New Methods for CBT Item Pool Evaluation

Lin Wang

ETS, Princeton, NJ

Paper presented at the annual meeting of the American Educational Research Association
(AERA) held between
April 12 to 16, 2004, in San Diego, CA.

Abstract

Evaluation of a CBT item pool's performance is critical to monitoring the quality of a computer based testing program. This report contains a new set of evaluation methods that can be applied to CBT operations. These methods evaluate (a) primary pool information, (b) item-related supporting information, and (c) examinee-related supporting information. Nine primary information methods evaluate the performance of CBT pools by monitoring several characteristics during a testing window. Five item-related methods evaluate item parameters, constraint violations, and item model fit statistics. The examinee-related category has six methods for person-fit evaluation and performance comparisons among subgroups of examinees. Three recommendations are offered for future work in order to fully realize the benefits of the new methods in this study.

Introduction

Rapid advance in both technology and psychometrics in recent years has made computer-based testing (CBT) a viable option for testing organizations (Way, 1997). A CBT program may also choose to use some adaptive methods, which leads to a computer adaptive testing (CAT). A CAT program potentially administers a unique test form to each examinee by selecting items that are most appropriate for the person's estimated ability while conforming to a set of constraints. On the other hand, a CBT program can use linear forms just as conventional paper-pencil testing programs. In this case, the same items are given in exactly the same order to all examinees. CBT is expected to demonstrate advantages over conventional paper-pencil tests in several areas that are crucial to the quality of testing. Such advantages include expanded possibility of measuring new constructs through new item types (e.g. multimedia components), more efficient measurement (in CAT), flexible schedules with more testing windows, and better control over testing conditions (Way, 1997).

The advantages offered by CBT are nevertheless accompanied by a variety of challenges. The challenges include creating and maintaining CBT item pools, designing and implementing appropriate psychometric models for item selections and ability estimation, securing active items against theft or overexposure, monitoring measurement scale stability over time and maintaining such stability through careful item calibration methods, etc. (Guo & Wang, 2003). In CBT operations, all individual tests are generated from one or more similar item pools, the quality of the tests is directly determined by the quality of the item pools. Much effort has been devoted to research on CAT pool development and maintenance (Kingsbury, 1997; Stocking & Lewis, 1995; Swanson & Stocking, 1993; Wang & Braswell, 2001; Way, Swanson, Steffen, & Stocking, 2001).

Relatively speaking, not much attention has been paid to the analysis of the performance of live CBT pools in the field. In CBT programs, newly developed item pools are evaluated through simulation to determine their fitness for operational service. A few pool analysis methods have been employed in some programs to evaluate item pools using observed data. These methods, however, present some problems. First, some current methods are borrowed from paper-pencil tests and are not appropriate for use in evaluating the quality of a computer-based test. Second, no current methods evaluate CBT-specific qualities such as pool security, regional performance, etc. Third, current evaluation relies almost solely on simulation data. Although evaluating pool quality with simulated data is useful, it is useful only to the extent that these simulations predict how the pool will perform when used operationally. Observational data is still needed to verify such predictions and to evaluate the actual performance of the pools.

The purpose of this study was to accomplish two important broad objectives: (a) to review current methods and to remove inappropriate ones and (b) to develop new evaluation methods. The approach of study was to conduct an in-depth review of the current methods to determine what methods should be discontinued and to develop new methods that may evaluate new aspects of pool performance. The outcome of the study is a new generation of methods for item pool evaluation. The presentation of the study is framed for CAT pools. However, CBT is used more often than CAT in this report because CBT covers CAT and many methods in this report apply to CBT in general.

Two objectives of the study

Objective One: Review current item pool analysis methods

The current pool evaluation methods were reviewed using two pool analysis reports from a large scale testing program. These two pool evaluation reports describe twelve methods that are used to evaluate the performance of a number of operational pools. A brief description of the twelve methods is summarized in Table 1.

Reviews of the current methods focused on three aspects: (1) functionality, (2) CBT-relevance, and (3) comparability.

1. Functionality. The functionality of a method refers to what information is obtained by applying this method in pool evaluation. For example, Method 1 in Table 1 obtains between-measure correlation coefficients from operational data, whereas Method 5 provides psychometric properties of the pool based on simulations. Reviewing the functionality of current methods was a very important step of this study in that the review helped to identify what item pool information has been collected in the past. Based on the results of these reviews, decisions can

be made on retaining, modifying, or eliminating existing methods, and on adding new methods.

Table 1 Current CAT pool evaluation methods

Method	**Description of Method**
1	Correlations between measures.
2	Pool composition characteristics: item parameters and item distribution across content areas and item types
3	Selected demographics of the examinees
4	Distribution of reported scores
5	Conditional standard errors of measurement and pool reliability estimate
6	Distribution of exposure rates from simulation and observed data.
7	Model-data fit (difference between an item's expected proportion correct and observed proportion correct)
8	Constraint summary of simulation and observed data
9	Distribution of reported scores for gender and ethnic/racial subgroups.
10	Distribution of section times (in minutes) for gender and ethnic/racial subgroups.
11	Distribution of number of items answered for gender and ethnic/racial subgroups.
12	Distribution of number of items answered for score level subgroups.

2. CBT-relevance. Relevance pertains to whether a method is appropriate for evaluating CBT pools and the performance of the tests that are constructed from the pools. Many methods that have been established to evaluate the quality of educational tests are originally designed for paper-pencil tests. These methods may not be appropriate for judging the quality of a CBT, especially a CAT program. Using a method that is developed for paper-pencil tests to evaluate the quality of a CAT can be misleading. This is simply because a paper-pencil test is typically assembled in a form that contains a fixed set of items. Anyone who takes this form sees the same set of items. Therefore, such criteria as score reliability and standard error of measurement can be applied to all scores of the test.

In a CAT, however, each examinee may see a different 'form' of the test that is comprised of a unique set of items from a CAT operational pool. This unique set of the selected items is related to an examinee's estimated ability level. Therefore, it is not appropriate to use the conventional reliability or standard error of measurement as the criteria of test quality evaluation. Instead, conditional standard error of measurement is used to describe amount of measurement error at a particular ability level, not across all the levels. The purpose of reviewing the CBT-relevance of current methods was to identify methods that were developed for paper-pencil tests and would not be appropriate for CBT applications.

3. Comparability. This pertains to whether evaluation of a current CBT pool can be compared to data from other pools that can be used as a baseline for evaluation. The importance of this aspect of the current reviews cannot be over-emphasized. CBT applications are relatively new in educational testing, and this is particularly true of operational CBT programs.

In paper-pencil tests, the quality of tests across forms and time can be maintained and monitored through equating. In CBT, however, there is little to rely on to keep track of the quality of the operational pools. Simulations are routinely used as a means of evaluating pool quality. There are, however, limitations in using simulations only for the purpose of tracking pool quality. One limitation is that such routine simulations assume normal testing situations in terms of test administration and examinee behavior. Another limitation is that the number of simulated data points is usually smaller than the actual number of examinees that take CBTs from a pool. Consequently, there can be discrepancies between simulated results and observed results.

A practical way of monitoring the quality of operational CBTs across pools and time is perhaps accumulating historical CBT data and using the data as a baseline. Subsequently, the operational data from a recently administered CBT pool can be compared to the historical or baseline data for evaluating consistency or identifying deviation from baseline patterns. It is certainly helpful if such information is compiled on some key methods and is available for consultation. This consideration was kept in mind in reviewing and adding methods.

The current pool evaluation methods in Table 1 were reviewed and evaluated with regard to these three aspects, and decisions were made on whether to retain, eliminate, or modify the methods. The results of the evaluation are presented in Table 2.

Table 2. ***Summary of the current method reviews***

Method	Functionality	CBT relevance	Comparability	Decision
1	Between-measure correlation	Yes	No	Modify
2	Pool composition	Yes	No	Retain
3	Sample demographics	Yes	No	Retain
4	Score distribution	Yes	No	Modify
5	Pool quality	Yes	No	Retain
6	Exposure rates	Yes	No	Modify
7	Model-data fit	Yes	No	Modify
8	Constraint summary	Yes	No	Retain
9	Scores of subgroups	Yes	No	Retain
10	Completion time for subgroups	Yes	No	Retain
11	Completion rates for subgroups	Yes	No	Modify
12	Completion rates for ability levels	Yes	No	Retain

The results of reviewing the current methods are summarized as follows:

1. The current methods are deemed adequate in providing relevant information about the performance of CBT pools. The information that the methods yield covers a variety of aspects of the current operational CBT pools and the tests that are generated from the pools. Some of the information deals with CBT operations

exclusively, such as exposure rates, conditional standard errors of measurement for different ability levels, etc. Other information can be seen in both CBT and paper-pencil applications, such as demographic information, descriptions of item characteristics, completion rates, etc.

