Computer-Assisted Test Assembly Using Optimization Heuristics

Richard M. Luecht
National Board of Medical Examiners

Numerous algorithms and heuristics have been introduced that allow test developers to simultaneously generate multiple test forms that match qualitative constraints, such as content blueprints, and quantitative targets, such as test information functions. A variation of a greedy algorithm is presented here that can be used in a wide range of test assembly problems. The algorithm selects items to have a locally optimal fit to a moving set of average criterion values. A normalization procedure is used to allow the heuristic to work simultaneously with numerous qualitative and quantitative constraints. A complex sample application is demonstrated. *Index terms: automated test assembly, computerized testing, optimal test assembly, optimization, pre-equating, test construction, test information functions.*

Research on optimal test assembly methods has tended to focus on the design of algorithms and/or heuristics that can be implemented as computer software. Some algorithms or heuristics attempt to maximize reliability or some other specific statistical function of the test scores without serious regard to or restrictions placed on the content (e.g., Armstrong, Jones, Li, & Wu, 1996; Luecht & Hirsch, 1992). Other methods attempt to meet varied and complex test specifications, including multiple quantitative targets and multiple categories of qualitative constraints, to increase the likelihood of statistically and content-parallel test forms over time (e.g., Stocking, Swanson, & Pearlman, 1993).

The approach presented here is especially suited for large, complex problems. This approach, called the normalized weighted absolute deviation heuristic (NWADH), is a greedy algorithm (Moret & Shapiro, 1991; Nemhauser & Wolsey, 1988; Zanakis, Evans, & Vazacopoulos, 1989) that uses a series of locally optimal searches to build one or more simultaneous test forms from an item bank.

The NWADH can incorporate large numbers of content or other categorical dimensions, multiple quantitative targets, multiple test forms, and otherwise complicated test construction challenges such as linked item sets or "enemy" items (item sets that are mutually exclusive) on the same test form. Heuristics like the NWADH have gained popularity in a variety of disciplines because of their capability to produce near-optimal solutions to complex, difficult, or time-consuming optimization problems. The most typical use of heuristics and greedy algorithms is to maximize improvement sequentially, beginning at some starting point and systematically determining a localized solution to some objective function. Each step yields more information for decision-making based on the previous selections. A new objective function or set of objective functions is generated at each step based on prior decisions (Moret & Shapiro, 1991). Luecht & Hirsch (1991, 1992), Stocking, Swanson, & Pearlman (1993), Swanson & Stocking (1993b), and van der Linden & Luecht (1994, 1996) demonstrated successful implementations of greedy algorithms specific to matching item response theory (IRT) item information functions to a target test information function (TIF). The

Applied Psychological Measurement, Vol. 22 No. 3, September 1998, 224–236

heuristic implementation presented here is similar to test assembly heuristics developed by Luecht & Hirsch (1991, 1992) and Swanson & Stocking (1993b).

The NWADH

The NWADH is a series of local optimization models that are sequentially solved item-by-item or for sets of items (e.g., test assembly using item sets or item bundles; see Rosenbaum, 1988). The NWADH combines all the categorical and quantitative specifications for a test into a composite objective function to be met for selection of the current item (or set of items). Following each selection, the NWADH updates the composite objective function based on the attributes of the selected items or item sets. This process is repeated until the test is completed.

Quantitative and Categorical Attributes, Targets, and Constraints

Most well-designed tests have documented specifications for the measurement properties and statistical characteristics of the test, as well as other features and attributes that are considered important for building new test forms. It is convenient to refer to the quantitative and categorical *attributes* of items as opposed to the test-level quantitative *targets* and the test-level *constraints* that constitute the test specifications.

Quantitative attributes include classical item statistics, such as percent correct difficulties (i.e., *p* values); item-test correlations; IRT item parameter estimates; IRT functions, such as the information or response functions; and word counts. These item attributes are usually related functionally to one or more targeted measurement properties for the test, such as an IRT target TIF (Luecht & Hirsch, 1992; Stocking, Swanson, & Pearlman, 1993; van der Linden & Boekkooi-Timminga, 1989). In most cases, the item statistics or functions of those item statistics are summed to determine the targeted quantity of the test. For example, an IRT TIF is merely the sum of item information functions. Therefore, it should be possible to select a particular number of items for which the sum of the item information functions is very close to the target TIF. In addition, the specifications may call for other nonpsychometric quantities, such as an acceptable range for the total word count on the test.

