Employing New Ideas in CAT to a Simulated Reading Test[1]

Tony D. Thompson

ACT, Inc.

This study attempted to improve upon the design of a CAT reading comprehension test. The goal was to develop an ideal design through simulation of the test prior to conducting an expensive series of real data studies. The specifications of the test required comparability with a paper-and-pencil version of the test, equal use of reading passages in the pool, and a match of content constraints.

Thompson and Davey (2000) describe a previous design of the reading test that was based on a routing test methodology. At that time, a fully adaptive design was not considered as content staff demanded the opportunity to preview all possible test form combinations that might be administered by the test, and a fully adaptive design would produce too many possible combinations to allow preview. A routing test design was devised that met the required specifications listed above and also permitted content specialists the opportunity to review all possible test form combinations. A series of simulation studies, however, showed that the routing design was not sufficiently adaptive to allow measurement efficiency to be improved over the paper-and-pencil version of the test.

The results of these simulation studies prompted a complete rethinking of the test specifications and design. After several discussions with content staff, it was concluded that preview of test forms was not absolutely essential and furthermore preview was an option unlikely to be used in practice due to demands it would impose on staff time. Because preview of forms was not necessary or practical, a fully adaptive CAT design that could maximize test efficiency was considered. The other test specifications (comparability with the paper-and-pencil test, equal pool usage, match of content constraints) were still considered essential. It quickly became apparent that standard methods for passage selection, constraining test content, and exposure control would not be well suited for the proposed CAT. The rest of the paper describes how recently developed CAT methodologies were employed in order to increase measurement efficiency while still meeting the test specifications.

Description of the Test

The test simulated is a passage-based reading comprehension test. Each passage on the test has ten multiple-choice questions associated with the passage. The passages selected to comprise the test must satisfy various content constraints.

On the CAT, content requirements specify that each examinee answer multiple-choice questions associated with four reading passages. Each CAT passage will have 10 items associated with it that will be administered as a set. Passages are divided into four content types, and specifications require that each examinee‘s test contain passages from the four content types in a

[1] Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April, 2002.

specified order. Thus, a passage from content type I is administered first, then a passage from content type II, etc. Although each of the four passages administered are from a different content type, they are related in such a way that the scores from passages 1 and 3 are combined into a subscore, and the scores from passages 2 and 4 are likewise used to form a subscore. The scores reported thus consist of an overall score and the two subscores. While the test contains two subscores for content purposes, the subscores are related to such a degree that a single ability measure is estimated from the test.

While item selection is a primary design consideration of any CAT, other general design issues need to be addressed that do not come up in paper-and-pencil testing. Two such issues are item preview and item review by examinees. For our test we chose to allow both preview and review for items within a passage, but not for items across passages. Allowing item preview/review requires that the items of a passage be selected and administered as a set.

Another design decision was to fix test length rather than to allow a variable test length. Although variable length tests have the advantage of being able to precisely match a desired measurement precision level for each examinee, fixed length tests are probably more appropriate for CATs with time limits. A timed test where some examinees answer fewer items that others would be unfair due to speededness concerns. On the other hand, an argument could be made that imposing almost any time limit will create test speededness, and speededness and an IRT driven CAT do not mix well. Time limits may be necessary from a practical standpoint, but the test developer should recognize that a fixed time limit CAT creates fairness and parameter estimation issues that are not easily resolved.

For our test, sample size limitations for pretesting will probably mandate that the algorithms that drive the CAT's item selection routines be based on a unidimensional IRT model. For this reason, the 3PL model was chosen to represent the interaction between examinees and items. However, we have found that when conducting simulation studies it is more realistic to generate simulated data using a multidimensional model (Davey, Nering, & Thompson, 1997). For this reason, two models were developed for the simulation—one based on a unidimensional model (3PL IRT) and the other a high dimensional MIRT model. Using a multidimensional model for the data allows one to test the robustness of the unidimensional model to realistic violations of model assumptions. The data used to calibrate the multidimensional item parameters for our simulation consisted of item responses from randomly equivalent groups of approximately 5000 examinees each, each group taking one of eight operational fixed forms from an existing paper-and-pencil test of reading comprehension. A complete description of the data generation process can be found in the series of papers Nering, Thompson, and Davey (1997), Reckase, Thompson, and Nering (1997), and Thompson, Nering and Davey, (1997).

It should be pointed out that the 3PL model is not ideally suited for a passage-based test, as local dependence concerns arise whenever items are administered in a set. While this is true in general, extensive prior experience with the paper-and-pencil version of the test has shown little or no local dependence within items of a passage. Also, any local dependence would be modeled by the high dimensional MIRT model (see above) and would show up as a difference in results between the unidimensional and multidimensional simulations. Although local dependence does not seem to be an issue with the reading test, alternative models are now being considered to replace the 3PL in the CAT.

