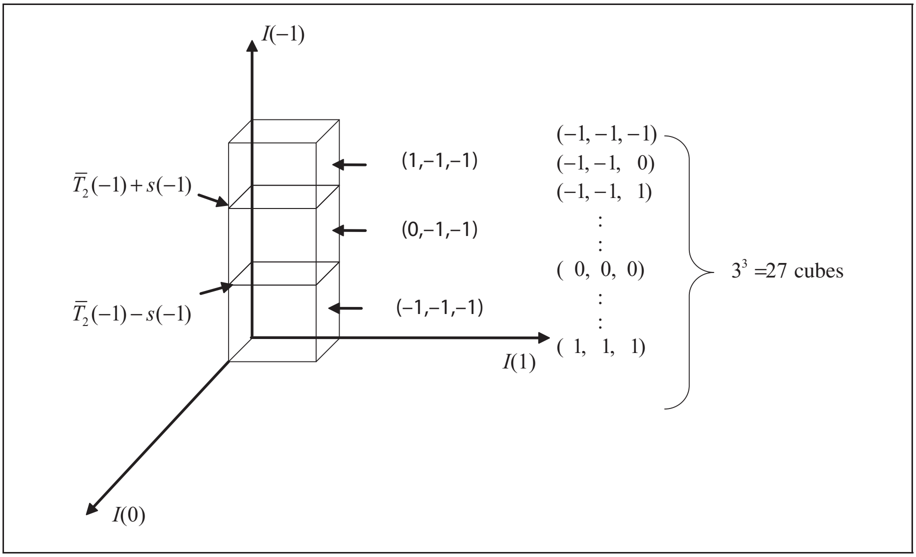

Figure 2. Graphic representation of 27 cubes in the subpool

Step 2: Select item y such that its address, denoted $[d_{y,-1}=1,\ d_{y,0}=0,\ d_{y,1}=1]$, satisfies $(d_{x,-1}+d_{y,-1},\ d_{x,0}+d_{y,0},\ d_{x,1}+d_{y,1})=\ (0,\ 0,\ 0)$.

Step 3: Repeat Steps 1 and 2 until all n_2 items are selected.

Figure 3 shows a representation of the selection for each pair of items. The purpose of using the algorithm is to select each pair of items so that the sum of these n_2 information functions equals the target test information $T_2(\theta_t)$ at the cube stage, by keeping the average of the two items in proximity to the average of the target information per item $(\bar{T}_2(\theta_t))$. If the number of required items in the cube stage (n_2) is odd, the last item will be selected based on the difference between the current sum of the item information functions $\sum_{i=n_1+1}^{n-1} I_i(\theta_t)$ of the n_2-1 items from the cube stage and the target test information for the cube stage $(\bar{T}_2(\theta_t))$ at each ability level θ_t. If for any selected item x there cannot be found a corresponding item y that satisfies the requirement that the sum of the addresses of item x and y is zero $(d_{x,\theta_t}+d_{y,\theta_t}=0)$ for each ability point θ_t, the selected item x will not be included in the constructed test. Another item will be selected until a corresponding item can be found. In this way, the average item information values of each pair of selected items at each ability level will be close to the average required test information function per item $(\bar{T}_2(\theta_t))$ at ability point θ_t $(t=1,\ \ldots,\ T)$. As a result, the sum of the selected items is close to the required test information function at ability level θ_t $(t=1,\ \ldots,\ T)$.

Method

One 540-item retired item pool from a large-scale assessment with three content areas was used. The reference form was composed of 30 items—12 items from Content

