Journal of Computerized Adaptive Testing

Volume 5 Number 1

February 2017

Uniform Test Assembly: Concepts, Problems, Solvers, and Applications for Adaptive Testing

Dmitry I. Belov

DOI 10.7333/1702-050101

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing

www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 5, Number 1, February 2017

DOI 10.7333/1702-050101

ISSN 2165-6592

Uniform Test Assembly: Concepts, Problems, Solvers, and Applications for Adaptive Testing

Dmitry I. Belov

Law School Admission Council

This paper presents the latest developments since the publication of the seminal book by van der Linden (2005) on general types of test assembly (TA) problems, major automated test assembly (ATA) methods, and various practical situations in which a TA problem arises. With the power of modern combinatorial optimization (CO) methods, multiple practical tasks in test development and design that were previously intractable can now be solved. The TA problem is, therefore, no longer a central issue for test development but rather a subproblem embedded in different practical tasks, where two major approaches are currently exploited: mixed-integer programming (MIP) and uniform test assembly (UTA). In the world of ATA, the MIP approach is dominant. However, UTA has multiple advantages over MIP. This paper concentrates on UTA and enumerates its multiple applications for adaptive testing.

Keywords: *automated test assembly, item pool analysis, item pool extension, item pool design, combinatorial optimization, mixed-integer programming, monte-carlo methods, uniform test assembly*

Testing organizations periodically produce test forms for assessments in various formats: paper-and-pencil (P&P), computer-based testing (CBT), multistage testing (MST), and computerized adaptive testing (CAT). Each test form includes items selected from an item bank to optimize a given objective function and/or to satisfy given test specifications in terms of both statistical and content constraints. Assembling such test forms can be formulated as a combinatorial optimization (CO) problem, referred to here as a test assembly (TA) problem.

CO is concerned with searching for an element from a finite set (called a *feasible set*) that would optimize (minimize or maximize) a given objective function. Numerous practical problems can be formulated as CO problems, where a feasible set is not given explicitly but is represented implicitly by a list of inequalities and inclusions.

Psychometric researchers started to apply CO to TA in the early 1980s. Theunisen (1985) reduced a special case of a TA problem to a knapsack problem (Papadimitriou & Steiglitz, 1982). Van der Linden and Boekkooi-Timminga (1989) formulated a TA problem as a maximin problem. Later, Boekkooi-Timminga (1990) extended this approach to the assembly of multiple nonoverlapping test forms.[1] Soon after, the TA problem attracted many researchers, whose major results are discussed in van der Linden (2005). Currently, the importance of CO in psychometrics is growing due to its recent applications that go beyond TA, such as identification of cognitive models (Cen, Koedinger, & Junker, 2006), resource management (van der Linden & Diao, 2011), optimal learning (van der Linden, 2012), optimal linking (van der Linden & Barrett, 2016), and test security (Belov, 2014, 2017), with more such applications on the horizon.

This paper discusses the latest automated test assembly (ATA) developments from a CO standpoint, specifically two approaches for ATA: mixed-integer programming (MIP) and uniform test assembly (UTA). Although the MIP approach is currently dominant, this paper demonstrates that UTA has great potential for ATA applications. In CAT, for example, the UTA-based approach by Belov, Armstrong, & Weissman (2008) is more robust to the aberrant behavior of examinees (Figure 5 and p. 437) than the MIP-based approach by van der Linden & Reese (1998). This result should also extrapolate to robustness against an uncertainty in item parameters. The uncertainty is due to estimation error in item parameters; and it exists in P&P, CBT, MST, and CAT, where items are selected from an item bank based on item characteristics (e.g., Fisher information) dependent on their modeling parameters.

Multiple applications of UTA for adaptive testing are demonstrated in this paper. In CAT, the UTA approach is applied in the following areas: CAT with content constraints, cognitive diagnostic CAT with content constraints, CAT bank assembly, assembly of multiple nonoverlapping (or partially overlapping) CAT banks, identification of a population distribution matching a given master bank and CAT specifications, and identification of item proper-

[1] Two tests are called *nonoverlapping* if they do not have items in common; otherwise, they are called *overlapping*. If the number of items in common (size of the overlap) is less than a specified number, these tests are called *partially overlapping*.

ties that would increase the usability of a given master bank for CAT. In MST, the UTA approach is applied in assembly of an MST form; assembly of multiple nonoverlapping (or partially overlapping) MST forms, identification of item properties that would increase the number of nonoverlapping (or partially overlapping) MST forms available from a given bank, and estimating IRT targets for MST.

Throughout this paper, the following notation is used:

- Lowercase letters $a, b, c, \ldots$ denote scalars;
- Bold lowercase letters $\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$ denote vectors;
- Capital letters $A, B, C, \ldots$ denote sets. The number of elements in a set S is denoted by $|S|$; $\varnothing$ denotes an empty set; and
- Bold capital letters $\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$ denote functions.

General Types of Test Assembly Problems

Van der Linden (2005) described specific types of TA problems for different types of assessments (i.e., P&P, CBT, MST, CAT) in his seminal textbook *Linear Models for Optimal Test Design*. Real instances of TA problems have been studied by Ariel, Veldkamp, and Breithaupt (2006); Armstrong, Belov, and Weissman (2005); Belov and Armstrong (2005); Belov and Armstrong (2008); Belov et al. (2008); Breithaupt, Ariel, and Veldkamp (2005); De Jong, Steenkamp, and Veldkamp (2009); Veldkamp (2002); and Veldkamp and van der Linden (2002).