The categorical attributes of items or sets are taxonomically coded features stored in the item bank (e.g., content codes, cognitive level codes, item type codes, and item author identification codes). These features are typically controlled at the test level by introducing constraints as part of the test specifications used in automated test assembly. For example, the required frequencies or proportions of items for a test form covering various subject areas typically taught in a high school mathematics course could be stated as either constraints on the range of items to include in each subject area or as exact frequencies (e.g., 10 to 15 items of intermediate algebra or 12 items of geometry). For some tests, the content specifications may include only a few general categories. In other cases, the content specifications may cover a very long content outline with numerous levels for each category and many auxiliary classification taxonomies with additional constraints (e.g., item types or formats, cognitive levels, and item authors).

Perhaps the most common constraint in test assembly is the test length. The test length can be constrained to be fixed (e.g., 100 items) or variable (e.g., 90 to 120 items) in which minimum and maximum test length constraints could be specified. Variable test lengths are most typically encountered in computerized adaptive testing (CAT), in which different examinees can receive different length tests (e.g., Swanson & Stocking, 1993a).

Two additional possible types of constraints are limits on the item reuse frequencies and item exclusions. Constraining item reuse frequency is important because it controls the exposure of test materials to examinees. If items are used too frequently, examinees may cheat by memorizing and sharing the items. Item exclusions are usually rules relating to the relationships among items. For example, an all-or-none exclusionary rule may be used so that any selected item set is used in its entirety or not at all. Another example governs the use of item enemies. An exclusionary rule may be used so that pairs or clusters of enemies (items that cue one another) are mutually exclusive on the same test form.

A General Version of the NWADH

Let u_i denote some quantitative attribute of $i = 1, \ldots, I$ items in an item database. For a particular test form comprised of $n \leq I$ items, the corresponding test attribute can be expressed as

$$\sum_{i=1}^{n} u_i \,, \tag{1}$$

where the attributes $u_i, i = 1, \ldots, n$, are assumed to be algebraically additive for all items in the test. Quantitative test attributes such as the mean number-correct score as the sum of the item proportion-correct difficulties, or the IRT TIF as a sum of item information functions, satisfy this assumption of additive quantitative attributes across items. The function, u_i, can likewise denote subsets of items in which additivity is only assumed across the item subsets as units, not necessarily within subsets (e.g., Rosenbaum, 1988).

A corresponding quantitative target test function is defined as T. That is, T denotes some corresponding test function that should be met (e.g., a TIF to be matched for building parallel test forms over time). An objective function could be defined as

$$\left| T - \sum_{i=1}^{I} x_i u_i \right| \,, \tag{2}$$

subject to two simple constraints:

$$\sum_{i=1}^{I} x_i = n \tag{3}$$

and

$$x_i \in \{0, 1\}, \quad i = 1, \ldots, I \,. \tag{4}$$

In Equation 4, x_i is the decision variable for selecting the n items. This constraint is stated in Equation 3. That is, $x_i = 1$ if the item is to be included in the test; $x_i = 0$ if the item is not to be included. The absolute difference function in Equation 2 can be replaced by a least squares criterion or similar function amenable to minimization. For many practical applications, the absolute difference function is adequate and sometimes preferable (e.g., Luecht & Hirsch, 1991, 1992).

To implement the NWADH algorithm, the absolute deviation minimization problem must be transformed into a maximization problem and additional notation must be introduced. The item selection process can be managed at the unit level, where $j = 1, \ldots, n$, objective functions are to be maximized. That is, for a series of n optimization models,

$$\text{Maximize} \sum_{i=1}^{I} e_i x_i \,, \tag{5}$$

subject to

$$\sum_{i=1}^{I} x_i = j \,, \tag{6}$$

$$x_{i_1} = x_{i_2} = \ldots = x_{i_{j-1}} = 1 \,, \tag{7}$$

and

$$x_i \in \{0, 1\}, \quad i = 1, \ldots, I \,, \tag{8}$$

where e_i is a variable coefficient,

$$e_i = 1 - \frac{d_i}{\sum_{i \in R_{j-1}} d_i} \,, \quad i \in R_{j-1} \tag{9}$$

and

$$d_i = \left| \left(\frac{T - \sum_{k=1}^{I} u_k x_k}{n - j + 1} \right) - u_i \right| ; \quad i \in R_{j-1} \,. \tag{10}$$

In Equations 9 and 10, R_{j-1} is defined as a set of indexes for the remaining items in the item bank after excluding the selected $j - 1$ items.