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

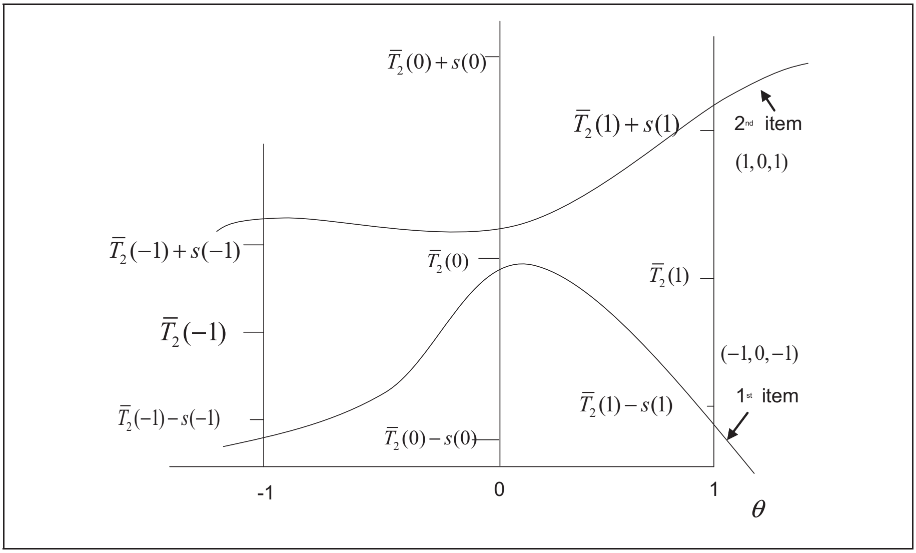

Figure 3. An example of selecting one pair of items in the Cube stage

Note: The first item has an address of $(-1, 0, -1)$, and the second item (corresponding item) has an address of $(1, 0, 1)$.

Area A, 9 items from Content Area B, and 9 items from Content Area C. Five forms were constructed by means of each test assembly method (MIP, Cell Only, and Cell and Cube) based on the reference form.

Item Pool and Reference Form Classification

Using the Cell Only method and executing the cell stage in the Cell and Cube method require the classification of the item pool and the reference form according to their item discrimination (a) and item difficulty (b) parameters. Because there were three content areas, the item pool and reference form were first divided into three subpools and three target subtests by content. Both the a- and b-parameters in the target subtest and item subpool for each content area were then classified. Preliminary experimentation determined that the use of only four a-groups and four b-groups was inadequate to provide a scatter plot similar to that of the reference form. Therefore, the scatter plots of the a- and b-parameters for the subpools and the target subtests were divided into four times eight—that is, 32—cells for the sampling and classification approach. First, the a-parameters were categorized into four groups: (a) $a \leq 0.7$, (b) $0.7 < a \leq 1.0$, (c) $1.0 < a \leq 1.3$, and (d) $a > 1.3$. Similarly, the b-parameters were divided into eight groups: (a) $b \leq -3.0$, (b) $-3.0 < b \leq -2.0$, (c) $-2.0 < b \leq -1.0$, (d) $-1.0 < b \leq 0.0$, (e) $0.0 < b \leq 1.0$, (f) $1.0 < b \leq 2.0$, (g) $2.0 < b \leq 3.0$, and (h) $b > 3.0$.

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

Item Selection Method

MIP method. To determine whether a sampling and classification approach could produce results the same or better than results obtained with MIP, a Minimax model (van der Linden, 2005) of the MIP method was used as a baseline. The Minimax model ensures that the largest deviation between the information function of the assembled test and the target test is minimized by means of a fixed test information function. That is, the Minimax model builds test forms by minimizing the absolute maximum deviation from the points of the specified target test information function. The software GAMS (GAMS, 2007) and the CPLEX solver (IBM, 2009) were used to solve the Minimax model. Considered in the formula of the Minimax model were two different sets of reference form values at 5 and 13 ability points ($\theta = (-2, -1, 0, 1, 2)$ and $\theta = (-3, -2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3)$), along with three content constraints of 12, 9, and 9 items for each content area. Variables were set up to link item-level and test-level constraints. Five 30-item forms were constructed simultaneously for each set of ability points to fit the target test information function, with the constraints of number of forms, content coverage, and no overlap among forms. Separate objective functions with random coefficients were used to deal with the simultaneous approach, multiple-forms test assembly problem, as suggested by Luecht (1998).

Cell Only method. The item pool and reference form were divided into three subpools and three target subtests according to the content areas. Five subtest forms from each content area were randomly selected sequentially without replacement from three subpools according to the number of items per cell in the target subtest, to construct five 30-item test forms.

Cell and Cube method. The first stage of the Cell and Cube method randomly selected 10, 7, and 7 items from the tests assembled using the Cell Only method from Content Areas A, B, and C, respectively. The remaining items in each content area created one new subpool for the Cube stage. The values of the item information function at the three ability points $(-1, 0, 1)$ were calculated for each item in each new subpool. The test information function was obtained for each subtest form from each content area constructed using the Cell Only method. By similar means, the sums of the selected 10-, 7-, and 7-item information values from the cell stage in the three content areas for each subtest form were also computed. The required subtest information function values for each content area at the three ability points $(-1, 0, 1)$ for the cube stage were obtained using the test information function from the Cell Only method for each subtest form, minus the sum of the selected items at the cell stage.

