Graduate Management Admission Council®

Implementing the Graduate Management Admission Test® Computerized Adaptive Test

Lawrence M. Rudner

Graduate Management Admission Council®

Presented at the CAT Models and Monitoring Paper Session, June 7, 2007

2007 GMAC® Conference on Computerized Adaptive Testing

Abstract

This paper discusses a host of practical issues that were encountered in converting the Graduate Management Admission Test to computerized adaptive format and in maintaining the GMAT® CAT program since 1997. Issues with regard to meeting the content specifications, item exposure, item banks, bias review, and drift are identified and discussed in the context of evolutionary changes in the GMAT® CAT program.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®. This paper is based on a chapter to appear in van der Linden, W. J. and Glas, C. A. W. (in press). *Computerized Adaptive Testing: Theory and Practice*. New York: Springer.

Citation

Rudner, L. M. (2007). Implementing the Graduate Management Admission Test® computerized adaptive test. 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

Lawrence M. Rudner, Vice President for Research and Development, Graduate Management Admission Council®, 1600 Tysons Blvd., Ste. 1400, McLean, VA 22102, U.S.A.
Email: lrudner@gmac.com

Implementing the Graduate Management Admission Test® Computerized Adaptive Test

Wise and Kingsbury (2000) argue that the success of a computerized adaptive testing (CAT) program is a function of how well the various practical issues are addressed. Decisions must be made with regard to test specifications, item selection algorithms, item bank design and rotation, ability estimation, pre-testing, item analysis, database design, and data security. The test sponsor is ultimately responsible for each of these decisions and must work closely with the vendor to assure that the sponsor interests are met.

This paper draws on the ten years of experience of the Graduate Management Admission Council® (GMAC) in implementing a CAT-driven large-scale assessment. The paper begins with an overview of the Graduate Management Admission Test® (GMAT®), outlines the conversion to CAT in 1996, and then presents a range of practical issues. For each issue, we outline several options that are available and, to the extent possible, the approaches taken by GMAC®.

Overview of the GMAT®

The GMAT® is a standardized assessment intended to help business schools assess the qualifications of applicants for advanced study in business and management, and is composed of three main components—the Analytical Writing Assessment (AWA), the Quantitative section, and the Verbal section. More than 200,000 examinees take the examination annually and GMAT® scores are reported to more than 3,000 different programs. The test is continuously available, by appointment, through more than 400 testing centers worldwide.

An analysis of the results from 273 validity studies involving 41,338 students conducted during the calendar years 1997-2004 has shown the GMAT® to be a good predictor of first-year grades (Talento-Miller & Rudner, 2008). The interquartile range of the predictive validity of the GMAT® total score, AWA score and undergraduate grade-point average is 0.448 to 0.626, with a mean of 0.530. Of special note is that the test is a much better predictor of performance in the 1st year MBA program than prior grades, perhaps because of the wide diversity of students pursuing a degree in management.

The GMAT® relies on the 3-parameter (3-PL) item response theory model. Items are calibrated and evaluated, in part, based on item parameters. Item banks are formed to meet target conditional errors based on the model. The testing algorithm uses 3-PL item parameters in selecting items to be adaptively administered.

Content

Table 1 provides an overview of GMAT® content, allotted time, and scoring. Total examination time is 2.5 hours, not including a short questionnaire and optional breaks. Although the content titles might appear to be similar to those of a general purpose admissions test, the GMAT® test emulates business-like conceptualization through its emphasis on logical reasoning in both the verbal and quantitative sections, and its use of business-related content.

Table 1. Overview of GMAT® Content, Allotted Time, and Scoring

GMAT® Section	Number of Questions	Allotted Time	Scoring
Analytical Writing Assessment		60 minutes	0 – 6
Analysis of an Issue	1	30 minutes	(half-point increments)
Analysis of an Argument	1	30 minutes	
Quantitative	37	75 minutes	0 – 60
Problem Solving			(1-point increments)
Data Sufficiency			
Verbal	41	75 minutes	0 – 60
Sentence Correction			(1-point increments)
Critical Reasoning			
Reading Comprehension			

With data sufficiency, an item type that is unique to the GMAT®, the examinee is required to determine whether there is enough information to solve the problem; the examinee is not asked to solve the problem. These questions are designed to measure the examinee's ability to analyze a quantitative problem, to recognize which information is relevant, and to determine at what point there is sufficient information to solve the problem. An example is shown in Figure 1.

Figure 1. Sample Data Sufficiency Problem

If a real estate agent received a commission of 6 percent of the selling price of a certain house, what was the selling price of the house?

(1) The selling price minus the real estate agent's commission was $84,600.
(2) The selling price was 250 percent of the original purchase price of $36,000.

(A) Statement (1) ALONE is sufficient, but statement (2) alone is not sufficient.

(B) Statement (2) ALONE is sufficient, but statement (1) alone is not sufficient.

(C) BOTH statements TOGETHER are sufficient, but NEITHER statement ALONE is sufficient.

(D) EACH statement ALONE is sufficient.

(E) Statements (1) and (2) TOGETHER are NOT sufficient.