The localized optimization model must be solved at each item selection, $j = 1, \ldots, n$, because Equations 5 through 10 only relate to the selection of the current item, j. As each new item is selected, Equation 5 is incremented through Equations 9 and 10. Finally, the expression in Equation 10,

$$\frac{T - \sum_{k=1}^{I} u_k x_k}{n - j + 1} \,, \tag{11}$$

provides the current value of the target function after removing previously selected items; thus, Equation 11 is the target value for selection of the next item. Note that if upper or lower bounds on the target test function are used in place of the explicit target test function, T, simple modifications to the heuristic can be made to accommodate those boundaries using tolerances (e.g., Stocking, Swanson, & Pearlman, 1993).

It is advantageous to normalize the coefficients as shown in Equation 9 for a number of reasons. Dividing the d_i variables by their sum over all eligible items (i.e., the unselected items in the set R_{j-1}) transforms the absolute difference function into a proportional quantity. Therefore, many different types of criteria can be treated simultaneously, minimizing any potential effects due to the scaling or magnitude of the functions T or u_i. However, because the overall objective still involves meeting the target, T, the normalized coefficients (proportions) must be subtracted from 1, as shown in Equation 9. This reverses the magnitude of the fitting function (i.e., Equation 10) so that larger values of e_i imply closer fit to the interim target for selecting item $j = 1, \ldots, n$.

Incorporating Content and Other Categorical Attributes

Under the NWADH, content or other categorical attributes, often specified as frequencies in the test specifications, require a slightly different treatment than quantitative attributes. The NWADH can incorporate multiple content dimensions or facets and levels within those dimensions (e.g., content outline sublevels). For convenience and clarity, only a single content dimension is considered here. The sample application described below involved a large number of categorical constraints spanning several content taxonomies.

Let G denote the total number of content categories for a particular content or other taxonomic dimension with categories indexed $g = 1, \dots, G$. Let v_{ig} $\{0,1\}$ denote the binary incidence of an item that has a particular categorical content attribute. $v_{ig} = 1$ if the item belongs in the category and $v_{ig} = 0$ if the item does not belong. Finally, let $Z_g^{[\min]}$ represent some minimum constraint quantity and let $Z_g^{[\max]}$ represent some maximum constraint quantity for each content category. For any particular categorical content attribute, the sum

$$\sum_{i=1}^{I} v_{ig} \, , g = 1, \dots, G \, , \tag{12}$$

provides the availability of items that have that attribute. Note that it is assumed that the availability of items is greater than 0 for all specified categories and that if

$$\sum_{i=1}^{I} v_{ig} < Z_g^{[\min]} \, , \tag{13}$$

then the constraint $Z_g^{[\min]}$ will be adjusted so that

$$\sum_{i=1}^{I} v_{ig} \geq Z_g^{[\min]} \, . \tag{14}$$

Recall that x_i is a binary decision variable denoting whether each item was selected. The count of items selected for each content category in the NWADH sequence (i.e., up to the preceding item selection, $j - 1$) can, therefore, be computed as

$$\sum_{i \in R_{j-1}} v_{ig} \, , g = 1, \dots, G \, . \tag{15}$$

This sum can be used to empirically determine a set of weights, W_g, $g = 1, \dots, G$. These category weights can take on user-assigned (e.g., integer weights or points) or empirically determined values (e.g., proportions based on remaining availabilities of items in the bank after each item selection). A simple and effective weighting scheme follows. Assume that $Z_g^{[\min]} < Z_g^{[\max]}$. Using a single item point assignment scheme, the weights for each category, W_g, could be assigned to take on one of three values:

$$\text{if } \sum_{i \in R_{j-1}} v_i \geq Z_g^{[\max]} \, , \text{ then } W_g = 0 \, ; \tag{16}$$

$$\text{if } Z_g^{[\min]} \leq \sum_{i \in R_{j-1}} v_i < Z_g^{[\max]} \, , \text{ then } W_g = 1 \, ; \tag{17}$$

or

$$\text{if } \sum_{i \in R_{j-1}} v_i < Z_g^{[\min]} \text{, then } W_g = 2(g = 1, \ldots, G) . \tag{18}$$

At each iteration of the NWADH, the weights are accumulated by each unselected item. Therefore, items belonging to categories not meeting the minimum constraint, $Z_g^{[\min]}$, receive more weight than those that meet the minimum and do not exceed the maximum constraint, $Z_g^{[\max]}$. Items in categories at or in excess of the maximums receive no weight.