Furthermore, the averages of each assembled item that contributed to the cube stage of the target information function for the five forms in each content area were obtained. The standard deviations of the item information function at the three ability points $(-1, 0, 1)$ in each new subpool were also calculated. The cutoff points for dividing the cubes in each new pool for five forms differed, because the required test information functions at the cube stage for the five forms were not identical. Finally, the three new subpools were divided into $3 \times 3 \times 3 = 27$ cubes using three item information functions at the three ability levels $(-1, 0, 1)$.

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

Evaluation Criteria

Several criteria were used to evaluate the results of applying the different methods for constructing forms. The first criterion was whether the test information functions produced by the three methods were similar. The test information functions of all the assembled tests and the reference form were plotted on the same scale, because one characteristic of the Minimax model (van der Linden, 2005) is that, if the two test information curves in the Minimax model are as close as possible, the tests are treated as statistically equivalent. That is, neither positive nor negative deviations from the target information are desired under this model. Therefore, the test characteristic curves of the three methods were compared visually to determine whether they were close to one another. Test characteristic curves relate ability to number-correct true scores, and they are widely used in test equating and scaling. Therefore, the test characteristic curve was also included in the evaluation criteria.

The second criterion was a mean square deviation (MSD) statistic, which was used to evaluate the closeness of fit between the assembled tests and the reference form, both for the test information functions and test characteristic curves. The value of the mean square deviation was determined by calculating the deviations of points from their desired target value, summing the measurements, and then dividing by the number of points:

$$\frac{\sum_{i=1}^{n}[(TI_{\text{obs}}(\theta_i)-TI_{\text{tgt}}(\theta_i)]^2}{n}, \tag{6}$$

$$\frac{\sum_{i=1}^{n}[(TC_{\text{obs}}(\theta_i)-TC_{\text{tgt}}(\theta_i)]^2}{n}. \tag{7}$$

In Equation 6, $TI_{\text{obs}}(\theta_i)$ indicates the test information of an assembled test at ability level θ_i, for every $i=1,\ldots,n$. Similarly, $TI_{\text{tgt}}(\theta_i)$ indicates the target test information function values at ability level θ_i, for every $i=1,\ldots,n$. In Equation 7, $TC_{\text{obs}}(\theta_i)$ indicates the test characteristic curve of an assembled test at ability level θ_i, for every $i=1,\ldots,n$. Similarly, the $TC_{\text{tgt}}(\theta_i)$ indicates the test characteristic curve of the reference form. In the present study, 61 ability points, ranging from −3 to 3 in increments of 0.1 units, were considered for both the test information function and the test characteristic curve.

Results

Figures 4 and 5 show the test information functions resulting from the MIP method, constraining on 5 and 13 ability points. Figure 4 shows that the information functions fit almost perfectly around the 5 ability point constraints. The item information functions deviated from the targets, both positively and negatively, between ability levels ranging from 0 to 1 and 1 to 2. This pattern indicates that only specifying three to five

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

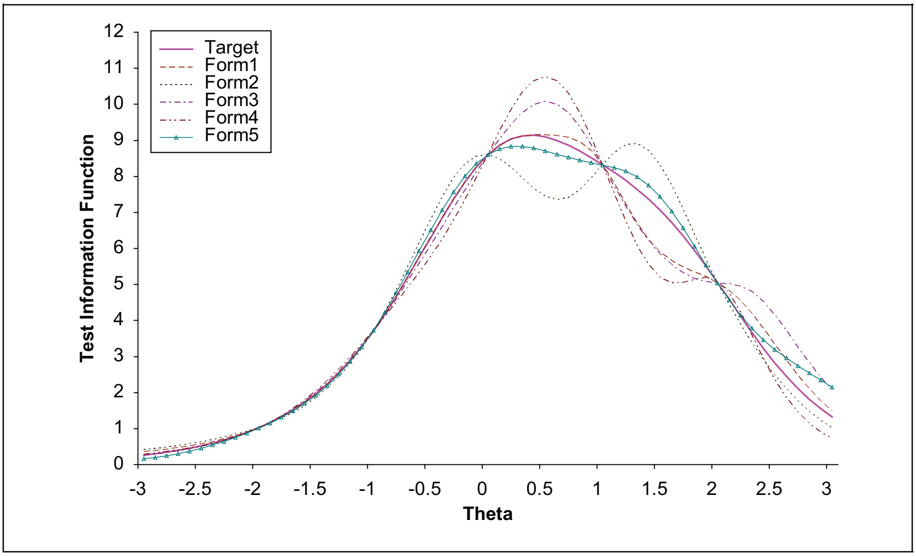

Figure 4. Test information functions of the 5-point mixed-integer programming method

target test information function values in the MIP method may be insufficient. Figure 5 shows that almost every assembled curve of the test information function was close to the target information function at every ability point.

