DOCUMENT RESUME

ED 432 614		TM 029 997

AUTHOR	Thompson, Tony D.; Davey, Tim
TITLE	CAT Procedures for Passage-Based Tests.
PUB DATE	1999-04-00
NOTE	19p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 20-22, 1999).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Adaptive Testing; College Entrance Examinations; Comparative Analysis; *Computer Assisted Testing; Higher Education; Reading Comprehension; *Reading Tests; Test Format; Test Items
IDENTIFIERS	Paper and Pencil Tests

ABSTRACT

Methods to control the test construct and the efficiency of a computerized adaptive test (CAT) were studied in the context of a reading comprehension test given as a part of a battery of tests for college admission. A goal of the study was to create test scores that were interchangeable with those from a fixed form paper and pencil test. The first approach to controlling the test construct is to require the CAT to balance the item content type by constraining the amount of information obtained from each content area through algorithms developed by T. Miller and T. Davey (1999). A second approach is to allow content constraints to vary across ability levels. A third approach is to allow a variable standard error across the ability scale. Preliminary results from a simulation study show that a CAT with fixed passages and adaptive items has the potential to produce interchangeable scores with a fixed form version of the test. (Contains 2 tables, 4 figures, and 13 references.) (SLD)

CAT Procedures for Passage-Based Tests

Tony D. Thompson
Tim Davey

This paper is prepared for the:
Annual Meeting of the National Council on Measurement in Education
Montreal Canada, April 1999

CAT Procedures for Passage-Based Tests[1]

Tony D. Thompson
Tim Davey
ACT, Inc.

Over the past few years, we have engaged in several research studies designed to create a CAT with test scores that are interchangeable with those from a fixed form paper and pencil test (Davey & Nering, 1998; Davey, Nering & Thompson, 1997; Hsu, Thompson & Chen, 1998; Nering, Thompson & Davey, 1997; Parshall, Davey & Nering, 1998; Reckase, Thompson & Nering, 1997; Thompson, Davey & Nering, 1998; Thompson, Nering & Davey, 1997). Our definition of interchangeable is the same as what Lord (1980) called *equity*, i.e., two tests that have identical conditional score distributions. In order to achieve equity, we have investigated some sophisticated CAT methods designed to improve control of the test construct being measured and the CAT's efficiency in measuring it. We will only briefly describe these approaches here, as they are fully detailed in Miller and Davey (1999). Greater emphasis is given instead on how these methods can be implemented in a passage-based CAT setting. The specific test of interest for this paper is a reading comprehension test given as part of a battery of tests for college admission. We note that the methods described in Miller and Davey (1999) apply whenever it is desirable to have a high degree of control over the construct being measured by a CAT, not just in the specific case of matching the conditional score distributions of a fixed form test. As our goal in this paper is to match a fixed form test, however, it is perhaps useful to examine some of the differences between CAT and paper and pencil testing.

Fixed form tests differ in several important practical ways from CAT tests. Whether these differences favor CAT depends to some extent on the perspective one takes. Two of the many examples of differences that are usually seen to favor CAT are the following: First, CAT permits custom-built tests for each examinee, allowing shorter tests and an increase in measurement precision. Second, CAT allows for the possibility of immediate score reporting. But these seeming advantages for CAT have their downside as well. While a custom-built test has a host of attendant advantages, a critical assumption made is that all tests constructed by the CAT algorithm will measure the same construct. This is a potential problem for us at ACT, since our tests tend to be somewhat multidimensional in nature. We must therefore remain vigilant to ensure that new forms measure the same unidimensional reference composite as previous forms. When creating fixed forms we strive to ensure that the test construct remains stable from form to form by invoking the expert judgment of test content specialists during form construction, and trust that any residual inconsistencies between forms are further mitigated by posttest equating. These precautions are not generally possible with CAT, as forms cannot be reviewed ahead of time given the nearly infinite number of potential CAT tests, and posttest rescoring of the test through an equating would prevent the immediate feedback of test scores. Thus, it seems that some of CAT's greatest potential advantages present us with some of our greatest challenges in maintaining control of the test construct for each examinee. We now review the methods described in Miller and Davey (1999) in order to address these challenges.

The first approach we have used in controlling the test construct is to require the CAT to balance the item content types in a test by constraining the amount of information obtained from each content area using algorithms developed by Segall and Davey (1995). This is in lieu of what is more commonly done in CAT, which is to balance the percentages of items administered

[1] Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada, April, 1999.

from the different content classifications (e.g., Stocking & Swanson, 1993). A problem with balancing the percentage or number of items administered is that ability estimates are not influenced so much by the *number* of items selected from a content domain, but by the *amount of information* that those items provide toward estimation (Thompson et al., 1998).

The second approach we have used in controlling test content across examinees is to allow content constraints to vary across ability levels. This would seem to be a natural characteristic of most achievement tests, as content types are likely to be correlated with ability. Fixed form tests measure different content areas with differing precision across the ability scale due to the correlation between content and ability. This should be a characteristic we require of the CAT as well, if we are to insure that the CAT administers a test of similar content to examinees of similar ability.

