AICPA

AMERICAN INSTITUTE OF CERTIFIED PUBLIC ACCOUNTANTS

Technical Report

SERIES TWO

Test Information Targeting Strategies for Adaptive Multistage Testing Designs

Chicago, IL – April 2003

Manuscript for presentation at NCME

Richard M. Luecht
William Burgin

May 2004 Number 6

Technical Report

SERIES TWO

Test Information Targeting Strategies for Adaptive Multistage Testing Designs

Chicago, IL – April 2003

Manuscript for presentation at NCME

Richard M. Luecht
William Burgin

May 2004 Number 6

Adaptive multi-stage testlet (MST) designs appear to be gaining popularity for many large-scale computer-based testing programs. In contrast to item-level computerized adaptive testing (CAT) designs, these adaptive MST designs use a modularized configuration of preconstructed testlets and embedded score-routing schemes to essentially prepackage different forms of an adaptive test. The adaptive nature of a MST offers the usual psychometric advantage(s) of improvements in testing efficiency (Luecht, Hadadi, and Nungester, 1996; Luecht and Nungester, 1998). However, there are other practical advantages (Luecht, 2000). First, by preconstructing and identifying the MST units as "data objects", test developers gain the capability to directly control the quality of the test forms and to verify the integrity of post-administration response data. Second, given the simplicity of scoring and routing mechanisms for MST, and the data management and processing loads are typically minimized, especially when using the Internet to interface between centralized servers and local test administration workstations. Finally, MST designs inherently provide many straightforward ways of dealing with item and test exposure risks.

An adaptive MST presents as a series of testlets or multi-item modules. The testlets are bundled, together with scoring routing rules, in a data object called a "**panel**" (Luecht & Nungester, 1998). Examinees can preview or review the test items within a testlet and change answers. Scoring and adaptive routing occurs between stages, after the examinee has "submitted" his or her testlet. The between-stage adaptive routing algorithm selects the testlet for the next stage using cumulative performance up to that point in the test.

Every panel is an instantiation of a particular MST **panel design template**. Panel design templates can vary with respect to five attributes: (1) the number of adaptive testing stages (i.e., two, three, or more stages), (2) the number of testlets per stage, (3) the size of the testlets per stage, (4) the statistical characteristics of the testlets within and across stages (e.g., average difficulty and amount of information provided by each testlet); and (5) the nature and extent of content, other categorical item-attribute requirements, and relevant quantitative test specifications required for the testlets at each stage. Figure 1 presents a general panel design that has three stages with one, three, and five testlets per stage, respectively. This would be called a "1-3-5" panel design (Luecht and Nungester, 1998). The solid arrows denote the primary pathways taken by the majority of examinees (i.e., examinees whose responses fit well the underlying IRT model). The dotted arrows represent auxiliary pathways that allow the panel to adapt to the examinee's ability after the first stage. In general, having more stages and using testlets of more varied difficulty per stage allow for greater adaptation (Luecht, Hadadi, and Nungester, 1996; Luecht and Nungester, 1998; Xing and Hambleton, 2001; Jodoin, Zenisky, and Hambleton, 2002).