CAT Procedures

The three major determinants of item selection on a CAT are measurement precision, content constraints, and exposure control. For the reading test studied, these factors apply to the selection of passages rather than items. To avoid confusion, in describing these procedures rather than referring to either passages or items the more generic term "unit" will be used. For the reading test a unit will refer to a passage, but the procedures described could just as easily be applied in settings where units refer to items. The following sections describe the rational for choosing the particular procedures used for the reading test.

Controlling Measurement Precision

In a traditional CAT measurement precision is seen as a quantity to be maximized subject to meeting content constraints and exposure control. Davey and Fan (2000), however, proposed an alternative idea in a procedure they called Specific Information Item Selection (SIIS). As opposed to selecting the optimal unit with respect to measurement precision, SIIS selects units so as to best match an information target. This is done by choosing an intermediate information target each time a unit needs to be selected for administration. The selected unit will come closest to matching the intermediate target subject to content constraints and item exposure. Intermediate targets can be chosen by multiplying the final information target by the percent of the test that has been completed. For example, the target halfway through the test should be half the final information target.

SIIS can be thought of as a generalization of maximum information item selection. If the intermediate targets are set to infinity, SIIS will select the unit that is most informative. Thus, even CATs that use maximum item selection operate under the SIIS principle. With SIIS, however, the developer can specify an information target if there is an advantage to doing so.

As described by Davey and Fan (2000), selecting units by maximum information has at least three potential disadvantages. First, measurement precision for examinees of the same ability is determined entirely by the random component in exposure control. Some examinees of a certain ability will get only the best units available, while other examinees of the same ability may receive many suboptimal units depending on the vagaries of exposure control. SIIS, however, updates the current intermediate target of the examinee to reflect the amount of information already obtained. This feedback mechanism inherent in SIIS allows for more consistent precision for examinees with the same ability.

The second drawback of maximum information item selection is that test measurement characteristics are unduly dependent on the composition of the pool of units. If pool quality varies across time it can be expected that measurement precision will also vary. In SIIS, however, measurement precision is more tightly controlled and the targeted precision will be achieved so long as the pool is of sufficient quality.

A third disadvantage of maximum information item selection is that it is likely to result in unbalanced pool use. This effect can only be counteracted by making exposure control more stringent. Of course, exposure control can only be made more stringent to the degree that changes in test length and measurement precision are possible. Furthermore, as noted above,

increasing the randomness of unit selection through exposure control also increases the variability of measurement precision for examinees of the same ability. As described later in the paper, the use of SIIS allowed a new exposure control procedure to be developed that better balanced pool use without an undue effect on measure precision or test length.

For the design of the reading test in this study, choosing passages by maximum information would be inappropriate. To achieve comparability with the paper-and-pencil version of the test, it is necessary to match an information target, and it is important that examinees of the same ability are measured to the same level of precision. Factoring in the desire to balance pool use, it is clear that SIIS presented a more useful approach for selecting passages on this test.

Because only four passages are administered on the reading test, implementation of the SIIS algorithm is fairly simple. The first and third passages administered to examinees comprise Subscore 1, while the second and fourth passages comprise Subscore 2. Two information targets were used, one for each subscore. The subscore information targets were formed by averaging the information functions from the passages contained in the pool. The intermediate information targets for passages one and two were one-half the total information targets for Subscores 1 and 2, respectively. The information targets for passages three and four were simply the subscore information targets.

Matching Content Constraints

Matching an information target is complicated by the need to satisfy content constraints. A problem that often crops up when dealing with content constraints stems from the requirement of the unit selection algorithm to meet the information target and content requirements simultaneously. Near the end of a CAT it quite often happens that no units exist in the pool to permit both requirements to be satisfied. This can be a result of the CAT algorithm backing itself into a corner through the selection of units early in the test. What is needed is an algorithm that looks ahead to ensure that the current unit being selected will not cause problems in satisfying constraints later on in the test.

A relatively simple idea was employed to solve this problem on the CAT reading test. Later, it was discovered that Van der Linden (2000) had already formalized this idea in a unit selection algorithm he called shadow tests. The idea behind shadow tests is that when a unit is selected to be administered, the algorithm also selects provisional units to fill out the remainder of the examinee's test. Thus, one can be assured that the current unit being selected is unlikely to later cause an unresolvable conflict in meeting content constraints. Only if content constraints vary by ability level will the shadow test algorithm possibly fail to meet content constraints, as the current ability estimate may differ from the final ability estimate. As shown later, when the shadow test idea is applied to the reading test simulation content constraints are easily met while still allowing for the matching of the target information function.