The correct answer is D.

Although Data Sufficiency and Problem Solving tap high-order skills, the content specifications call as well for a balance of items requiring basic arithmetic, algebra, and geometry skills. In addition, there are specified numbers of items that are applied mathematics

problems and problems that are principally formula driven. Within each of the three basic skills, there are upper bounds to the numbers of items that tap specific skills. For example, no more than a certain percentage of items can include triangles or percentages. There are also lower and upper bounds regarding gender content and a correct answer location. In total, the GMAT® Quantitative section has 27 constraints; the Verbal section has many more. The problem in Figure 1 can be classified as a data sufficiency, algebra, percentage content, applied answer "D" problem. It does not count toward the gender limits.

Becoming a Computerized Adaptive Test

The GMAT® first became an adaptive examination in October 1997, five years after the idea was first presented to GMAC® management. The principal issue for the GMAC® at the time was access. The paper-and-pencil (P&P) GMAT® examination was offered only four times each year. Test-taking volume was growing and prospective examinees were having an increasingly difficult time obtaining a seat, especially in locations outside of the U.S. The second issue was that the more selective schools were having a more difficult time discriminating among the large number of examinees at the upper end of the score scale.

The first presentation to the GMAC® Board of Directors was made in 1992 by a vice-president at Educational Testing Service (ETS). At the time, ETS provided comprehensive test development, administration, and scoring and reporting services for the GMAT®, and ETS was interested in moving several of their clients (including GRE and TOEFL) to computer-based adaptive testing. Presumably, a larger client base would mean more tests being administered and would make computer-based delivery economically feasible. The key advantages presented to the Board were increased access, opportunities for new item types, and the possibility of adding new assessments to the GMAC® portfolio at some point in the future.

The first formal presentation to the GMAC® Board in 1993 addressed the potential benefit of transitioning the GMAT to CAT. CAT promised to address both of GMAC®'s principal issues—better access would be provided by more frequent testing opportunities worldwide and converting the test to adaptive format offered the promise of better discrimination at the upper end of the score scale. The expected costs of transitioning to an adaptive format would be principally additional infrastructure costs for changes in registration systems, item banking, score reporting, and the like. Because GMAC® already had a fairly extensive item bank, there would be no need for an appreciable increase in item production. The expected bill for conversion was between US$4 and US$7 million.

The move was approved by the GMAC® Board in 1995. GMAC® then proceeded to communicate the plans to its membership and other GMAT® score users. Because GMAC® had no resident psychometric expertise (the entire staff was only 10 people) and the GMAC® Board was comprised mostly of Deans and admissions directors, none of whom had measurement expertise, an independent third party was brought in to advise GMAC® on the merits of the plan and to review the migration of the test from P&P to CAT format. One of the consultant's major contributions was insistence on a study to compare the results of CAT administration onto the P&P scales that the GMAC® knew so well.

In mid-1996, well after GMAC® had told its clients of all the benefits and the need for the pending changes, ETS came to understand that it had substantially underestimated the need for additional, new item development and communicated that to the GMAC® Board. GMAC® was already committed and reaffirmed its desire to implement GMAT® CAT. The risk to GMAC®

was enormous. The final bill for the CAT transition, new item development, and infrastructure changes came in at nearly US$11.7 million – almost the entire cash reserves of GMAC®. Improved access was needed and the CAT transition was viewed as essential to attaining that objective.

In October 1996, twelve months before launch, the comparability study was conducted. Details of the comparability study and a subsequent equating study are documented in Bridgeman, Wightman & Anderson (undated). The intent was a balanced design with examinees taking both P&P and CAT, with randomly assigned order. Test registrants were invited to participate in the first study and were offered free examinations with only the higher score getting reported. Of the 10,196, invitees, 4,300 examinees accepted, 3,606 satisfactorily completed the CAT version, and 2,545 took both versions. The members of the P&P-first group in the usable sample were notably different than the members of the CAT-first group on several important measurable variables, and the groups as a whole were different than all other people historically taking the P&P version.

The study concluded that P&P results were not comparable to CAT results and that sizable equating adjustments would be required. "Between scores of 290 and 600, the equated scores *(from the first equating study)* were within plus or minus 10 points of the original scores. However adjustments of 20 to 30 points were needed at the lower end of the scale and 20 to 40 points at the high end of the scale" (Bridgeman, Wightman & Anderson, undated; italics added). In other words, the results were not comparable at the tails and differential adjustments were required.

Part of the issue was that the CAT test was unexpectedly speeded. About 18% of the examinees failed to answer the last two quantitative items; many additional examinees clearly applied guessing strategies without reading the final questions. In an attempt to remedy this situation, ETS decided to add five minutes to the CAT Quantitative section and to shorten the test by two operational items.

A second study to equate results was conducted in April 1997, a scant six months before launch. Because of the time constraint, a P&P first only design was used. Three thousand registrants were invited to participate, but only 773 who took the P&P version also took the CAT version. Apparently, many examinees were well satisfied with their P&P scores, and they did not return for the CAT administration.