Test Assembly as a Problem of Combinatorial Optimization

Generally, a TA problem can be formulated as the following CO problem:

$$\begin{aligned} &\text{maximize } \mathbf{F}(\mathbf{x}) \\ &\text{subject to } \mathbf{x} \in X \end{aligned} \tag{1}$$

$\mathbf{x} = (x_1, x_2, ..., x_n)^T$ is a binary decision vector defining a test, such that if $x_i = 1$, then item i is included in the test; otherwise (i.e., $x_i = 0$), item i is not included in the test.

n is the number of items in the item bank.

Set X contains all binary vectors, each defining a feasible test. Therefore, this set is called a *feasible set*. In practice, a feasible set is not given explicitly but is represented implicitly by a list of inequalities and inclusions constraining the decision vector. This list is constructed directly from test specifications. For example, the following represents a feasible set with all tests containing 5 to 10 items:

$$5 \le \sum_{i=1}^{n} x_i \le 10$$
$$x_i \in \{0,1\}$$

where the constraint $x_i \in \{0,1\}$ is included in any CO problem (i.e., each feasible solution $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ has to be a binary vector).

$\mathbf{F}(\mathbf{x})$ is an objective function (possibly a vector function; for a multiobjective TA problem, see Veldkamp, 1999). For example, in CAT, the following linear objective maximizes the Fisher information of a test at a current ability estimate $\hat{\theta}$:

$$\text{maximize} \sum_{i=1}^{n} \mathbf{I}_i(\hat{\theta}) x_i \ , \tag{2}$$

where $\mathbf{I}_i(\hat{\theta})$ is the Fisher information of item i at ability level $\hat{\theta}$ (Lord, 1980).

Test Assembly as a Problem of Constraint Satisfaction

A TA problem can also be formulated as the following constraint satisfaction problem:

$$\mathbf{x} \in X. \tag{3}$$

Many practical tasks can be reduced to the analysis of the feasible set X. For example, in P&P and CBT modes, each item can be administered only once. Therefore, it is crucial for item bank maintenance to have an estimate of the maximum number of nonoverlapping tests available from an item bank, given the test specifications. An approximate solution can be found by sampling from the feasible set and then solving the maximum set packing problem (given a collection of subsets, find a maximum subcollection with mutually disjoint subsets). For the sampling, the problem in Expression 3 (hereinafter referred to as Problem 3) can be solved multiple times such that each vector from X has an equal probability of being a solution.

In other words, every test from the feasible set X has $1/|X|$ probability of being assembled. This process is called *uniform test assembly* (UTA). Additional details on UTA and its applications are presented later in this paper.

Often a good lower bound for the objective function is known or can be easily computed (Belov & Armstrong, 2009). Subsequently, Problem 1 can be approximated by Problem 3. For example, the following represents a feasible set with all possible tests containing 5 to 10 items and having Fisher information at ability estimate $\hat{\theta}$ above the lower bound 3:

$$\begin{aligned}&5 \le \sum_{i=1}^{n} x_i \le 10\\ &\sum_{i=1}^{n} \mathbf{I}_i(\hat{\theta}) x_i \ge 3 \\ &x_i \in \{0,1\}.\end{aligned} \tag{4}$$

Interestingly, Problem 3 can be approximated by Problem 1 as well:

$$\begin{aligned}&\text{maximize } \sum_{i=1}^{n} \alpha_i x_i\\ &\text{subject to } \mathbf{x} \in X,\end{aligned} \tag{5}$$

where $\alpha_1, \alpha_2, ..., \alpha_n$ are independent and uniformly distributed on [0, 1). Vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_n)^T$ is resampled each time before Problem 5 is solved, thus allowing the assembly of different tests.

Intuitively, it would be expected that because vector $\boldsymbol{\alpha}$ is uniformly distributed, the resultant sample of assembled tests should be uniform. However, Belov (2008) proved that, in general, a sequence of optimal solutions to Problem 5 does not provide a uniform sample from the feasible set. Only if a feasible set contained pairwise nonoverlapping tests (which hardly ever happens in practice) would a sequence of optimal solutions to Problem 5 provide a uniform sample. In general, therefore, UTA cannot be formulated as Problem 5. The question as to whether UTA can be achieved via a sequence of optimal solutions to a certain instance of Problem 1 is still open (Belov, 2008).

Test Assembly Problem Under Uncertainty

Usually, inequalities defining the feasible set X can be grouped into content constraints, (i.e., the first inequality in Problem 4) and statistical constraints (i.e., the second inequality in Problem 4). Content constraints are known precisely. Statistical constraints usually include the parameters of item response theory (IRT) models (Lord, 1980), which are calibrated from the response data and therefore subject to error. Thus, the assembled test might not actually satisfy the statistical constraints, and/or the objective function can be overestimated or underestimated. All real-life instances of TA problems are under uncertainty due to estimation errors in the statistical parameters of items.

Optimization under uncertainty is a well-studied field. There are two major approaches: stochastic optimization (Birge & Louveaux, 1997) and robust optimization (Bertsimas, Brown, & Caramanis, 2011).

Consider a common TA problem in CAT:

$$\begin{aligned}&\text{maximize} \sum_{i=1}^{n} \mathbf{I}_i(\hat{\theta})x_i\\ &\text{subject to } \mathbf{x} \in X\end{aligned}, \tag{6}$$

where feasible set X is defined by content constraints (known precisely), but each coefficient $\mathbf{I}_i(\hat{\theta})$ has an error associated with it as a result of item parameter estimation using a given procedure such as expectation maximization (EM) or Markov chain monte-carlo (MCMC).