With only four passages being administered in each test, the shadow test approach is implemented easily for the reading test. Passages in the test are administered in order, first content type I, then content type II, etc. In addition, further content constraints are place on the passages selected. To create a balanced use of the pool, the first passage was selected at random from the available passages of content type I. The fourth passage administered (content type IV)

was found to have little influence on content constraints and thus could be effectively ignored for content balancing purposes. So only the second and third passages needed to be considered for the shadow test approach. The solution to the content constraint problem was to select (but not administer) the third passage prior to selecting the second passage. Now when passage two was selected, it was known that a passage three existed in the pool that would enable the test to meet content and information constraints (as determined by the current ability estimate). After passage two was administered, the selection process for passage three was repeated. This allowed the updated ability information from passage two to be used in the selection of passage three. After passage three was administered to the examinee, the fourth passage was selected and administered.

Controlling Unit Exposure

The third major element of unit selection is exposure control. Many exposure control procedures have been proposed in the literature. Reviews of these procedures can be found in Kingsbury and Zara (1989), Davey and Nering (1998), Revuelta and Ponsoda (1998), and Stocking and Lewis (2000). Many of the most popular exposure control methods are based on a procedure developed by Sympson and Hetter (1985). The Sympson and Hetter method orders units by desirability according to a suitable criterion and then the unit that first passes exposure control is administered. Exposure control parameters are determined through simulation, where the units most likely to be selected are given parameters that limit their chances of being administered. The Sympson-Hetter procedure has been extended by several researchers over the years with the main goal of controlling unit exposure conditionally on examinee ability (Davey & Parshall, 1995; Thomasson, 1995; Nering, Davey, & Thompson, 1998; Stocking & Lewis, 2000).

Although the Sympson-Hetter methods work well for preventing the over use of desirable units, balanced pool use is not usually achieved. No mechanism increases the chances a less desirable unit will be selected, and thus, it is typical for large amounts of the pool to be unused. This situation is not acceptable for practical testing applications where units are expensive to develop.

In this study an alternative exposure control procedure is used that has many features in common with earlier proposed procedures. The method is referred to as a new exposure control procedure, as I have never seen it mentioned in the literature, but it is really just a combination of previously developed procedures. The main advantage of this exposure control procedure is that it enables less frequently administered units to be used whenever they satisfy content and measurement constraints. Thus, it allows for a more even use of the pool while guaranteeing that only units appropriate for the examinee are administered.

The method first determines which units are acceptable with respect to content constraints and measurement precision. Measurement precision acceptability is determined by setting bounds around the intermediate information target. A unit is selected for administration from the set of acceptable units with a probability inversely related to its administration rate. Thus, if the acceptable set contains two units, the less often used unit is more likely to be selected.

The exposure control parameters are simply the administration rates of the units as determined through simulation. The need for simulation to estimate administration rates is similar to the Sympson-Hetter methods. In this study, the following rule was used to determine a unit's probability of being selected given its inclusion in a particular acceptable set:

$$\frac{A_i}{\sum_i A_i},$$

where A_i is the administration rate (exposure parameter) for unit i, and the denominator represents the sum of administration rates for all units contained in the acceptable set.

Unfortunately, there was not sufficient time to complete a simulation study to compare the results of the new procedure with other exposure control methods. Such a study is currently being planned. In the absence of a comparison study, one can only speculate on the relative merits of the new procedure. It can be presumed that a more balanced pool use is likely to be achieved, but it is not known whether this effect is apt to be large or small. On the negative side of the ledger, the procedure might be more likely to result in high overlap rates between examinees of similar abilities, as compared to methods based on Sympson-Hetter. Also, the current method does not condition on ability, although this feature might be added.

Results and Discussion

Two simulation studies were conducted to examine the success of the implemented CAT procedures. In one simulation the true model underlying the data was based on a MIRT model, in the other simulation the true model was based on the 3PL. Because both simulations gave very similar results and identical conclusions, only the results of the unidimensional simulation are presented here.

The success of the implemented CAT procedures was judged by examining various aspects of the simulated tests. First, comparability with the paper-and-pencil (P&P) test was evaluated, as comparability was a requirement for the CAT. Second, the effectiveness of the individual components of the passage selection process was examined. That is, the SIIS algorithm, the shadow test methodology, and the exposure control procedure were individually evaluated.

The majority of results of the simulation study are presented conditional on a unidimensional approximation of true ability, with 5000 simulated examinees being simulated for each scale score point. Each of the simulated examinees took both the CAT and a randomly determined form of the P&P test. For the multidimensional simulation, the true ability approximation was constructed by first finding the unidimensional 3PL ability with response probabilities that best matched the response probabilities corresponding to the MIRT model that represented truth in the simulation (see Thompson, Davey, & Nering, 1998). This was done for the overall score and for both subscores using all of the items in the pool. The true ability approximations were then rescaled to true scale scores, using the same transformations that would be used for an operation test. For the unidimensional simulation, the true scale scores were also only an approximation of true ability, as the transformation of ability-to-scale score is a many-to-one mapping.