Figures 6 and 7 show the test information function curves from the Cell Only and Cell and Cube methods. In the case of the Cell Only method, one form of the test information curves deviated from the target moderately in the positive direction, ranging from ability points −0.5 to 0.5, whereas another form also deviated positively between ability points 1.6 and 2.5. Additionally, two forms of the test information curves deviated slightly from the target in the negative direction between ability points −0.4 and 2. Figure 7 shows that two forms from the Cell and Cube method deviated from the targets in the positive direction, predominantly between ability points 0 and 1. One form deviated slightly from the target in the negative direction, ranging from ability points −0.2 to 1. Two forms deviated from the target, ranging from ability points 1 to 3 in both the positive and negative directions.

Figures 8 and 9 show the test characteristic curves for the MIP method with 5- and 13-point constraints of the test information function. In Figure 8, most of the test characteristic curves shifted to the left in the positive direction, whereas three forms deviated from the target in the negative direction at high ability levels. Unexpectedly, the test characteristic curves of the forms with the 13-point constraints on the test information function shown in Figure 9 were not close to each other. The deviations increased in the positive direction when ability levels decreased, starting from the ability point of 1. For the Cell Only method, most of the test characteristic curves in Figure 10 were close to the target, except for one form that slightly deviated in the

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

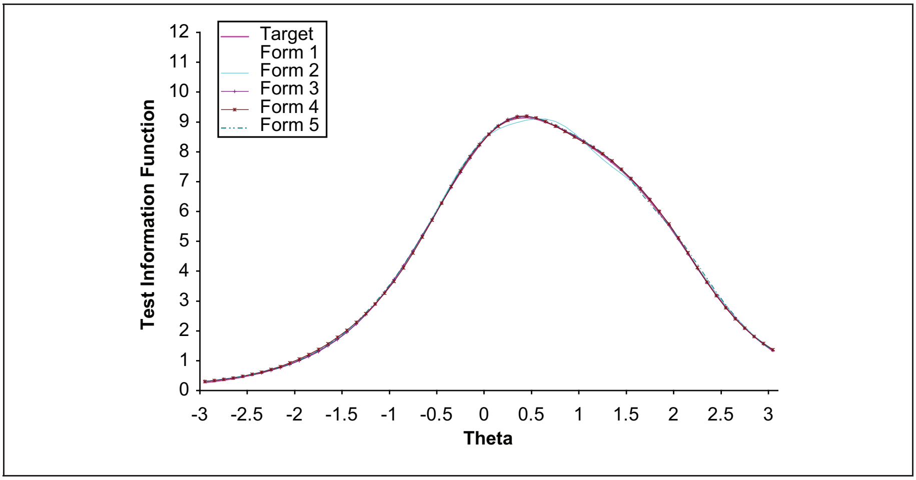

Figure 5. Test information functions of the 13-point mixed-integer programming method

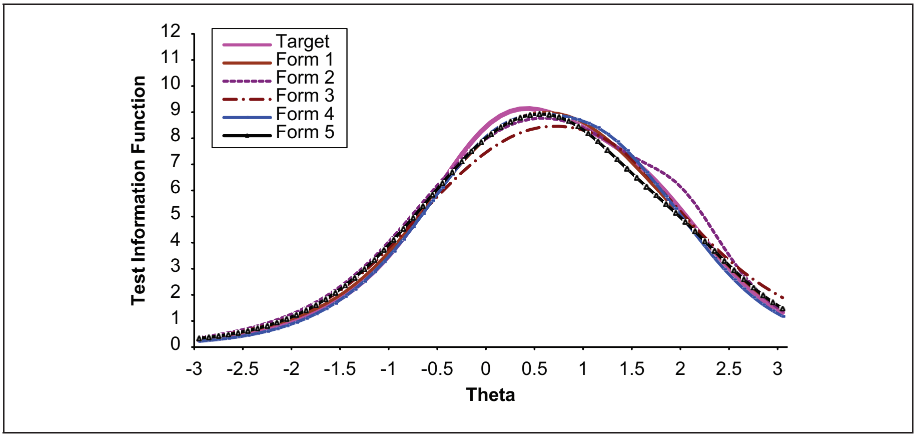

Figure 6. Test information functions of the Cell Only method

negative direction for ability points between −0.6 and −3. For the Cell and Cube method, the test characteristic curves in Figure 11 slightly deviated from the target in both the negative and positive directions. The test characteristic curve results show, overall, that the Cell Only and Cell and Cube methods outperformed the MIP methods. The 13-point MIP method did not generate closer test characteristic curves, as expected. The 5-point MIP method yielded the largest deviations from the target.

