Graduate Management Admission Council®

Exploring Potential Designs for Multi-Form Structure Computerized Adaptive Tests with Uniform Item Exposure

Michael C. Edwards
The Ohio State University
and
David Thissen
The University of North Carolina at Chapel Hill

Presented at the Item Exposure Paper Session, June 7, 2007

Abstract

This paper describes research regarding the performance of different designs for uniform multi-form structure (uMFS) computerized adaptive tests (CATs). The uMFS CAT is an extension of the MFS structure (Armstrong, Jones, Berliner, Pashley, 1998) that incorporates exposure control. In an MFS-based CAT, the adaptation occurs at the level of blocks of items, rather than individual items. These blocks of items may be related to a common stimulus, or may be unrelated beyond measuring the same construct. The amount of flexibility in the adaptation of the MFS CAT is controlled largely by the number of stages and the number of levels. The performance of uMFS CATs was explored with 2, 3, or 4 levels and 1, 2, 3, 4, 6, or 9 stages. Performance was evaluated using simulation with a focus on reliability of scores and standard error curves.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Edwards, M. C. & Thissen, D. (2007). Exploring potential designs for multi-form structure computerized adaptive tests with uniform item exposure. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Michael C. Edwards, Department of Psychology, 1827 Neil Avenue, Columbus, OH 43210.
Email: edwards.134@osu.edu

Exploring Potential Designs for Multi-Form Structure Computerized Adaptive Tests With Uniform Item Exposure

Computerized adaptive testing (CAT) has played an increasingly prominent role. The most common form of CAT used today is an "item-by-item" CAT, in which adaptation occurs after each item response in the choice of which item to administer next. The Armed Services Vocational Aptitude Battery (ASVAB) and the Graduate Record Exam (GRE) are both examples of this kind of CAT. Early in the history of CAT, research on item selection algorithms focused on maximizing measurement precision and meeting content requirements. Early practical experiences demonstrated the need for some system to control the number of times an item was used (Wainer, 2000). This is commonly referred to as *exposure control* and has become an extremely active subfield of research.

There are numerous exposure control methods available (Stocking & Lewis, 2000), many of which are extremely complicated and/or require large simulation studies prior to being used. Most of these methods focus on constraining the selection algorithm so that it does not over-use highly discriminating items. When combined with the often complex requirements for content balancing, the amount of "on-the-fly" computation required to field an item-by-item CAT has continued to increase. These mounting complexities have led several researchers to consider other types of adaptive tests.