Recognizing that the design and sample size of the April administration was not adequate for a defensible equating study, the final equating was based on a combination of data from the October and April data collections.

There were numerous design and implementation issues. The comparability study was conducted in October 1996—a month with historically documented significantly higher mean GMAT® scores. The second equating study was conducted in April 1997—a month with historically lower GMAT® scores. Most important, the April administration used a P&P-first only design. Participation rate was low and it is highly unlikely that the samples were representative of the GMAT® test-taking population. Most of these issues had been pointed out by the consultant in her critique of the design document.

It is worth noting that 10 years later a different approach to assessing comparability was employed as GMAC® transitioned test contractors. Rather than a common group design,

individuals taking the GMAT® under the new contractors, ACT and Pearson/VUE, were matched to individuals having taken the test under ETS using propensity score analysis (Rosenthal & Rubin, 1985; Rubin, 1997; Rudner & Peyton, 2006). This rigorous methodology overcomes the issues encountered in the 1996 comparability study.

Given the data and design of the 1996 and 1997 studies, the new scale was as similar as possible to the old scale. Nevertheless, (1) CAT-based scaled scores were not truly equivalent to the P&P scores; (2) mean quantitative scores climbed dramatically once CAT was introduced; and, (3) the new test failed to meet the goal of better differentiation in the upper end of the score scale.

Admissions officers and GMAC® were quite pleased with the results. The major goals were achieved. There was no discernible difference in scores from P&P and CAT administrations and access was, in fact, greatly improved. Nine years later, focus groups were held to discuss the desirability of normalizing and extending the scale on the upper end. The overwhelming response was this would be an unnecessarily disruptive refinement that would have very little practical advantage. Scores that are in the top 20th percentile are treated equally by almost all admissions representatives using the GMAT®.

Implementation Issues

The following sections discuss several implementation issues that have arisen and the approach taken by GMAC® to address those issues.

Meeting content specifications. Because the content specifications define the test and the construct being measured (Sireci, 1998), meeting the content specifications is of critical importance. The issue, then, is how to draw items from a larger pool and meet the specifications, given a large number of desired specifications and the limited number of operational test item slots.

Kingsbury and Zara (1989) outlined a constrained (C-CAT) that provides content balancing by selecting the item within the content area that has the largest discrepancy and which provides the most information at the examinee's momentary achievement level estimate. A major disadvantage of this approach is that the item groups must be mutually exclusive. In this case, as the number of item features of interest increases, the resulting number of items per partition decreases.

Wainer and Kiely's (1987) testlet approach can provide excellent content balancing, as each testlet can cover specific parts of the desired test specifications. However, Wainer, Kaplan and Lewis (1992) have shown that when the size of the testlets is small, the gain to be realized in making the testlets themselves adaptive is modest.

Swanson and Stocking (1993) and Stocking and Swanson (1993) describe a weighted deviations model (WDM) which selects the subsequent item for which a weighted sum of deviations from the projected test attributes are minimized. WDM seeks to satisfy all the conditions by treating some as desired properties and moving them to the objective function (Stocking & Swanson, 1993, p.280). For a highly constrained CAT, WDM assures that the test specifications will be met on the margin. That is, on average, a given group of examinees will meet the specifications; however, certain individual examinees in the group might not meet the test specifications. To GMAC®, this is not acceptable. All examinees should receive the same content mix.

Van der Linden and Reese (1998) describe the shadow test approach (STA) in which the items are not selected directly from the item bank but from a sequence of full tests (i.e. shadow tests) assembled in real time. With STA, large sets of content specifications can be met along with other desired constraints, such as item cloning, item-exposure control, and control of speededness. The relative importance of each constraint can be specified and tradeoffs of objectives can be evaluated.

The approach taken for the GMAT® is to separate the specifications for the individual and the specifications for the item bank. At the broadest level, GMAT® Quantitative items can be classified using three variables: skill area (data sufficiency or problem solving), content base (algebra, arithmetic skills, or geometry), and application (applied or formula-based). GMAC® specifies that each individual must receive a certain number of items in each of the seven categories just mentioned. We do not specify any bounds on the numbers of items in the cross-classification categories, e.g. the number of items that are data sufficiency, algebra, and applied. These specifications for individuals are implemented by having pre-specified content for each item position. A separate set of less important specifications is provided for the item banks. Banks must contain the desired balance in terms of answer location, gender, within-subject content, and other desired test characteristics. This way we can assure that the critical content specifications are always met and permit the less critical specifications, e.g., answer location, to vary.

Test specifications for P&P tests typically call for minimum reliabilities. For CAT, we specify a target conditional standard error curve and call for a minimum conditional reliability, which we evaluate based on simulated data. Rather than have the target conditional standard error curve follow the U shape typical of observed data, the GMAT® targets are completely flat across the center of the θ scale. Rather than using the *mean* errors of estimation conditioned on θ, we use *medians* conditioned on θ. This way, once the target error of estimation, computed as the square root of the inverse of the information function, for an examinee is met, the algorithm is free to select from all the items that meet the prescribed maximum standard error rather than items that maximize information. The use of median rather than mean target values provides an opportunity to broaden the use of items within the pool.