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

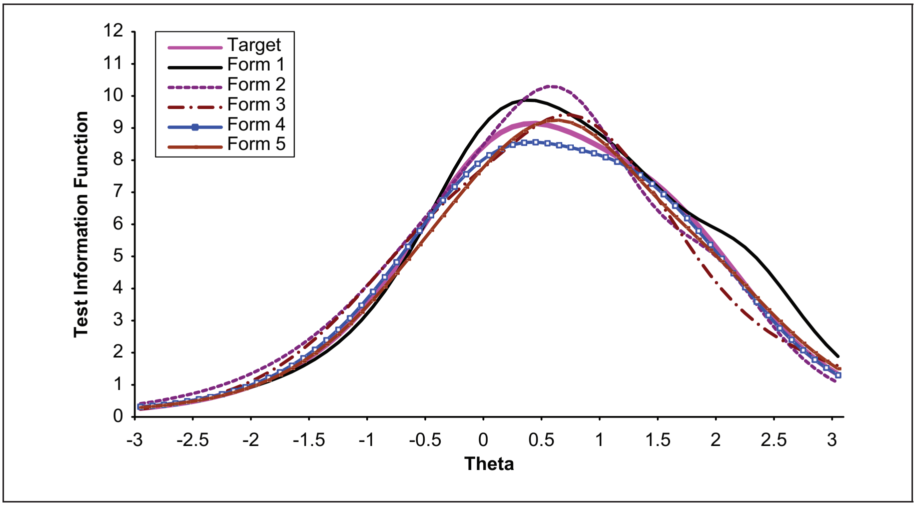

Figure 7. Test information functions of the Cell and Cube method

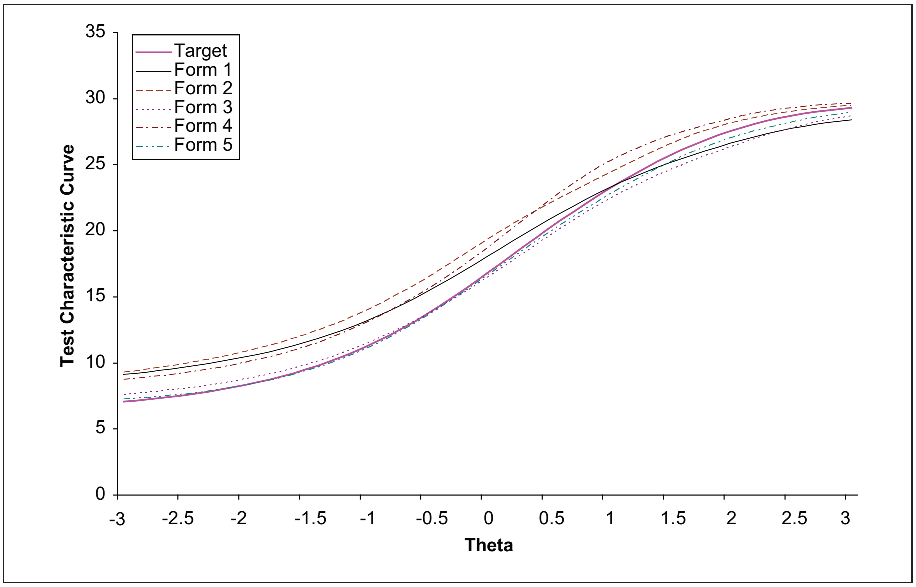

Figure 8. Test characteristic curves of the 5-point mixed-integer programming method

Tables 1 and 2 list the mean, the standard deviation, the maximum and minimum values for the mean square deviation of the five forms from the target test information, and the test characteristic curve at 61 ability points for all test assembly