It may seem conspicuous that there are no negative or penalty weights assigned for exceeding any maximum, $Z_g^{[\max]}$. Instead of penalizing items that violate particular upper bound constraints, an alternate approach was developed that rewards all items not having the categorical attribute that is at or exceeds $Z_g^{[\max]}$.

Let $W^{[\max]}$ represent the maximum value of the weights across all G categories. An approximate complement to W_g, denoted $\underline{W}_g$, can be computed as

$$\underline{W}_g = W^{[\max]} - \frac{1}{G} \sum_{i=1}^{G} W_g . \tag{19}$$

As the constraints in a particular category are met, the average weight term approaches $W^{[\max]}$ and $\underline{W}_g$ approaches 0. Items not belonging to any of the specified (i.e., constrained) categories are awarded bonus points for not contributing to categories at or in excess of the maximums.

Now, let c_i be the accumulated content weights for each unselected item in R_{j-1}. The weights, W_g and $\underline{W}_g$, are used to compute c_i:

$$c_i = v_{ig} W_g + (1 + v_{ig}) \underline{W}_g \ ; \quad i \in R_{j-1}, \ g = 1, \ldots, G . \tag{20}$$

This new item-level variable can be normalized for all unselected items (i.e., items remaining in the set of unselected items, R_{j-1}). The normalized variables can then be used with the normalized coefficient in Equation 9 to define a new variable coefficient to be maximized in the objective function,

$$e_i^* = \left(1 - \frac{d_i}{\sum\limits_{i \in R_{j-1}} d_i} \right) + \frac{c_i}{\sum\limits_{i \in R_{j-1}} c_i} , \quad i \in R_{j-1} , \tag{21}$$

where e_i^* can now be substituted for e_i in Equation 5. For some applications, user-assigned, proportional weight coefficients can be incorporated into the composite function in Equation 21 to reflect the importance of meeting statistical versus categorical or content specifications. By adding new terms to Equation 21 to accommodate multiple categorical or quantitative targets or constraints, the NWADH can be extended to handle very large test assembly problems.

Application-Specific Modifications to the NWADH

Improving speed of the NWADH. When content or other categorical attributes are fixed as primary test construction requirements having exact quantities on each test form, it is possible to significantly improve the speed of the NWADH by implementing prioritized searches for the items within particular categories. For example, Z_g (a constraint's minimum or maximum value) can be treated as a fixed quantity so that every test form must have exactly Z_g items in the G categories. A *need-to-availability* ratio for each category can be computed as

$$A_{(j-1),g} = \frac{Z_g - \sum\limits_{i \in R_{j-1}} v_{ig}}{\sum\limits_{k=1}^{I} V_{kg} - \sum\limits_{i \in R_{j-1}} V_{ig}} \,, \quad g = 1, \ldots, G, \ j = 1, \ldots, n \,. \tag{22}$$

If the denominator of Equation 22 is 0, then there are no items remaining in the category and $A_{(j-1),g} = 0$. This need-to-availability ratio can be updated after each item selection, with the largest values indicating the highest priority. At each iteration, the NWADH is applied only to items in the category with the maximum value of $A_{(j-1),g}$ (i.e., the greatest need-to-availability ratio). This can significantly increase the speed of the overall solution because the computationally intense NWADH is applied only to a small set of items.

When prioritized in this manner, items in categories with the greatest need and smallest availability tend to be selected earlier than items in categories with low demand or a surplus of items. When the demand is high and the supply is small (e.g., the specifications call for five items in a subject area and only five items exist), there is little choice of selection. This prioritization mechanism forces high-priority items into the solution early, giving the NWADH more flexibility to build around them later.

Creating multiple test forms. The NWADH can also concurrently select items for multiple test forms by implementing separate objective functions for each form. The starting form for each item selection in the series, $j = 1, \ldots, n$, can be randomly assigned to prevent the same form from always having the first choice. In practice, the form-specific objective functions will probably diverge enough after several items have been selected to prevent all the forms from competing for the same items.