Several simulations were conducted to find the appropriate test length for the CAT. Eventually, a test length of 32 items was the minimum test length judged to meet all constraints.

Comparability

Comparability with the P&P test was assessed by examining first and second order equity (Lord, 1980). First order equity addresses whether two tests give the same average scores (is one test biased relative to the other). Figure 1 displays the results of the conditional average difference between the CAT and P&P tests. For Subscore 1, Subscore 2, and the total score, the average difference between the tests was less than one score point across the ability scale.

Second order equity addresses whether the conditional standard errors of two tests are the same. Figure 2 presents the conditional standard errors of the total score and the two subscores. In all cases, the differences were small or favored the CAT.

In addition to examining first and second order equity, the match-to-target information functions for the CAT and P&P tests were plotted. Figure 3 displays the information functions of the two tests and the target information function. These curves match reasonably well. A slight departure from the target was observed for Subscore 1. No explanation was found for this result.

Figures 1-3 show that the CAT was comparable to the P&P test. The obtained equity was as good or better than that obtained from a previous computer design that used a routing test procedure (see Thompson & Davey, 2000). Comparability was obtained despite the fact that the CAT had a test length of 32 items, as compared to the 40 item P&P test.

Information Control

Results from the comparability section indicated that both the CAT and P&P test matched the information target on average. A separate question is how much variability in obtained information is present for the two tests. Figure 4 presents the conditional standard deviation of obtained information for the two tests. The figure shows that the CAT resulted in tests that were much more consistent in the amount of information obtained for a large part of the ability scale. This was particularly true for the total score and Subscore 2.

Further evidence of the smaller degree of variance shown by the CAT is given in Figure 5. This figure displays the percentage of simulated examinees with tests that did not meet information constraints for Subscores 1 and 2. Percentages were calculated from a simulation of 60000 administrations with a Normal (0,1) ability distribution. Information constraints were defined as ±10% of the target information function. This is the same rule that determined if a unit was to be admitted to an acceptable set during unit selection. The figure shows that the tests of over 30% of P&P examinees were too low in information, while at least another 30% exceeded the target information range. The results were the same for both subscores. The CAT produced much fewer tests that failed information constraints, although in the case of Subscore 1 about 14% of the CAT tests had too little information. The fact that the CAT was much more consistent in the amount of information obtained could in part be attributed to the use of the SIIS algorithm.

Content Constraints

The shadow test procedure was very successful in meeting content constraints. In every test administered, all content constraints were met. Another algorithm had been tried early on in the research to control content. It failed to meet content requirements in a satisfactory manner, and so was discarded in favor of the shadow test method.

Exposure Control

Figure 6 presents the administration rates of the 32 passages in the pool for a simulation sample of 60000 administrations with a Normal (0,1) distribution of ability. The black dots in the figure show the administration rates with the exposure control procedure implemented. The rates varied from approximately .08 to .16. The ideal administration rate of a perfectly balanced pool is graphed as a solid black line. As a baseline of comparison, another simulation was conducted with equal exposure control parameters for each passage. As might be expected, without exposure control pool use is very uneven. The results show that the CAT's pool use was acceptably balanced for operational use, although the administration rates for a couple of the passages were lower than desired.

Conclusions

The adaptive nature of the CAT simulated in this paper was greatly constrained. This was due to the desire to achieve various goals, such as providing comparability with the P&P version, meeting content constraints, and balancing pool usage. Innovative CAT methods were employed in an effort to enable the CAT to significantly reduce the 40-item test length of the P&P test while simultaneously operating under of the various constraints.

The goal of meeting the constraints was achieved. The results showed a high degree of comparability between the CAT and the P&P test. In addition, the CAT improved upon the P&P test by reducing the variability of obtained information, particularly for examinees in the middle of the ability distribution. The CAT had much fewer examinees whose obtained information fell outside of the information target range. Also, content constraints were met for every CAT test administered and the relative use of passages was acceptably balanced.

Although constraints were successfully satisfied, the price paid was an only moderate reduction in test length. The shortening of the test by eight items, while a substantial amount, may not be enough to justify the costs of developing a separate CAT test. It is clear that it is difficult to design a highly constrained CAT that is able to significantly reduce test length.