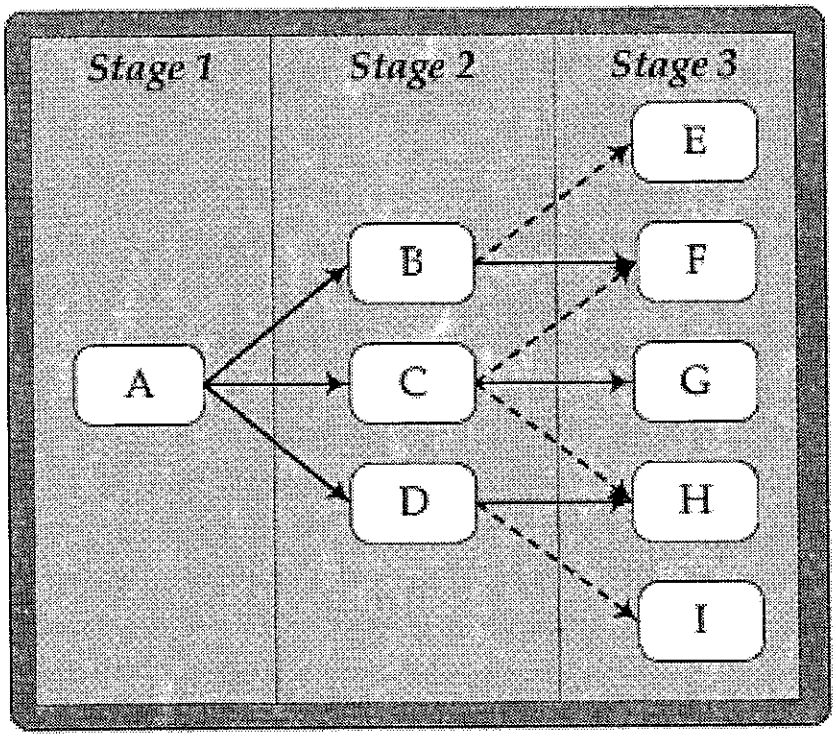

Figure 1. A General "1-3-5" MST Panel Design Template

Building MST panels can be a complex process. Each panel needs to meet a variety of test specifications (i.e., constraints and statistical objectives) at the testlet level, and possibly, at the total-test level. Automated test assembly (ATA) procedures (e.g., van der Linden, 1998; Luecht, 1998, 2000) are well suited for constructing multiple panel instances from an active item bank, using the panel design template as a model. That is, the items are assigned to multiple testlets and the testlets are assigned to panels. The inventory of items on hand, item overlap restrictions, and other ATA constraints largely determine the number of panels that can be constructed from an existing item bank.

Central to building MST testlets and panels, and a primary focus of this paper, is the choice of statistical targets for the testlets. It is common to use test information functions (TIFs) as the preferred statistical targets in ATA (Van der Linden, 1998-; Luecht, 1992, 1998). A target TIF is a specified curve that indicates the amount of test information required across the latent proficiency scale, θ. In the present context, a target TIF also indirectly helps control the distribution of the item difficulties for each testlet (i.e., the average and range of item difficulty). Because each panel has testlets of varied difficulty, different target TIF curves are required, one for each testlet position within a particular panel design template. For example, the panel design implied by in Figure 1 would require nine target TIFs, one for each testlet. Once these target TIFs are determined, ATA procedures can be used to replicate those targets and generate a subpool of testlets.

A target TIF, denoted $T(\theta)$, is an idealized amount of measurement precision required per testlet. For a given testlet of length nj the basic goal in ATA can be expressed by a simple mathematical programming goal, to

minimize $$\left|T(\theta) - \sum_{i=1}^{I} x_i I_i(\theta)\right| \quad (1)$$

subject to $$\sum_{i=1}^{I} x_i = n_j \quad (2)$$

$$x_i \in \{0,1\} \quad (3)$$

for i=1,...,I items in the item bank, where $Ii(\theta)$ is the IRT item information function for each of the items in the bank, computed at a single θ value[1]. Equation 2 constrains the number of items to match the desired length of the testlet. In practice, additional constraints for content and other attributes would be included. Equation 3 implies that xi is a binary decision variable, where $x_i = 1$ if the item is included in the testlet or $x_i = 0$, otherwise.

Because the testlets are part of a larger panel, each target TIF needs to reflect three goals: (1) to help guarantee that the IRT test information functions provide measurement precision where it is most needed for critical decisions and score-reporting purposes; (2) to derive targets that make it feasible to actually produce large numbers of content-balanced MST testlets[2]; and (3) to achieve a desired level of conditional exposure of test materials in the examinee population for each constructed panel.

For illustrative purposes, we can consider a very basic "1-2" MST panel design as shown in Figure 2. This 1-2 panel design has one testlet at Stage 1 (testlet A) and two testlets at Stage 2 (testlets B and C). A separate TIF is shown for each testlet. A cumulative normal distribution (represented by the short dotted curve) has been superimposed on the 1-2 design; the right-hand scale shows the cumulative proportions for the population.

---

[1]In operational ATA, the information functions are typically computed for a vector of points spanning some region of the proficiency scale.
[2]A highly informative target that greatly exceeds the average item information in the bank can lead to dramatic variability among testlets.