Assume that distributions of each parameter estimated under an IRT model are known, where these distributions are outcomes of a specific MCMC procedure that was used to estimate the item parameters. Then, a stochastic counterpart of Problem 6 is formulated as follows:

$$\begin{aligned}&\text{maximize} \sum_{i=1}^{n} \mathrm{E}[\mathbf{I}_i(\hat{\theta}, h_{i1}, h_{i2}, ...)]x_i\\ &\text{subject to } \mathbf{x} \in X\end{aligned}, \tag{7}$$

where the expectation is taken over parameters $h_{i1}, h_{i2}, ...$ of item i. In this case, Problem 7 can be solved directly.

Assume that $\mathbf{I}_i(\hat{\theta})$ has an error such that $\mathbf{I}_i(\hat{\theta}) \in [u_i - d_i,\ u_i]$ with high probability, where u_i and d_i are estimated by EM or MCMC procedures. From Bertsimas and Sim (2003), it follows that a robust counterpart of Problem 6 can be formulated:

$$\begin{aligned}&\text{maximize}\left[\sum_{i=1}^{n} u_i x_i - \max_{\{S|S\subseteq N,\ |S|\leq g\}} \sum_{j\in S} d_j x_j\right],\\ &\text{subject to } \mathbf{x} \in X\end{aligned} \tag{8}$$

where $N = \{1, 2, ..., n\}$ and g is a parameter chosen beforehand. An optimal solution to Problem 8 defines a test with Fisher information at ability level $\hat{\theta}$ above a certain threshold. This inequality holds under uncertainty in at most g items. Clearly, Problem 8 cannot be solved directly. However, Bertsimas and Sim (2003) developed a method to solve Problem 8 by directly solving $n+1$ problems:

$$\begin{aligned}&\max_{l=1}^{n+1}\left[-gd_l + \max\left[\sum_{i=1}^{n} u_i x_i - \sum_{j=1}^{l} (d_j - d_l)x_j\right]\right],\\ &\text{subject to } \mathbf{x} \in X\end{aligned} \tag{9}$$

where, without loss of generality, $d_1 \ge d_2 \ge ... \ge d_n \ge d_{n+1} = 0$ is assumed. Additional details on the application of robust optimization for ATA can be found in Veldkamp (2012a).

An alternative approach to accommodate the uncertainty in item parameters is to state the TA problem as Problem 3 but use narrower bounds for statistical constraints. New bounds should be computed (e.g., by a monte-carlo method) such that the probability of a feasible test violating the original bounds is below a given significance level. This approach can be implemented within existing ATA methods.

Automated Test Assembly Methods

From a geometrical standpoint, the TA problem is solved by searching through the vertices of the hypercube $\{\mathbf{x} = (x_1, x_2, ..., x_n)^T \mid 0 \le x_i \le 1,\ i = 1, 2, ..., n\}$ until a vertex $\mathbf{x}_0 \in X$ that optimizes the objective function $\mathbf{F}(\mathbf{x})$ is found (see Problem 1) or until a vertex $\mathbf{x}_0 \in X$ is found (see Problem 3). The number of vertices of the hypercube is 2^n, where n is the number of items in the item bank. Therefore, the search can run for some time, exponentially dependent on the number of items in the bank. In practice, this problem is often solvable by modern ATA methods in a reasonable amount of time on a personal computer.

Branch-and-Bound

The branch-and-bound (B&B) method solves Problem 1 by performing an intelligent search through vertices of the hypercube. It starts by finding an optimal solution to the relaxation of Problem 1 without the key constraint $x_i \in \{0,1\}$, $i = 1, 2, ..., n$. The relaxation can often be solved in polytime, which means that the running time of the solver is bounded by a polynomial in size of the problem (see Garey & Johnson, 1979), resulting in a fast convergence. An optimal solution to the relaxation provides a choice of branching decisions and an upper bound for Problem 1. More precisely, a coordinate, $1 \le j \le n$, is selected, where an optimal solution to the relaxation has a fractional value. Then two new subproblems are added to a list of subproblems, which is initially empty: (1) relaxation with additional constraint $x_j = 0$ and (2) relaxation with additional constraint $x_j = 1$.

Each subproblem on the list is solved, where one of the following cases is possible:

1. The subproblem is infeasible; that is, the corresponding feasible set $X = \varnothing$.
2. An optimal solution to the subproblem is binary, which provides a feasible solution to Problem 1; this solution is used to update the global solution.
3. An optimal solution to the subproblem is not binary, and its objective function is less than or equal to the global objective found so far.
4. An optimal solution to the subproblem is not binary, and its objective function is greater than the global objective found so far.

In Cases 1–3, the subproblem is removed from the list and the next subproblem on the list will be analyzed. In Case 4, branching of the subproblem is applied (see above) and then the subproblem is removed from the list. When the list is empty, it can be claimed that an optimal solution to Problem 1 has been found. For more details, see Papadimitriou and Steiglitz (1982) and Nemhauser and Wolsey (1988).

With the B&B method, optimality of a feasible solution to Problem 1 can be proved. The success of applying B&B depends on how well a solver adapts to each instance of Problem 1 or, more precisely, how well the structure of an instance is taken into account to organize effective branching and bounding.

When Problem 1 is linear and its matrix of the system of inequalities is totally unimodular (Nemhauser & Wolsey, 1988), the relaxation of Problem 1 has a binary optimal solution. In addition, several fast polytime algorithms are available to solve the relaxation (Ahuja, Magnanti, & Orlin, 1993). If a large submatrix of the matrix of the system of inequalities is totally unimodular, then the assembly of linear tests can be performed efficiently (Armstrong, Jones, & Kunce, 1998; Armstrong, Jones, & Wu, 1992) by a combination of network flow programming, Lagrangian relaxation, and B&B.