2. One important feature that is absent in the current methods is some type of baseline information against which a new pool can be compared. As was mentioned earlier, such comparative information helps to monitor the quality of new pools for abnormal behaviors. Although simulations are routinely employed in CBT applications to set up pool configurations and to investigate possible outcomes under certain testing circumstances, some criteria are still needed for evaluating actual operational data. Also, because CBT pools are built and put to use in the field on a continuous basis, it is important to ensure that not only the pools are 'parallel' or comparable to one another, but also, and more importantly, the tests that are generated from the pools are comparable in terms of both content and statistical characteristics. To achieve this goal, it is necessary to accumulate information from previous CBT administrations and use such information as a baseline for future comparisons.
3. Current methods provide no information on several aspects that are important to the quality of CBT operations. For example, although the item selection algorithm ensures that each examinee sees a set of items that are comparable in content coverage and statistical properties, this does not necessarily guarantee that the items of the same content categories will contribute the same amount of information to the tests. This is simply because an item contributes different amount of information to a test at different ability levels.

   Another example is that the current methods do not have information about speededness of all CBTs that have been administered. Method 11 in Table 1 and Table 2 reports percent of examinees that completed number of test items. It is, however, not known how many of those completed items were in fact completed in a hurry or by random guessing as time was running out. It is possible that such speededness may vary from pool to pool. Pool speededness affects the quality of the pool, and this information must be available for evaluating pools. Additional new methods are, therefore, needed to enhance the quality and usefulness of CBT pool evaluation in operations.

4. The current format of presenting findings from pool evaluation can be improved upon. The current methods are applied to each measure of each CBT pool in a package. The results are all presented in tables that are cluttered with numbers. This makes it difficult to read, or to retrieve important information from all the numbers. In fact, not all tables or numbers are equally important or interesting. Some (e.g. exposure rates) are more important than others (e.g. subgroup score distributions). Therefore, new formats are needed to organize results and to present findings

Objective 2: Develop a new generation of pool evaluation methods

In order to have a dynamic, timely, and straightforward view of the CBT pool performance when operational data becomes available, it is necessary to develop a new generation of evaluation methods that incorporate both current methods, modified where necessary, and new methods. The new set of the methods is compiled and developed to fulfill these expectations:

1. The new methods will provide a full range of information that depicts performance of both items and examinees that see the items. ‘Full range’ means that it will not only contain the information about item content, item exposure, score distribution, etc., but also information about when a pool is released to and withdrawn from the field, what score or item distributions look like on different testing dates, and what is the quality of all CBTs with regard to individual’s test information.
2. The new methods will establish a baseline for evaluating each individual pool. The baseline will be from simulations and historical data. A baseline provides a meaningful reference point for evaluating data from a newly administered pool, and will help to keep track of score trends or to spot abnormal phenomena across pools and administrations.
3. The new methods will have sufficient flexibility for customized analyses in addition to routine analyses. Specifically, the new methods will be capable of not

only generating standard or routine results of pool evaluation as part of CBT operations, but also allowing for special analyses that are needed at times.

4. The new methods will be organized according to the importance of the information that they provide. As was mentioned earlier, although one may want to evaluate a CBT pool as thoroughly as possible by looking at a variety of the pool performance, not all information is equally important. Therefore, it is necessary to arrange the findings into primary and supporting categories. It is also a good idea to use charts rather than tables for primary information because a chart communicates information in a straightforward and focused manner.

A summary of the new methods

With the four expectations in perspective, a new generation of pool evaluation methods has been developed by modifying the current methods and by designing additional methods. The new methods are arranged in three parts (Part I, Part II, and Part III) according to their perceived importance and criticality and whether a method is in nature related to item or examinee properties. Part I contains ten methods for primary information for monitoring pool performance in operations. Part II includes five methods for supporting information related to items in a pool. Part III collects six methods for examinee-related supporting information.

These new methods are outlined in Table 3. In the table, the column for 'Functionality' describes the type of analysis that a method requires and variables of interest. The column for 'Focus' indicates whether the focus of a method is on the performance of items or examinees. The last two columns identify whether a method is classified as 'Primary' or 'Supporting' in terms of importance of the information, and how the results will be presented.

Part I. Evaluation methods for primary information

The methods in this part are designed to evaluate the performance of pools, items and examinees. The methods will be used both during a live CBT administration period and after the administration period is concluded. These methods are expected to provide informative and timely snapshots of the performance of pools, items and examinees. These snapshots can be used for making necessary adjustments, investigation, and other decisions about the CBT operations. Each method in this part will produce a chart to present related performance information. The charts in this report are for illustration purposes and are constructed with mock data.

Table 3. ***New methods for CBT pool evaluation***

Method	Functionality	Focus	Category	Format
Part I. Methods for primary information				
1.1	Planned and actual pool release dates	Pool	Primary	Chart
1.2	Item completion rates and pool speededness	Item	Primary	Chart
1.3	Score distribution comparison	Examinee	Primary	Chart
1.4	Total information by content and ability	Item	Primary	Chart
1.5	Distribution of observed exposure rates	Item	Primary	Chart
1.6	Total information by estimated abilities.	Examinee	Primary	Chart
1.7	Item information at all item positions	Item	Primary	Chart
1.8	Item latency for administration periods	Item	Primary	Chart
1.9	Subscores by days of administration	Examinee	Primary	Chart
1.10	Average regional scores for administrations	Examinee	Primary	Chart
Part II. Methods for item-related supporting information				
2.1	Content and psychometric composition	Item	Supporting	Table
2.2	Psychometric characteristics	Item	Supporting	Table
2.3	Constraint summary	Item	Supporting	Table
2.4	Correlations between measures	Item	Supporting	Table
2.5	Model-data fit: the Z_c flags	Item	Supporting	Table
Part III. Methods for examinee-related supporting information				
3.1	Person-fit: the L_z statistic	Examinee	Supporting	Table

3.2	Item completion by gender and ethnic/raciality	Examinee	Supporting	Table
3.3	Summary of selected demographic variables	Examinee	Supporting	Table
3.4	Reported scores by gender and ethnic/raciality	Examinee	Supporting	Table
3.5	Test completion by gender and ethnic/raciality	Examinee	Supporting	Table
3.6	Scores for Pool XXXX and all past pools	Examinee	Supporting	Table

Method 1.1. Planned and actual days that pools are available in the field

The purpose of this method is to identify the planned or scheduled pool release dates and the actual dates that a pool is released in the field. Past experience tells us that sometimes there can be a discrepancy between the two types of dates. Typically, a pool is released before its scheduled release date. It is necessary to keep track of such information for both operational and analytical needs.

In the sample output chart (Figure 1.1), 'X' marks the planned or scheduled release dates for a pool. The numbers next to the Xs are the numbers of examinees taking the test. On a date where there is no 'X' but a number is found, this means the pool was released on a date it was not supposed to be released.

	July 2003											Monthly
	1	2	3	…	15	16	17	…	29	30	31	total
Pool 1	**X**	**X**	**X**	**X**	**X**							
	360	410	56	…	123							
Pool 2						**X**	**X**	**X**	**X**	**X**	**X**	
			23		15	230	250	310	200	98	87	
Pool 3				**X**	**X**	**X**	**X**	**X**	**X**			
		6	12	115	210	142	100	100	110	3	7	

Figure 1.1. **Planned and actual days that the listed pools are used in the field**

Method 1.2. Pool speededness: Rates of completed items and rushed items

This method compiles data on two things. One is the item completion rate, which means what percent of examinees responded to *K* or fewer items of their CBTs. Cumulative frequency and percentage is used here. The other part of the method looks at the speededness of the pool. When looking at how many or what percent of examinees responded to an item, one does not know whether the examinees' responses were generated in a normal manner (meaning after working out a solution) or a random guess (or even a random click on a choice) due to running out of time. When the item latency for an item is less than 10 seconds (or some other criteria), the response to this item can be operationally defined as a 'rushed' response. The percent of rushed items are an indicator of the speededness of a pool.

Figure 1.2 gives an example of a test that has 35 items in a test. The numbers on the horizontal axis refer to items. There are two columns for each item. One column (shaded) is for the percent of total examinees who completed the number of items as is marked on the axis. The other column is for the percent of the examinees who responded to the item in less than 10 seconds. For example, at '33' on the axis, the shaded column shows that 66% examinees responded to up to 33 items in the test. The other column indicates that 27% of the examinees 'rushed' their responses to their 33rd item in less than 10 seconds.