Nonetheless, the CAT procedures investigated in this study show much promise and may prove useful in a variety of other CAT settings. A future simulation study is planned to more directly compare the utility of these newer methods compared with more traditional CAT methods. For example, the purpose of the present study was not to directly compare the tradeoffs of using SIIS versus maximum information item selection on the reading test, but such a study would be interesting. Additionally, the new exposure control procedure is likely to need some refinement and testing before implementation in an operational CAT could be advised. In particular, a version that is conditional on ability might need to be developed.

Although only a first step, the present paper has shown that there is still a need for further research in CAT. Test development using CAT is a balancing act where the developer seeks large gains in measurement precision without damaging test validity or increasing developmental costs. It is likely that we will need new and better item selection procedures in order succeed in this delicate balancing act.

References

Davey, T., & Fan, M. (2000). *Specific Information Item Selection for Adaptive Testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA, April.

Davey, T., & Nering, M. (1998). *Controlling item exposure & maintaining item security.* Paper presented at the ETS colloquium, *Computer-Based Testing: Building the Foundation for Future Assessments*, Philadelphia.

Davey, T., & Parshall, C. G. (1995). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359-375.

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

Nering, M. L., Davey, T., & Thompson, T. (1998). *A hybrid method for controlling item exposure in computerized adaptive testing.* Paper presented at the annual meeting of the Psychometric Society, Urbana, IL.

Nering, M., Thompson, T. D., & Davey, T. (1997). *Simulation of realistic ability parameters*. Paper presented at the Psychometric Society meeting, Gatlinburg, TN, June.

Reckase, M. D., Thompson, T. D., & Nering, M. (1997). *Identifying similar item content clusters on multiple test forms.* Paper presented at the Psychometric Society meeting, Gatlinburg, TN, June.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 311-327.

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*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

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.

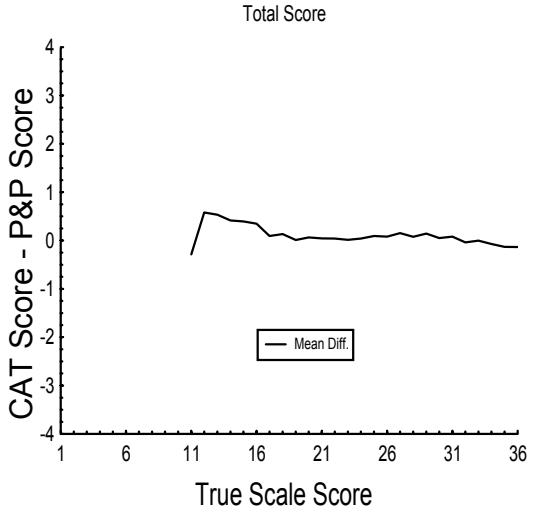
Thomasson, G. L. (1995). *New item exposure control algorithms for computerized adaptive testing.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis.

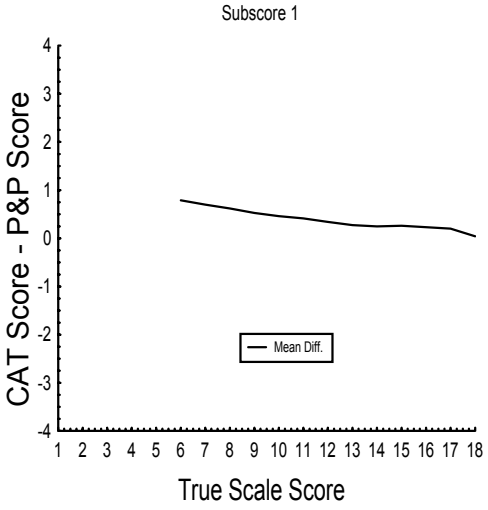
Thompson, T. D., & Davey, T. (2000). *Applying specific information item selection to a passage-based test.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA, April.

Thompson, T. D., Davey, T., & Nering, M. (1998). *Constructing adaptive tests to parallel conventional programs.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA, April.

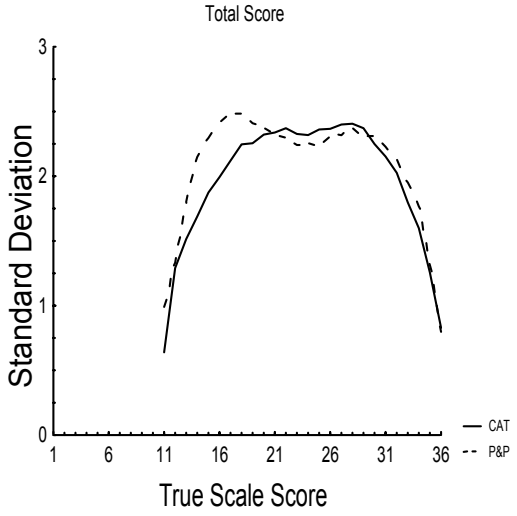
Thompson, T. D., Nering, M., & Davey, T. (1997). *Multidimensional IRT scale linking without common items or common examinees*. Paper presented at the Psychometric Society meeting, Gatlinburg, TN, June.