The B&B method is a core of the MIP approach to large practical problems of item bank analysis and design. Typically, a real-life problem is formulated as an instance of Problem 1 and then solved directly with the B&B method.

Heuristics

Heuristic methods (heuristics) provide a relatively fast search through vertices of the hypercube that are likely to discover a near solution. In the case of Problem 1, it is a suboptimal solution; in the case of Problem 3, it is a subfeasible solution. A comprehensive review of ATA heuristics is given by van der Linden (2005).

Some heuristics (Swanson & Stocking, 1993) move the constraints to the objective function, which essentially is a Lagrangian relaxation (Nemhauser & Wolsey, 1988). Then set X is no longer a feasible set because some vectors from X may violate constraints that were incorporated into the objective function $\mathbf{F}(\mathbf{x})$.

For example, consider the following TA problem:

$$\begin{aligned} &\text{maximize} && \sum_{i=1}^{n} \mathbf{I}_i(\hat{\theta})x_i \\ &\text{subject to} && 5 \le \sum_{i=1}^{n} x_i \le 10, \\ & && x_i \in \{0,1\} \end{aligned} \tag{10}$$

where the feasible set only contains vertices of the hypercube with 5 to 10 positive coordinates (corresponding to tests with 5 to 10 items). By applying Lagrangian relaxation, the TA Problem 10 is transformed into the following:

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{n} \mathbf{I}_i(\hat{\theta})x_i + \lambda_1\left(5-\sum_{i=1}^{n} x_i\right) + \lambda_2\left(10-\sum_{i=1}^{n} x_i\right) \\ \text{subject to} & x_i \in \{0,1\} \\ & \lambda_1 \le 0 \\ & \lambda_2 \ge 0 \end{array} , \qquad (11)$$

where the feasible set contains all vertices of the hypercube.

Most heuristics in the ATA literature are based on sequential item selection: One item is selected at a time until the required number of items is reached, where each selection minimizes the current value of a residual. There are numerous types of residuals (Ackerman, 1989; Leucht, 1998; Swanson & Stocking, 1993) driven by various TA constraints and/or TA objectives. These heuristics minimize the current value of the residual, expecting that when the required number of items are selected, these items should satisfy the constraints and/or optimize the objective. Such heuristics belong to a class known in CO literature as *greedy heuristics* (Papadimitriou & Steiglitz, 1982). For example, consider the following TA problem:

$$\begin{array}{l} \sum_{i=1}^{n} \mathbf{I}_i(\hat{\theta})x_i = t \\ \sum_{i=1}^{n} x_i = 10 \\ x_i \in \{0,1\} \end{array} , \qquad (12)$$

where t is a target value of the Fisher information at the current θ estimate. Assume that three items $S = \{i_1, i_2, i_3\}$ were already selected. Then, according to Leucht (1998), the fourth item, i_4 should minimize the following residual:

$$\left| \mathbf{I}_{i_4}(\hat{\theta}) - \left(t - \sum_{i \in S} \mathbf{I}_i(\hat{\theta})x_i\right) \Big/ 7 \right| . \qquad (13)$$

While greedy heuristics are fast, their solutions are only locally optimal and therefore might violate some of the constraints (e.g., see TA Problem 11). At the same time, in high-stakes testing, violation of certain or all constraints is not acceptable (Ariel et al., 2006;

Armstrong et al., 2005; Breithaupt et al., 2005; De Jong et al., 2009; Veldkamp, 2002; Veldkamp & van der Linden, 2002). Several CO approaches have been applied to avoid getting stuck in a local optimum while solving a TA problem, such as simulated annealing (van der Linden, Veldkamp, & Carlson, 2004) and genetic algorithms (Verschoor, 2004).

Monte-Carlo Test Assembler

The monte-carlo test assembler (MCTA) was introduced by Belov and Armstrong (2004, 2005) to solve TA Problem 3. It is straightforward in concept and consists of two steps:

Step 1: Generate a random vector of items.

Step 2: If this vector satisfies test specifications, save it as a new test and stop; otherwise, return to Step 1.

The biggest challenge with the monte-carlo technique is avoiding generating many "useless" vectors at Step 1. Belov and Armstrong (2004, 2005) have developed several strategies to reduce the search space such that it still has a nonempty intersection with the feasible set. They exploited properties of the constraints, using a divide-and-conquer principle and tabu search, and prioritized constraint checking based on their computational complexity. MCTA has been applied for P&P (Belov & Armstrong, 2004, 2005), MST (Belov & Armstrong, 2008), and constrained CAT (Belov et al., 2008). The performance of MCTA is surprisingly fast. For example, Belov et al. (2008) reported that the monte-carlo CAT performed 20 times faster than the shadow CAT (van der Linden & Reese, 1998).

The major advantage of MCTA is its ability to perform uniform sampling from the feasible set X. This advantage is useful in practice. For example, due to its random nature, the convergence rate of MCTA determines how large the feasible set is: The higher the rate, the larger the feasible set. The size of the feasible set directly indicates how given test specifications match a given item bank. Other potential approaches to produce a uniform sampling from the feasible set are analyzed by Belov (2008). MCTA is a core of the UTA approach to large practical tasks in item bank analysis and design, where properties of a feasible set of a given instance of Problem 3 are explored and exploited via uniform sampling from the feasible set.