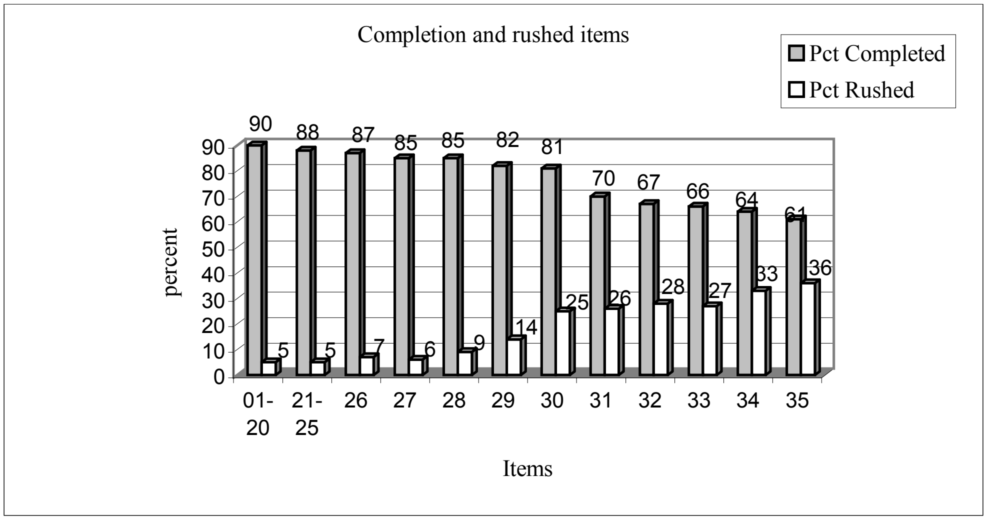

Figure 1.2. **Pool 1 speedeness: Rates of completed items and rushed items**.

Method 1.3. Distributions of reported scores: Current pool vs. the baseline

This method is designed to compare the distribution of reported scores of the examinees that see the current pool to the distribution of the reported scores in the same month of the past years and to the distribution of the reported scores in the past years as a whole. Comparison to historical data tells whether the current pool yields comparable score distributions. This method can be readily adapted for scores of subgroups.

The rationale for this method is that it is necessary and very important to monitor score distributions for each administered pool and to spot unusual patterns in a timely manner. The best criterion against which a pool can be evaluated is the cumulative observed data from past pools. This cumulative data serves as a baseline once it is established. There is no such baseline information for current CBT data. Therefore, all available CBT data should be used to create such an initial baseline, which will then be updated with the aggregate data from each pool thereafter.

Two types of baseline data are considered in this context. One is the total data baseline, or data of all the past pools. The other is the monthly data baseline, the data from the same months in the past years. For example, the baseline for the monthly data may include data from all the pools that were used in the month of July in the past years. Inclusion of the scores in the same month in the past allows one to evaluate possible seasonal effects.

The sample chart is in Figure 1.3. In the title of the chart, ‘Current’ can be replaced with the actual pool number. In the legend box, ‘July 02’ is for the current pool, ‘July 97-01’ represents all the pools that were used for the month of July in the past, and ‘All’ includes all the past data that are available.

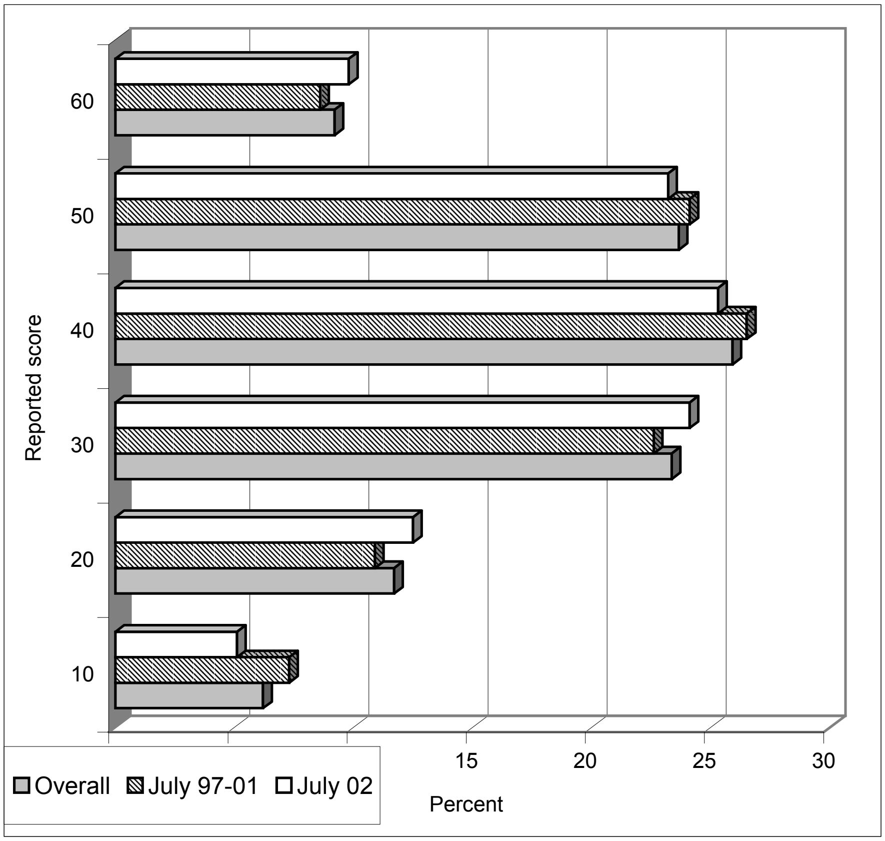

Figure 1.3. **Pool 1 reported scores: Current vs. baseline**

Method 1.4. Distribution of information contributed by content and ability groups

In the current CBT operations, content constraints define the number and characteristics of the items to be selected for a particular adaptive test that is administered to an examinee. The scoring weights of all items in a test are not the same. Since each examinee can potentially see a unique test, it is possible that the content contribution to the final scores can be different. This means that examinees may not have tests of the same content composition even if items of similar content categories are administered (Payton & Golub-Smith, 2001).

The influence that an item exerts on a final test score can be evaluated using a statistic called Fisher's information, or information for short. Method 1.4 describes the distribution of such information that is contributed by items of a content category to the total information of the items that are administered to examinees in different ability groups (marked by score ranges). The results may indicate how much information items of a content category contributes to the total information at a given ability level or range. Such results help in the understanding of the actual composition of the contents in the CBTs when examinees' final ability estimates are determined.

Specifically, item information is calculated at each examinee's final ability (θ) estimate using

$$I_i(\theta) = \frac{[P_i^{'}(\theta)]^2}{P_i(\theta)Q_i(\theta)} = \frac{D^2 a_i^2 (1-c_i)}{[c_i + EXP(Da_i(\theta - b_i))][1 + EXP(-Da_i(\theta - b_i))]^2} \qquad \text{(Formula 1.5.1)}$$

where *a*, *b*, *c* are item parameters (Birnbaum, 1968, chapter 17). Examinees are put into a number of ability groups according to their reported scores. Within each ability group, the content categories of all the items are identified and the item information is computed for all the items and summed. The percentage of total information that is contributed by items of each content category is computed and presented in the chart in Figure 1.4. The ability groups and content categories are for illustration purposes only and will be defined by each testing program.

Method 1.5. Distribution of observed exposure rates for the current pool and all pools.

This method is to monitor unusual or unexpected changes in observed exposure rates from field data. Although item exposure rates are established through runs of simulation before a pool is packaged for field use, it is still necessary to evaluate the actual exposure rates for the items of a particular pool, and to find out if the observed exposure rates are very similar to some baseline data. Therefore, Method 1.5 compares the exposure rates of the items in the current pool with the exposure rates of the items of all the past pools. Again, the data from the past pools serve as the baseline data or criterion of comparison. Figure 1.5 shows what the displayed results may look like and the results should facilitate spotting unusual changes that might occur in the exposure rates of the items in the current pool.