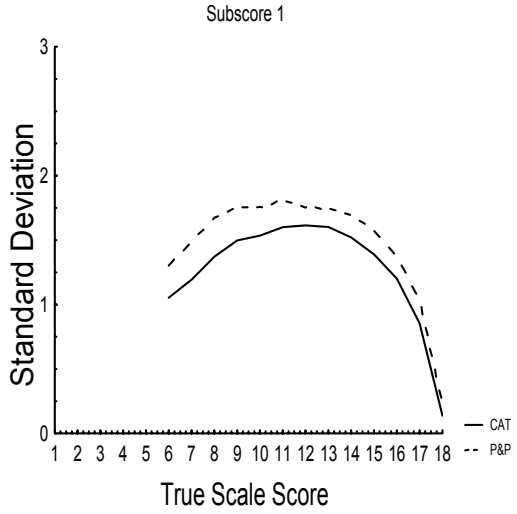
van der Linden, W.J. (2000). Constrained Adaptive Testing with Shadow Tests. In W.J. van der Linden & C.A.W. Glas (Eds.). *Computerized Adaptive Testing: Theory and Practice.* Dordrecht, The Netherlands: Kluwer Academic Publishers.


Total Score
CAT Score - P&P Score
4
3
2
1
0
-1
-2
-3
-4
1
6
11
16
21
26
31
36
Mean Diff.
True Scale Score


Subscore 1
CAT Score - P&P Score
4
3
2
1
0
-1
-2
-3
-4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Mean Diff.
True Scale Score


Subscore 2
CAT Score - P&P Score
4
3
2
1
0
-1
-2
-3
-4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Mean Diff.
True Scale Score


Total Score
Standard Deviation
0
1
2
3
1
6
11
16
21
26
31
36
True Scale Score
CAT
P&P


Subscore 1
Standard Deviation
0
1
2
3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
True Scale Score
CAT
P&P

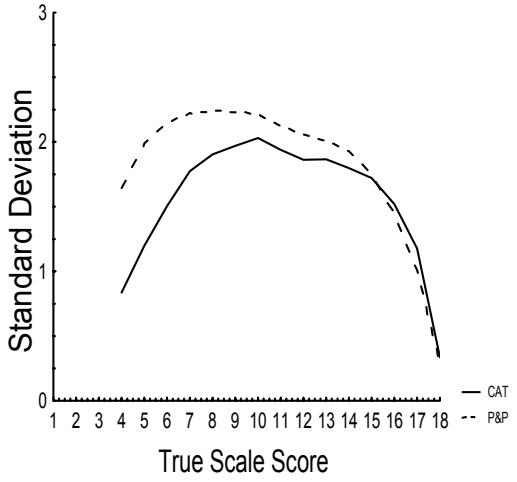

Subscore 2
Standard Deviation
0
1
2
3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
True Scale Score
CAT
P&P