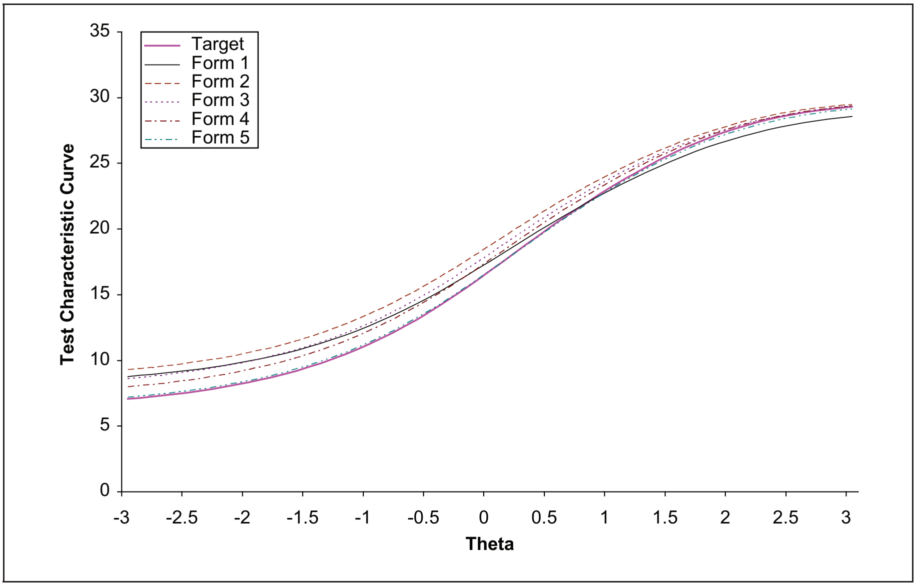

Figure 9. Test characteristic curves of the 13-point MIP method

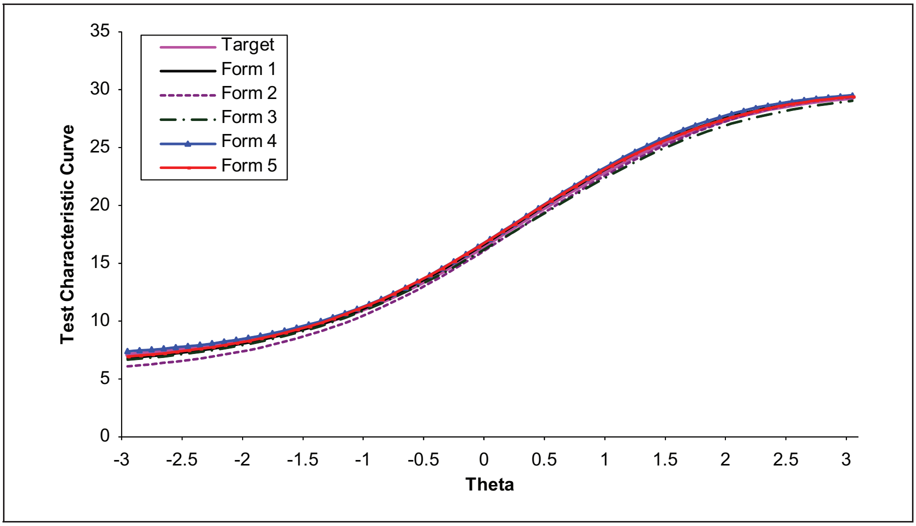

Figure 10. Test characteristic curves of the Cell Only method

methods, respectively. Table 1 shows that the 13-point MIP method yielded the smallest mean (0.002) and standard deviation (0.003). Although the mean for the Cell

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

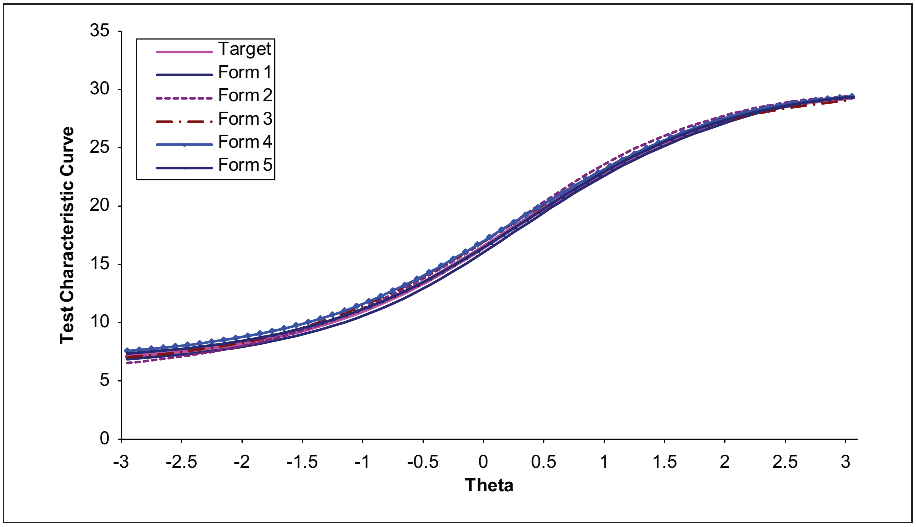

Figure 11. Test characteristic curves of the Cell and Cube method

Table 1. Descriptive Statistics of the Mean Square Deviation From the 61 Target Information Values of Five Forms