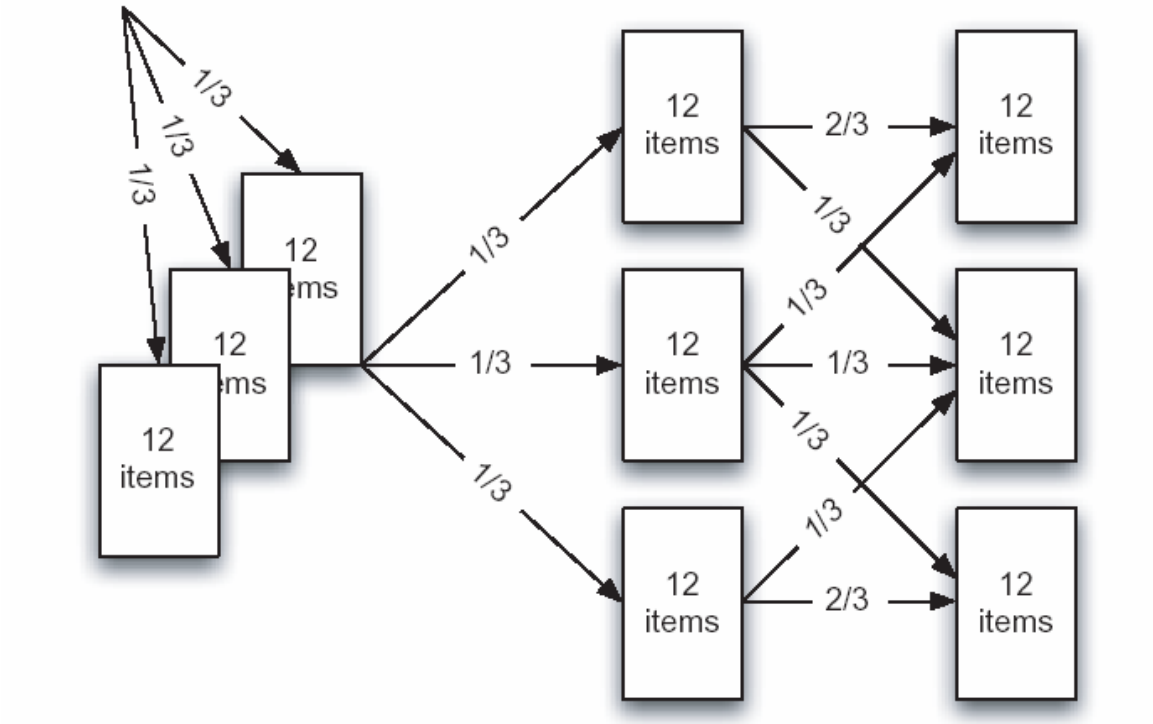
Another Kind of CAT

Building on early observations by Lord (1971, 1980), Armstrong, Jones, Berliner, & Pashley (1998) proposed a multi-form structure (MFS) CAT. In an MFS-based CAT, the adaptation occurs at the level of blocks of items (forms, in the parlance of Armstrong et al.), rather than individual items. These blocks of items may be related to a common stimulus, or may be unrelated beyond measuring the same construct (e.g., testlets, Wainer & Kiely, 1987). The original motivation for the MFS CAT was that the blocks could be assembled prior to administration of the CAT. This permits a more leisurely pace for content balancing and human review of forms. In an item-by-item CAT there are typically far too many possible instantiations for a human to review them all. Human review is useful to identify items that provide information about other items (called "enemy items"). Such information can be included in the selection algorithm of an item-by-item CAT, but this assumes it is possible to compare each new item to all other items currently in the item pool.

Luecht and Nungester (2000) and Luecht (2003) have extended a class of designs similar to the MFS to take into account the issue of exposure control. The computer-adaptive sequential testing (CAST) and bundled multistage adaptive testing (BMAT) developed by Luecht and colleagues represent a compromise between the demands of the item-by-item CAT and the simplicity of the MFS approach. A variation of the MFS CAT is the uniform MFS (uMFS), described in detail Edwards, Flora, & Thissen (2008). A sample uMFS design is illustrated in Figure 1. This uMFS CAT has three stages (from left to right) and three levels (for the second and third stage). This sample uMFS design has 12 items in each block for a total of 108 items. An examinee would be randomly assigned to one of the three first-stage blocks (routing blocks). This insures that each of the 36 items in the first stage is seen by approximately the same number of examinees. Based on their performance on the routing block and a previously determined set

of cut scores, examinees are routed to a level for the second stage of the uMFS CAT. In very general terms there will be one “difficult” block, one “medium” block, and one “easy” block. Another scoring and routing occurs at the end of Stage 2 to move examinees into their final block of questions. Adhering to the branching fractions specified in Figure 1 insures that each item is seen by roughly one-third of the examinees. Each examinee sees only 36 items (3 stages with 12 items per stage), with adaptation occurring in the choice of the block given in the second and third stage.

Figure 1. Illustration of a Three-Stage uMFS CAT, With Three (Equivalent) Routing Tests Comprising 12 Items, Followed by Two Stages Each Comprising High-, Medium-, and Low-Difficulty blocks of 12 items (Path Branching Probabilities Shown on the Arrows are Pre-Specified)



The amount of flexibility in the adaptation of the uMFS CAT is controlled in part by the number of stages (which controls the frequency of adaptation) and the number of levels (the number of differentially difficult blocks within a stage). There are many possible designs a uMFS CAT can assume, but this discussion considered only a subset. The designs explored below were obtained by crossing 2, 3, or 4 levels with 1, 2, 3, 4, 6, or 9 stages. Performance was evaluated using simulation with a focus on reliability of scores and standard error curves.

Method

A series of simulations were conducted to evaluate the performance of 17 different uMFS designs (excluding the 9-stage 2-level design). The simulations assumed a standard normal ability (θ) distribution and that the items in question were multiple choice with four response alternatives. A 3-parameter logistic (3PL) model was used and the items were simulated by drawing item parameters from the following distributions:

$$\log(a) \sim N(-0.3, 0.3), \quad (1)$$

$$b \sim N(0.2, 1.2), \quad (2)$$

$$\text{logit}(c) \sim N(-1.4, 0.5). \quad (3)$$

These parameters were chosen based on experience with large, high-stakes testing programs in a variety of contexts. All draws were made independently and draws more extreme than four standard deviations above the mean were replaced.

In the interest of comparability, in each design the examinee responds to 36 items. This choice, when paired with the different designs, dictated the number of items necessary to fill all the blocks. In the 3-stage, 3-level design shown in Figure 1, each block contains 12 slots for items, so a total of 108 items are required (9 blocks × 12 items per block). After the simulated items are created, they must be placed in the uMFS design. The items are initially randomly seeded into slots; then threshold accepting (Dueck & Scheuer, 1990) is used to find a configuration of items that is best in terms of an objective function. For this simulation, the objective function consists of four parts:

1. The error variance should be small,
2. The error variance should be uniform over θ,
3. The routing blocks should be equivalent, and
4. The expected branching fractions should be very near the specified values.