Applications of UTA for Adaptive Testing

The major purpose of ATA is to assemble one test at a time. The specifics of a particular assessment, however, might influence the methods described in the previous section. In CAT, the shadow CAT method (see the MIP approach by van der Linden & Reese, 1998) selects the next item maximizing Fisher information at the current θ estimate, such that the administered sequence of items satisfies content constraints. Monte-carlo CAT (Belov et al., 2008) allows a balance between the maximization of Fisher information and the robustness of the θ estimate to possible mistakes made by the examinee during a test or to an uncertainty in item parameters. In MST, each path in an MST form must be assembled, taking into account common testlets between paths (for more details, see Belov & Armstrong, 2008).

In each assessment, there are multiple tasks in which ATA is a crucial subproblem (van der Linden, 2005). Belov (2008) demonstrated that, from a mathematical standpoint, many of these tasks can be reduced to the analysis of properties of the feasible set X .

Analyzing Properties of the Feasible Set

For real-life item bank and test specifications (e.g., in P&P, CBT, MST, or CAT), computing the whole feasible set is intractable. The analysis of the matrix of the system of inequalities is very limited and possible only for linear systems. In general, the only way to study the properties of a feasible set is therefore to construct and analyze a uniform sample from the feasible set.

Assume that there is a way to assemble tests such that each element of the feasible set has an equal probability of being selected (as in UTA) and multiple tests can be assembled without withdrawing their items from the bank. Since the resulting sample of tests is drawn uniformly, it can be considered representative of the feasible set. The statistical inference about properties of the feasible set thus can be acquired from this sample. For example, an item usage frequency can be calculated (its applications are demonstrated below). Given a set of tests, the *usage frequency* of an item is the number of tests that include this item, where an item with the highest usage frequency is called the *most usable item* and an item with the lowest usage frequency is called the *least usable item*. The computation of item usage frequency is straightforward:

Step 1: Assemble multiple tests uniformly without withdrawing their items from the bank.

Step 2: For each item in the bank, count how many assembled tests include the item.

Analyzing Feasibility of a Test Assembly Problem

This is a crucial task for any testing program, including P&P, CBT, MST, and CAT. It should answer the following questions: Is the feasible set empty? If yes, which constraints make it empty? If no, how large is the feasible set?

Belov and Armstrong (2005) used uniform sampling from embedded feasible sets (where each embedding corresponds to an additional subset of constraints) to identify the most difficult constraints. Difficult constraints dramatically reduce the size of the feasible set and might even cause the feasible set X to be empty, which makes the corresponding TA problem infeasible. Therefore, a larger drop in performance of MCTA, after another subset of constraints is added to the TA problem, indicates a more difficult subset of constraints (for more details, see

Belov & Armstrong, 2005). An alternative approach based on MIP is presented by Huitzing, Veldkamp, and Verschoor (2005).

Due to the nature of MCTA, its rate of convergence characterizes the strengths and weaknesses of a given bank in relation to the test specifications. The larger the number of potential tests (meaning the larger the feasible set), the faster the MCTA finds tests. Thus, the performance time of MCTA can be used to compare and evaluate different item banks and/or test specifications.

Constrained CAT Based on UTA

Test assembly in the context of CAT frequently shares with its P&P counterparts the need to select items satisfying content constraints. Added to this problem, however, are the following requirements: (1) to assemble tests adapted to examinee θ, (2) to assemble these tests while administering them, (3) to monitor and control the exposure of items in the bank, (4) to provide estimates of θ robust to uncertainty both in item parameters (see above) and in examinee behavior [e.g., when a high (or low) ability examinee performs poorly (or well) at the beginning of the test].

The shadow CAT by van der Linden & Reese (1998) automatically meets requirements (1) and (2). However, to meet requirement (3), it needs an additional mechanism for item exposure; for example, van der Linden & Veldkamp (2004) added exposure control through item ineligibility constraints. Requirement (4) calls for an additional mechanism provided by methods of robust optimization (Veldkamp, 2012b) or stochastic programming.

Monte-carlo CAT (MCCAT) was developed by Belov et al. (2008) to meet all of the above requirements. As with multiple other applications of UTA, the concept of MCCAT is simple. Given the current estimate of θ and l already administered items, perform the following:

Step 1: Uniformly assemble multiple test forms satisfying all content constraints, where each form has l already administered items.

Step 2: Select an item most informative at the current θ estimate from the sample of not-yet-administered items randomly drawn from the test forms assembled at Step 1, where the size of the sample gradually increases with each administered item. This gradual increase helps to find more informative items closer to the end of the test, when the current θ estimate is near the true θ and to avoid administering highly informative items at the beginning of the test, when the current θ estimate might be far from the true θ.

Belov et al. (2008) demonstrated that MCCAT, which has a slightly larger estimation error, provided much better item exposure and robustness in comparison to shadow CAT. This was due to UTA and the gradual increase of the sample of items. The authors did not evaluate how the uncertainty in item parameters might affect shadow CAT and MCCAT. A positive correlation between an item discrimination parameter and its estimation error was examined by Veldkamp (see Veldkamp, 2012a, Figure 2, p. 599). More informative items thus should have larger uncertainty in their parameters. Therefore, the shadow CAT should be more affected by this uncertainty because, in contrast to MCCAT, it always selects an item most informative at the current θ estimate.

MCCAT is governed by just one parameter (the size of the sample) to balance between the precision and the robustness of the θ estimate. Belov et al. (2008) demonstrated how different changing rules for this parameter can influence the outcome of the MCCAT. For example, a rule described in Proposition 2 (pp. 434-435) will cause the MCCAT to be equivalent to the shadow CAT.