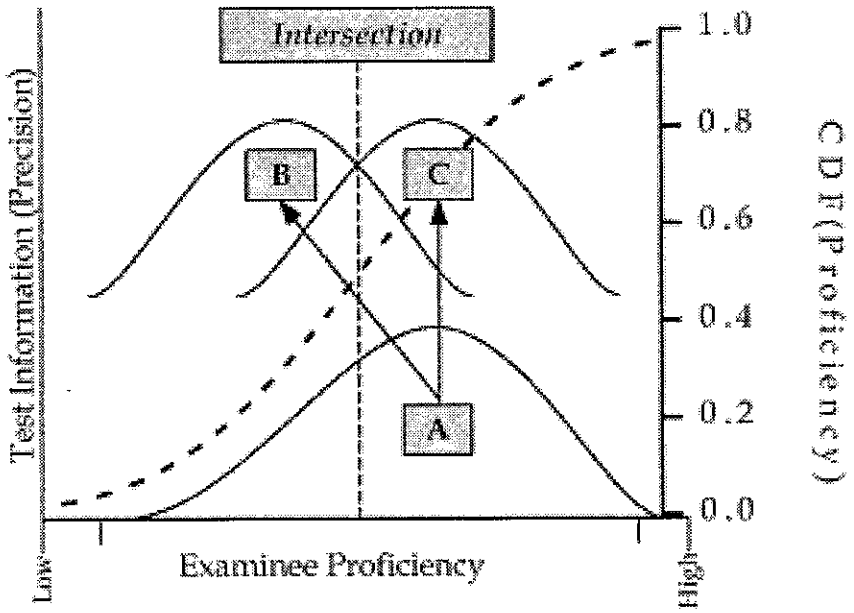

Figure 2. Three Target TIFs for a 1-2 Panel, with CDF (Proficiency)

Figure 2 has three important, if subtle, features. First, for the two possible routes through the panel, there are two corresponding test information peaks: one implied for testlets A + B; the other more directly discernible for testlets A+C. The rightmost peak (testlets A+C) might represent an attempt to maximize the test information at a pass/fail cut point for mastery decisions (e.g., in a certification or licensing testing situation). The leftmost peak might attempt to [somewhat] maximize the score precision for failing examinees in order to provide diagnostic feedback to help them study for a retest. Second, the curves for testlets B and C **intersect.** If the shape and/or location of those two testlet-level curves were to change, the intersection point will likewise change. Third, the intersection point for the test information curves occurs at approximately the 50th percentile of proficiency distribution (referring to the cumulative proficiency distribution curve). The implication is that the examinees in the population with proficiency below the median would be administered a maximally informative test by getting testlets A+B (given only these testlets); testlets A+C would be maximally informative for the examinees above the median.

This rather simple example suggests a straightforward way of generating target TIFs for MST panels. In the context of Figure 2, the goals should be clear. For the two possible routes (i.e., branching to testlets B or C), find feasible testlet TIF targets for testlets B and C: (1) that explicitly route a specified proportion of the population to either testlet B or testlet C and (2) that make the targets for testlet B and C as informative as possible, considering the quality of the items in the item bank, content, and other test specifications. We call our solution the **conditional information targeting (CIT) strategy.** Although the

CIT strategy generalizes fairly easily to panels having multiple routes and stages, we restrict our description to the simple 1-2 design shown in Figure 2 for purposes of illustration. We further ignore the issue of targeting the first-stage testlet (see Luecht, 2000) and focus on deriving target TIFs for the second-stage testlets, B and C.

The CIT strategy depends on three values along the proficiency scale that we refer to as "posts". There is a **left post** (θ_L), a **right post** (θ_R), and a **center post** (θ_C). These posts are related to each other as follows: $\theta_L < \theta_C < \theta_R$. The center post, θ_C, is always fixed and controls the proportion of the population that the test developer intends to be routed left and right. For example, assuming a normal distribution of proficiency, if θ_C =0.0, then 50 percent of the population would be routed to testlet B and 50 percent to testlet C. If we wanted only the lower 30 percent of the population routed to B (e.g., to less-secure testlets used to improve diagnostic scoring), we would set the center post value to θ_C = -0.524. The left and right posts are used to move (or fix) the provisional target TIFs for testlets B and C. Either post (but not both) may be constrained to have a fixed position on the scale—for example, θ_R might be the pass/fail cut point on the proficiency scale.