Part 1 is an obvious consideration for any test maker: The test should be reliable. Part 2 is a choice, but one that is of interest to many test makers. The objective function could very easily be altered if some other shape of the information function was preferred. The third and fourth parts are unique to the uMFS CAT. The third part is an attempt to keep the routing blocks identical so that no person is placed at a disadvantage by the random choice of the routing block. The fourth part is important to keep the item exposure nearly uniform across the test. For further details on this optimization procedure see Edwards, Flora, & Thissen (2008).

For convenience all routing was based on summed scores within block and the final θ estimates were based on patterns of summed scores. For each shape, two unique sets of items and were generated and the optimization procedure was “shuffled” twice for each set. This was done because the threshold accepting procedure is not guaranteed to find the global optimum of the objective function. The optimization is performed twice to reduce the likelihood it selects a particularly bad local solution. This resulted in four “trials” per uMFS shape. Performance of the various shapes was evaluated using reliability and standard error curves.

Results

Reliability (computed as the squared correlation of the θ estimates with the true values) for the various uMFS designs are shown in Table 1. The 1-stage results amount to a linear test that used the same number of items as uMFS CATs with the same number of levels (e.g., a 1-stage, 2-level uMFS comprises two nearly equivalent 36-item blocks to which examinees are randomly assigned). Not surprisingly, moving from one to two stages (which is the point at which the uMFS structure gains some adaptability) increased reliability for any number of levels.

Table 1. Reliability Results from the Simulations

	Stages					
Levels	1	2	3	4	6	9
2	.806	.838	.84	.84	.839	--
3	.82	.841	.854	.856	.855	.853
4	.815	.829	.84	.85	.853	.854

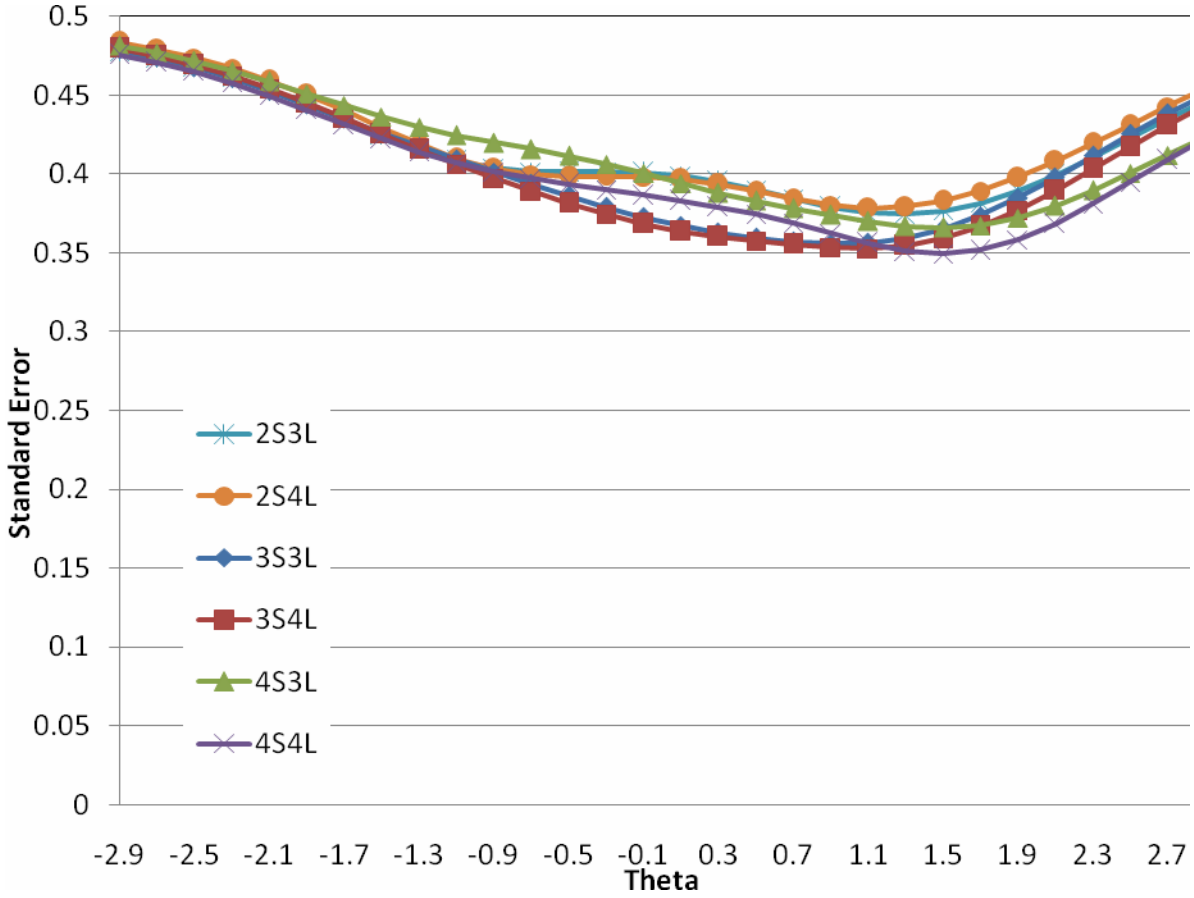
The 3-level uMFS shapes were uniformly more reliable than their 2-level counterparts. This was also true for five of the six numbers of stages when comparing 3-level with 4-level designs. On the basis of reliability, this suggests that the 3-level solutions are optimal. Among the 3-level solutions, there was a noticeable increase from one to two stages and another increase from two stages to three. There were smaller increases when the number of stages was further increased to four or six. Figure 2 shows standard error curves for six of the designs, focusing on comparing three and four levels. The 3-stage and 4-stage designs provided better measurement over most of the range considered. The 4-stage, 3-level uMFS design appears to be worse than its 4-design, 4-level (4s4l) counterpart at all levels of θ. The 3-stage designs performed similarly to one another and provided similar standard errors (in terms of magnitude) to the 4s4l shape. The standard error was lower for higher scores in the 4s4l uMFS shape, but to achieve this it lost some precision near the mean of the distribution.