The third approach we have used is to allow a variable standard error across the ability scale. This is important to us because it allows an increased control in the precision with which the test construct is measured. In order to match the fixed form test, which implicitly has a variable standard error, it is necessary for the CAT to have a standard error that can vary across ability as well. The goal of a variable standard error is met most easily with a variable length test. However, we are constrained to use a fixed length test for the CAT reading test in order to reduce the possible effects of test speededness that might occur if every examinee is given a fixed time limit but a variable number of items. Nevertheless, we believe that a fixed length CAT can match a variable standard error target information function, as long as the CAT selects items in such a way as to match the target without exceeding it. The standard CAT approach of selecting items to maximize information does not work here, as the target is likely to be exceeded by the end of the test. In a later section we describe the approach we are taking in trying to match the target precisely.

As mentioned previously, a key characteristic for our CAT reading test is for the CAT scores to be interchangeable with scores from the fixed form reading test. We define interchangeable to mean that examinees with the same ability have the same expected score distribution regardless of which test they take, and therefore, have no reason to prefer either the CAT or the fixed form test on the basis of their expected score distribution. Thompson et al. (1998) provide further detail about score interchangeability and why it is necessary for our CAT program. For the current paper, however, we focus on the restrictions that score interchangeability forces on the design of our CAT reading test. How to best design the CAT to comply with these restrictions has not been completely finalized. For this study, we focus on one form that our CAT reading test may take.

The simulated CAT reading test is based on the paper and pencil fixed form reading test. The fixed form test requires each examinee to answer multiple choice questions associated with four reading passages. Each passage has 13-15 items associated with it that are pretested, with ten of those items being used on the operational fixed form. Each passage is technically its own content type, but 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 consist of an overall score and the two subscores.

Several four-passage sets are created each year, with content specialists carefully examining each set to assure that a large number of formal and informal test construction rules are adhered to. In order to match content constraints, the CAT version of the reading test must also have four passages. To allow examinees the opportunity to review items before answering we plan to administer the items of a passage as a set, which would require the CAT algorithm to select all of the items corresponding to a passage prior to administering that passage. In addition, because of speededness concerns we prefer to administer an equal number of items to examinees.

Although our CAT's item selection routines are driven by algorithms based on a unidimensional IRT model, we have found that when conducting a simulation study it is more realistic to generate the simulated data using a multidimensional model (Davey et al., 1997). The data used to calibrate the multidimensional item parameters for our simulation consisted of item responses from randomly equivalent groups of approximately 3000 examinees each, each group taking one of eight operational fixed forms. A complete description of the data generation process can be found in the series of papers Davey et al. (1997), Nering et al. (1997), Reckase et al. (1997), and Thompson et al. (1997). Two CAT item pools were developed from the actual fixed form items. The first was simply the 320 items that made up the eight fixed forms. In the second pool, three items were cloned for each passage by duplicating their multidimensional item parameters. Items were cloned to increase the pool size to something more similar to what would be observed in practice. Although only 10 items per passage are used operationally, 13-15 items are actually pretested and would be available for use by the CAT. The items selected to be cloned in each passage were the item with the greatest multidimensional discrimination parameter, the item with the least multidimensional discrimination parameter, and a random item. As the cloned pool has more realistic numbers of items per passage, it was the primary pool used in the study.

In choosing a CAT design to simulate, we considered three alternatives. One was to have the test administer preselected fixed forms, which is not a CAT at all of course, but would satisfy the score interchangeability issue rather easily. The second option would be to select passages and items in real time. In this case, each examinee would receive a set of passages that was best suited for their ability, and within each passage, a set of items also tailored to their ability. A major disadvantage of this approach is that it prevents content specialists from reviewing forms ahead of time, and so would necessitate an extensive set of test construction rules to be encoded into the CAT passage selection algorithms. As noted before, the current fixed form test construction rules are complicated and in some cases not even fully formalized. Consequently, encapsulating these rules into computer code would be difficult, if not impossible. This leads to the third alternative for the CAT, which would be to fix the passage sequence but to have the items associated with each passage selected in real time. This was the option we chose to simulate, as it allows greater flexibility than using preselected fixed forms, and seems from the point of view of constructing a test to be more practical than selecting passages in real time.

For purposes of controlling the frequency with which passages appear together, which may be thought of as a form of exposure control, we may choose to have anywhere from several dozen to more than one hundred passage combinations. Even with this many passages we still will be able allow our content specialists to carefully review the suitability of each combination prior to the test administration. Our pool has eight passages of each of the four content types so there are 8^4 (= 4096) possible passage combinations—many more than we need for purposes of passage exposure control. It makes sense then to choose passage combinations that are in some way optimal. One useful property we could require of our passage combinations would be for each set to contain a sufficient amount of information across the entire ability range, so that any examinee's ability parameter could be well estimated regardless of the examinee's true ability. Passage sets that are low in information for certain parts of the ability continuum are not so desirable.

To operationalize the idea of meeting information constraints across the entire range of ability, the average information value for each of the two subscores was calculated across all of the 32 passages that formed our fixed form pool. Because the fixed form tests are scored by number right rather than using an IRT ability estimate, it is most appropriate to use the information function for the number right score, given in Lord (1980 Equation 5-13), to calculate the information obtained by the fixed form tests. These average information values represent our target information values for the CAT. Ideally, every passage combination would contain enough information to match the average fixed form subscore information at every ability level.