Item reuse and exposure. It is also possible to allow items to appear on multiple forms. An upper bound can be placed on the number of reuses allowed per item or on all items in the item bank. This constraint can even be converted into a proportion of maximum allowable use and incorporated into the variable coefficient term, so that items having no or smaller amounts of reuse across test forms are more likely to be selected for a particular form (all other considerations being equal). In certain test assembly situations in which test forms are constructed as the examinee is taking the test (i.e., CAT), exposure controls can be implemented as part of the item selection heuristic to govern the expected rate of item exposure within the examinee population.

Incorporating item sets. Item sets are multi-item units (such as several items associated with a reading passage, vignette, or other common stimulus). They are easily incorporated in the NWADH. The objective function for the heuristic can be modified to locally optimize the selection of multiple items as easily as a single item. In fact, item sets may carry their own class-level categorical attributes (e.g., type of reading passage) that can be entered as constraints.

There are two recommendations for using item sets with the NWADH. Similar recommendations have been made elsewhere (e.g., Luecht & Hirsch, 1991, 1992; Swanson & Stocking, 1993b). First, all items belonging to a set should be treated on an all-or-none basis. Either all of the items in the set are selected or none of the items are selected. This restricts the potential number of combinations (of items within sets and sets on test forms) that would otherwise need to be evaluated. Second, item set size can be included as a categorical attribute of the items in each set, and test-level constraints on the number of sets at each viable set size can be included. For example, test developers might specify that each test form must contain 10 two-item sets, 15 four-item sets, and 4 eight-item sets.

This method of restricting set sizes reduces the overall number of combinations to be considered. Implementing both of these recommendations can improve the heuristic's performance.

Using IRT item and test information functions. In IRT, the probability of observing item response patterns is modeled as a function of an underlying latent trait, θ, and characteristics of the items, $\boldsymbol{\xi}_i$,

$$\text{Prob}(y_i = 1|\theta, \boldsymbol{\xi}_i) \equiv P_i(\theta) \ , \quad i = 1, \ldots, I \ , \tag{23}$$

where

y_i is a dichotomous response score,

$P_i(\theta)$ is a logistic or cumulative normal probability function, and

$\boldsymbol{\xi}_i$ is a vector of item parameters from a dichotomous IRT model (e.g. Lord, 1980;).

One of the principal functions used in IRT for test assembly is the TIF,

$$\text{TIF} = I(\theta) = \sum_{i=1}^{n} I_i(\theta) \ , \tag{24}$$

where

$$I_i(\theta) = \frac{\left[\dfrac{\partial P_i(\theta)}{\partial \theta}\right]^2}{P_i(\theta)[1 - P_i(\theta)]} \tag{25}$$

is the item information function (Birnbaum, 1968; Lord, 1980). From Equation 24, it is clear that the TIF is comprised of additive item information functions at any value of θ. This satisfies the earlier assumption about the algebraic additivity of the item functions in the test function, i.e., $I_i(\theta)$ is substituted for the generic function, u_i. The TIF governs the distribution of estimation errors, and tests with matching TIFs are considered to be at least approximately parallel (e.g., Samejima, 1977).

In the context of automated test assembly, a target TIF can be used to ensure statistical parallelism across multiple test forms. That is, if the target function is defined as $I(\theta)$ and all constructed test forms meet that target, each form will satisfy what Samejima (1977) called "weak parallelism."

Because θ is real-valued, it is not practical in test construction to simultaneously consider all values of θ. Instead, it is convenient to select some smaller number of discrete points, $\theta_q, q = 1, \ldots, Q$, to represent θ over a reasonable range of values. For example, $Q = 25$ equidistant quadrature points for $-2.0 \leq \theta_q \leq +2.0$ might be used to represent $P(\theta)$ and $I(\theta)$ for a population of examinees, as $P(\theta_q)$ and $I(\theta_q), q = 1, \ldots, 25$.

Let $I(\theta_q)$ denote the target TIF. Then, from Equations 9 and 21, the variable coefficients can be modified as necessary to

$$e_i = 1 - \frac{\sum_{q=1}^{Q} d_{iq}}{\sum_{i \in R_{j-1}} \sum_{q=1}^{Q} d_{iq}} \ , \quad i \in R_{j-1} \ , \tag{26}$$

where

$$d_{iq} = \left| \left[\frac{I(\theta_q) - \sum_{k=1}^{I} I_k(\theta_q)x_k}{n-j+1} \right] - I_i(\theta_q) \right| , \quad i \in R_{j-1},\ q = 1, \ldots, Q\ . \tag{27}$$

Density weights or any other desired user-assigned weight functions can also be incorporated into the summation over the Q coefficient values of d_{iq}, e.g., $\approx \sum_q d_{iq} f(\theta_q), q = 1, \ldots, Q$.