The CIT strategy also requires a mechanism to compute robust, provisional target TIFs that reflect the actual properties of the items in the item bank as well as relevant test assembly specifications (e.g., content constraints). These provisional target TIFs can be computed by merely selecting items, without replacement, to have maximum information at either the left or right post. However, to ensure that the targets are *feasible* to each use in constructing multiple testlets and panels, it is helpful to generate several maximally informative testlets at each post and average the unique information functions. That is, we want to determine some number, M > 1, of non-overlapping testlets to build at both the left and right poles. If the item bank is large (e.g., more than 500 items) five to ten testlets can be constructed per pole. For smaller item banks, the number of replications per pole may need to be restricted to less than five. One series of testlet replications will be constructed to maximally informative at θ_L. The second series of testlet replications will be constructed to be maximally informative at θ_R. All testlets must meet the content (and other relevant) constraints. A simple heuristic to select items that maximize the testlet information is as follows. Let $I_i(\theta)$ denote the item information at θ. Further, let i_k be the item in the bank administered as the k^{th} item in a test or testlet and where R_k, is the current set of unselected items in the bank. The maximum information criterion for selecting item i_k can be expressed as

$$i_k = \arg\max_j \{I_j(\theta) : j \in R_k\}. \quad (4)$$

An adaptive heuristic essentially solves a series of optimization models that satisfy this criterion, also subject to relevant content constraints. By carrying out the sequence of item selections with respect to θ_L and θ_R, and replicating the process without replacement, we are able to compute an **average information function** that is feasible and robust enough to use in building multiple testlets and panels from the existing item bank. The average information function (maximally informative at either the left or right post) can computed as

$$T^*(\theta) = \frac{\sum_{m=1}^{M}\sum_{i=1}^{n} I_i(\theta)}{M} \qquad (5)$$

Given our center post (θ_c), which determines the desired proportion of population to be routed left or right and a mechanism for generating provisional average testlet information functions at the left and right posts (θ_L and θ_R), we can use a two-part numerical strategy to meet our two previously stated goals.

In part one of the CIT strategy, we need to numerically find the value of θ corresponding to the intersection of the provisional average testlet information functions, $T^*(\theta_L)$ and $T^*(\theta_R)$, based on the current left and right posts. That is, we want to find the point of the θ scale where the curves intersect. We can denote the value corresponding to the TIF intersection point as θ_{int}. Although multiple intersection points are theoretically possible with test information curves, such problems are unlikely if the effective numerical search range is kept between -2.0 and +2.0. A modified bisection algorithm works well in practice and is reasonably easy to implement. By experimenting with modifications to a bisection algorithm, we found that it is sometimes helpful to specify the **minimum** amount of information needed at θ_L, θ_R, and at the intersection point, θ_{int}. These minimum information constraints will insure that each target TIFs has sufficient information near its peak as well as at the point where the two TIF curves intersect.

In part two of the CIT strategy, we move the left and/or right posts until the intersection point of the provisional average testlets TIFs (θ_{int})aligns with the center post (θ_c). As alluded to earlier, either the left post (θ_L) or right post (θ_R) can be fixed to ensure that the testlet information is maximized at that point. However, both posts cannot be simultaneously fixed (i.e., one must be free to move). An obvious example where fixing the left or right post would apply is in a mastery testing context, where the fixed post corresponds to the pass/fail cut point. As our results subsequently show, fixing one or the other post actually helps in convergence. To carry out part two of the CIT strategy, and depending on whether the left or right pole is fixed, a second numerical algorithm—similar in concept to a bisection algorithm, is needed to move one or both poles toward θ_c, until $\theta_{int} = \theta_c$ at some acceptable level of convergence.

A Sample Study

A small study was carried out to illustrate the CIT strategy and to compare outcomes in terms of two sets of conditions: (i) the position of the center post and (ii) allowing the left and right posts to vary versus fixing one post. In this study, two 20-item target TIFs were sought under each set of conditions, one TIF corresponding to testlet B and the other corresponding to testlet B (see Figure2).

Data Source

The item bank for this study consisted of a bank of 443 operational items from a large-scale, professional certification examination. The items were previously used on paper-and-pencil tests and therefore represent a more restrictive set of characteristics than might ordinarily be expected for building MST forms. The items were calibrated to a common scale using the three–parameter logistic model (with D=1.702 to approximate the normal ogive). The descriptive item statistics for the bank were as follows: a-parameters, $m(a)$ = 0.716, $s(a)$ = 0.252; b-parameters, $m(b)$ = -0.233, $s(b)$ = 0.845; c-parameters, $m(c)$ = 0.176, $s(c)$ = 0.107. Four content areas were also employed in building each of the target TIFs. The full-test requirements were proportionally reduced for purposes of establishing content constraints on 20-item testlets.