The amount of unidimensional information that is obtained by a passage depends upon the particular items that are used with the passage. Using the larger pool of 416 items, up to 13 items may be used with each passage. As we hope to make the CAT more efficient than the fixed form test, we would prefer to use fewer than 10 items per passage for the CAT. As stated previously, the CAT will administer a fixed number of items to each examinee and these items will have to be selected by the CAT algorithm prior to administering the passage in question. Information values were obtained for each of the 8^4 passage combinations based on 8-11 items per passage. Seven ability values that spanned the range of ability were used to calculate the information, with the items being selected so as to maximize the possible information at each ability level. These values represent the amount of information that could be obtained if the ability parameter were known *a priori*, i.e., with perfect item selection. The number of passage combinations that meet or exceed the average information value for *all* ability levels is given in Table 1.

Table 1: Number of four-passage combinations that meet or exceed the average information values at all ability levels.

Number of items in passage	Number of combinations meeting restrictions
8	0
9	0
10	36
11	147

As can be seen from the table, 10 items per passage are required before any of the passage combinations meet the constraints for all ability levels, and even with 10 items per passage only 36 combinations exist.

Since the fixed form test uses 10 items per passage no efficiency is gained by using a CAT to select items for a predetermined four-passage test, given the constraints we are placing on the CAT. To increase the number of passage combinations that meet the information constraints across all ability levels, we added a small amount of passage adaptivity to the CAT to allow the examinees' responses to influence the passage sequence administered. The method is a multistage test, wherein the first passage is administered to the examinee at random from the pool of passages eligible to be administered in the first position. Then the examinee's ability parameter is estimated and compared to a cut score. The cut score determines which of two three-passage sets the examinee will receive to complete their test. We refer to the passage combinations in the multistage method as seven-passage sets, as each combination is composed of a starting routing passage and two three-passage sets, only one of which would be administered to the examinee. The seven-passage sets increase the number of possible passage combinations to 8^7 (= 2,097,152), and the number of these that meet or exceed the average information values at all ability levels is given in Table 2.

Table 2: Number of seven-passage combinations that meet or exceed the average information values at all ability levels.

Number of items in passage	Number of combinations meeting restrictions
8	59040
9	107442
10	199248
11	345548

Notice that with seven-passage sets there is a large number of combinations meeting the information constraints, even with only eight items per passage. Yet, using seven-passage sets will still allow content specialists to preview the passage sets prior to administration.

Although it not difficult to find all possible seven-passage sets that meet the information constraints, a harder task is to form a pool of 40-50 passage sets so that every passage is equally likely to be selected. This is necessary to prevent the more informative passages from being overused. For purposes of the simulation, passage sets were selected by trial and error with only a minor effort being made to equalize the relative frequency of occurrence. If the results from the simulation seem promising, however, an algorithm will need to be developed to sort through the thousands of acceptable combinations and find a pool of passage sets that most nearly equalizes the frequency of passage occurrence. One finding that came out of the trial and error process was that no combination of forms was found that allowed all of the passages to be used when each passage contained eight items. The best that could be done was for the first passage administered to contain nine items, and the last three administered passages eight items each.

The following seven steps outline the process of administering a CAT reading test using the multistage method.

1. A seven-passage set is selected at random from a pool of eligible combinations. The passage sets included in the pool would be carefully examined by content specialists to ensure that all content criteria were fully met. Also, the pool of passage sets would need to contain an equal representation of all of the available passages. This would build in a kind of exposure control for each passage.

2. A predetermined number of items from the first passage are selected at random. The items would be selected at random since the CAT algorithm would have no knowledge of the examinee's ability before the test begins. Another option would be to administer a predetermined set of items. Due to exposure control concerns, however, a better idea might be to use a number of predetermined item sets in such a way so as to ensure that all the items from the passage were used with equal frequency across examinees. The use of predetermined item sets would also reduce the possibility of a test being poorly matched to an examinee's ability, which would be more likely with random selection of items.

3. An ability estimate is determined from the examinee's responses to the items from the first passage and is compared to a cut score to select which of two three-passage sets is used to complete the test. We operationalized this in the simulation by numerically integrating the examinee's posterior ability estimate against the information functions of the two three-passage

sets. The three-passage set with the greater potential information for the examinee's posterior is selected.

4. After the administration of each passage, the target information function for the subscore in question is updated. The update consists of reducing the target to account for the information already obtained. Then, an information target value for each subscore is obtained by numerically integrating each target information function over the posterior estimate of ability for the examinee. Thus, each subscore information target value is a scalar number that is essentially the expected target information for the examinee's ability estimate.

5. The item information function for each item associated with the second passage is numerically integrated over the posterior estimate of ability for the examinee. This gives an item information value (a scale number) that represents the expected information of the item for the examinee in question.