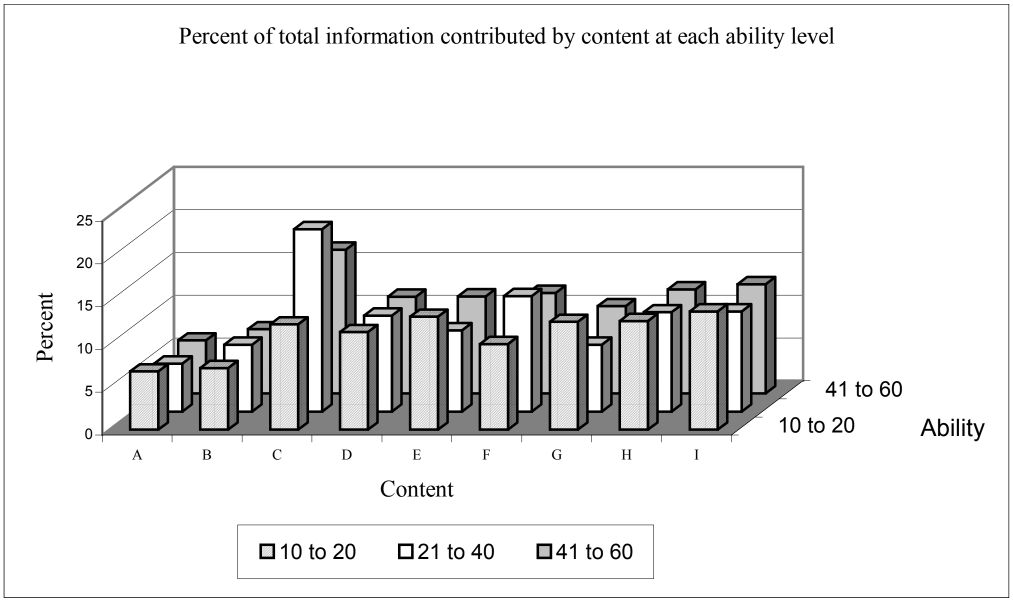

Figure 1.4. **Percent of total information by content at each ability level**

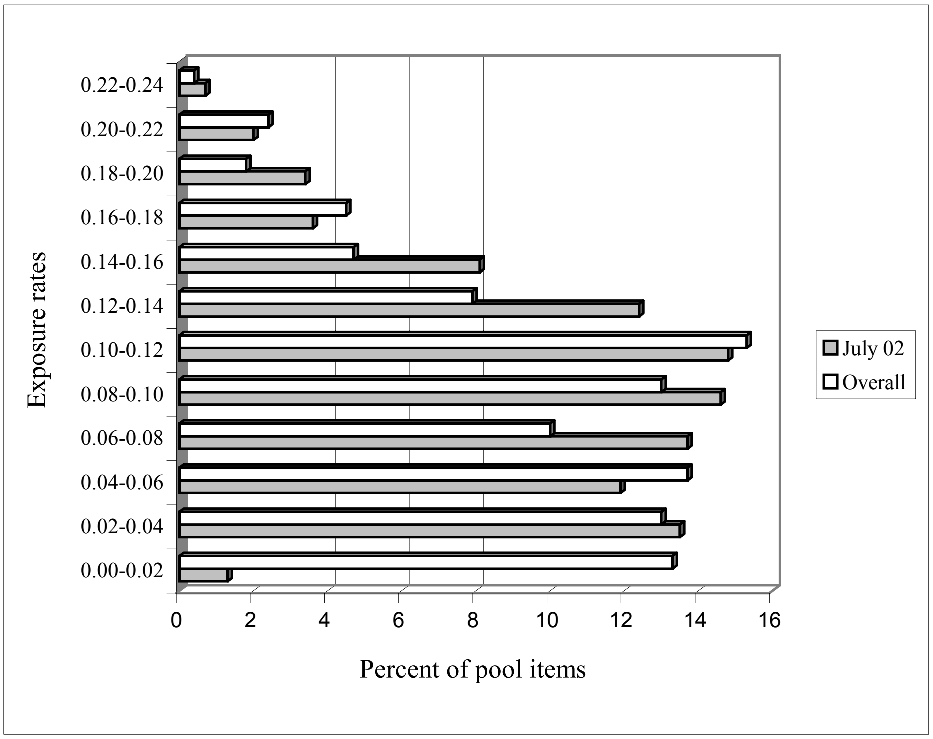

Figure 1.5. **Distribution of observed exposure rates for the current pool and all pools**

Method 1.6. Distribution of the total information for the current pool and all pools

Unlike the conventional paper-and-pencil linear tests where examinees see the same items in a particular test form, examinees in a CAT administration see different items that are configured uniquely in real time for individual examinees. The statistical quality of a CAT administration may vary from person to person. In CAT, this quality is measured by measurement precision that can be expressed in the form of test information within the IRT framework. For each examinee who takes *n* items in a CAT session, the test information is calculated as:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \qquad \text{(Formula 1.6.1)}$$

where $I_i(\theta)$ is given in Method 1.4. Higher amount of test information means better measurement precision or less measurement error. As the calculation suggests, the value of test information depends on both θ, an examinee's final ability estimate, and the parameters of the items that are administered to the examinee.

One of the promoted advantages of CAT is better measurement precision across ability levels because a CAT administration is tailored to each examinee by targeting his/her estimated ability. This is different from a paper-and-pencil test, which targets only the middle section of ability distribution where most examinees are located. In reality, however, not everyone gets the same or similar measurement precision in CAT because there may not be adequate items for all ability levels. Therefore, it is both necessary and important to evaluate pool performance in operational settings by monitoring observed pool information distributions across administrations and time.

Method 1.6 was developed to evaluate whether the distribution of the current pool's information conforms to the distribution of the past pools. This distribution is constructed at discrete points on the ability or θ scale by grouping examinees whose final ability estimates are adjacent to a chosen θ point. At each chosen θ point, the mean test information over the examinees *N'* is calculated as:

$$\bar{I}_{N'}(\theta) = \frac{I(\theta)}{N'} \qquad \text{(Formula 1.6.2)}$$

where $I(\theta)$ is already defined in Formula 1.6.1.

In addition to the mean test information at θ, the minimum and maximum information values are also used to indicate the degrees of variation in the test information for the examinees of this ability level. The mean, minimum, and maximum information values of the current pool are compared with the mean, minimum, and maximum information values of (a) past pools for the same month as the current pool, and (b) all the pools in the past. Figure 1.6 illustrate what the comparisons may look like by displaying the mean, minimum and maximum test information values for the current pool and the past pools together.

All the information values of the current pool are expected to be similar to the values of the past pools to show consistency. If the current pool displays unusual departure from the past pools, investigation of the performance of the current pool is necessary to find out what may have contributed to such a departure.

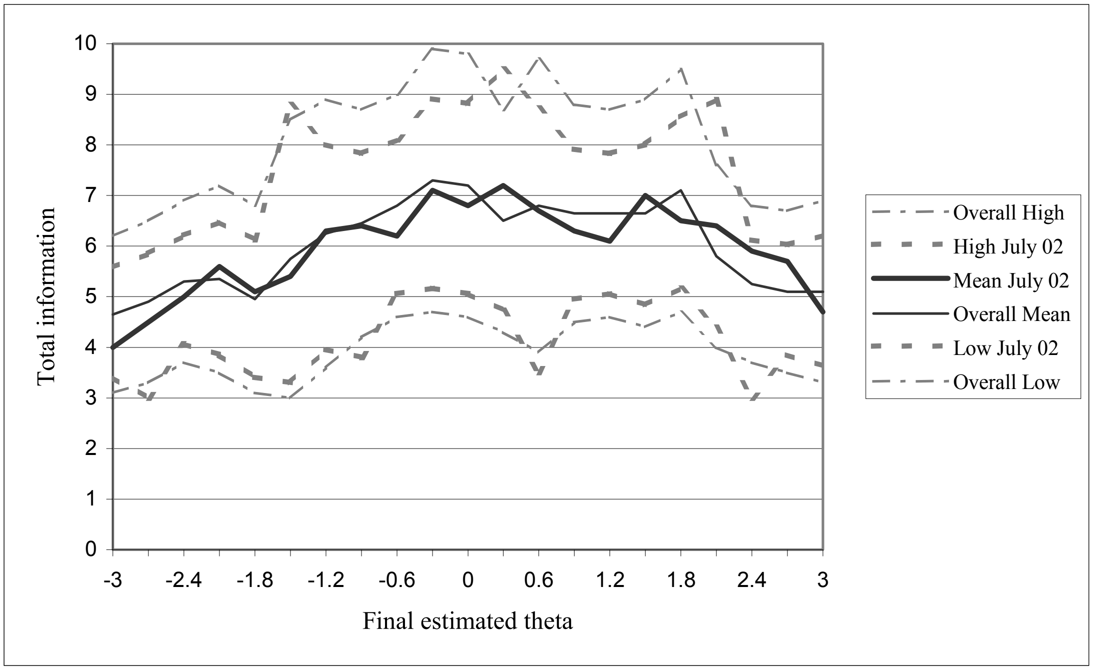

***Figure 1.6*. Distribution of the total information for the current and all pools**

Method 1.7. Information contributed by items by administration positions

In CAT, examinees' final scores are derived from maximum likelihood estimations of theta (ability) using pre-calibrated item parameters and examinees' responses to these items (Wainer & Mislevy, 2000). Because of the adaptive process, a final ability estimate depends on not only the items that are administered, but also indirectly on the sequence in which the items are administered. Some items count more than others towards the final scores do. Responses to the items that are administered earlier influence the selection of the items to be administered next. For example, a wrong response possibly leads to the selection of a less difficult item, and vice versa, assuming other item selection criteria (content constraints and exposure control parameters) being held constant. Hence, the administration positions of items in a CAT test can influence the final score that an examinee receives (Wang & Gawlick, 2001).

Because the current CATs have fixed length, all examinees are given the same number of items. The items that are administered at each position (the 1st item, 2nd item, … $(n-1)^{\text{th}}$ item, and n the last item) can be different from person to person, with possible overlap, of course. For each administration position, the item information can be aggregated over all examinees. A comparison of this aggregated item information at all administration positions may reveal at what position(s) items are administered that yield more information. Items that yield more information are the items that have more discriminating power. Therefore, such a comparison also point out where more discriminating items are administered as a whole.

