Graduate
Management
Admission
Council®

Investigating CAT Designs to Achieve Comparability With a Paper Test

Tony Thompson and Denny Way
Pearson

Presented at the Applications and Issues Paper Session, June 8, 2007

2007 GMAC® Conference on Computerized Adaptive Testing

Abstract

This study simulated three different CAT designs based on a large statewide assessment and evaluated the degree to which psychometric comparability was achieved compared to a paper test. The CAT designs included a variable-length test with maximum information item selection, and a variable-length and a fixed-length CAT that used targeted information selection. In addition to simulating designs for comparability, a fixed-length optimal precision CAT design was simulated for comparison purposes. The general findings from the study indicate that CAT designs using information targets can successfully obtain comparability, although comparability for classification accuracy was weaker than desired. The study also highlighted the large improvement in measurement precision that could be potentially obtained using a CAT when comparability is not a concern. Studies on CAT for statewide assessment will grow in importance as adaptive testing becomes more seriously considered by state departments of education and as online testing becomes commonplace in statewide assessment programs. Because at least some paper testing will continue well into the future, the design of CAT programs that are comparable to paper tests will be of interest to testing professionals. Conclusions from the study will also be of general interest to practitioners of CAT due to the focus on selection of items with information targets.

Acknowledgment

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

Copyright © 2007 by the Authors

Citation

Thompson, T. & Way, D. (2007). Investigating CAT designs to achieve comparability with a paper 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

Tony Thompson, Pearson, 2510 North Dodge Street, Iowa City, IA, 52245, U.S.A.
Email: tony.thompson@pearson.com

Investigating CAT Designs to Achieve Comparability With a Paper Test

Adaptive testing applications in statewide tests are somewhat constrained by the requirement that all of the items administered in an assessment used for No Child Left Behind (NCLB) accountability purposes must be developed to measure grade-level standards. However, this stipulation does not prohibit the use of adaptive testing in state assessment programs for NCLB purposes, as long as items in the computerized adaptive test (CAT) banks developed for each grade and content are written to measure the relevant grade-level standards (Trotter, 2003). For statewide tests composed of all multiple-choice items, even a grade-level specific CAT will have appeal if it can reduce student testing time and make better use of limited computer resources. An appealing feature of adaptive testing is the capability of obtaining uniform measurement precision across the proficiency scale, meaning that each student is measured to the same precision. For students outside the central region of the proficiency scale, a CAT will likely greatly increase obtained measurement precision compared to a typical paper-and-pencil exam. The increase in measurement precision might not only increase classification accuracy, but might also allow vertical scale scores based on difference measures to be reasonably precise (Kang & Weiss, 2007). This is in sharp contrast to difference scores based on paper-and-pencil testing which are notoriously unreliable. In addition, a CAT might be a more efficient way of controlling the exposure of test items than alternating linear test forms or using some other scheme of mixing the items administered to students testing electronically.

A challenge to implementing a CAT in K-12 settings is that many schools are not fully prepared to test students by computer. This suggests that adaptive testing could not be the only method of testing for a statewide program; that is, the CAT and the paper test would have to co-exist until all schools could test all of their students electronically. This raises comparability issues, not only based on testing mode but also based on psychometric differences in testing procedures. Because the technical characteristics of adaptive and paper tests are not the same, it is difficult to establish and maintain comparable scores and test results between a CAT and a paper test (Wang & Kolen, 2001). Any state interested in implementing adaptive testing as part of their K-12 program must address this issue.

A conventional CAT design that might achieve comparability with a paper test is one that employs a variable-length stopping rule. In this approach, stopping rules can be specified based on varying precision levels that are equivalent to the precision of a paper test. One potential drawback to a variable-length CAT is that equity issues arise when the adaptive tests are timed (Parshall, Spray, Kalohn, & Davey, 2002). However, many statewide assessments are untimed.