6. A predetermined number of items, let us say x, are then selected to be administered. The set of x items selected is the one with item information values (step 5) that sum most closely to the subscore information target value (step 4) associated with that passage. In addition, some item level content constraints may have to be satisfied as well. This step required an integer programming problem to be solved.[2]

7. Repeat steps 4-6 for the remaining two passages.

CAT Algorithms

Although the term CAT is often used rather generically, the performance of a CAT can vary greatly depending upon the particular algorithms used. The algorithms we used in the simulation follow those described in Thompson et al. (1998), in which a discrete item math test was simulated. For this simulation, the following options were employed using the 3PL model. The number of items administered was fixed, with the first passage containing nine items and the last three containing eight each. The estimation algorithm for the provisional ability estimate was EAP, and the final ability estimate was computed by maximum likelihood. No exposure control was used at the passage level, as we plan to enforce exposure control by the random assignment of passage sets from the passage set pool. This assumes that the frequency of passage use is equally distributed throughout the passage set pool. No item exposure control was used either, except in the case of the first passage administered where the items were selected randomly. For the last three passages administered some form of exposure control will probably need to be employed eventually. However, first we plan on examining relative frequencies of items being administered without exposure control.

Results

A simulation study is currently underway to fully examine the potential of the approach outlined above for the CAT reading test. At this time we have only some preliminary results, but they generally show that the method has promise.

Before discussing the results, we make a couple of notes concerning the figures. The results in the figures described below are all conditional on a unidimensional approximation of true ability. 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 et al., 1998). The true ability approximation was then rescaled. First, it was transformed to an expected number right score for each of the fixed forms, for both the overall score and the two subscores. These were then transformed to scale scores for each form using the operational fixed

[2] Ron Armstrong of Rutgers University wrote a computer program for us to solve the integer programming problem.

form equating tables. The scale scores were then averaged to find an approximate true scale score for the overall test and each of the two subscores. At each scale score level for the total score, 5000 simulees were administered a random fixed form and a CAT. The scores from the fixed forms were equated and scaled as would normally be done in an operational administration. It should also be noted that a replication of the study found very little change in the plots. Thus, 5000 examinees per scale score level seems to be a sufficient sample size.

Figure 1 presents the average conditional difference between the simulees' CAT score and their fixed form score. As we are trying to match the CAT to the fixed form test, ideally these graphs would be flat and show zero difference. The graphs show little difference for the middle of the ability distribution, with the bias being less than one scale score point. A small positive bias does exist for the lower end of the ability scale, which may be due to the CAT having a lower limit on scores because of the guessing parameter. For purposes of the graphs, the fixed form scores below chance level were truncated to the chance score. Although this makes the fixed form scores more compatible with the CAT, it may not totally mitigate the effect as the CAT still has a greater lower limit because the lower asymptote parameters tend to be above the chance level.

Another critical measure of the degree of interchangeability of the CAT and fixed form scores is the conditional standard errors of the scale scores, which is presented given in Figure 2. The conditional standard error plots show the CAT scores and fixed form test scores to have similar standard errors for most of the ability scale, and this is especially true for the two subscores. That the subscores match better than the overall score is not surprising as the item selection algorithm only used subscore information targets. We had hoped that matching the subscores would also match the overall score, however, this seems not to be the case. Our only explanation of why the total score standard errors were not matched as well as the subscores is that the multidimensional nature of the test is preventing it from doing so. The graphs also show that the subscore match could be improved at the lower end of the ability scale. We discuss how we might solve these problems in the discussion section.

While Figure 2 addresses how interchangeable the CAT and fixed form scores are, it does not indicate how closely the information obtained in the CAT matches the target information functions. This comparison is made in Figure 3. In addition to the target information function, which in the case of the overall score is simply the sum of the subscore information functions, the plots also give the obtained 3PL information function of the CAT items using the best unidimensional ability based on the true MIRT ability. For the most part, the target information function and the unidimensional approximation of the true information function were quite similar, indicating that the CAT matched the target on average.

Figure 4 shows graphs for the conditional standard deviation of the unidimensional approximation of true information. Small standard deviation values indicate that the CAT is providing a consistent amount of information for all examinees within that ability level. Large values indicate that the CAT is inconsistent in the information obtained for that ability level. Ideally, the standard deviations would be near zero throughout the ability scale. However, this goal is rather unrealistic given that the CAT reading test essentially makes only four decisions throughout the test—the first being which three-passage set to administer based on the responses from the first passage, and the other three decisions being which items to administer within the final three passages. A discrete item CAT would have an easier time matching the information target precisely, as the selection of each successive item would be carried out with progressively more precise ability estimates. The results in Figure 4 seem to indicate that some improvement might be possible to increase the consistency of the obtained information. On the other hand, it may be that Figure 4 simply reflects the best that can be done given the restrictive nature of the CAT we simulated.

A replication of the simulation study was also conducted with one change. Instead of using the larger augmented item pool of 13 items per passage, the new simulation was conducted using the original fixed form item pool that had 10 items per passage. The results found were very similar to those obtained with the larger pool.

Discussion

The preliminary results from this study show that a CAT with fixed passages and adaptive items has the potential to produce interchangeable scores with a fixed form version of the test. Although there were differences found between the CAT and the fixed form simulations, overall the results looked quite good considering it was an initial attempt. We had a similar experience with the discrete item math test, where considerable fine-tuning was necessary even after finding initially favorable results (Fan, Thompson, & Davey, 1999; Thompson et al., 1998). Despite the promising start, however, much work lies ahead of us. We wish both to improve the interchangeability of the scores and also to add to the realism of the simulation. Our future plans are discussion below.