This method examines the distribution of the item information (see Formula 1.5.1) by item administration position (or item order) across all the examinees for a CAT pool. The purpose is to find out whether the more discriminating items are administered in a consistent manner across all the CBTs from a pool, and how the distribution from the current pool compares to the overall distribution from the past pools. A sample chart is given in Figure 7 for illustration of this method. The aggregated item information at each administration position is calculated as:

$$I_K = \frac{\sum_{j=1}^{N} I_{jK}}{\sum_{K=1}^{n}\sum_{j=1}^{N} I_{jK}} \qquad \text{(Formula 1.7.1)}$$

where j = 1, 2, … N for examinees, K = 1, 2, … n for the n_{th} item that is administered. In effect, I_K is the proportion of the grand total information that is contributed by the items at the n_{th} position.

Theoretically, given the maximum information selection method used in current CATs, test developers would like to use less discriminating items in the earlier part of a CAT where little is known about an examinee's ability. As a CAT proceeds to the end, more becomes known about an examinee's ability and more discriminating items that are available for this ability level can be selected for administration. This scenario implies that one would expect to see a gradual ascend of the aggregated information to the end of a CAT. Namely, one would expect to see the columns in Figure 1.7 to go higher and higher for the later positions. In the real CAT administrations, because of the requirements of content balance and exposure control, it is unlikely to realize the theoretically expected pattern. Instead, what is in Figure 1.7 is probably what is found most of the time. Consequently, consistency becomes more important and interesting in evaluating pool performance in this aspect. Thus there is the need to compare the current pool data with historical data of the past pools.

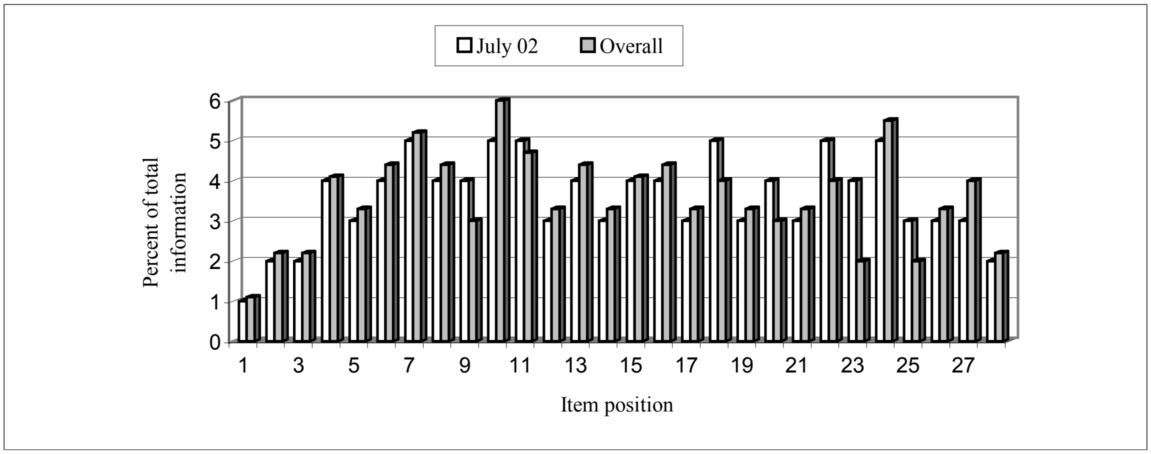

***Figure 1.7*. Information contributed by items at various positions**

Method 1.8. Distribution of item latency over administration periods

Because of the continuous administration mode of CBT, item and test security is always a great concern. Unauthorized disclosure of active operational items by individuals may provide enough information for other examinees to enhance their test performance unfairly. It is absolutely necessary to watch out for such possible cheating practices in CBT administrations. Monitoring pool security may be conducted at both the individual and aggregate levels.

This method is designed to monitor pool security at an aggregate level by evaluating item latency patterns over a number of administration periods. Item latency is the amount of time that an examinee spends on a particular item. This information is part of this examinee's CBT record that is automatically generated in a CBT session. This method of item latency is based on an assumption that the average time to complete an item is monotonically increasing with an item's difficulty level. In other words, it is expected that examinees on average spend longer time on difficult items than on easy items over the entire administration period under normal circumstances.

With this method, average item latencies from items that are answered correctly in each of the administration periods (three by current design and can be changed if necessary) are calculated and evaluated with regard to the difficulty levels of the items. The latency distributions over the administration periods are expected to follow the same pattern. A systematic deviation of the item latency pattern of a later administration period from the earlier periods might be an alert to possible problems in pool security. An illustration is given in Figure 1.8 with heuristic data.

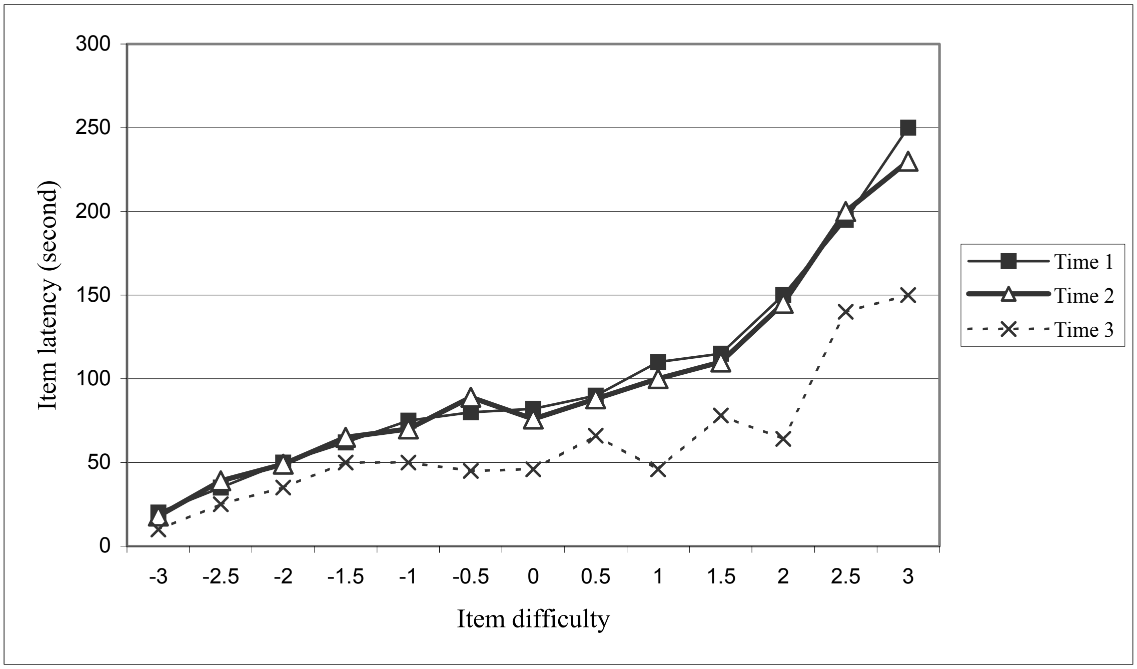

Figure 1.8. **Distribution of item latency over administration periods**

Method 1.9. Comparing reading and non-reading subscore distributions

A typical case of unauthorized disclosing of test items by individuals is that individuals memorize some items they see during their test sessions and share with others what they can recall about the items. Naturally, the set or passage-based items, such as those in reading tests, are much harder to recall and are less likely to be disclosed in meaningful ways. In contrast, some non-reading, short discrete items types are prone to memorization.

Because a CBT pool contains a limited number of items and a testing occurs for several days, the probability increases that some people may have seen certain items before they take the tests. If this happens to a CBT pool in a sizable scale for a particular administration period (this would typically occur at the later segment of the period), one may expect to see some systematic divergence of the scores (subscores) between reading items and non-reading items for this segment of the administration period.

The purpose of this method is to monitor possible compromised items in a pool by comparing the distribution of aggregated performance on reading and non-reading items. Figure 1.9 portrays a possible scenario for illustration. In this example, the reading scores are used as a baseline for the entire administration period when a CBT pool is active in the field. The non-reading scores are the focus of interest. If the non-reading scores show a systematic upward trend away from the baseline of the reading scores for the later segment of the administration period, this can be an indication that some non-reading items may have been compromised.