Figure 3: Target Information

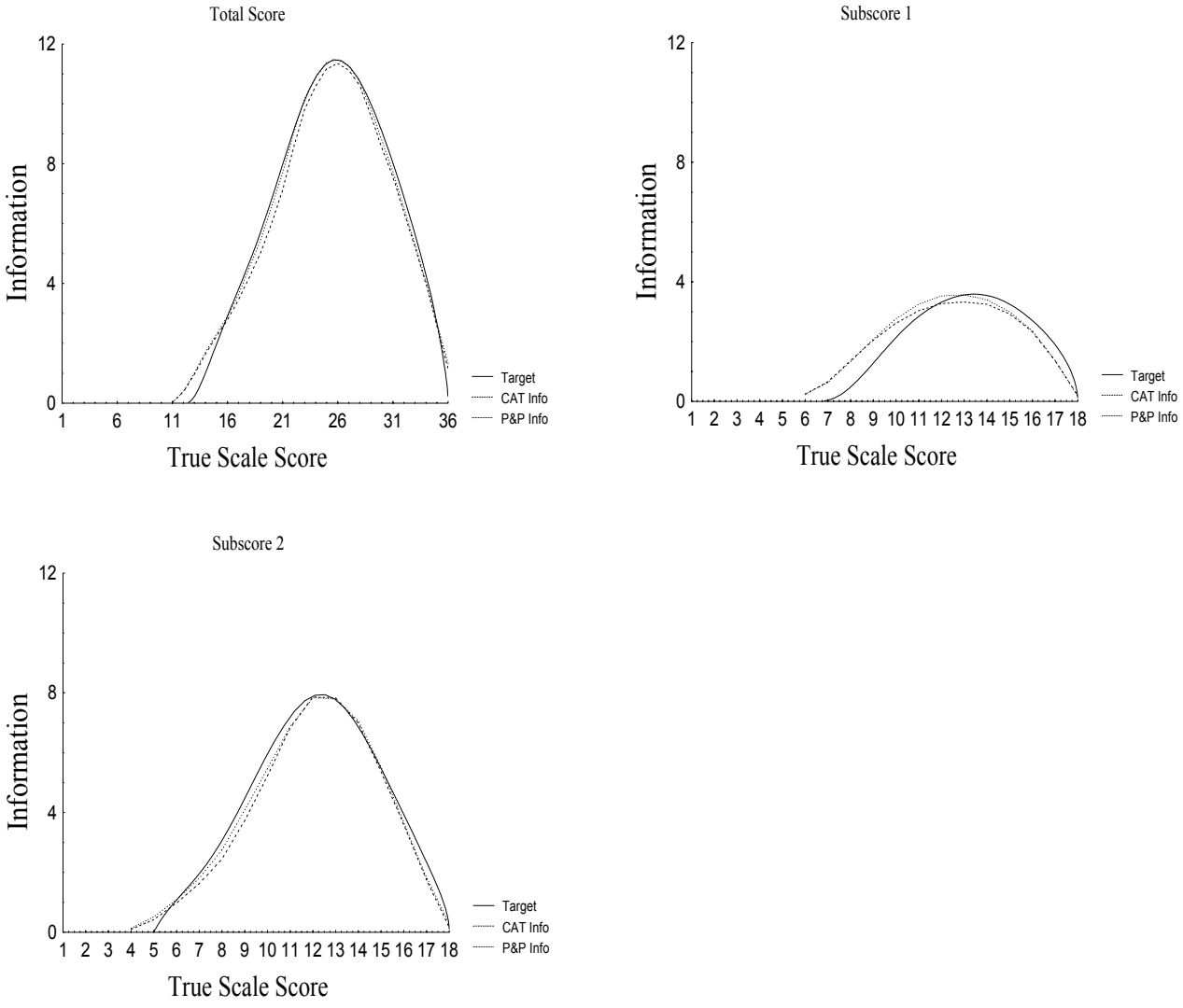
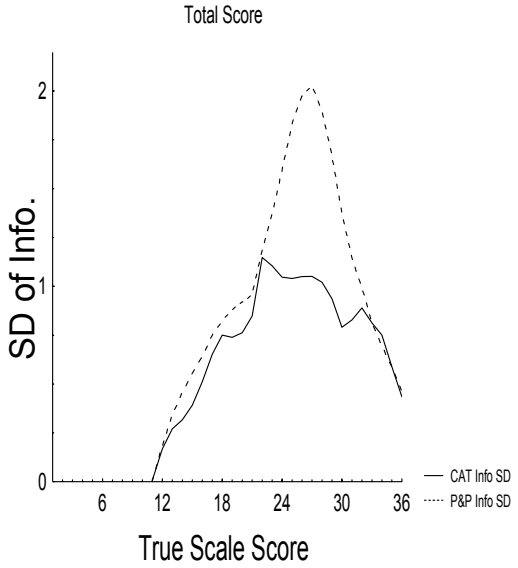
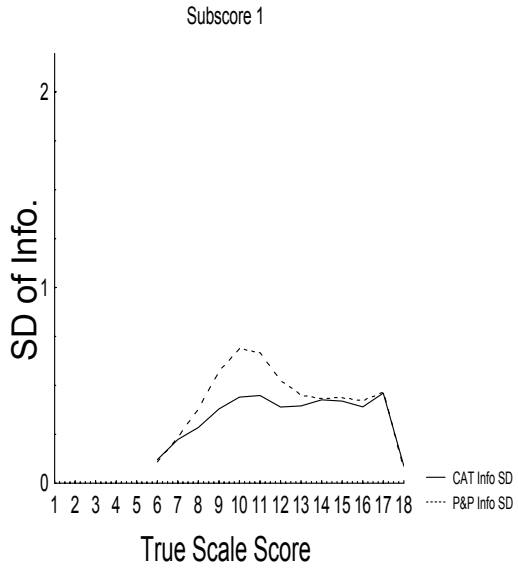