Some research has addressed comparability to a paper test when a fixed-length CAT is desired. Davey and Fan (2000) described an item selection procedure that holds promise for better controlling the measurement characteristics based on a CAT. This procedure, here referred to as *targeted information selection* (TIS), selects items that best match a series of intermediate information targets. Depending on the difference between the current estimated level of information for the examinee and the intermediate target, the TIS algorithm might seek out an item that has the maximum information potential for that examinee or an item with a suboptimal level of information. If the current estimated information is much greater than the intermediate target, the algorithm might even seek out a very suboptimal item, as the goal is to match the final target as closely as possible. Comparability simulation studies by Davey and Fan (2000) and by Thompson (2002) concluded that use of the TIS procedure could achieve a high degree of comparability between the CAT and paper versions of a test. In addition, the CAT using TIS was found to reduce the variability of obtained information, particularly for examinees in the middle of the ability distribution. The CAT also had many fewer simulated examinees whose obtained information fell outside of the information target range.

As described by Davey and Fan (2000), TIS is applied in CAT applications where all students are administered the same number of items. However, conceptually a targeted information approach can be applied to a variable-length CAT setting. A variable-length TIS CAT was studied by French and Thompson (2003), where the goal was to improve the item bank usage of an operational CAT. Using an exposure control procedure specifically designed for TIS (see Thompson, 2002), the French and Thompson study found that bank usage could be markedly improved over the procedure implemented in the operational CAT.

The purpose of this study was to use computer simulation to investigate CAT designs that have the goal of achieving measurement precision that is comparable to the measurement precision found on a conventional test. The two main approaches (TIS and variable-length stopping rules) were evaluated separately and in combination (that is, variable-length CAT that employed TIS). A secondary motivation for the study was to investigate the potential improvement in measurement precision that could be gained through CAT as opposed to a traditional paper test typically used in statewide tests. Toward this end, an optimal precision CAT was also simulated.

Method

Sources of Data

Simulations were based on data from a statewide grade 11 mathematics test administered in Spring 2003. The 60-item operational test consisted of discrete four-option multiple-choice items and a small number of grid-in items (about 9%). There were 60 different sets of 10 field test items embedded in different versions of the test.

The initial item bank for the CAT simulations was comprised of the field-test items from the paper, a total 600 of items. The 60 operational questions comprised the conventional test form to which the CAT results were compared. Table 1 provides the numbers of items in each content objective for the operational test and CAT item bank .

Table 1. Numbers of Mathematics Items by Objective Areas

Mathematics Test Objective	No. Items in Paper Test	No. Items in CAT Bank
Objective 1	5	47
Objective 2	5	59
Objective 3	5	61
Objective 4	5	61
Objective 5	5	48
Objective 6	7	61
Objective 7	7	71
Objective 8	7	81
Objective 9	5	48
Objective 10	9	63
Total Number of Items	60	600

Three-parameter logistic (3PL) calibrations, carried out using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1999), were conducted on the item bank and served as both the true parameters and as the parameter estimates for the CAT and paper test simulations. That is, estimation error of the model parameters was not considered in the simulation.

CAT Simulation Methods

A CAT simulation program was developed for the study. The details of the procedures and algorithms implemented in the study are described below.

Item selection algorithm. As described previously, the intention of the study was to examine a variable-length CAT with maximum information item selection (MI) and both fixed- and variable-length CAT with TIS item selection. In addition, a fixed-length CAT with MI was included as a baseline condition, which was not expected to achieve comparability with the paper test. Table 2 gives the test lengths for the four CAT conditions in contrast to the 60-item paper version.

Table 2. CAT Test Lengths

Item Selection Method	Variable Length	Fixed Length

Maximum Information	Minimum = 20 Items
	Maximum = 60 Items
		35 Items

Targeted Information	Minimum = 20 Items
	Maximum = 60 Items
		35 Items

The test length of the fixed-length CAT for both item selection methods was determined by running a simulation with variable-length MI. Simulated examinees near the center of the distribution were found to require about 35 items on average to obtain approximately the same information as simulated examinees taking the paper version. Away from the center of the distribution, the average test length was reduced (see results section). Thus, the fixed-length CAT test length was set to 35 items.

To simplify coding, the same item selection algorithm was used for both MI and for TIS. The algorithm to select the next item (item k) for the CAT was as follows:

1. For each eligible item (see content constraints below) perform each of the following steps.

2. Read (or compute) the intermediate target information, $I_{k,TAR}(\theta)$, at 41 θ values from –4.0 to +4 with step size 0.2. The type of item selection used (e.g., MI or TIS) was determined by the values in the intermediate target information matrix. The values used for the different methods are given below.