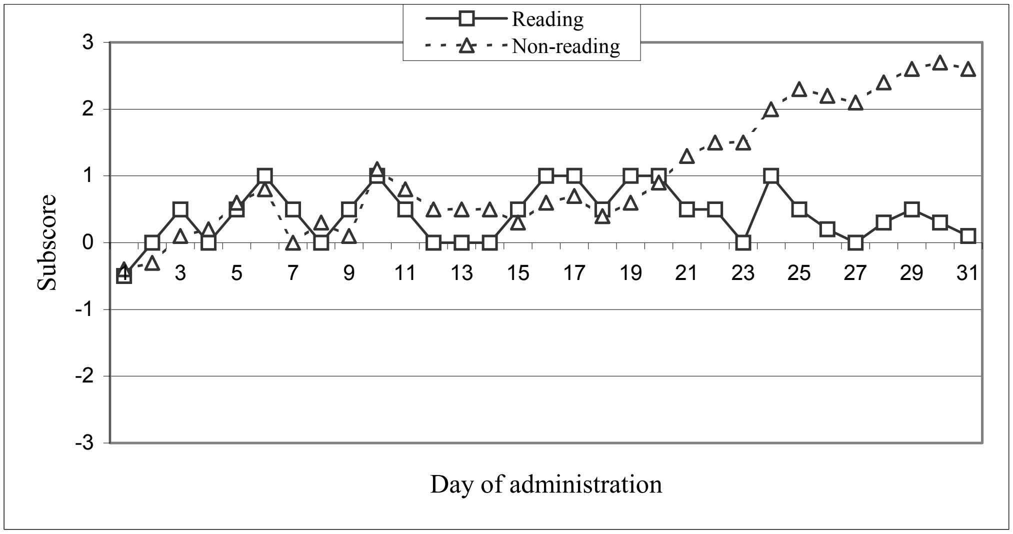

Figure 1.9. **Distribution of reading and non-reading subscores* by days**

* The subscores are averaged theta values for the items that are in reading or non-reading categories.

Method 1.10. Comparing mean scores by regions and by time

One concern of testing programs is unauthorized disclosure of active items through internet by individuals in certain geographic regions. It can be reasoned that if a considerable number of people have the opportunity to see some active items before they take their CBTs in the same administration period, these examinees' scores are likely to be unduly inflated, and the average scores of this group of examinees would be higher than the scores of those who take the tests earlier. A simple and quick evaluation method is needed to identify such a situation in operations so as to alert us to possible security breach incidents.

This method is designed to provide such information to monitor the CBT operations worldwide. Specifically, the method compares the average scores of the examinees who are grouped by their geographic regions (or even test centers) and when they take the tests (say, during the first ten days, or the last ten days, etc.). Comparison of the group mean scores is between time points for the same administration month (or a period), or is for the same administration month against this month in the past and against the overall means (all administrations).

Figure 1.10 contains two charts that illustrate what one might see as the results of the two types of comparisons. The top chart shows mean scores for each of three administration periods at regions of A to F. The average performance at Region B and Region E appears to be related to the time and gives an indication of abnormal patterns worth investigation. The lower chart, on the other hand, compares the average scores of the people who took the tests in July 02 to the scores of the July in the past several years as well as the mean scores of all the administrations. Again, Region B shows an unusual sign of better performance this time than before. With the information from the two charts, one certainly wants to look further into the scores in this region to find out what is happening there and take necessary actions if warranted.

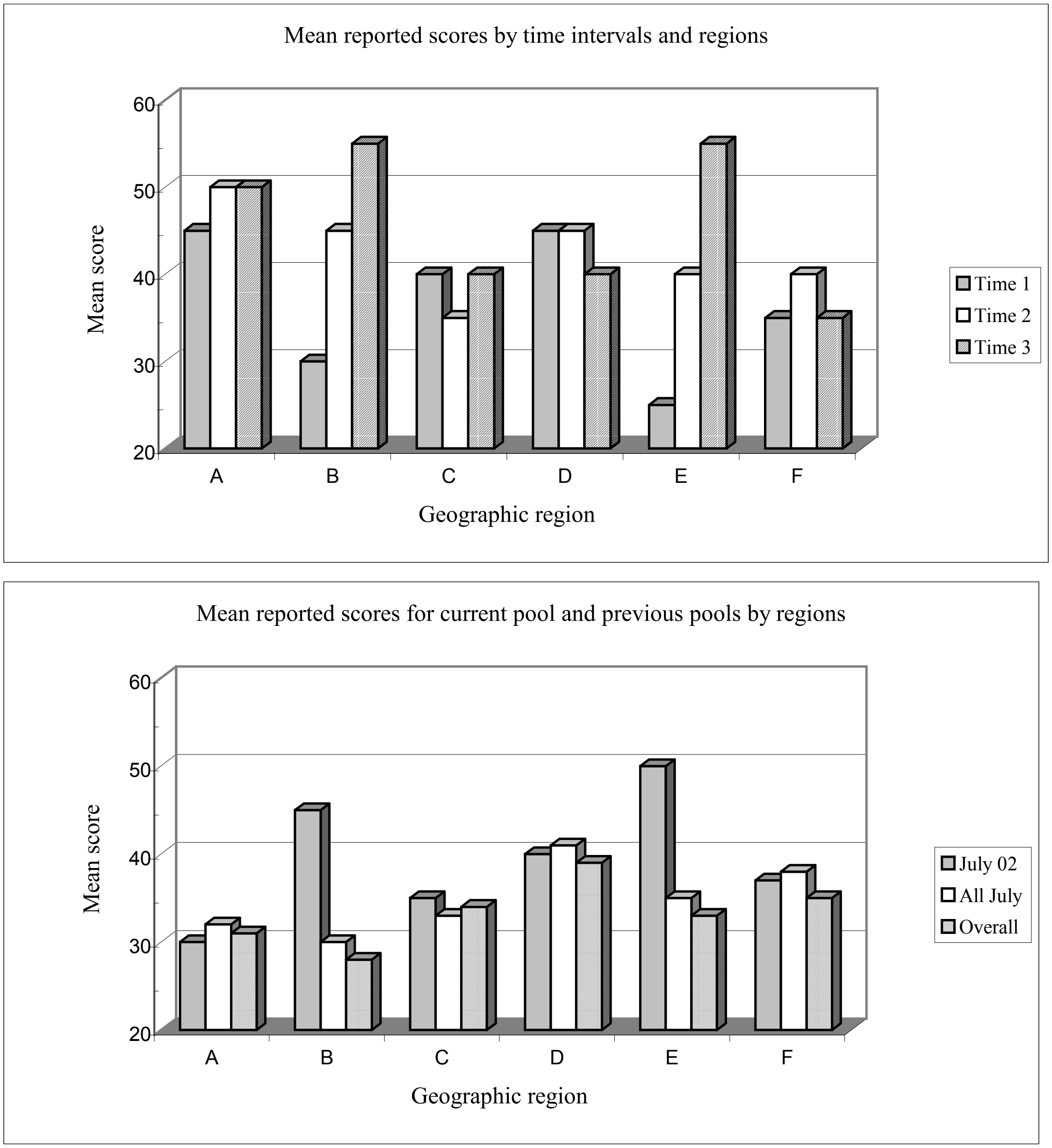

Figure 1.10. **Mean scores by regions in three time intervals**

Part 2. Evaluation methods for item-related information

Method 2.1. Content and psychometric composition

In current CBT operations, an operational pool is created from a large collection of items. Although each pool is created according to the same set of specifications, every pool contains unique items. The content and psychometric characteristics of the pools may also vary within allowed limits. Therefore, it is necessary to keep track of all pools for their item composition.

This method is an existing one. The method summarizes a pool's item composition in terms of major content categories, item types and calculates descriptive statistics for the IRT item parameters (*a*, *b*, *c*) by content categories. Table 2.1 in the appendix illustrates the cross-tabulated results.

Table 2.1 ***Content and Psychometric Composition***

		Major Content Classification						Total
		I	II	III	IV	V	VI	
Number of								
Discrete Items		32	43					75
Set Members				93				…
Stimuli				14				…
Total Items		32	43	93				…
Total Elements		32	43	107	…	…	…	…
a - Parm	Mean	0.83						
	Std	0.21						
	Min	0.47						
	Max	1.58						
b - Parm	Mean							
	Std							
	Min							
	Max							
c - Parm	Mean							
	Std							
	Min							
	Max							

Method 2.2. Psychometric characteristics of Pool 1

This method summarizes conditional statistics of a pool and a reliability estimate. The statistics are all based on simulation results because only in simulations can researchers define what ‘true’ abilities are and then use these ‘true’ abilities with observed ability estimates to calculate conditional statistics.

Conditional statistics presented by this method include mean scores, scores of the 25th and the 75th percentile of the group seeing this pool, and the conditional standard error of measurement (CSEM). These statistics are given for each score point that corresponds to a defined ‘true’ ability level in the simulation that generates the conditional statistics. An example is given in Table 2.2.

Table 2.2 ***Psychometric characteristics of Pool XXX***

Scale Score	Conditional Statistics			
	Mean	P25	P75	SEM
20	…	…	…	…
25	23	19	26	6.4
30	…	…	…	…
35				
40				
45				
50				
55				
60				

The reliability estimate for the pool is calculated using an internal consistency formula:

$$r_{xx} = 1 - \frac{V_E}{V_X} = 1 - \frac{\sum [p_j \times CSEM_j^2]}{VAR(SS_{est})} \qquad \text{(Formula 2.2.1)}$$

where p_j is the proportion of simulees expected to have a particular scale score SS_j that is converted from a defined ‘true’ ability level, and VAR(SS_j) is the variance of the distribution of estimated scaled scores weighted for a year’s test-taking population.