Discussion

Considering the results show in Table 1 and Figure 2, the 3-level designs performed best in terms of reliability and standard error. While there were slight increases in performance when more than three stages were used, simplicity is desirable. The increased complexity necessitated by adding additional stages does not seem warranted beyond a third stage. Taken as a whole, these results lead us to recommend a 3-stage, 3-level (3s3l) uMFS shape.

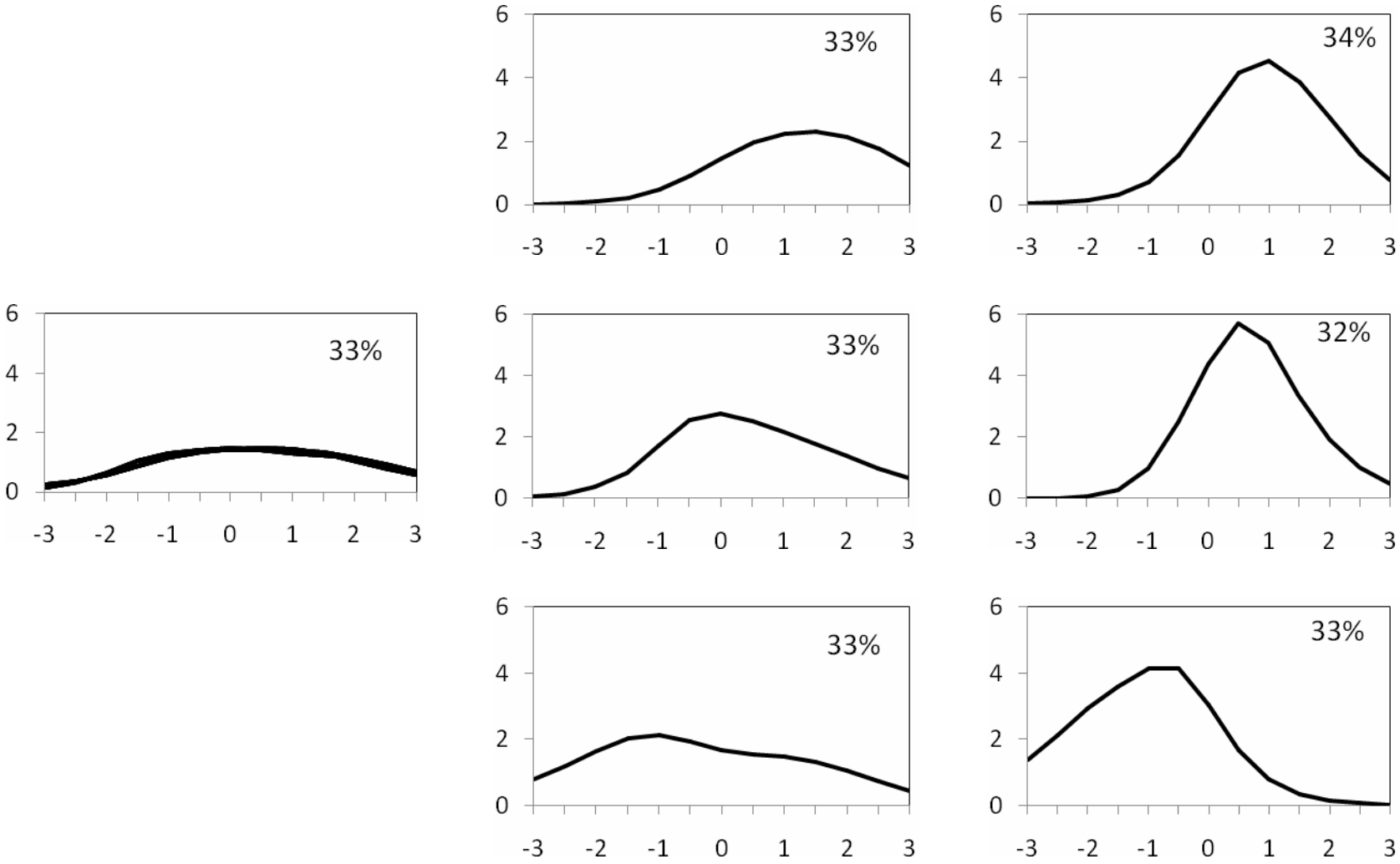
The 3s3l uMFS CAT improved the reliability over a linear test of comparable length from .82 to .854. Although this might not seem to be a major improvement, the shape of the reliability function as reliability increases must be considered. As a scale becomes more reliable it becomes increasingly difficult to further improve upon reliability. Using the Spearman-Brown prophecy formula allows framing this increase in reliability in terms of how much longer the linear test would have to be to achieve the same increase. To match the .82 to .854 improvement in reliability from a linear test to a 3s3l uMFS (both with 36 items), a linear test would require an additional 10 items.