One of the first steps will be to ensure that all of the passages are used with a relatively equal frequency. This would make exposure control at the passage level easy to implement. In order to achieve equal passage frequency, a computer program would need to be written that could determine all the passage combinations that meet the information constraints, and then systematically choose from the eligible combinations so as to distribute the passages equally. The program may also have to vary the number of items to be administered per passage in order to find the best pool of passage sets. In the simulation described in this paper, the number of items administered was nine for the first passage, and eight for the last three. In order to match information constraints and equalize passage frequency of occurrence, however, these numbers may have to be manipulated to some degree.

Although the pool of seven-passage sets may implicitly provide exposure control at the passage level, we still need to investigate the possible need for exposure control at the item level. It is likely that certain items will be administered more frequently than others across examinees, and in that case some form of item exposure control within each passage will be have to be imposed. Any implementation of item exposure control will further degrade the information potential of the passages, and thus further adjustments to the CAT may be necessary in order to reach the information targets.

Another issue regarding the information targets is that in the simulation discussed in this paper, the CAT had a target information function for the subscores but not the overall score. We hoped that by explicitly forcing the CAT to match the two subscore targets, it would implicitly match the total score target. However, Figure 2 shows that the CAT does a better job of matching the fixed form standard errors for the subscores than for the total scores. One option to correct this problem is to add a target for the total score. How to best implement this option is now being considered.

Finally, after all of the above issues have been addressed, we should have a fully functional CAT that will hopefully be comparable to the fixed form test. At this point our simulations could investigate the effects of changing some of the characteristics of the CAT. For example, we could examine the effects of various provisional and final ability estimate algorithms, or of using maximum information vs. weighted information in selecting items.

References

Davey, T., & Nering, M. (1998). *Evaluating and insuring measurement precision in adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA, April.

Davey, T., Nering, M., & Thompson, T.D. (1997). *Part 1: Realistic Simulation Procedures for Item Response Data.* Paper presented at the Psychometric Society meeting, Gatlinburg, TN, June.

Fan, M., Thompson, T.D., & Davey, T. (1999). *Constructing adaptive tests to parallel conventional programs.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada, April.

Hsu, Y., Thompson, T. D., & Chen, W. (1998). *CAT item calibration.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, April.

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

Miller, T., & Davey, T. (1999). *Principles for administering adaptive tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada, April.

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

Parshall, C. G., Davey, T., & Nering, M.L. (1998). *Test development exposure control for adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA, April.

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.

Segall, D. O. & Davey, T. C. (1995). Some new methods for content balancing adaptive tests. Paper presented at the Psychometric Society meeting, Minneapolis, MN, June.

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

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.