Simulation studies are used to evaluate pools for adherence to the specifications prior to their use. For simulated data, the reliability target for the GMAT® CAT is defined as

$$\rho_{xx} = 1 - \frac{\sigma_e^2}{\sigma_e^2 + \sigma_t^2}. \qquad (1)$$

True score variance, σ_t^2, is computed from simulated data as

$$\sigma_t^2 = \sum_k^K w_k (\theta_k - \bar{\theta})^2 \qquad (2)$$

where K is the number of generated discrete θ values, w_k is the expected proportion of examinees at each simulated θ value θ_k and $\bar{\theta}$ is the grand mean of the true θs

$$\bar{\theta} = \sum_{k}^{K} w_k \theta_k / K \tag{3}$$

Error variance, σ_e^2, is computed over the individual θ estimates as the mean squared difference between true and estimated θs.

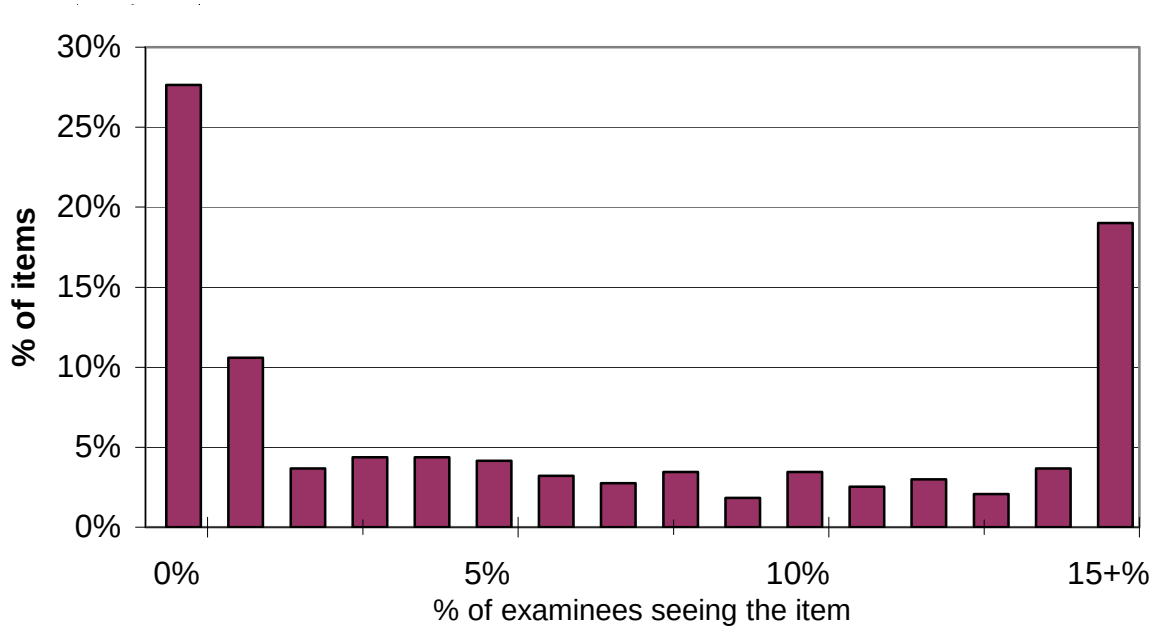
$$\sigma_e^2 = \sum_{k}^{K} \sum_{j}^{N_k} w_k (\theta_k - \hat{\theta}_j)^2 \Big/ N_k \ . \tag{4}$$

θ could be replaced with the corresponding scale scores to derive a more meaningful observed score reliability measure.

Item exposure, item use, and the CAT algorithm. Test items are costly to develop, often in the range of US$1,500 to US$2,500 per item. Given that expense, the test publisher is interested in assuring that all items are used and that no items are over-used. An unconstrained greedy algorithm can cause a severe problem in that respect. Wainer (2000) described an item bank consisting of 822 items. Upon repeat administrations of an exam utilizing this item bank with an information greedy algorithm, 14% of the item pool, or 113 items, accounted for 50% of the items administered to examinees. If one considers a hypothetical situation in which the average ability of examinees is very high and the standard deviation of test scores is very low, an information greedy algorithm would reduce the effective size of the item pool even further.

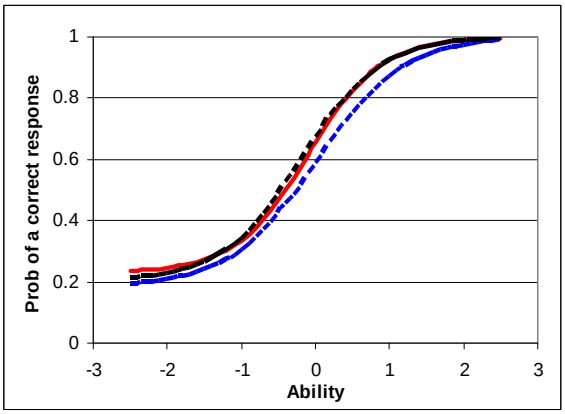
Figure 2 shows the observed item exposure distribution for items in a past operational bank of GMAT® items using a constrained algorithm based on maximum information. Approximately 28% of the items were never used, and 18% of the items in the bank were seen by more than 15% of the examinees. For this bank size, number of operational items, and test length, the ideal exposure (ignoring content constraints, the examinee ability distribution, and the item difficulty distribution) would have been 100% of the items being seen by slightly less than 4% of the examinees.