Figure 2. Average Conditional Standard Errors for the IRT Scale Score Associated With Each Pattern of Summed Scores for 3- and 4-Level Shapes With 2, 3, or 4 Stages



Sample block-wise information functions for a 3s3l uMFS are given in Figure 3. The three lines in the first plot are nearly overlapping, because the three routing blocks are nearly equivalent. The information function is not peaked, but rather spread out over a large range of the construct. This is a useful distribution of information for the routing blocks, as their goal is to attempt to sort examinees into one of the next three blocks. The three blocks in the second stage have information functions with varying peaks that reflect “difficult”, “medium”, and “easy” sets of items. The level of information at the second stage is slightly greater than the first stage blocks. The third stage mirrors the second, however at this stage all three blocks provide a clearly visible increase in information over the first two stages. The fact that the information functions are similar across the three levels (along with the standard error curves in Figure 2) suggests that the optimization algorithm is doing a reasonable job creating a configuration that provides relatively uniform measurement precision. The percentages shown in each plot are the percent of examinees that were routed to a given block. In the 3s3l configuration a third of the examinees (33%) should see each block of items. Seven of the nine blocks had 33% exposure and the remaining two were only off by 1% (32% & 34%).

Figure 3. **Illustrative Information Curves for 12-Item Blocks From a 3-Stage, 3-Level uMFS CAT. The Percentages in Each Plot Indicate the Percentage of Examinees That Were Administered That Block. (The *y* Axis is Information and the *x* Axis is θ)**



The uMFS procedure summarized here has all the benefits of the MFS procedure (e.g., form pre-construction, human review, etc.) with a simple scoring and routing procedure that results in nearly uniform item exposure. Although the block-wise adaptation of the uMFS CAT will not provide the same increase in reliability as an item-by-item CAT, the decrease in "on-the-fly" calculations might be beneficial in some situations. That exposure control can be implemented in a very thorough way during the construction of the test is an added benefit that makes the uMFS CAT an attractive alternative to a traditional CAT.

References

Edwards, M. C., Flora, D. B., & Thissen, D. (2008). Multi-stage computerized adaptive testing with uniform item exposure. *Under review.*

Dueck, G., & Scheuer, T. (1990). Threshold accepting: A general purpose optimization algorithm appearing superior to simulated annealing. *Journal of Computational Physics, 90*, 161-175.

Luecht, R. M. (2003, April). *Exposure control using adaptive multi-stage item bundles*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

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

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163-182). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Wainer, H. (2000). Rescuing computerized adaptive testing by breaking Zipf's law. *Journal of Educational and Behavioral Statistics, 25*, 203-224.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.