Recently, MCCAT was applied by Mao & Xin (2013) for cognitive diagnostic CAT with content constraints. They demonstrated via computer simulations that (1) MCCAT satisfied test specifications and produced satisfactory measurement precision and item exposure rates and (2) MCCAT outperformed the modified maximum global discrimination index method when MCCAT utilized item selection methods based on Kullback–Leibler divergence. Overall, the recovery rate of the knowledge states, the distribution of item exposure, and the utilization rate of the item bank were improved when MCCAT was used

Assembly of a CAT Bank

Usually, in CAT there is a large master bank from which a smaller CAT bank is assembled for the next administration. Any realistic method of CAT bank assembly should guarantee the following two objectives of CAT bank design: (1) the existence of at least one feasible test form (i.e., a sequence of items satisfying all content constraints) and (2) bounded values of mean squared error and bias for the estimated θ. Exploiting the MIP approach, a CAT bank was assembled by van der Linden, Ariel, and Veldkamp (2006) as a set of nonoverlapping feasible forms, where each form maximized information at a certain point, and points were distributed according to the expected population. Van der Linden et al. (2006) demonstrated a satisfaction of the two CAT design objectives via computer simulations. However, their heuristic is information greedy, causing each subsequent CAT bank assembled from the master bank to be less and less informative. A modification of this method (based on the UTA approach) by Belov and Armstrong (2009) enables the assembly of multiple (information-parallel) CAT banks that guarantee the two CAT design objectives.

CAT Bank Analysis

Exploiting the UTA approach, Belov and Armstrong (2009) computed a distribution of examinees most suitable for a given item bank and test specifications in two stages:

Stage 1: Sample from the feasible set.

Stage 2: Compute the distribution based on test information functions from tests assembled in the previous stage (see Belov & Armstrong, 2009, Algorithm 4, p.541).

Extracting Multiple Nonoverlapping (or Partially Overlapping) Tests and CAT Banks

Given test specifications and an item bank, the number of available nonoverlapping tests is a critical indicator of bank usability for testing organizations producing P&P, CBT, and MST because each corresponding test form can be administered only once. In the case of CAT, this is equivalent to the number of available nonoverlapping CAT banks assembled from a large master bank. All methods described below are immediately applicable for extracting multiple nonoverlapping (or partially overlapping) CAT banks.

The simultaneous assembly of multiple nonoverlapping tests suggested by Boekkooi-Timminga (1990) is often intractable, given real test specifications and an item bank. A simple heuristic (Boekkooi-Timminga, 1990) is to assemble a test, withdraw its items from the bank, then assemble another test, and so on, until the TA problem becomes infeasible (or until a TA solver cannot assemble a test within given period of time). Such an approach is known as *sequential assembly*. However, it is easy to demonstrate (Belov, 2008) that this approach might often assemble only a few nonoverlapping tests. Subsequently, alternative methods have been developed that greatly outperform sequential assembly by utilizing properties of the feasible set.

Belov and Armstrong (2006) suggested a set packing approach. They assembled multiple nonoverlapping tests in two stages:

Stage 1: Sample from the feasible set.

Stage 2: Solve the maximum set packing (or, equivalently, the maximum clique) problem (Garey & Johnson, 1979) for the resulting sample.

They applied this approach for P&P, in particular for the LSAT (Belov & Armstrong, 2005) and for MST (Belov & Armstrong, 2008).

Belov (2008) developed a modified sequential assembly (MSA) approach by exploiting item usage frequency in order to keep more usable items for later assemblies. The set packing approach and the MSA both demonstrated twice the speed and resulted in the same number of nonoverlapping tests when the least usable items were withdrawn from the bank before implementing the two approaches (Belov, Williams, & Kary, 2015). In addition, Belov et al. (2015) studied a mixed method where the utilization of item usage frequency is combined with the set packing approach:

Step 1: Remove the most usable items from the bank.

Step 2: Apply the set packing approach.

Step 3: Withdraw items of the assembled tests from the bank.

Step 4: Add the most usable items to the bank.

Step 5: Apply the set packing approach.

An alternative approach based on MIP modeling to improve the sequential assembly, called *shadow test assembly*, was developed by van der Linden and Adema (1998). More specifically, to assemble m nonoverlaping tests, at each iteration i, $i = 1, 2, ...m$, the following two steps are performed:

Stage 1: Assemble test i.

Stage 2: Assemble a shadow test that satisfies relaxed inequalities with original bounds multiplied by $m - i$. A major issue with this approach, however, is that some constraints cannot be relaxed this way.

The above methods can be used for assembling nonoverlapping testlets for MST. A special MIP model was developed by Ariel et al. (2006) for assembling multiple nonoverlapping testlets. Their approach assembles a given number of nonoverlapping testlets simultaneously while maximizing a lower bound for testlet information functions. This simultaneous assembly was tractable because the corresponding MIP model induced the matrix of the corresponding linear system of inequalities with a special structure (e.g., totally unimodular).

Obviously, all of the above methods are also applicable when a partial overlap between tests is allowed. For example, in the set packing approach, a graph can be built (where each vertex corresponds to a test from the uniform sample) based on the size of the overlap between tests. Precisely, given two different tests, if the size of their overlap is less than a given threshold (Threshold 1 means no overlap), then the edge is created between corresponding vertices of the graph; and the maximum clique (Garey & Johnson, 1979) in the graph is the solution of the problem.

Efficient Item Bank Maintenance

In all assessments, test developers need to identify the properties of future items that would help maintain their item bank efficiently. In particular, in P&P, CBT, MST, and CAT bank assembly, test developers need to minimize the number of new items required in order to maximize the number of nonoverlapping test forms (or CAT banks) available from an existing bank. This minimax problem can be solved by exploiting item usage frequency (Belov & Armstrong, 2005, 2008), which is computed from a uniform sample of tests. In computer experiments with an LSAT item bank and constraints, Belov and Armstrong (2005) demonstrated that adding just a few new items that have properties similar to those of the most usable items dramatically increases the number of nonoverlapping tests that can be assembled. Alternative approaches for designing and maintaining item banks are based on MIP modeling (Ariel, van der Linden, & Veldkamp, 2006; Ariel, Veldkamp, & Breithaupt, 2006; Ariel, Veldkamp, & van der Linden, 2004).