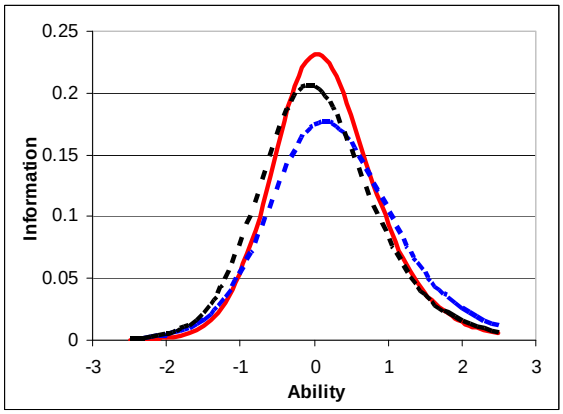
Figure 2. Distribution of Item Exposure Under a Maximum Information Algorithm



Without adequate exposure control, item selection based on maximum information will force some items to be underutilized and others to be over-utilized. An example is shown in Figure 3, which presents item response functions (IRFs) and the corresponding information functions for three items. The IRFs are nearly identical. Each of the three items would perform comparably if administered to an examinee with near average ability. However, if item are being selected based only on maximum information, one item would supersede the others nearly every time. The end result is an exposure distribution similar to that shown in Figure 2.

Figure 3. Three IRFs and Their Information Functions





An additional problem is that in the beginning of a testing session items are being selected that are targeted to the current, poor θ estimate. Because the θ estimate is poor, the difficulty of the selected items are often far from the examinee's true ability. An algorithm that selects the most informative, discriminating items at this stage is wasteful. The best items are being exposed while contributing little to the examinee's final ability estimate.

Overriding the item selection process to limit exposure will better assure the availability of item level information and enhance test security. However, overriding also degrades the quality of the CAT. Thus, it is likely that a longer test would be needed. However, if (1) an item bank is made of sufficiently high quality items, (2) the test is of sufficient length, and (3) the goal is to meet a target standard error rather than to minimize each examinee's standard error, then degradation is not an issue.

Sympson and Hetter (1985) developed an approach that controls item exposure using a probability model. The approach seeks to assure that the probability the item is administered, $P_{(A)}$ is less than some value r—the expected, but not observed, maximum rate of item usage. If $P_{(S)}$ denotes the probability an item is selected as optimal, and $P_{(A|S)}$ denotes the probability the item is administered given that it was selected as optimal, then $P_{(A)} = P_{(A|S)} \times P_{(S)}$. The values for $P_{(A|S)}$, the exposure control parameters for each item, can be determined through simulation studies. In aggregate, Sympson-Hetter addresses both the over- and under- exposure problems. However, large numbers of items remain under-exposed.

Another approach to control exposure is to randomly select the item to be administered from a small group of best-fitting (i.e., most informative) items. Various randomization rules can be applied. For example, McBride and Martin (1983) suggest randomly selecting the first item from the five best-fitting items, the second item from the four best-fitting items, the third from a group of three, and the fourth from a group of two. The fifth and subsequent items would be selected optimally. After the initial items, the examinees would be sufficiently differentiated and would optimally receive different items. Kingsbury and Zara (1989, p 369) report adding an option to Zara's CAT software to randomly select from two to ten of the best items. The randomization rule now used with the GMAT®, which was developed by ACT and is a mixture of these two approaches, yields item exposures that are closely distributed around the ideal.

At GMAC®, exposure risk is gauged by examining the probability that examinees with similar θ values will be administered items in common. Given a fixed number of examinees, as item bank size gets larger, all items are exposed less and the conditional exposure rates will decrease. Another approach to reduce conditional exposure rates used for the GMAT® has been to randomly select from multiple banks in the field at any one time. We have also staggered our bank rotation, have rotated banks frequently, and have used different banks in different regions. The closer the algorithm and bank are to achieving the ideal of administering a totally independent set of items, the less likely a given examinee can benefit from compromised items.

Item bank characteristics. In preparation for converting from P&P to CAT in 1997, GMAC® built up its item bank to include more than 9,000 quality items, and there has been a steady increase in the size of the available bank since that time. The challenge is to partition the item bank into pools that meet the specifications and to allow examinees to receive items that yield satisfactory standard errors.

The ideal item pool for a CAT would be one with a large number of highly discriminating items covering each content requirement at each ability level. The information functions for

these items would appear as a series of peaked distributions across all values of θ. Another way to look at an item bank is to look at the sum of the item information functions. This item bank information function shows the maximum amount of information the item bank can provide at each level of θ.