3. The likelihood function after $k-1$ items have been administered, $L_{k\text{-}1}(\theta|\underline{\mathrm{U}})$, was calculated (and normalized) at the same 41 θ values based on the current estimate of θ.

4. Item information was calculated for the item being considered for selection, $I_i(\theta)$, at the same 41 θ values.

5. The cumulative item information, $I_{current}(\theta)$, was computed at each of the 41 θ values by summing $I_i(\theta)$ across the k–1 items already administered.

6. The criterion function, C, was calculated for the considered item (i) as

$$C=\sum_{j=1}^{41} L_{k-1}(\theta_j \,|\underline{U})\ \left[I_{k\_TAR}(\theta_j)-I_{current}(\theta_j)+I_i(\theta_j))\right]^2 \qquad (1)$$

The algorithm essentially compares the predicted total information that would be obtained about the examinee if item *i* were selected versus the intermediate target information for that point in the test. The eligible items are then ranked from most desirable to least desirable based on which items give the lowest criterion. The most desirable item that passes exposure control is administered.

It was mentioned above that the values in the intermediate target information matrix determine the item selected method that is used. To use the item selection algorithm with MI, each value in the target matrix is set to an impossible to reach number. In the simulation, each value in the matrix was set to 99. As no item can achieve the targeted value, the effect of this is to cause the eligible items to be ranked according to their information value, as the items with the highest information will come closest to the target. The same target matrix can be used for both a variable-length and a fixed-length MI CAT.

For TIS, the goal is to make steady progress toward the information target and reach the final target as the last item is administered, so that the target is reached but not exceeded. Because the ultimate aim is comparability with the paper test version, the final target for the fixed-length TIS for each of the 41 θ values was set equal to the sum of the item information for the items on the paper test. Intermediate targets were set to approach the final target linearly. At any point in the test, the intermediate target was found with the following rule:

$$I_{k,TAR}(\theta_j)=\frac{k}{N}I_P(\theta_j), \qquad (2)$$

where *k* is the current position in the CAT, *N* is the total number of items administered in the CAT, $I_P(\theta)$ is the final target (the paper test total information), and $I_{k,\_TAR}(\theta)$ and θ_j are as defined previously. Other rules for the intermediate targets could be formed, but the linear approach follows what was done in Davey and Fan (2000) and Thompson (2002).

Although the above rule works well for a fixed-length CAT where *N* is a constant value for all examinees, it does not seem as appropriate for a variable-length TIS CAT. For this study, Equation 2 was used for the variable-length TIS CAT with one change. The constant *N* was changed from a fixed value to $N(\theta_j)$, which varied depending upon the θ category. $N(\theta_j)$ was determined by first running the variable-length MI CAT and setting $N(\theta_j)$ to the average test length for each θ category. This was hoped to result in a TIS variable-length CAT that had the same average test length as the MI variable-length CAT. If the CAT reached a point where $k > N(\theta_j)$, the intermediate target was set to the final target.

Stopping rules. The fixed-length CAT versions terminated after 35 items. For the variable-length CAT versions, however, a stopping rule was needed. After *k* items were administered, the CAT stopped if *k* was greater than or equal to 20 (the minimum test length) and if

$$\sum_{j=1}^{41}\left[L_k(\theta_j \,|\underline{U})\times I_k(\theta_j)\right]\geq\sum_{j=1}^{41}\left[L_k(\theta_j \,|\underline{U})\times I_P(\theta_j)\right], \qquad (3)$$

where $I_k(\theta)$ is the sum of the item information for all items taken in the CAT administration to that point, $I_P(\theta)$ is the sum of the item information for all items on the paper test version, and the other terms are defined as described previously. The CAT would also stop if the maximum number of items (60) was reached.

Content balancing method. Content was balanced for the 10 objective score areas described in Table 1. The goal was for each objective to be proportionally represented in the CAT the same as the paper test. The algorithm selected selected the “most needy” content area at each point in the CAT (ties resolved randomly). Most needy meant the objective whose proportional representation was most dissimilar to the paper test content distribution. The items from the content area defined as “most needy” were the only items eligible for use by the item selection method.