Software

For purposes of this study, the first author programmed the CIT strategy algorithms, using modified bisection numerical routines for parts one and two (see previous section for a description). All analyses were conducted on a 1.3MHz notebook computer. Solution times ranged from 4.9 seconds to 30.8 seconds.

Study Conditions

Three routing conditions were investigated. In the 30:70 condition, 30 percent of the population was expected to route to testlet B and 70 percent were expected to route to testlet C. The two other routing conditions were 40:60 and 50:50. The corresponding values of the center post were determined using inverse cumulative normal equivalents, respectively, of θ_c = {-0.524, -0.255,0.000}.

The routing conditions were repeated across two other conditions: (i) allowing the left and right posts to vary or (ii) fixing the right post at θ_R = 0.60. The fixed value was slightly higher than the actual pass/fail point for this certification examination.

Results and Discussion

The solutions are summarized in Table 1. The leftmost column shows the ratio of the population respectively routed left or right. The next two "Intersection" columns show the value of the intersect point for the final TIF targets as well as the amount of information at that point. A convergence criterion of 0.01 was used. As indicated in the previous section, the corresponding center point values—the target values for the intersection points—were θ_c = {-0.524, -0.255, 0.000}. The two "Left Post" columns and the two "Right Post" columns in Table 1 show the final values at those outer posts and the associated total testlet information at each point.

Table 1. Results of the CIT Strategy by Condition

Routing	*Intersection*		*Left Post*		*Right Post*	
Ratio	Value	TIF	Value	TIF	Value	TIF
			Variable Left & Right Posts			
30:70	-0.46	7.7	-0.93	7.1	-0.44	-0.4
40:60	-0.26	9.4	-0.71	8.3	-0.07	10.1
50:50	0.01	10.3	-0.52	9.3	0.01	10.3
			Fixed Right Post			
30:70	-0.53	5.3	-1.84	3.9	0.60	8.8
40:60	-0.26	7.0	-1.38	5.4	0.60	8.8
50:50	0.00	8.4	-0.80	7.9	0.60	8.8

One of the solutions did not fully converge. The 30:70 routing condition with variable left and right posts produced a final intersection point of θ_{Int} = -0.46, even though the center post criterion was set at -0.524. This result was traced back to the limited information in the item bank near the tails of the distribution. Several minor algorithmic modifications were implemented to produce a more exact solution, however, we decided to report this original finding to highlight the complications of producing rather extreme targets from a fairly restricted item bank.

Figure 3 shows the target information curves for the variable-post condition (i.e., left and right posts free to vary). The variable labels indicate the "L" or "R" TIF, the routing condition plotted, and "V" for "variable posts". The plots for the three routing conditions (30:70, 40:60, and 50:50) are respectively displayed, top to bottom. The intersections are distinguishable across the routing condition. However, it is not just a shift in the targets. The target TIFs actually change to reflect the characteristics of the inventory in the item bank.

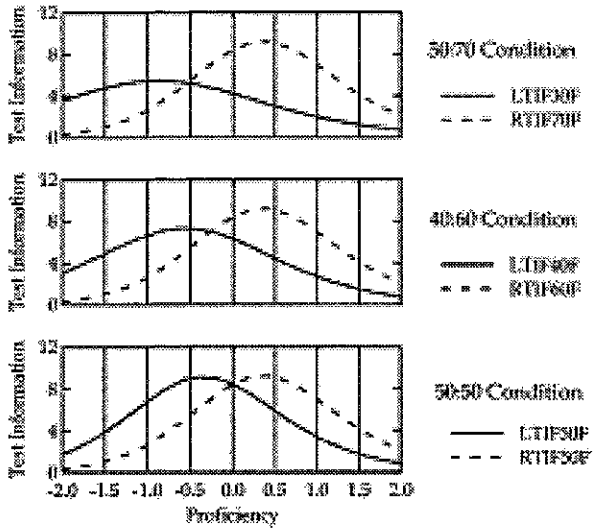

Figure 3. Generated Stage 2 Targets for Three Routing Conditions, Variable Left/Right Posts

Figure 4 shows the target TIFs for the fixed [right] post condition. The right-hand target TIF is fixed, as intended. The left-hand TIF changes so that the intersection point of the curves aligns with the criterion center post value for each run (i.e., θ_c equal to -0.524, -0.255, or 0.000). Compared to Figure 3, this figure indicates somewhat better defined targets. That makes sense, when we consider that fixing one of the posts helps to constrain an otherwise challenging solution, where aligning the intersection to the center post is the primary criterion. In practice, fixing one of the outer posts seems to be a worthwhile thing to do.