Figure 1: Average Difference in CAT and Fixed Form Test Scores

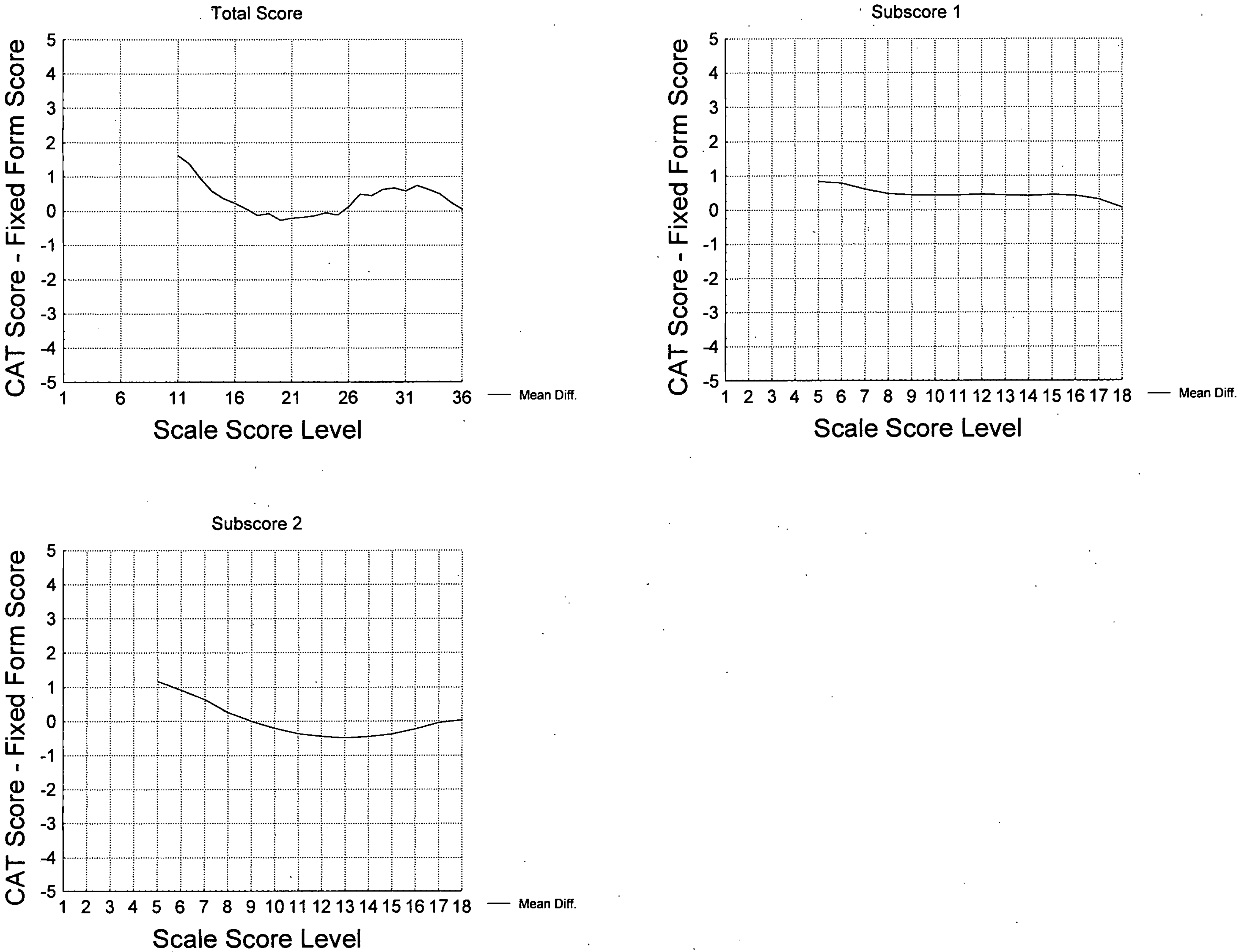
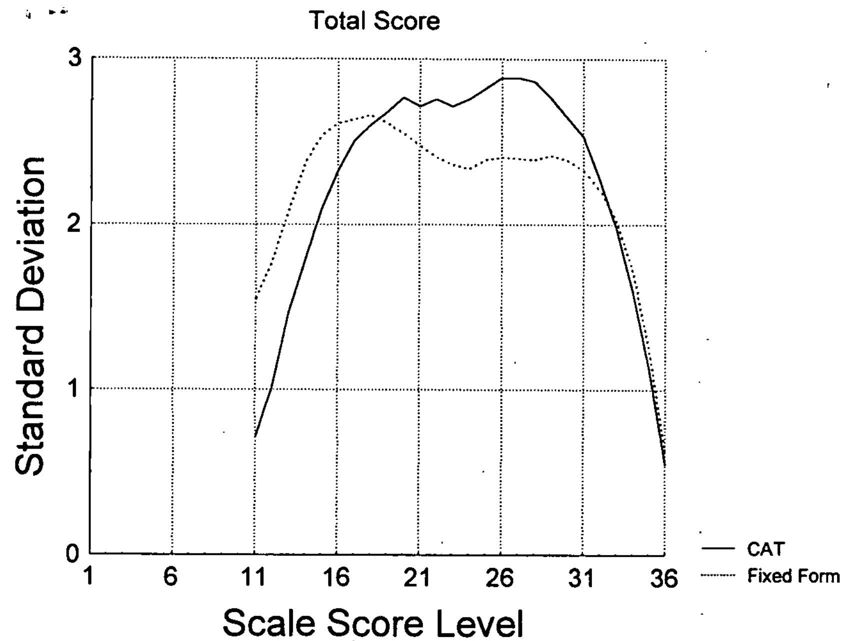
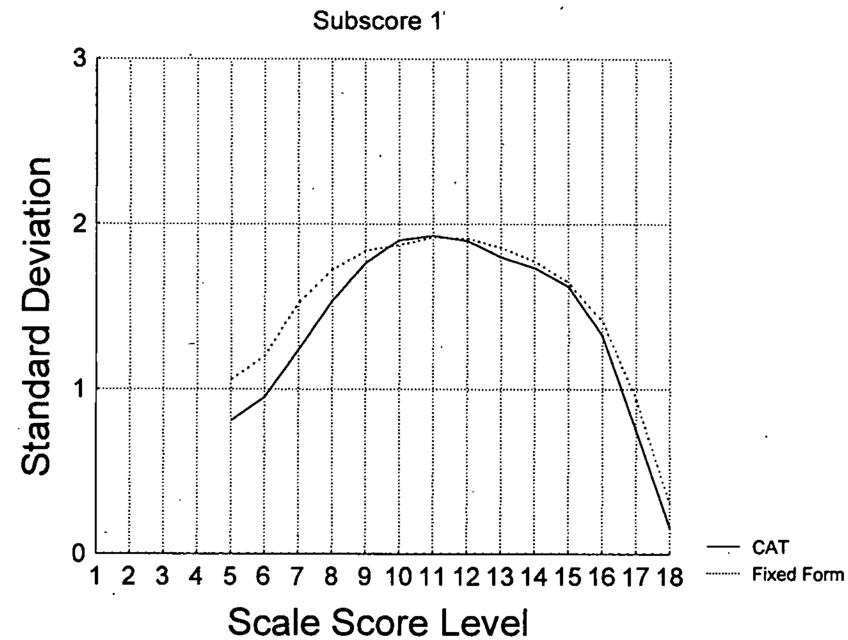