Method 2.3. Constraint summary

In CBT operations, each pool is created to satisfy numerous constraints that control the content coverage for each adaptive test the pool produces. The high complexity of item selection processes makes allowances for a very limited number of constraint violations in some cases where such violations are deemed not a risk to the quality of an adaptive test in terms of content coverage. It is, however, very important to monitor the frequency of occurrences in constraint violations. This monitoring process is conducted at pool simulation time and on observed data after administration. The results for each pool are kept on record.

This method summarizes several key statistics of the constraints for a pool using both simulation and observed data. Table 2.3 describes the types of constraints that are included. Each constraint is defined by a weight, a lower bound and an upper bound for the number of items of an item type to be included in an adaptive test. To see whether a constraint incurs any violations, the mean number of administrations of an item is calculated in the simulation process and for the observed data. Similarly, the proportions of simulees or real examinees whose adaptive tests involve violations of a constraint are calculated and presented in Table 2.3. If the discrepancy between the simulation and observed data is sizable, an investigation is needed to determine the possible cause for the discrepancy.

Table 2.3 ***Constraint Summary***

Constraint Label				Mean Number Administered		Mean Number Administered	
	Weight	Lower bound	Upper bound	Simulated	Observed	Simulated	Observed
A	25.0	01	01	1.00	1.00	0.0000	0.0001
B							
…							
…							
F	20.0	00	03	2.77	2.79	0.0000	0.0000

Method 2.4. Observed correlations between tests

In some CBT operations, multiple pools are created at the same time and sent together to the field. The pools are used according to some pre-defined rotation rules. Between-test correlations are calculated where data are available. The pools are supposed to be 'parallel' in terms of content and psychometric characteristics. This 'parallelism' is guaranteed mostly in the pool creation process by means of constraint controls and simulation evaluations.

Table 2.4 ***Correlations between measures***

	A1	**A2**	**A3**	**A4**
B1	0.50 (650)	0.41 (56)	0.48 (200)	0.42 (320)
B2	0.44 (34)	0.41 (46)	0.47 (360)	0.43 (300)
B3	-	0.44 (210)	0.45 (200)	0.46 (140)
B4	0.50 (140)	0.49 (210)	-	0.47 (220)

The between-test correlations that are calculated in this method provide us with important information to evaluate how comparable the pools are in field performance. Table 2.4 shows Pearson correlation coefficients between each pair of two tests from different pools (A1 for Pool 1 of Test A, B3 for Pool 3 of Test B, etc.). The degrees of comparability of the pools for a test can be evaluated by looking at the coefficients in each row or each column in the table. If the coefficients in a row or a column are like one another, this is evidence of the expected 'parallelism' or comparability.

In the table, each coefficient is accompanied by the sample size in the parentheses. The sample size information is important for evaluating the meaningfulness of the coefficient. If a very small sample (say, 20) is used in getting a coefficient, there may be considerable sampling error in this coefficient as compared to another coefficient that is calculated from 200 examinees' data. In short, cause should be exercised in viewing the results in this table.

Method 2.5. Model-data fit summary

When items in a CBT program are calibrated using an IRT model, each item can be described probabilistically using the calibrated item parameters and graphically through an ICC curve. Expected probabilities of getting the item can be computed across different ability levels. If some people have acquired some knowledge of an item before taking their tests, they are likely to get this item correct regardless their abilities and the difficulty level of the item. This situation will likely change this item's performance pattern compared to its expected or predicted behavior according to its IRT model, and will lead to a model fit problem. Checking for such model fit problems can be used as a means to monitor the behavior of items in the CBT operations to identify possible security breaches and compromised items.

Steffen et al (2001) has suggested a method to evaluate model fit problems. This method compares the actual number correct to the predicted number correct standardized by the standard deviation of the predicted number correct, resulting in a ZC* index. The formula for this *ZC** is:

$$ZC_{ig}* = \frac{O_{ig} - \hat{E}_{ig} - 0.5 \cdot N_g \cdot sign(O_{ig} - \hat{E}_{ig})}{\sqrt{\hat{V}_{ig}}} \qquad \text{(Formula 2.5.1)}$$

where $O_{ig} = \sum_{j=1}^{N_g} u_{ij}$, where u_{ij} is the response of examinee *j* to item *i*, and N_g is the number of examinees in group *g*.

$\hat{E}_{ig} = \sum_{j=1}^{N_g} P_i(\hat{\theta}_j)$, where $P_i(\hat{\theta}_j)$ is the 3-PL function and $\hat{\theta}_j$ is the CAT estimated ability for examinee *j*.

$$\hat{V}_{ig} = \sum_{j=1}^{N_g} P_i(\hat{\theta}_j)(1 - P_i(\hat{\theta}_j)) .$$

In practice, the abilities included in computing the statistic for each item is set to be within the range of $\theta_{\max} - \xi$ and $\theta_{\max} + \xi$, where ξ can be determined on the data at hand. $\theta_{\max}$ (thetamax) is the point on the ability distribution of maximum information for item *i* and is calculated as:

$$\theta_{\max} = b_i + \frac{1}{1.702 \cdot a_i} \ln \frac{1 + \sqrt{1 + 8c_i}}{2} \qquad \text{(Formula 2.5.2)}$$

An item is flagged as 'bad' or 'questionable' if the *ZC** exceeds 3 or 2, respectively. To avoid sample size problems, 400 or more examinees are recommended for calculation.

For the model fit summary in this evaluation method, all the items of a pool are grouped into difficulty intervals by their *b* values. Within each group, the flagged items (both bad and questionable, denoted as P and Q in Table 2.5) are counted. The frequencies of the flagged items and their proportions are reported in Table 2.5. This summary serves as a piece of evidence of the quality of the CBT pool.

Table 2.5 ***Model-Data Fit***

Item b value	Number of Items	Number of Z*c Flags (P and Q)	Percent of Flagged Items
< -2	50	1	2
[-2 - -1)	70	2	2.9
[-1 – 0)	85	5	5.9
[0 – 1)	80	3	3.8
[1 – 2)	75	2	2.7
> 2	40	0	0.0

Part 3. Evaluation methods for examinee-related information

Method 3.1. Person-fit statistic L_Z.

In CBT operations, some data screening measures are needed to identify examinees' responses that are not consistent to implemented IRT models. This type of measures is called person-fit measures. A number of person-fit indices have been developed (see McLeod & Lewis, 1999 for references). McLeod & Lewis (1999) suggested a l_Z index, which is a standardized function of the maximum of the likelihood (ML) function:

$$l_Z = \frac{\ln[L(\hat{\theta})] - E\{\ln[L(\hat{\theta})]\}}{\sqrt{Var\{\ln[L(\hat{\theta})]\}}} \qquad \text{(Formula 3.1.1)}$$

where

$$\ln[L(\hat{\theta})] = \sum_{i=1}^{n} \{u_i \ln[P_i(\hat{\theta})] + (1-u_i)\ln[1-P_i(\hat{\theta})]\}\,, \qquad \text{(Formula 3.1.2)}$$

$$E\{\ln[L(\hat{\theta})]\} = \sum_{i=1}^{n} \{P_i(\hat{\theta})\ln[P_i(\hat{\theta})] + [1-P_i(\hat{\theta})]\ln[1-P_i(\hat{\theta})]\}\,, \qquad \text{(Formula 3.1.3)}$$

$$Var\{\ln[L(\hat{\theta})]\} = \sum_{i=1}^{n} \left\{ P_i(\hat{\theta})[1-P_i(\hat{\theta})] \left(\frac{\ln[P_i(\hat{\theta})]}{1-P_i(\hat{\theta})} \right)^2 \right\}, \qquad \text{(Formula 3.1.4)}$$

where

i is for the item (I = 1, …, n),

θ is the continuous latent trait, or true ability,

u is a response to an item in the test (1 = correct, 0 = incorrect),

$P_i(\hat{\theta})$ is a 3-PL function for $\hat{\theta}$ which is the ML estimate of θ.

A large <u>negative</u> value of l_Z indicates a misfit. In practice, an operational cutoff value of –2.5 might be used to flag response patterns that have l_Z values at –2.5 or lower.

Table 3.1 illustrates the frequencies and proportions of l_Z statistics in both observed data and the simulation data from the same pool. With–2.5 as the operational cutoff, the cumulative percents of flagged cases between simulation and observed data can be compared for a pool. This comparison can also be used to compare a current pool with the aggregated data of the past pools as a whole.

Table 3.1 ***Person-Fit Statistic Lz***

Lz statistic	Simulation		July 02	
	Freq	CPct	Freq	CPct
0 or above	2,000	100.0	2,400	100.0
-0.5	1,000	46.9	1,100	46.7
…	…	…	…	…
-2.0	80	4.4	85	4.6
-2.5	**40**	**2.3**	**52**	**2.8**
-3.0	25	1.2	34	1.6
…	…	…	…	…

Method 3.2. Item completion rates for Pool 1.