	Mean Square Deviation From the Target Information			
Method	Mean	*SD*	Maximum	Minimum
Five-point MIP method	0.309	0.621	3.454	0.000
Thirteen-point MIP method	0.002	0.003	0.025	0.000
Cell Only method	0.100	0.172	1.060	0.000
Cell and Cube method	0.190	0.318	1.733	0.000

Note: MIP = mixed-integer programming.

Only method was similar to that for the Cell and Cube method, it had a smaller standard deviation. The results for the 5-point MIP method exhibited the largest mean and standard deviation among the four methods. In Table 2, the mean square deviations from the test characteristic curve of the reference form show that the Cell Only and Cell and Cube methods yielded similar results, with the smallest means and standard deviations of MSD. The 13-point MIP method did not yield better results than did the Cell approach, whereas the MIP with a 5-point constraint yielded the largest mean and standard deviations of test characteristic curve MSD.

Discussion and Conclusions

Current ATA methods have lightened the burden of the labor-intensive work of item selection; however, most of them are essentially based on similar concepts of

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

Table 2. Descriptive Statistics of the Mean Square Deviation From the 61 Target Test Characteristic Curves of Five Forms

	Mean Square Deviation From the Target Information			
Method	Mean	*SD*	Maximum	Minimum
Five-point MIP method	1.934	2.232	7.698	0.000
Thirteen-point MIP method	1.246	1.565	5.277	0.000
Cell Only method	0.115	0.178	0.961	0.000
Cell and Cube method	0.115	0.117	0.480	0.000

Note: MIP = mixed-integer programming.

constrained combinatorial optimization. Various methods belonging to the optimization approach have been applied to real testing programs, such as the Law School Admission Test, and they have shown promising results. Commercial discrete optimization solvers and modeling systems make it convenient to model and specify complicated test assembly problems through syntax. However, the educational measurement community is generally unfamiliar with such software and their associated platforms (Cor et al., 2009). Psychometricians usually require additional optimization-related knowledge and programming competence. Therefore, it will be interesting to see how this field develops as more alternatives become available, which may provide simpler solutions.

The results show that the 13-point MIP method performed the best among the four item selection methods, in terms of hitting target test information curves and possessing the smallest mean square deviation of test information function curves. However, the test characteristic curve results show that the Cell Only and Cell and Cube methods yielded more closely matching test characteristic curves than did the 13-point MIP method. Therefore, by optimizing test information function curves with respect to several well-chosen test information function points, the MIP method could provide the most test information function curves well-fitted to the reference form. However, it appears that constraining test information function points does not exclusively guarantee the generation of similar test characteristic curves, also known as true score, which is typically the basis for determining statistical equivalence in test equating. The Cell Only and Cell and Cube methods did not yield better results than did the 13-point MIP method in terms of test information functions, yet they provided relatively consistent results in hitting both the target test information functions and test characteristic curves.

The principal contribution of the present study has been to develop a new sampling-and-classification method as a protocol to construct parallel forms from a seed test. It has also contributed to the development of a concept that is simple to comprehend and an easy implementation procedure, particularly for tests with simple test specifications. Writing MIP programs to solve the problem of ATA without commercial software is usually very costly in terms of time and effort. Large-scale

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

assessments with complicated test constraints could benefit from the power of an optimization solver to build desired forms very quickly. The sampling and classification method is suitable for testing programs with simple test specifications and for testing agencies with limited budgets. This method mimics the joint distribution of *a*- and *b*-parameters and is more intuitive, making it easier for most educators and researchers to understand.

The Cell Only and Cell and Cube methods presented here are simplified versions. This approach could be further modified to make it more practical. Before applying the Cell Only and Cell and Cube methods to operational item banks, other issues must be addressed. First, the number of cells and cubes, and the cutoff points to be used, is intuitive. When specifying the cutoff points of cells and cubes, test constructors should consider several factors, including item pool size, the composition of item parameters in the item bank, and test specifications. Second, although this study incorporated simple nonstatistical characteristics, such as content specifications, real tests might include complicated test specifications, such as item type, enemy items, or set-based items. The present study should be viewed as the first of a series of studies in the development of a sampling and classification approach to ATA. Further research is underway on how to incorporate content balancing techniques or to satisfy complex test specifications based on this approach.