Example Application of the NWADH

The NWADH was applied to a medical licensure test assembly problem. Four test forms were selected from a common item bank, a large number of categorical constraints covering multiple content dimensions was used, and item sets were included in the final test forms. In addition, different TIF targets were used to vary the mean test difficulty of some of the selected test forms. One form was designed to be easy, two parallel forms were designed to be moderately difficult, and one form was designed to be difficult. However, all four forms had to meet the same categorical test specifications. This application demonstrated how the target TIFs could be manipulated to systematically vary the test form difficulty and how quickly the NWADH can produce test forms. Item selection was done by the computer program BUILDER (Luecht, 1994, 1995, 1996).

The Test Assembly Problem

The construction of the United States Medical Licensing Examination (USMLE) Step 2 (Federation of State Medical Boards & National Board of Medical Examiners, 1996) was simulated, using actual items from the Step 2 item banks. Step 2 measures medical knowledge and understanding of clinical science. Although current Step 2 test forms usually contain approximately 600 items and are administered over two days, this application involved the assembly of 300-item forms. Each of the four 300-item test forms was selected from the Step 2 item bank.

The content specifications for each form were designed to be approximately proportional to the general content requirements on the USMLE Step 2 outline (Federation of State Medical Boards & National Board of Medical Examiners, 1996). This outline requires 10% to 15% coverage of growth and development and general principles of care, and 85% to 90% coverage of organ systems and types of disorders. These primary content areas are represented by primary group categories, which can include three to four levels for purposes of test construction. The sublevel coding of the Step 2 content includes a second content dimension covering general physician's tasks (health and health maintenance, diagnosis, and principles of patient management) and mechanisms of disease.

All Step 2 items are patient vignettes. Therefore, in addition to meeting the primary content specification, the item selections were constrained for features of the vignettes, including site of care, patient age, and gender. Finally, item selections were constrained according to item types (one-best-answer discrete items versus extended matching sets sharing a common option list) and the allowable sizes of the sets. Table 1 shows that there were eight groups of constraints in which each constraint group contained some larger number of individual categories for which constraints were specified. The total number of categories and associated constraints used for assembling each test form for each group are also shown. Items were allowed to satisfy multiple categories and many of the constraints were nested in other constraints because of the hierarchical structure of the Step 2 content outline.

Table 1
Number of Categories and Constraints Used in Test Assembly (Constraints Count Includes Minimum and Maximum Constraints for Most Groups)

Constraint Group	Categories	Constraints
USMLE Step 2 content outline levels	353	706
Site of care	4	8
Patient gender	2	4
Patient age group	6	12
Adult patient age ranges	5	10
Specific disease categories	11	22
Item types	2	2
Item set sizes	2	2

Item set constraints called for 33 sets to be included in each test form, so that 100 of the 300 items were part of extended matching multi-item sets. This group of constraints called for 29 two-item sets and 4 three-item sets in each test form. Alternate combinations of set sizes were not analyzed.

The item bank consisted of 3,165 items that had been previously used in a Step 2 form, at least as experimental pretest items. All of the items were fully coded for all of the categories specified as constraints. The item bank included 263 extended matching sets and each set contained two or three items.

All 3,165 items in the bank were calibrated on a common scale using the Rasch IRT model (Wright & Stone, 1979),

$$P_i(\theta) = \left[1 + \exp(b_i - \theta)\right]^{-1} , \quad i = 1, \ldots , I . \tag{28}$$

with $b_i = \xi_i$ denoting a difficulty parameter estimate for each item. Under this model, the item information function given in Equation 25 simplifies to

$$I_i(\theta) = P_i(\theta)[1 - P_i(\theta)] , \quad i = 1, \ldots , I . \tag{29}$$

The item difficulty parameter estimates in the bank were standardized to have mean 0 and standard deviation (SD) 1.

The NWADH was required to simultaneously build one easy form, two moderately difficult forms, and one difficult form, all subject to the specified categorical constraints, each including 300 items and 100 extended-matching sets per form. No item overlap was allowed.