One approach to pool formation is to put all the available items from the item bank into the pool. Certainly this would yield a pool with the best available items. However, there might be dire consequences should that massive pool become compromised. As a test sponsor, we would like to see the smallest possible pools that permit the content specifications to be met. Weiss (1985) pointed out that satisfactory implementations of CAT have been obtained with an item pool of 100 high quality, well distributed items. He also noted that properly constructed item pools with 150-200 items are preferred. If one is going to incorporate a realistic set of constraints (e.g., random selection from among the most informative items to minimize item exposure, or selection from within subskills to provide content balance) or administer a very-high stakes examination, then a much larger pool would be needed. Given content constraints and standard error targets, pools of 600 to 1,000 items for tests such as the GMAT® are not unrealistic.

Weiss (1985) was correct in that an item pool of 100 items can be used to produce a highly satisfactory CAT. In developing an on-line, 24-item, diagnostic CAT version of the GMAT®, we ran simulations to evaluate the needed pool size given our desired content balance, the quality of the item bank in terms of mean *a* parameter value, and our desire to permit examinees to use the same pool for up to three administrations with the constraint that an examinee would never see the same item twice. The criterion was the conditional reliability, computed as described earlier, over all simulated examinees as a function of the number of times they took the test. The results of the simulation are shown in Table 2 and Figure 4. With an item bank having a mean *a* parameter of 1.25, a quality CAT can be developed, i.e. one having a reliability of .90 or greater with as few as 96 items, provided that the tests will be administered only one time.

Although these results are appropriate when one can hand-pick items and have few constraints, forming relatively small, effective pools for the actual GMAT® administration is a more difficult task. The content specifications must be met, sufficient numbers of quality items at each score point are needed, and one wants to minimize the number of items that have been used extensively in the past. Because some items are clones of others or have similar content, pools are typically formed so that they do not contain any item enemies. In addition, because many examinees retake a test—Rudner (2005) reported that 61% of the GMAT examinees that retake the GMAT do so within 3 months—one would not want to use items that have appeared in recent pools. The GMAC® pool formation rules require a certain rest period before items are considered for reuse.

Table 2. Reliability as a Function of Testing Attempt, Item Bank Quality, and Bank Size

Bank Size	Mean *a*	Reliability				
		1st time	2nd time	3rd time	4th time	5th time
48	.50	.68	.59			
	.75	.78	.63			
	1.00	.84	.69			
	1.25	.86	.69			
96	.50	.70	.65	.62	.50	
	.75	.81	.76	.66	.60	
	1.00	.88	.82	.71	.59	
	1.25	.92	.85	.74	.63	
144	.50	.69	.68	.67	.62	.58
	.75	.82	.80	.76	.70	.65
	1.00	.88	.85	.80	.74	.70
	1.25	.92	.89	.81	.74	.74
192	.50	.69	.70	.69	.66	.64
	.75	.81	.81	.80	.77	.72
	1.00	.89	.87	.85	.79	.77
	1.25	.92	.90	.88	.83	.81
240	.50	.72	.72	.72	.69	.69
	.75	.83	.82	.82	.79	.76
	1.00	.88	.88	.86	.84	.81
	1.25	.92	.91	.89	.86	.83
288	.50	.69	.68	.68	.67	.67
	.75	.83	.83	.82	.81	.78
	1.00	.88	.88	.87	.86	.83
	1.25	.92	.90	.89	.88	.85