θ estimation. The base θ estimation method used was maximum likelihood. Until at least one incorrect and one correct response occurred, θ was estimated through a step size value procedure. In this method, the initial θ was set at –1.0, and θ moved by +1.0 after each correct response or by –1.0 after each incorrect response until maximum (+4.0) or minimum (–4.0) θs were reached.

Exposure control algorithm. The Sympson-Hetter exposure control procedure was implemented (Sympson & Hetter, 1985). The maximum desired item administration rate was set to .15. The calibration of exposure parameters was performed for 20 cycles on samples of 4,000 per cycle. The θs used to generate the response data for each cycle were generated from a N(0,1) distribution.

Simulation and replication. Simulated response vectors using 41 true θ values from –4 to +4 were randomly generated. At each θ level, 200 simulated examinees were generated. After completing the initial simulation, two more replications were performed. For aggregate analyses not conditional on θ, results were approximated by weighting the conditional output by deviates of a N(0,1) distribution. These simulation conditions applied to all four CAT versions as well as the paper test versions.

Results and Discussion

The same pattern of results was found in each of the three replications. For the purpose of simplifying the presentation of results, the tables and graphs below report averages across the three replications.

Two general considerations are important to consider in evaluating the results. One is the degree of comparability achieved with the paper test. The main goal of the current study was to explore the best method of achieving comparability, and to that end several comparability comparisons are made. The second consideration is the degree to which measurement characteristics can be improved by replacing a paper test with a CAT version. Naturally, a single instantiation of a CAT cannot improve upon the psychometric properties of a paper test and at the same time remain comparable to it. However, the results highlight the flexible nature of CAT that allows for either comparability or psychometric improvement while reducing test length in either case.

The two variable-length CATs and the fixed-length TIS CAT were designed for comparability while the fixed-length MI CAT was designed for psychometric improvement. The success of each CAT version should be measured in terms of the goal it was intended to meet.

Table 3 presents average test lengths and correlations of estimated and true θ for the four CAT tests and the paper test. The two variable-length CATs had almost the same average test length and were over eight items shorter on average than the fixed-length CATs. A review of the correlations shows that the highest match with true θ occurred with the MI fixed-length CAT whereas the 60-item paper test had the lowest correlation. The two TIS CATs had correlations most similar to those of the paper test, but they were both slightly higher. Although the correlation differences were not large on an absolute scale, the differences were consistent across replications lending credibility to their interpretation. Although further study or more replications would be needed to fully support the conclusion, the findings indicate that no version of the CAT fully reproduced the paper test's correlation with true θ.

Table 3. Correlation with True θ and Average Test Length

Test	Correlation	Average Test Length
MI Selection Fixed-Length CAT	.99	35
MI Selection Variable-Length CAT	.98	26.6
TIS Variable-Length CAT	.97	26.7
TIS Fixed-Length CAT	.97	35
Paper Test	.96	60

An important criterion for either comparability or psychometric improvement is the classification accuracy of the test. NCLB tests classify students into proficiency levels and these classifications have can have crucial ramifications for schools. For the particular mathematics test this simulation was based on, three classifications are made: Below the Standard; Met the Standard; Advanced. Table 4 gives the classification accuracy of the tests simulated in terms of the percentage of simulated students correctly classified within the three categories. For example, out of all the simulated students truly in the Met category, the MI fixed-length CAT correctly classified 83% as Met. The table shows that all of the tests made very accurate Below and Advanced classifications. Substantial differences were found in the Met category, however. The MI fixed-length CAT was found to have substantially higher accuracy in the classification of true Met simulees than the paper test. The other CAT versions to a lesser degree also had higher classification accuracy of Met simulees than the paper test. As was the case with the correlation results, the classification results imply that three CAT versions designed for comparability were not completely successful in this goal.