Figure 1 shows the three target TIFs used for test assembly. Under the Rasch model, the maximum value of the TIF for a 300-item test was 75. The target information functions for the easiest and most difficult test forms were slightly higher than the center target for the moderately difficult test forms. This was done to achieve similar expected number-correct score distributions and is discussed below.

Results

The content specifications were matched for all 353 primary content categories on all four test forms. There were some minor constraint violations on particular forms for some of the secondary features. Table 2 shows the number of constraint violations, by form, for each of the eight constraint groups from Table 1. All violations involved exceeding the maximums, usually by no more than one item. Considered across forms, this simultaneous test assembly problem involved satisfying all but 7 of 3,064 constraints.

Table 3 summarizes the item statistics for each of the four forms, including the expected number-correct scores computed from population estimates of θ for a large sample of recent USMLE Step 2

Figure 1
Target TIFs for the Easy, Moderate, and Difficult Tests

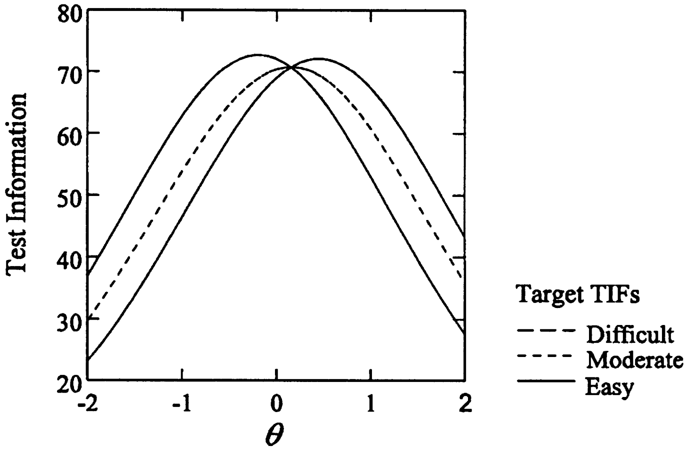


Table 2
Number of Constraints Violated in Each Constraint Group for Each Test Form

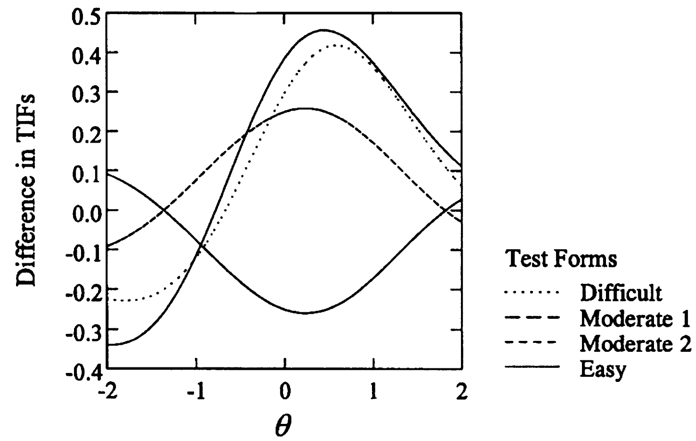
Constraint Group	Easy Form	Moderate Form 1	Moderate Form 2	Difficult Form
USMLE Step 2 content outline levels	0	0	0	0
Site of care	0	0	0	0
Patient gender	0	0	0	0
Patient age group	0	0	0	0
Adult patient age ranges	0	1	1	1
Specific disease categories	1	0	1	2
Item types	0	0	0	0
Item set sizes	0	0	0	0

examinees. The intended form-to-form variations in mean item difficulty are apparent. It is also clear that the easiest and most difficult forms had smaller item difficulty SDs than the two moderate forms. This is consistent with the narrower target information functions used for those forms (see Figure 1).

Table 3
Item Difficulties and Expected Number-Correct Score Statistics for Four Generated 300-Item Test Forms

	Rasch Item Difficulties		Expected Number-Correct Score		
Test Form	Mean	SD	Mean	SD	Coefficient α
Easy	−.201	.374	230.096	21.004	.886
Moderate 1	.111	.519	210.000	21.001	.866
Moderate 2	.119	.475	209.990	20.998	.867
Difficult	.432	.418	189.040	20.980	.850

Figure 2
Differences in TIFs (Target Minus Test) Between Four Fitted Test Forms and the Targets



The expected number-correct scores in Table 3 confirm the differences in mean item difficulty. The mean scores shown correspond to percent-correct scores of approximately 77%, 70%, and 63% for the easy form, the two moderate forms, and the difficult form, respectively. Also note the similarities in expected sample SDs. As noted earlier, this was intentional and one of the criteria used in determining the test information targets in Figure 1.