Figure 2: Conditional Standard Errors for CAT and Fixed Form Tests





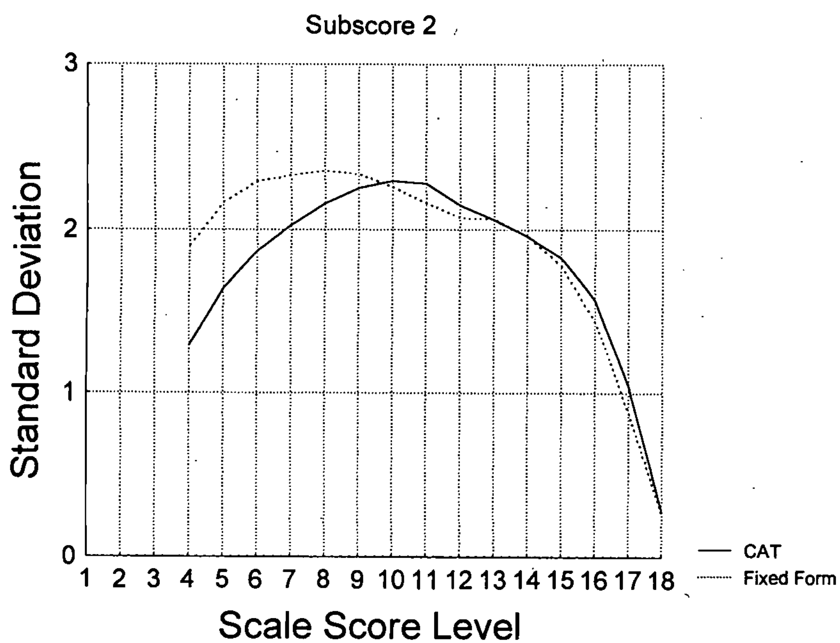


Figure 3: Target and Obtained Unidimensional Information Functions

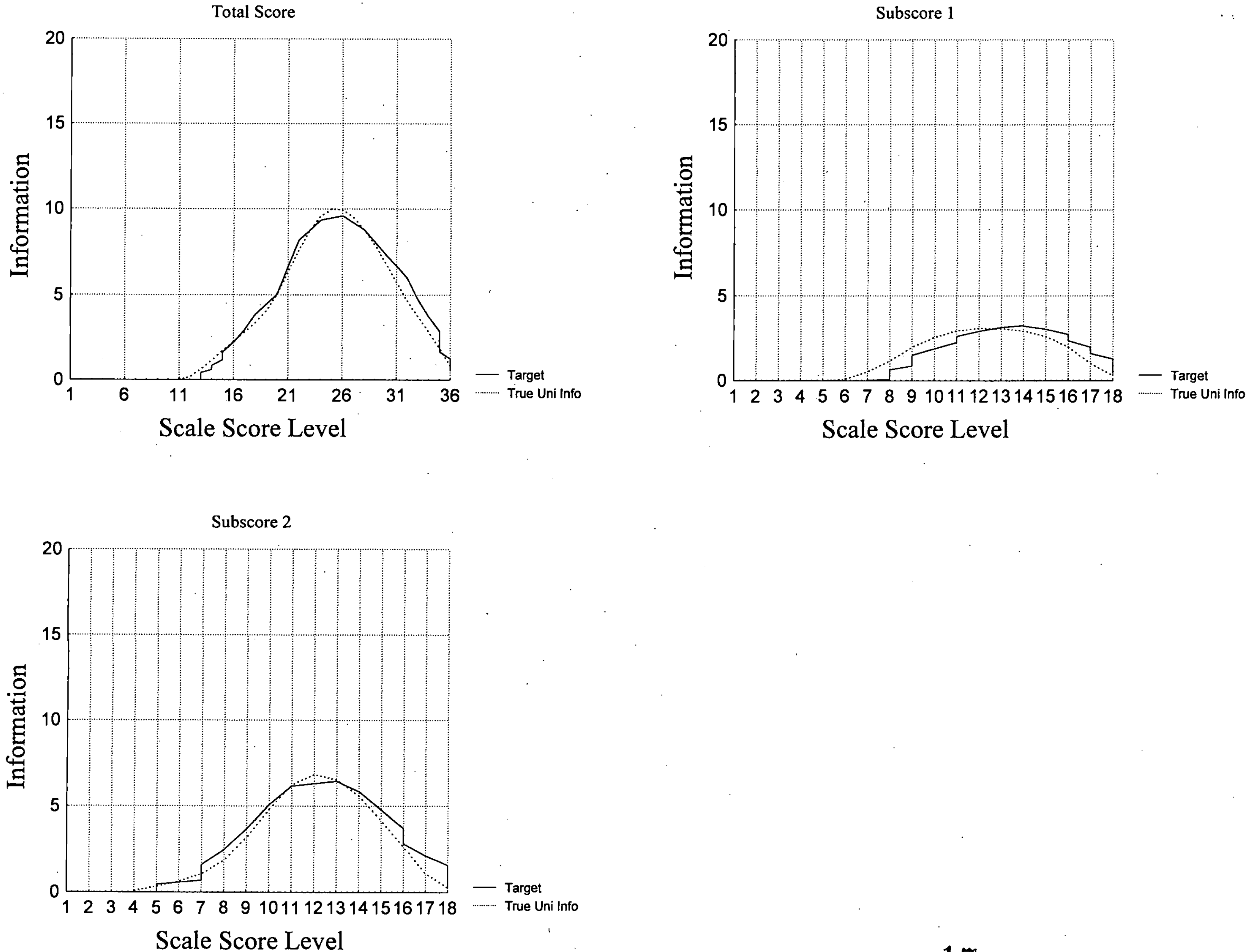
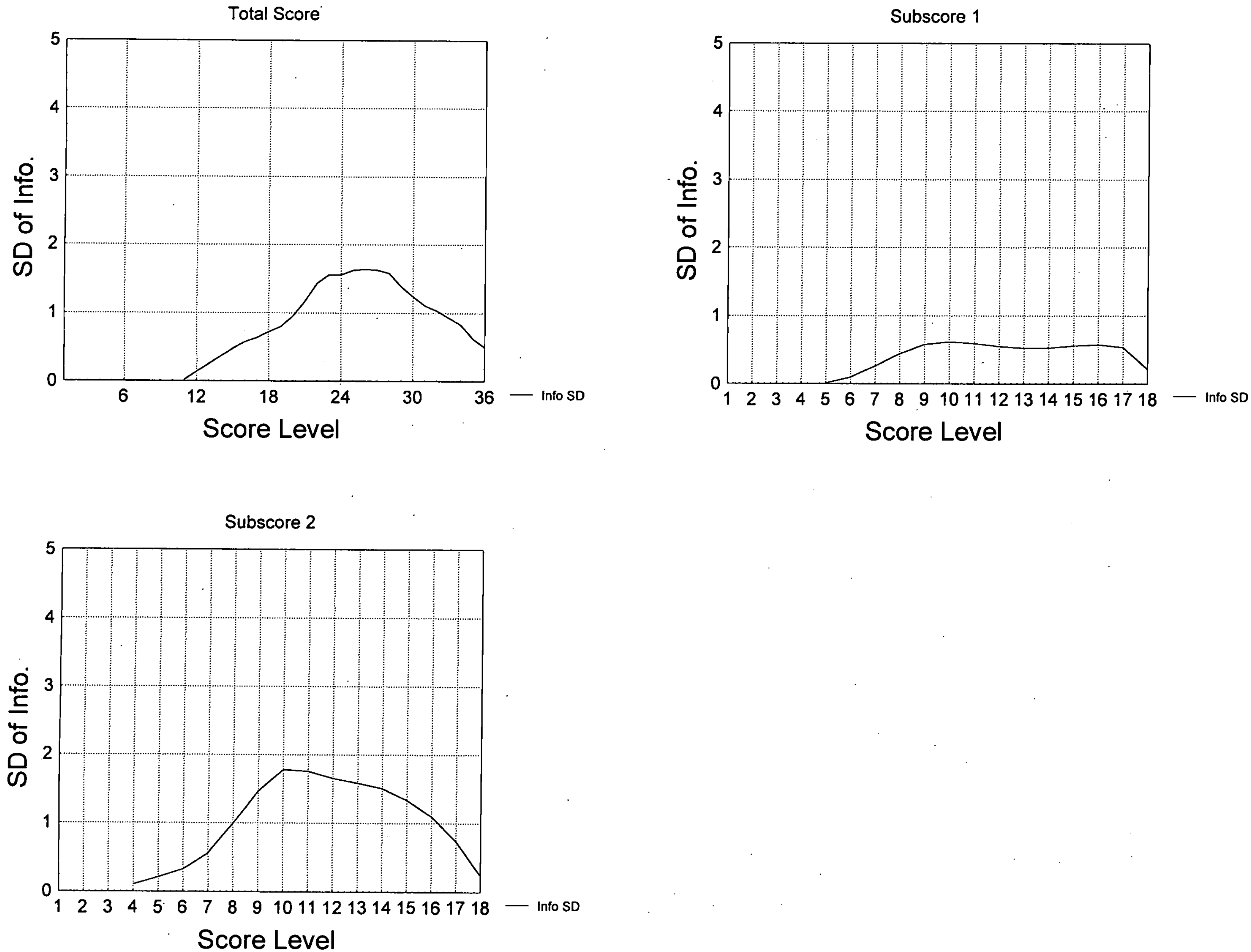


Figure 4: Standard Deviation of Unidimensional Approximate True Information



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM029997

REPRODUCTION RELEASE

(Specific Document)

NCME

I. DOCUMENT IDENTIFICATION:

Title: CAT Procedures for Passage-Based Tests[1]

Author(s): Tony D. Thompson, Tim Davey

Corporate Source:

Publication Date: NCME 1999 Presentations

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1	2A	2B
Level 1	Level 2A	Level 2B
[X]	[ ]	[ ]
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Tony D. Thompson

Printed Name/Position/Title: Tony D. Thompson/Psychometrician

Organization/Address: ACT, Inc.
P.O.Box 168, Iowa City, IA 52243

Telephone: 319/337-1213

FAX: 319/339-3021

E-Mail Address: ThompsoT@act.org

Date: 6/18/99

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com