Table 4. Percent Perfect Classifications for Three CATs and a Paper Test

Test	Below	Met	Advanced
MI Selection Fixed-Length CAT	98%	83%	99%
MI Selection Variable-Length CAT	98%	80%	98%
TIS Variable-Length CAT	98%	78%	98%
TIS Fixed-Length CAT	98%	78%	97%
Paper Test	98%	74%	98%

The next series of results are presented conditionally by the simulated true θ level. The first of these conditional results is given in Figure 1, which shows the conditional average test length of the four CAT versions. The two fixed-length CATs, of course, administer 35 items to each simulated student. Figure 1 shows that the two variable-length CATs had almost exactly the same expected test length regardless of θ level. This result was not unexpected, as the intermediate targets for the variable-length TIS CAT was specified so that final target would be reached at the same point in the test that the variable-length MI CAT terminated. It does show, however, that using TIS in a variable-length setting does not necessarily mean that test length needs to be sacrificed.

Figure 1. Average Test Length by θ

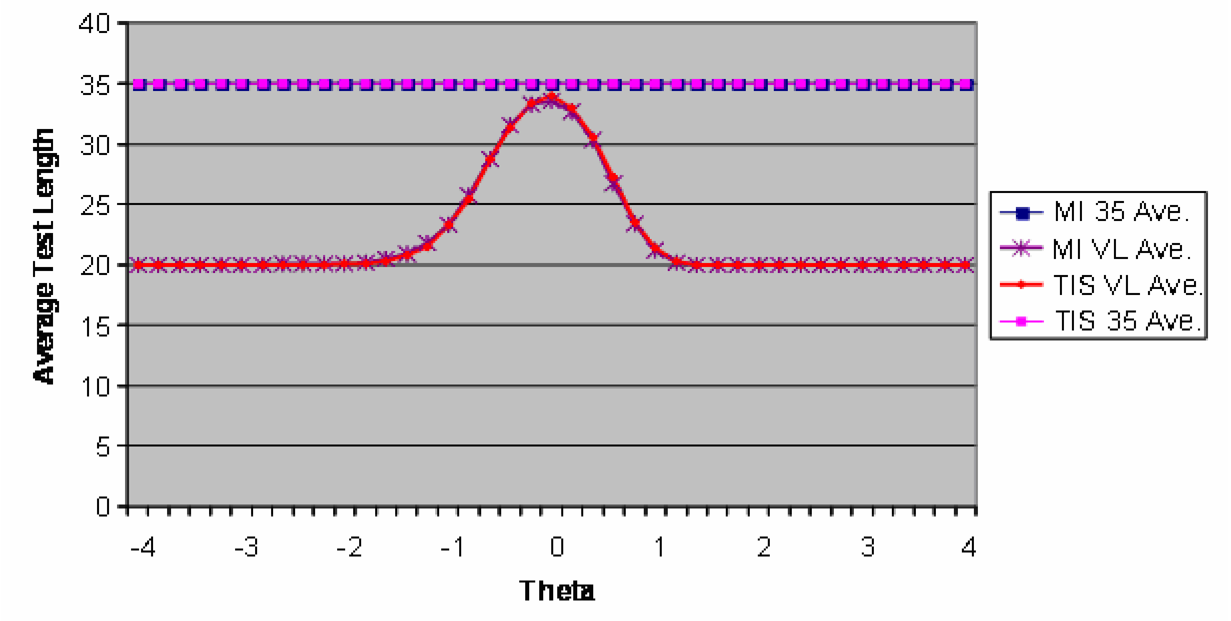


Figure 2 presents a comparison of the conditional bias of the tests simulated. Three main results are observed. First, the fixed-length MI CAT showed less bias in the extremes than the paper test, particularly in the regions from –3 to –2 and from +2 to +3. The second finding is that the two TIS CAT versions tracked the paper test results very closely. The third finding is the variable-length MI CAT was not as close to the paper version as the two TIS CAT versions.

Figure 2. Conditional Bias (Estimated θ Minus True θ)