Estimating IRT Targets for MST

When a testing organization migrates from P&P to an adaptive format such as MST, content constraints for each path in an MST form are the same as in a P&P form. However, IRT targets for each path (i.e., targets for the test response function and the test information function of each path in an MST form) should differ in order for the assembled MST form to adapt to examinee θ. Belov and Armstrong (2008) address this issue as follows:

Step 1: Build a uniform sample from the feasible set of linear forms, where each form is a vector of items satisfying the content constraints of the MST path.

Step 2: Administer the resultant sample to simulated examinees drawn from a given distribution.

Step 3: Use the resultant scores to partition the sample such that the target for each MST path is constructed from items most informative at the corresponding θ range.

This UTA-based method allows balancing between the measurement precision of assembled MST forms and the utilization of an item bank.

Summary

This paper has presented recent developments in general types of TA problems, major ATA methods, and various practical situations where a TA problem arises, with an emphasis on adaptive testing. Due to the latest achievements in CO theory and methods, multiple practical tasks in test development and design that were once intractable can now be solved. Therefore, it can be concluded that the TA problem is no longer a central issue for test development but is rather a subproblem embedded in larger practical tasks. This paper distinguished two major approaches to these larger tasks:

MIP: Treating a TA problem as Problem 1 and solving it with the B&B method.

UTA: Treating a TA problem as Problem 3 and solving it with a monte-carlo method, resulting in a uniform sampling from the feasible set.

Both approaches have been applied in practice. The MIP approach is a natural choice for testing programs in which the test is defined by constraints and an objective function to be optimized; see Ariel et al. (2006); Armstrong et al. (2005); Breithaupt et al. (2005); De Jong et al. (2009); Veldkamp (2002); and Veldkamp and van der Linden (2002). On the other hand, the UTA approach is appropriate for testing programs in which the test is defined by constraints only; see Armstrong et al. (2005); Belov and Armstrong (2005); and Belov and Armstrong (2008).

UTA is often undervalued in the world of ATA, where the MIP approach is dominant. However, UTA has a great potential for the ATA under uncertainty, which is always the case in P&P, CBT, MST, and CAT, where items are selected from an item bank based on item parameters estimated with error. For example, in CAT, the UTA approach by Belov et al. (2008) is more robust to aberrant behavior of examinees than the shadow CAT by van der Linden & Reese (1998), which should also extrapolate to robustness against the uncertainty in item parameters.

A major advantage of UTA is its conceptual simplicity and scalability. Once a UTA solver that assembles just one test form at a time is developed, UTA can be applied to multiple applications with ease. Most applications involve just two major steps: (1) building a uniform sample from the feasible set and (2) computing an estimate of interest from the resultant sample (e.g., item usage frequency, number of nonoverlapping tests, and sample of items to select the next item administered in CAT). This makes it easy for practitioners to interpret results. In contrast, the MIP approach often results in a complicated mathematical model and a special heuristic, both of which need to be explained to practitioners.

In CAT, the UTA approach was applied in the following areas: CAT with content constraints, cognitive diagnostic CAT with content constraints, CAT bank assembly, assembly of multiple nonoverlapping (or partially overlapping) CAT banks, identification of population distribution matching a given master bank and CAT specifications, and identification of item properties that would increase the usability of a given master bank for CAT.

In MST, the UTA approach was applied in the following areas: assembly of an MST form, assembly of multiple nonoverlapping (or partially overlapping) MST forms, identification of item properties that would increase the number of nonoverlapping (or partially overlapping) MST forms available from a given item bank, and estimating IRT targets for MST.

A major practical disadvantage of UTA is that for each particular testing program the UTA solver has to be developed first. This may involve substantial efforts from software developers, whereas multiple commercial and free MIP solvers are readily available.

References

Ackerman, T. (March, 1989). *An alternative methodology for creating parallel test forms using the IRT information function*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Ahuja, R. K., Magnanti, T. L., & Orlin, J. B. (1993). *Network flows: Theory, algorithms, and applications*. Englewood Cliffs, NJ: Prentice Hall.

Ariel, A., van der Linden, W. J., & Veldkamp, B. P. (2006). A strategy for optimizing item-pool management. *Journal of Educational Measurement*, *43*(2), 85–92. CrossRef

Ariel, A., Veldkamp, B. P., & Breithaupt, K. (2006). Optimal testlet pool assembly for multi-stage testing designs. *Applied Psychological Measurement*, *30*(3), 204–215. CrossRef

Ariel, A., Veldkamp, B. P., & van der Linden, W. J. (2004). Constructing rotating item pools for constrained adaptive testing. *Journal of Educational Measurement*, *41*(4), 345–359. CrossRef

Armstrong, R. D., Belov, D. I., & Weissman, A. (2005). Developing and assembling the Law School Admission Test. *Interfaces*, *35*(2), 140–151. CrossRef

Armstrong, R. D., Jones, D. H., & Kunce, C. S. (1998). IRT test assembly using network-flow programming. *Applied Psychological Measurement*, *22*(3), 237–247. CrossRef

Armstrong, R. D., Jones, D. H., & Wu, I. L. (1992). An automated test development of parallel tests from a seed test. *Psychometrika*, *57*(2), 271–288. CrossRef

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

Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing*, *2*(3), 37–58. CrossRef

Belov, D. I. (2017). Identification of item preknowledge by the methods of information theory and combinatorial optimization. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 164–176). New York, NY: Routledge.