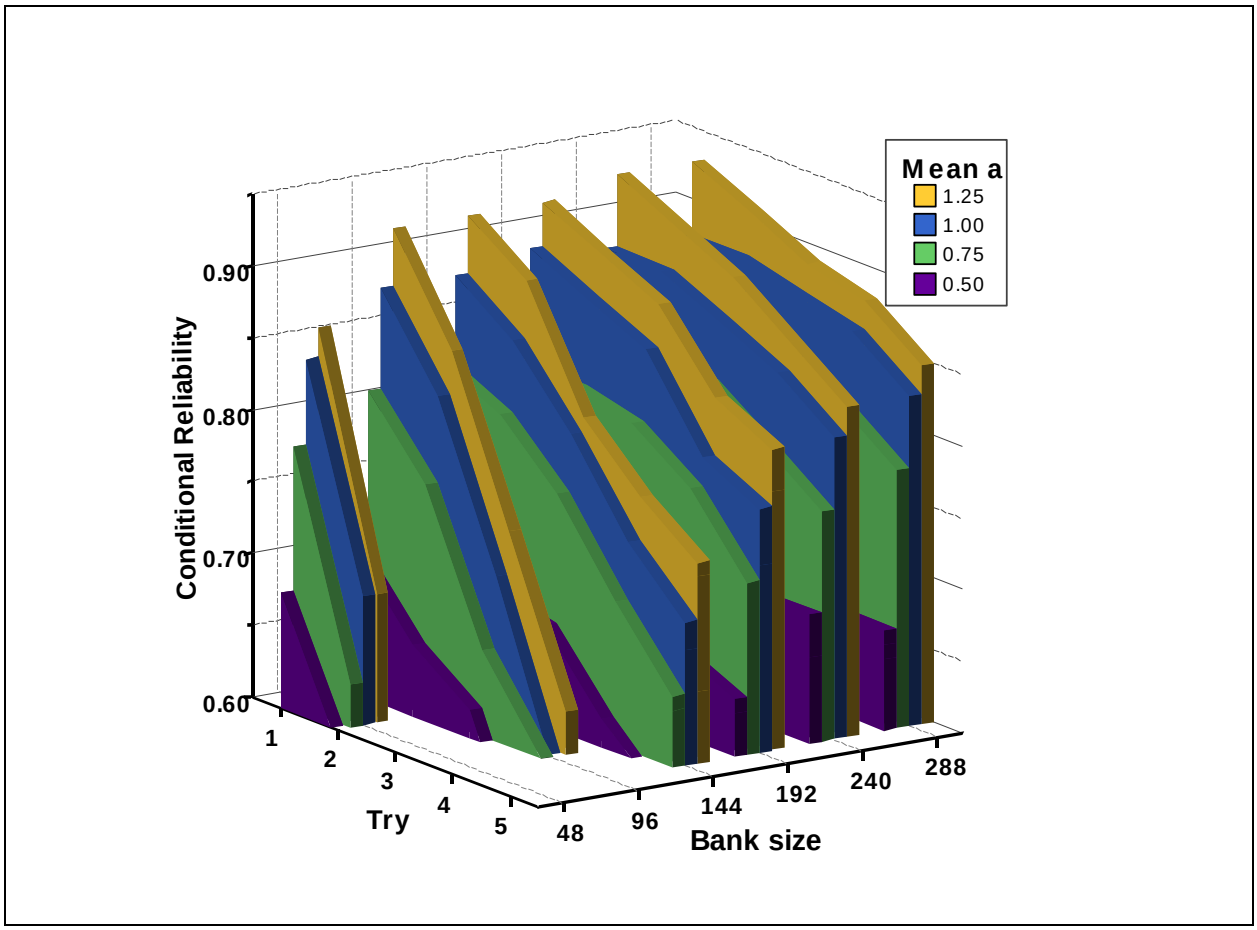
Note. Based on 100 iterations, 1,000 simulated examinees, a test length of 24 items, and 8 content constraints.

Given the constraints, software tools have been developed to help form initial GMAT® pools. Simulation studies taking into account the expected θ distributions are used to evaluate the pools. These simulations provide a host of information, including expected exposure rates, pool overlap, and expected conditional standard errors. Gaps in the conditional errors are then corrected by manually replacing or adding items. The process is iterated until targets are met. The GMAT® test development contractor, ACT, run these simulations for each pool. GMAC® reviews the results months in advance. Sporadically GMAC® compared the simulations against actual data. To date, the simulations have mimicked reality exceptionally well.

Given the work needed to formulate a pool, it is tempting to reuse pools or large parts of previous pools. However, brain dumps, illicit test preparation sites, and other groups making operational items available make reusing portions of pools a risky proposition. GMAC® pool formation rules also include a specification of the maximum percent of items that can overlap with any previous pool. GMAC® now devotes more than two full-time equivalent staff members

to monitor the Internet, document infringements, and bring civil and criminal action against individuals that infringe on GMAC's copyrighted material.

Figure 4. Reliability as a Function of Testing Attempt, Mean *a* Parameter, and Bank Size



Item bias. The common approach to investigating differential item functioning is to examine the item parameters resulting from group-specific calibrations. GMAC® does this on a routine basis as part of the item pre-test evaluation.

Pre-test data, however, are often limited in terms of the number of examinees in subgroups. Accordingly, we also investigate bias using operational items. For example, we were interested in whether GMAT® items show any bias when used in Europe (see Talento-Miller, 2008). Guo, Rudner, Owens, and Talento-Miller (2006) presented a method suitable for operational CATs. The IRFs estimated using responses from a subgroup are compared to the IRFs defined by the operational item parameters. An item is biased if examinees in a subgroup with the same ability do not have the same conditional probability of correct answers as the population of examinees, i.e. the total group used in calibrating the operational item parameters. We have had good success including pre-test items, simultaneously calibrating people and items, and then regressing sets of parameters. What is noteworthy about the approach is the change in the reference population. Rather than a comparison of groups, e.g. majority and minority examinees, all groups are compared to the operational parameters. Thus, the question changes from one of

group comparisons to one of impact. That is, do members of a group have an advantage or disadvantage when the operational parameters are used?

Item parameter and scale drift. The final practical consideration to be discussed is shifting parameter estimates. Once an item is calibrated and found to be of sufficient quality, there is little reason, other than being compromised, to retire the item, as long as it continues to function as it did when originally calibrated. Thus, there are the very real questions whether the individual item parameters have shifted beyond the standard error of calibration and whether that shift makes a difference. Guo and Wang (2005) presented a methodology used for the GMAT® based on a set of commonly administered items. Other methodologies used by GMAC® have included examining empirical IRFs from alternate administrations, calibrating operational items that have been placed in pretest slots, and calibrating CAT-administered items given examinee θs and prior *c* parameter values. We have had the most success simultaneously recalibrating an entire pool's worth of item response data, including the non-operational items. Given the relatively large number of examinees seeing collections of non-operational items, the resultant parameter estimates proved to be quite stable. Very few items had item parameters beyond the standard error of calibration. Parameters were updated for those that did.