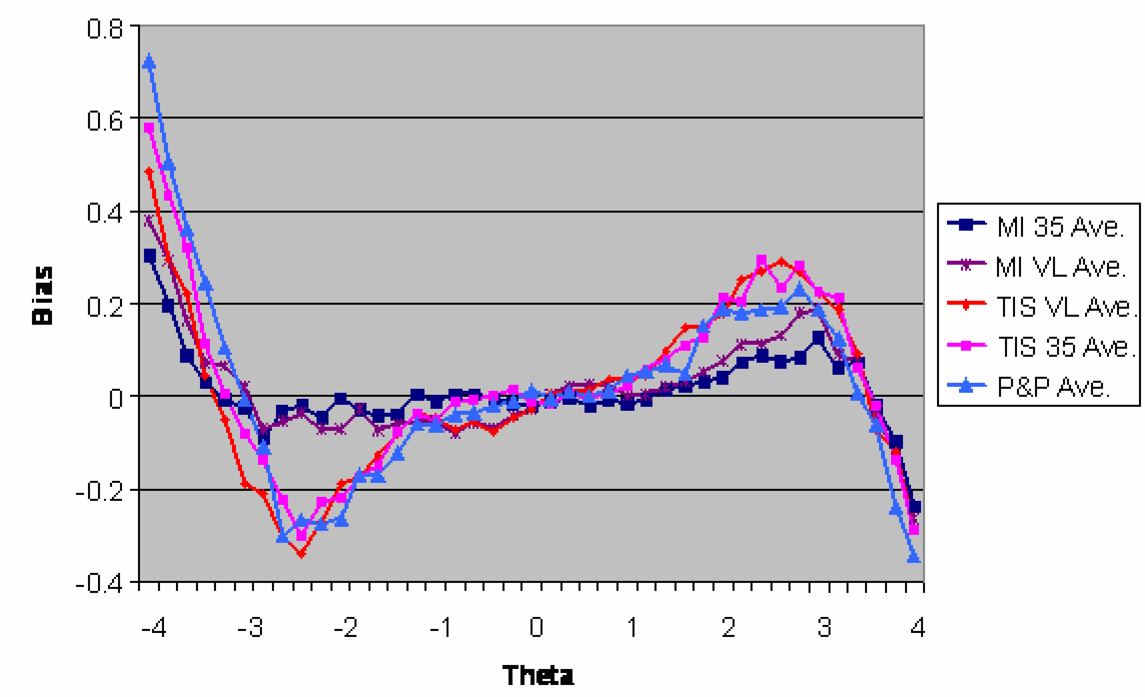


Figure 3 presents another aspect of comparability, namely the conditional standard error of measurement (CSEM) for the tests simulated. The pattern of CSEM results matches those found for bias. The fixed-length MI CAT CSEM was much better than the paper test, especially in the non-central regions below –1 and above +1. The two TIS CAT versions closely tracked the paper test and the variable-length MI CAT was somewhat in between—not as comparable as the TIS CATs and not as accurate as the fixed-length MI CAT. One possible reason for the lack of comparability for the variable-length MI CAT is the minimum test length that was selected. The minimum was 20 items, and for some θ levels 20 items was probably more than sufficient to reach the information target set by the paper test. For these cases, the variable-length MI CAT would be forced to continue to 20 items and thus measure more precisely than the paper test. To some degree then, comparability with the paper test could probably been improved with a shorter minimum test length, although this decision might have other implications for the testing program.