Figure 4: Standard Deviation of Obtained Information





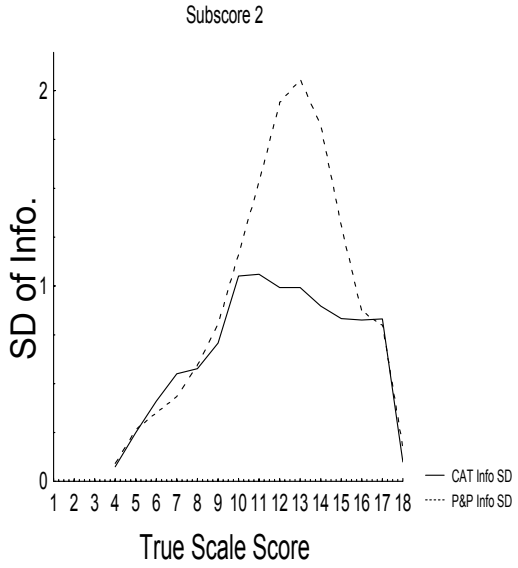


Figure 5: Percentage of Simulees with Tests Not Meeting Information Constraints

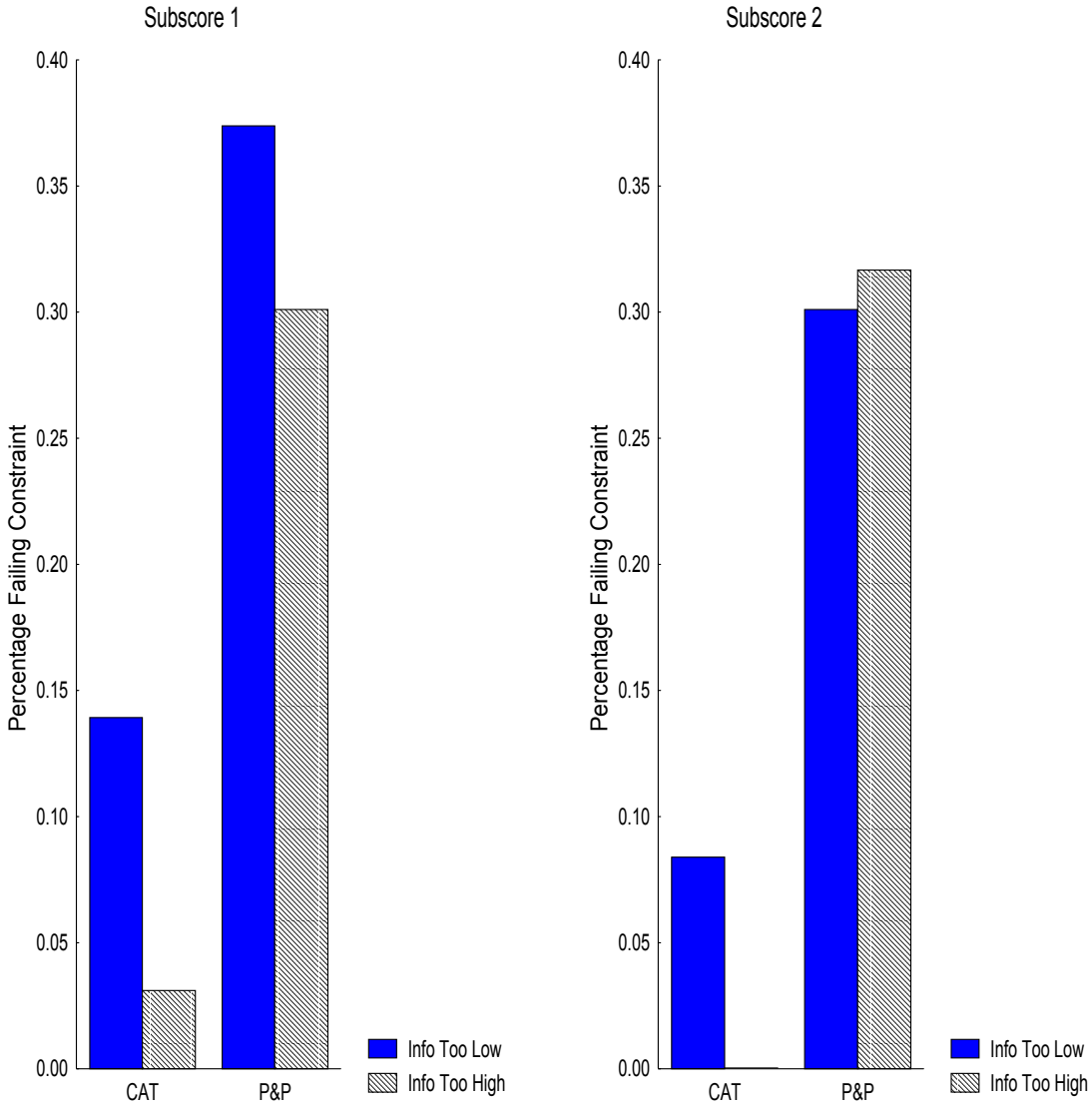


Figure 6: Passage Administration Rates with and without Exposure Control