Conclusions

A key component to any successful CAT program is the careful design and implementation of a system that provides quality information regarding examinee ability while minimizing item exposure and security risks. This paper presented some of the practical issues considered by the Graduate Management Admission Council® in the design and evaluation of the Graduate Management Admission Test®.

Some of the key considerations are:

1. *Test specifications.* Content specifications should assure similarity of content for every examinee, while balancing a wide range of considerations. GMAT® content specifications identify the items to be received by every examinee, requirements for the pools, specifications for the conditional errors, and a reliability target.
2. *Item exposure, item use, and the CAT algorithm.* Most of the work on item exposure has addressed the issue of over-exposure. We do not want the same items to be administered to large percentages of examinees. At the same time, the sponsor is interested in maximal use of the investment. That is, the test sponsor would like every quality item to be used. Quality test items are expensive to develop. Exposure and use issues associated with algorithms based only on maximum information were identified.
3. *Item pool characteristics.* Although there are psychometric advantages to placing all available items in a test pool, there are practical issues to consider as well, not the least of which is the consequences of a security breach. Small pools are attractive from the test sponsor's perspective, but small pools raise issues with regard to conditional exposure rates. A practical balance and an approach to evaluating pools are discussed.
4. *Item bias.* While traditional approaches to investigating item bias are employed for pre-test items, operational pools provide opportunities for investigations that are not possible for pre-test items. This paper presents an alternative—a practical viewpoint of differential item functioning focused on whether the operational item parameters are appropriate for each subgroup.

5. *Parameter shift.* Items administered via CAT can have an extremely long shelf life. Approaches employed to investigate the consistency of GMAT® item parameters over time are presented.

The papers by Georgiadou, Triantafillou, and Economides (2006) and by Green, Bock, Humphreys, Linn and Reckase (1984) provide excellent guidelines for evaluating CATs. The issues presented in this paper supplement these guidelines by examining practical concerns of test sponsors.

References

Bridgeman, B., Wightman, L., & Anderson D. (undated., circa 1997). *GMAT comparability study*. Internal GMAC® administrative report.

Georgiadou, E., Triantafillou, E., & Economides, A.A. (2006). Evaluation parameters for computer adaptive testing. *British Journal of Educational Technology*, 37, 261-278.

Green, B., Bock, R. D., Humphreys, L., Linn, R., & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21,* 347–360.

Guo, F. & Wang L. (2005). Evaluating scale stability of a computer adaptive testing system. *GMAC® Research Reports,* RR-05-12. McLean, VA: Graduate Management Admission Council. November 30, 2005.

Guo, F., Rudner, L., Owens, K., & Talento-Miller, E. (2006). *Differential impact as an item bias indicator in CAT*. Paper presented at the International Testing Commission 5th International Conference on Psychological and Educational Test Adaptation across Language and Cultures. Brussels, July 6-8, 2006.

Kingsbury, G., Zara, A. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education. 2*, 359-75.

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

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New Horizons in Testing* (pp. 223-236). New York: Academic Press.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Rosenbaum P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*, 33–38.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*(Suppl. 8), 757–763.

Rudner, L. M. (2005). Examinees retaking the Graduate Management Admission Test®. *GMAC® Research Reports*, Report Number RR-05-01. McLean, VA: Graduate Management Admission Council. March 17, 2005.

Rudner, L M. & Peyton, J. (2006). Consider propensity scores to compare treatments. *Practical Assessment Research & Evaluation, 11*(9). Available at http://pareonline.net/getvn.asp?v=11&n=9

Sireci, S. (1998). The construct of content validity. *Social Indicators Research, 45(3),* 83-117.

Sireci, S., & Talento-Miller, E. (2006). Evaluating the predictive validity of Graduate Management Admission Test® scores. *Educational and Psychological Measurement, 66,* 305-317.

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

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

Sympson, J. B., & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association*. San Diego, CA: Navy Personnel Research and Development Center.

Talento-Miller, E. and Rudner, L. (2008). The validity of Graduate Management Admission Test® scores: A summary of studies conducted from 1997 to 2004. *Educational and Psychological Measurement, 68,* 129-138.

Talento-Miller, E. (2008), Generalizability of GMAT® validity to programmes outside the U.S. *International Journal of Testing. 8,* 127-142.

van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints, *Journal of Educational Measurement, 42,* 283-302

Wainer, H & Kiely, G. (1987). Item clusters and computerized adaptive testing: The case for testlets. *Journal of Educational Measurement, 24,* 189-205.

Wainer, H. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H., Dorans, N., Flaughter, R., Green, B., Mislevy, R., Steinberg, L., & Thissen, D. (1990) *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, J., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. *Journal of Educational Measurement, 29,* 243-252.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologica, 21,* 135-155.