Figure 3. SEM Conditional on θ

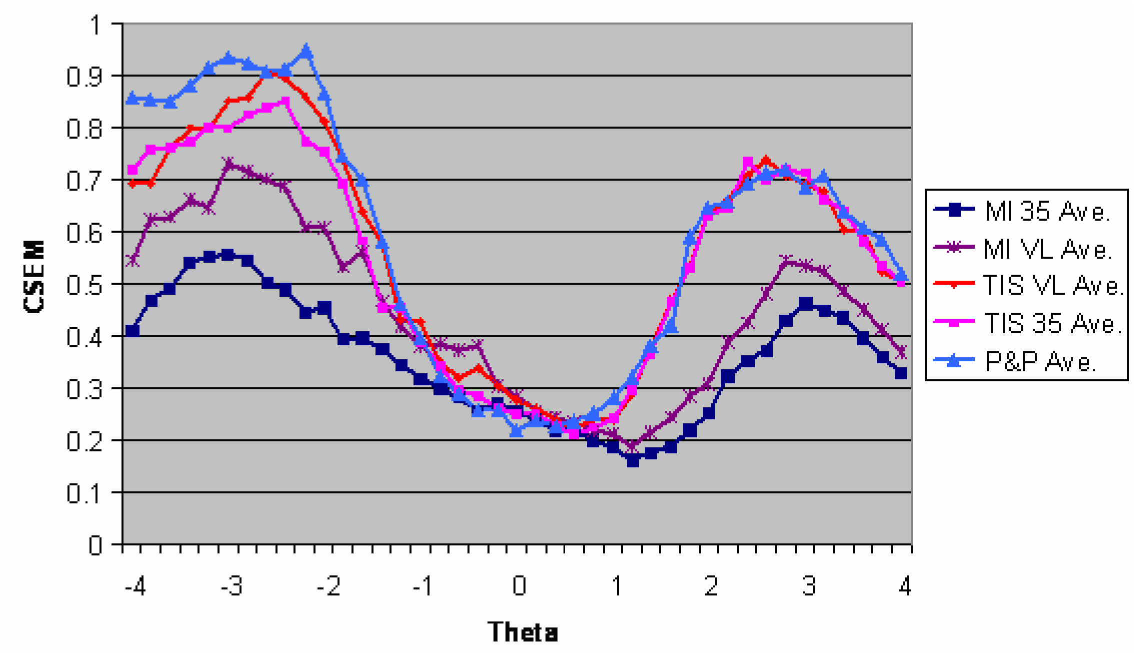
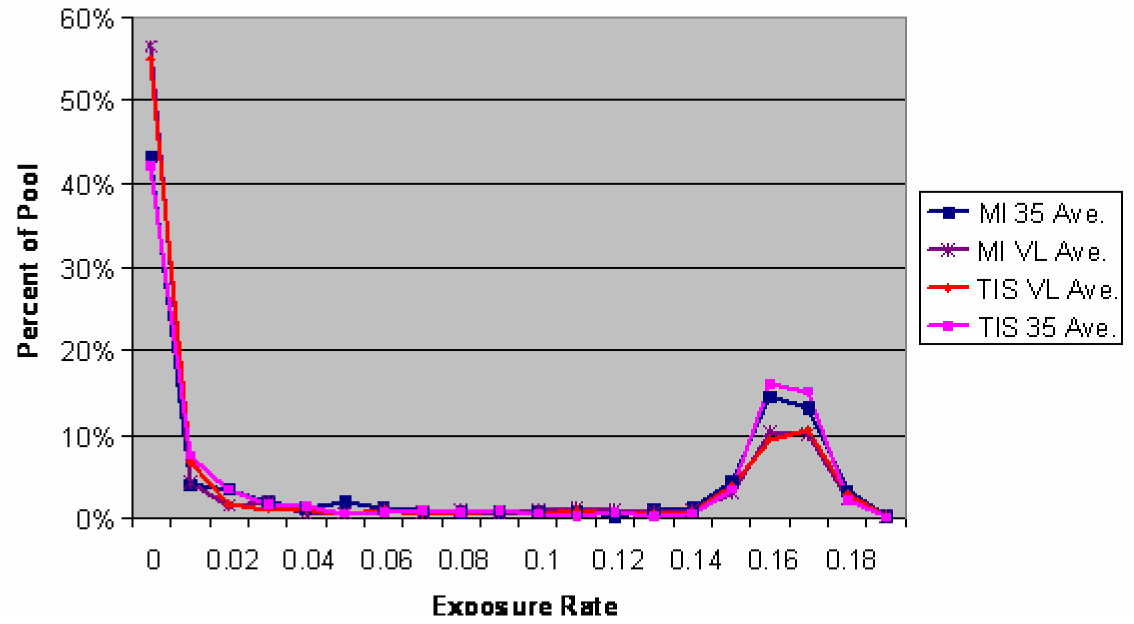


Figure 4 shows the administration rates of the items for the four CAT versions. In general, the four versions were similar in their exposure of items. In all cases, the maximum exposure was less than .2, which was judged to be satisfactory for this particular application. The percentage of items never administered was quite high in all cases, ranging from 43% to 57%. The CAT simulations can be said to have adequately controlled for exposure, but percentage of the bank unused might be troubling for testing directors considering the implementation of an operational CAT.