Belov, D. I., & Armstrong, R. D. (April, 2004). *A Monte Carlo approach for item pool analysis and design*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Belov, D. I., & Armstrong, R. D. (2005). Monte Carlo test assembly for item pool analysis and extension. *Applied Psychological Measurement*, *29*(4), 239–261. CrossRef

Belov, D. I., & Armstrong, R. D. (2006). A constraint programming approach to extract the maximum number of nonoverlapping test forms. *Computational Optimization and Applications*, *33*(2/3), 319–332. CrossRef

Belov, D. I., & Armstrong, R. D. (2008). A Monte Carlo approach to the design, assembly, and evaluation of multi-stage adaptive tests. *Applied Psychological Measurement*, *32*(2), 119–137. CrossRef

Belov, D. I., & Armstrong, R. D. (2009). Direct and inverse problems of item pool design for computerized adaptive testing. *Educational and Psychological Measurement*, *69*(4), 533–547. CrossRef

Belov, D. I., Armstrong, R. D., & Weissman, A. (2008). A Monte Carlo approach for adaptive testing with content constraints. *Applied Psychological Measurement*, *32*(6), 431–446. CrossRef

Belov, D. I., Williams, M., & Kary, D. (July, 2015). *Exploiting properties of a feasible set to improve item pool utilization*. Paper presented at the international meeting of the Psychometric Society, Beijing, China.

Bertsimas, D., Brown, D. B., & Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM Review*, *53*(3), 464–501. CrossRef

Bertsimas, D., & Sim, M. (2003). Robust discrete optimization and network flows. *Mathematical Programming*, *98*(1), 49–71. CrossRef

Birge, J. R., & Louveaux, F. (1997). *Introduction to stochastic programming*. New York, NY: Springer-Verlag.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics*, *15*(2), 129–145. CrossRef

Breithaupt, K., Ariel, A., & Veldkamp, B. P. (2005). Automated simultaneous assembly for multi-stage testing. *International Journal of Testing*, *5*(3), 319–330. CrossRef

Cen, H., Koedinger, K., & Junker, B. (2006). Learning factors analysis—a general method for cognitive model evaluation and improvement. In M. Ikeda, K. Ashlay, & T-W. Chan, (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 164-175). Heidelberg: Springer-Verlag. CrossRef

De Jong, M. G., Steenkamp, J. B. E. M., & Veldkamp, B. P. (2009). A model for the construction of country-specific, yet internationally comparable short-form marketing scales. *Marketing Science*, *28*(4), 674–689. CrossRef

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. New York, NY: W. H. Freeman and Company.

Huitzing, H. A., Veldkamp, B. P., & Verschoor, A. J. (2005). Infeasibility in automatic test assembly models: A comparison study of different methods. *Journal of Educational Measurement*, *42*(3), 223–243. CrossRef

Leucht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, *22*(3), 224–236. CrossRef

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

Mao, X., & Xin, T. (2013). The application of the Monte Carlo approach to cognitive diagnostic computerized adaptive testing with content constraints. *Applied Psychological Measurement*, *37*(6) 482–496. CrossRef

Nemhauser, G., & Wolsey, L. (1988). *Integer and combinatorial optimization*. New York, NY: John Wiley & Sons, Inc. CrossRef

Papadimitriou, C. H., & Steiglitz, K. (1982). *Combinatorial optimization: Algorithms and complexity*. Englewood Cliffs, NJ: Prentice-Hall.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17*(2), 151–166. CrossRef

Theunissen, T. (1985). Binary programming and test design. *Psychometrika*, *50*(4), 411–420. CrossRef

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

van der Linden, W. J. (April, 2012). *Key methodological concepts in the optimization of learning and educational resource availability*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

van der Linden, W. J., & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement*, *35*(3), 185–198. CrossRef

van der Linden, W. J., Ariel, A., & Veldkamp, B. P. (2006). Assembling a CAT item pool as a set of linear tests. *Journal of Educational and Behavioral Statistics*, *31*(1), 81–99. CrossRef

van der Linden, W. J., & Barrett, M. D. (2016). Linking item response model parameters. *Psychometrika*, *81*(3), 650–673. CrossRef

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for IRT-based test design with practical constraints. *Psychometrika*, *54*(2), 237–247. CrossRef

van der Linden, W. J., & Diao, Q. (2011). Automated test form generation. *Journal of Educational Measurement*, *48*(2), 206–222. CrossRef

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22*(3), 259–270. CrossRef

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, *29*(3), 273–291. CrossRef

van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement*, *28*(5), 317–331. CrossRef

Veldkamp, B. P. (1999). Multiple objective test assembly problems. *Journal of Educational Measurement*, *36*(3), 253–266. CrossRef

Veldkamp, B. P. (2002). Multidimensional constrained test assembly. *Applied Psychological Measurement*, *26*(2), 133–146. CrossRef

Veldkamp, B. P. (2012a). Application of robust optimization to automated test assembly. *Annals of Operations Research*, *206*, 595–610. CrossRef

Veldkamp, B. P. (2012b). Ensuring the future of CAT. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 35–46). Enschede, The Netherlands: Ipskamp Drukkers B.V.

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

Verschoor, A. (2004). *IRT test assembly using genetic algorithms*. Arnhem, The Netherlands: Cito B.V.

Author Address

Dmitry I. Belov, Psychometric Research, Law School Admission Council, 662 Penn Street, Newtown, PA 18940, U.S.A. Email: dbelov@lsac.org; belovd@mail.ru