Each of the four forms came quite close to its respective target information function. Figure 2 shows the difference between each of the targets and the corresponding observed TIFs for the four test forms. All differences were between −.5 and +.5. The figure indicates very good fit, especially if the misfit is considered in standard error units (i.e., taking the reciprocal of the square root of the information function as the asympotic standard error of a maximum likelihood θ estimate; see Wright & Stone, 1979; Lord, 1980).

The analysis took 88.3 seconds to complete on a Pentium 166Mz microcomputer. This indicates how quickly heuristics like the NWADH can build multiple parallel test forms in terms of statistical properties and content, and forms that differ in difficulty by design.

Discussion

The method demonstrated here illustrates some of the advantages of being able to simultaneously incorporate content and statistical test specifications in a test assembly algorithm. The use of IRT item statistics was meant to be illustrative, not prescriptive. This heuristic has also been successfully implemented using classical testing techniques (i.e., using proportion correct and point-biserial correlations to fit a target number-correct score mean and SD).

The NWADH simply matches the coded or numerical features or attributes of items to specified test level constraints. The heuristic cannot manage qualities of the test or items, other than as explicit codes or values that can be counted or summed. The process of selecting items for tests, at least those aspects that are manageable by computer, will always be restricted by three aspects of the test assembly problem: (1) the quality and size of the item bank (i.e., the item bank inventory); (2) the reasonableness of the constraints and targets (given the eligible item bank); and (3) the degree to which the coded content or other qualitative attributes of the items are salient, are reliably assigned, and can be clearly defined as minimax constraints.

No heuristic can implement the entire process of test construction, which often requires extensive interaction between test developers and item data. Sometimes tools are created that may streamline certain aspects of the process. The NWADH is an effective tool for managing large minimax problems and it seems particularly well-suited for a variety of test construction applications. The purpose of the heuristic is to help test development practitioners create parallel draft test forms in relatively short periods of time.

References

Armstrong, R. D., Jones, D. H., Li, X., & Wu, I-L. (1996). A study of a network-flow algorithm and a noncorrecting algorithm for test assembly. *Applied Psychological Measurement, 20,* 89–98.

Birnbaum, A. (1968). Estimation of ability. In F. M. Lord & M. Novick, *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Federation of State Medical Boards and National Board of Medical Examiners (1996). *USMLE 1997 Step 2 general instructions, content descriptions and sample items.* Philadelphia: Author.

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

Luecht, R. M. (1994, 1995, 1996). *Builder* [Computer program].

Luecht, R. M., & Hirsch, T. (1991, June). *Computerized test construction of parallel forms for problem-linked items.* Paper presented at the annual meeting of the Psychometric Society, New Brunswick NJ.

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

Moret, B. M. E., & Shapiro, H. D. (1991). *Algorithms from P to NP: Volume I, design and efficiency.* Redwood CA: Benjamin/Cummings Publishing.

Nemhauser, G. L., & Wolsey, L. A. (1988). *Integer and combinatorial optimization.* New York: Wiley.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53,* 349–359.

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

Stocking, M. L., Swanson, L., & Pearlman, M. (1993). Application of an automated item selection method to real data. *Applied Psychological Measurement, 17,* 177–186.

Swanson, L., & Stocking, M. L. (1993a). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292.

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

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A minimax model for test design with practical constraints. *Psychometrika, 54,* 237–248.

van der Linden, W. J., & Luecht, R. M. (1994). *An optimization model for test assembly to match observed score distributions* (Research Report 94-7). Enschede, The Netherlands: University of Twente, Department of Education.

van der Linden, W. J., & Luecht, R. M. (1996). An optimization model for test assembly to match observed score distributions. In G. Englehard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3). Norwood NJ: Ablex.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Zanakis, S. H., Evans, J. R., & Vazacopoulos, A. A. (1989). Heuristic methods and applications: A categorized survey. *European Journal of Operational Research, 43,* 88–110.

Author's Address

Send requests for reprints or further information to Richard M. Luecht, National Board of Medical Examiners, 3750 Market Street, Philadelphia PA 19104, U.S.A. Email: rluecht@mail.nbme.org.