Acknowledgments

The authors would like to thank Jiahe Qian for his telephone consultations. Thanks also to Barbara Dodd, George Marcoulides, Hsiao-Cheng Yu, and two anonymous reviewers for their useful comments on earlier versions of this article. Special thanks to Daniel Junglas and Ming Lei for their useful comments during the revision.

Authors’ Note

This article was in part prepared while the first author was Assistant Professor in the Department of Educational Psychology and Learning Systems at Florida State University, Tallahassee, FL, USA.

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) disclosed receipt of the financial support for the research, authorship, and/or publication of this article: The National Academy for Educational Research of Taiwan, NAER-100-24-B-2-02-00-2-02, and the National Science Council of Taiwan, NSC100-2511-S-009-005, partially supported the work of Pei-Hua Chen. The Department of Educational Psychology and Learning Systems partially supported the work of Pei-Hua Chen and Haiyan Wu.

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

References

Armstrong, R. D., Jones, H. J., & Wu, I.-L. (1992). An automated test development of parallel tests from a seed test. *Psychometrika, 57*, 271-288.

Belov, D. I. (2008). Uniform test assembly. *Psychometrika, 73*(1), 21-38.

Belov, D. I., & Armstrong, R. D. (2005). Monte Carlo test assembly for item pool analysis and extension. *Applied Psychological Measurement, 29*, 239-261.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics, 15*, 129-145.

Breithaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement, 67*, 5-20.

Chen, P.-H. (2005). *IRT-based automated test assembly: A sampling and stratification perspective* (Unpublished doctoral dissertation). University of Texas at Austin, Austin, Texas.

Chen, P.-H., & Chang, H.-H. (2005). *IRT based automated test assembly for multiple test forms: A sampling and stratification approach*. Paper presented at the Annual Meeting of National Council on Measurement in Education, Montreal, Quebec, Canada.

Chen, P.-H., & Chang, H.-H. (2006). *A statistical perspective of IRT-based automated test assembly: The Cell & Cube method*. Paper presented at the Annual Meeting of American Educational Research Association, San Francisco, CA.

Cor, K., Alves, C., & Gierl, M. J. (2008). Conducting automated test assembly using the Premium Solver Platform Version 7.0 with Microsoft Excel and the large-scale LP/QP solver engine add-in. *Applied Psychological Measurement, 32*, 652-663.

Cor, K., Alves, C., & Gierl, M. J. (2009). Three applications of automated test assembly within a user-friendly modeling environment. *Practical Assessment, Research & Evaluation, 14*. Retrieved from http://pareonline.net/pdf/v14n14.pdf

Davey, T. (2009). Linear models for optimal test design by van der Linden, W. J. *Journal of Educational Measurement, 46*, 470-473.

Finkelman, M., Kim, W., & Roussos, L. A. (2009). Automated test assembly for cognitive diagnosis models using a genetic algorithm. *Journal of Educational Measurement, 46*, 273-292.

GAMS (2007). The General Algebraic Modeling System (GAMS) (Version 23.7.3). Washington, DC: GAMS Development Corporation.

Glover, F., Taillard, E., & de Werra, D. (1993). A user's guide to tabu search. *Annals of Operations Research, 41*, 3-28.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: John Wiley.

Hillier, F. S., & Lieberman, G. J. (2001). *Introduction to operations research* (7th ed.). New York, NY: McGraw-Hill.

IBM. (2009). IBM ILOG CPLEX Optimization Studio (Version 12). Armonk, NY: International Business Machines Corporation.

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement, 22*, 224-236.

Luecht, R. M., & Hirsch, T. M. (1992). Item selection using an average growth approximation of target information functions. *Applied Psychological Measurement, 16*, 41-51. doi:10.1177/014662169201600104

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika, 42*, 193-198.

Sanders, P. F., & Verschoor, A. J. (1998). Parallel test construction using classical item parameters. *Applied Psychological Measurement, 22*, 212-223.

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014

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

van der Linden, W. J. (2005). *Linear models of optimal test design*. New York, NY: Springer.

Veldkamp, B. P. (2002). Multidimensional constrained test assembly. *Applied Psychological Measurement, 26*, 133-146.

Wightman, L. F. (1998). Practical issues in computerized test assembly. *Applied Psychological Measurement, 22*, 292-302.

Downloaded from epm.sagepub.com at NATIONAL CHIAO TUNG UNIV LIB on April 28, 2014