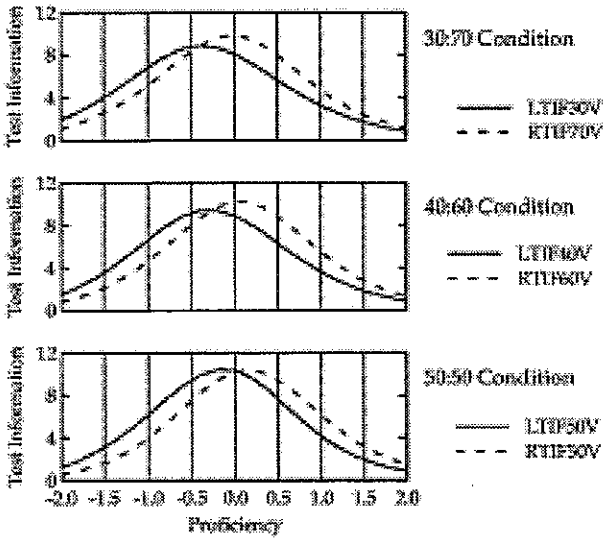

Figure 4. Generated Stage 2 Targets for Three Routing Conditons, Right Post Fixed (θ_R = 0.6)

Final Comments

This paper presented the CIT strategy as a straightforward way of achieving several goals in for generating feasible testlet TIF targets for MST designs: (1) to explicitly control the proportion of the population routed along various pathways in a panel; and (2) to make the targets as informative as possible, considering the quality of the items in the item bank, content, and other test specifications. This study illustrated the strategy for a simple 1-2 MST panel design.

Several issues should become topics of future research. One issue involves varying the number of TIFs averaged to produce the provisional targets. Generating very few TIFs per outer post could be expected to differentiate the targets somewhat. However, the targets may prove to be too informative (i.e., impossible to meet over time). In general, increasing the quantity of nonoverlapping testlet TIFs generated at the left and right posts will produce more robust targets. Yet, the indirect effect of using more TIFs will be to buffer the amount of information provided by the final average TIF. Content constraints and the inventory of items in the bank can also be expected to influence the results. The good news is that the CIT strategy allows these issues to be empirically investigated.

The second issue involves the extension of the CIT strategy to multiple stages or situations where there are more than two testlets per stage, each to be targeted as a different level of difficulty. It is relatively easy to conceptually discuss those extensions. The complications of implementation are another matter, especially with real item banks. That, too, is an issue for future research.

A final issue involves the matter of "auxiliary routes" (e.g., see the dotted line pathways in Figure 1). Although not discussed in-depth in this paper, the CIT strategy only works for the primary routes (solid line pathways in Figure 1). If most examinees fit the underlying IRT model, a very small percentage of the population should follow auxiliary routes, especially if the testlets are sufficiently long to provide stable proficiency estimates. However, gaining a better understanding of the actual impact on auxiliary routing for various MST panel designs may require extensive, future simulations.

References

Jodoin, M. G., Zenisky, A. L., & Hambleton, R. K. (2002, April). *Comparison of the psychometric properties of several computer-based test designs for credentialing exams.* Paper presented at the meeting of NCME, New Orleans.

Luecht, R. M. (1992, April). *Generating Target Information Functions and Item Specifications in Test Design.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

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

Luecht, R. M. (2000, April). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests.* Symposium paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2002, April). *A Testlet Assembly Design for the Uniform CPA Examination.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans.

Luecht, R. M., Hadadi, A., & Nungester, R. J. (1996, April). *Heuristic-Based CAT: Balancing Item Information, Content, and Exposure.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. New York, NY.

Luecht, R. M. & Nungester, R. J. (2000). Computer-adaptive sequential testing. In W. J. van der Linden & C. A. W. Glas (Eds). *Computer-Adaptive Testing: Theory and Practice*, pp. 117-128. Dordrcht, The Netherlands: Kluwer.

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

Pitoniak, M. J. (September, 2000). *Testlet-based Designs for Computer-Based Testing in a Certification and Licensure Setting.* Jersey City, NJ: AICPA Technical Report.

Van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, 22, 195-211.

Xing, D., & Hambleton, R. K. (2001, April). *Impact of several computer-based testing variables on the psychometric properties of credentialing exams.* Paper presented at the meeting of NCME, Seattle.