Figure 4. Item Exposure Rates



Conclusions

The primary goal of the study was to examine different approaches to creating a CAT that had comparable psychometric qualities to a paper test version. Of the three methods studied, the two based on targeted information selection gave the best comparability results. Little difference in comparability was found between the variable-length TIS CAT and the fixed-length TIS CAT. In terms of first- and second-order equity (bias and CSEM), targeted information selection gave results that closely paralleled the paper test. Comparability was less good for the classification accuracy of simulated students. Although the TIS CAT variants resulted in closer comparability than the maximum information selection methods, the TIS methods still had higher classification accuracy than the paper test. Although better classification would normally be seen as an advantage, the goal was to replicate the psychometric properties of the paper test as closely as possible. Further study is needed to uncover the cause of the discrepant classification accuracy results. Given that comparability was achieved in terms of first and second order equity, it seems likely that a CAT using targeted information should also be able to be used to match the classification accuracy as well.

The third CAT variant used to achieve comparability was using maximum information item selection with a variable test length. In general, comparability was not achieved with the variable-length MI CAT for either classification accuracy or for first or second order equity of scores. The hypothesized reason for the lack of comparability was that the minimum test length prevented the CAT from stopping for non-central θ levels. A different minimum test length might have allowed for better comparability. If comparability is the primary goal of a testing program and a variable-length MI CAT is being considered, then the possible effect of selection of the minimum test length must be taken into account. The minimum test length issue was not a problem for the variable-length TIS CAT, which had virtually the same average test length as the MI version, but achieved much better comparability.

Although examining CAT and paper test comparability was the primary aim of this study, a secondary focus was to highlight the potential gain in measurement precision that could be achieved with CAT. The fixed-length MI CAT was designed with this goal in mind. Results showed that CAT could attain much better classification accuracy and improve the measurement precision for non-central θ levels while greatly reducing test length (from 60 to 35 items). Content coverage and exposure of items were also adequately controlled. The study demonstrates that there is potential for CAT to significantly improve the measurement of students in state-wide testing.

One area that warrants further study is item bank usage. The CAT variants in this study all controlled the exposure of items while still significantly reducing test length over the paper test, but did so at a cost of using only approximately 50% of the item bank. Previous research had found targeted information procedures to improve bank use, so the findings of the current study were disappointing in that respect. A study by the authors has been planned to incorporate the exposure control procedure used by Thompson and French (2003). That exposure control procedure was designed specifically for targeted information methods and was found to greatly increase bank usage at the cost of slightly increasing test length. It would be interesting to see if this method or some other procedure would give the desired results for the current test.

Perhaps the most important message that can be taken from the current study is that targeted information item selection is a flexible tool that can be used to custom design a test to meet a wide variety of goals of test designers, whether those goals require a fixed- or variable-length CAT. Using the same CAT algorithm and only changing the information target matrix, the CAT be tuned to reproduce the measurement properties of an alternate version of the test; or, with an alternate target matrix the CAT can measure students as efficiently as possible. In either case, significant reduction in test length will be possible in many instances. Furthermore, it is likely that other target matrices could be devised to achieve other goals.

One example of an alternate goal might be to increase the classification accuracy at multiple decision points. While this is fairly straightforward CAT design with a variable-length test, it is more difficult to achieve with a fixed-length CAT that uses a conventional item selection procedure. Using a targeted information procedure, the targets near the cutpoints would be set higher than the other θ levels. In this way, the more discriminating items could be "reserved" for students near the cutpoints. While this is one example, other applications likely exist for targeted information selection.

References

Davey, T., & Fan, M. (2000, April). *Specific information item selection for adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA..

French, B. F., & Thompson, T. D. (2003, April). *The evaluation of exposure control procedures for an operational CAT.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kang, G. K., & Weiss, D.J. (2007, June). *Comparison of computerized adaptive testing and classical methods for measuring individual change.* Paper presented at the GMAC conference on computerized adaptive testing, Minneapolis, MN.

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

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 (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thompson, T. (2002, April). *Employing new ideas in CAT to a simulated reading test.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Trotter, A. (2004). A question of direction. *Educational Week, May 8*. Retrieved November 9, 2004 from http://counts.edweek.org/sreports/tc03/article.cfm?slug=35adaptive.h22

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement 38*, 19–49.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1999). *BILOG-MG: Multiple group IRT analysis and test maintenance for binary items* [Computer program]. Chicago: Scientific Software International.