In high-stakes tests, test speededness is carefully controlled and monitored to ensure both test fairness and measurement quality. In paper-pencil tests, a test form is evaluated for item completion rates. In CBT programs, each examinee sees a different combination of items, which in essence make up a unique test form. It is therefore inappropriate to evaluate the test speededness at the 'form' level. Instead, this evaluation should be conducted at the pool level by way of completion rates for the items in a pool.

This method generates a pool summary of item completion rates for the total group and some subgroups of examinees who have seen items from the pool. For the purpose of this method, the operational definition of item completion rate is the cumulative percent of examinees that responded to *K* items of their tests (*K* = 1 to the total number of items of a test) and no item was answered in less than 10 seconds. In CBT operations, if an item is answered in less than 10 seconds, the item is considered as a 'rushed' item because this typically happens when an examinee is running out of time and starting random guessing. Therefore, the record of an examinee's CBT test may show that all the items were answered, but each of the last five items was completed in less than 10 seconds.

Another feature of this method is to account for those ‘rushed’ items along with the completion rates. In other words, for each cumulative percent of completion up to the *K* item, the method also reports the cumulative percent of ‘rushed’ items that are found in those completed items. The ‘rushed’ items as well as unanswered items indicate the speededness of the pool in operations (also see Figure 1.2).

Table 3.2 presents a sample output of the method. In the second column, ‘Answered’ is for all the items that were responded to, whereas ‘Rushed’ is for the items that were answered in less than 10 seconds. The summary is provided for demographic subgroups by gender and ethnic/raciality, and for the total group.

Method 3.3. Summary of selected demographic variables for Pool 1

Just as in any other forms of test programs, the demographic compositions of the examinees may vary from pool to pool. It is necessary to keep record of such examinees’ characteristics in operations so that changes in a target testing population might be followed and monitored. This method has been applied to pool management in the past years and is included in this collection of new methods. The method looks at five demographic categories: gender, ethnic/raciality, academic major, GPA (undergraduate or high school), and age

The example in Table 3.3 describes the detailed information to be reported by this method.

Table 3.2 ***Item Completion for Gender and Ethnic/racial Subgroups for Pool 1***

			Items									
Subgroup			**01** | **20**	**21** | **25**	**26**	**27**	**…**	**30**	**31**	**…**	**35**	**N**
Female	***Answered***	Freq	73	196	46	51	…	138	101	…	2604	3799
		CPct	1.9	7.1	8.3	9.6	…	17.1	19.7	…	100.0	
	*Rushed**	Freq	0	0	12	24	…	44	49	…	84	475
		CPct	0.0	0.0	2.5	7.6	…	28.6	38.9	…	100.0	
…	**…**	Freq						…				
Afr Amer	***Answered***	Freq						13				
		CPct						19.9				
	Rushed	Freq						7				
		CPct						11.8				
…	**…**	Freq						…				

Table 3.3 ***Summary of Selected Demographic Variables for Pool 1***

		Total Group		Analysis Sample	
	Subgroup	N	Percent	N	Percent
Gender	Female	4436	55.2	3964	56.4
	Male	…	…	…	…
Ethnic/raciality	African American				
	…				
	White				
Undergraduate Major	Education				
	Humanities				
	…				
	Science				
UGPA	A				
	A-				
	…				
	D				
Age	18 – 24				
	25 – 30				
	…	…	…	…	…
	41 – 46	414	5.2	391	5.6
Total number of examinees		8036		7.30	

Method 3.4. Distribution of reported scores for Pool 1

This method describes the distribution of reported score ranges for gender and ethic/racial subgroups. Table 3.4 shows what information is reported with this method. The information can be used to compare the performances of the different groups for the current pool and to the past pools. The information can also be retained for trend evaluations.

Table 3.4 ***Distribution of Reported Scores for Gender and Ethnic/racial/Racial Subgroups for Pool 1***

Reported Scores	Gender				Ethnic/raciality/Race				Total	
	Female		Male		Afri. Amer		…			
	N	Cpct	N	Cpct	N	Cpct	N	Cpct	N	Cpct
60	47	100.0	102	100.0	2	100.0	…	…	141	100.0
50	103	98.8	138	96.2	6	96.1	…	…	221	94.1
…										
30										
20	8	0.2	7	0.1	5	0.9	…	…	12	0.2
N	3742									
Mean	30									
Median	35									
Std Dev	8									
Skewness	0.12									

Method 3.5. Test completion time for Pool 1

Although the current CBT tests are of fixed lengths and timed, the actual time that each examinee takes to complete a test may vary. This existing method yields cumulative number and percent of examinees for their completion time in minutes, and calculates descriptive statistics of the test completion time. The reports are done for the gender and ethnic/racial subgroups and for the total sample. This summary of test completion time can also be viewed as additional information about the speededness of the pool.

Table 3.5 is an illustration of the report. It is assumed that few examinees are expected to complete their tests in less than half of the given time (or some other amount of time that can be determined from existing data). Therefore, it is not necessary, nor reasonable to list every minute of the given test time in the report. Instead, one or two ranges are used for the time that most examinees are likely to need for their tests. One-minute intervals are applied only to the higher end of the time span for the tests. More examinees are expected for each one-minute interval.

Method 3.6. Distribution of scores for current Pool and past pools

This method produces a report of the score distributions from the current pool and the past pools. Table 3.6 illustrates the results of the analysis using this method. The score distribution of the current pool is compared with the scores of the past pools for the same month and with all the past pools. This table presents a concise snapshot for a comparison of the current pool with the past pools. Again, the scores are grouped, frequency, percent, cumulative percent of examinees in each score range are calculated. Descriptive statistics about the examinees are also provided.

Table 3.6 Score distributions for current pool and past pools

Reported Score	Total Group July 02			July Sample 1997-2001			All Sample 1997-2001		
	Freq	Pct	Cpct	Freq	Pct	Cpct	Freq	Pct	Cpct
60									
50									
…									
30									
20									
N									
Mean									
Median									
Std Dev									
Skewness									

Summary and suggestions for implementation

This study has reviewed common methods for item pool evaluation in ETS CBT operations, identified the needs for new methods, and developed a new set of methods for implementation. The new methods are divided into three categories to reflect both operational priority and the focus of the methods. Because CBT operations are on a sort of ‘continuous’ basis in that tests are given more than once a day or a month, and data flow is also continuous on a daily basis, it is important for the new methods to accommodate this dynamic nature of the CBT data and obtain timely information about the performance of the item pools as the pools are still being used in the field.

The primary methods are designed with this reality and need in mind and will have the capability of monitoring some aspects of live pools' performance. If there is any indication of abnormality in a pool, further investigation can be called upon immediately instead of having to wait until the pool is rotated out of use. This capability will certainly benefit operational CBT programs in an unprecedented way to monitor the quality of the programs. The methods for item-related and examinee-related information will provide additional information for evaluating a pool's performance after the pool is rotated out of an administration period. These methods will enable us to have a thorough evaluation of the pool and retain detailed information for future reference.

Proper implementation of the new set of evaluation methods is critical to fully realizing the benefits of the new methods. The following suggestions might help in designing the implementation phase. First, the methods must be automated. In operations, all the methods need to be applied to each CBT pool. Multiple pools may be needed for a testing window. Potential work load will make it difficult to implement these pool evaluation methods without some type of automation. Second, the implementation system should allow interactive control of the analyses to be carried out. Many methods can be set up to run on default settings for routine evaluation of item pools using all data. There are times, however, when a test developer wants to look at certain segment of the data for various purposes. Users should be given the freedom of defining the scope and type of data to be included in analyses. Third, a special database is needed to store aggregated data information. Each time when a method is applied, data from an item pool is analyzed and some aggregate information is generated. It is not only cost effective but also necessary to retain such aggregate information. Such aggregate information can be retrieved at any time and can be used as a baseline for comparisons over time and across pools.

To conclude, the methods from this study are built on the current experience in CBT operations and reflect some of the research that has been done in this regard so far. It goes without saying that there must be some new research and development in CBT that is not identified nor included in this study. More appropriate and feasible methods will certainly emerge to improve or even replace the methods from this study in future. The field of CBT, especially in large-scale testing programs, is still relatively young and is growing fast. New technology and new development in CBT will provide new opportunities to measure things that can not be measured today. Future use of CBT pools can be quite different from what it is now. Accordingly, new evaluation methods will also be developed to support the future CBT operations.

References

Guo, F., & Wang, L. (2003, April). Online calibration and scale stability of a CAT program. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kingbury, G. G. (1997, March). Some questions that must be addressed to develop and maintain an item pool for use in an adaptive test. Paper presented at the annual meeting of the National Council on Measurement in education, Chicago, IL.

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurementt*, *23*(2), 147-160.

Steffen, M., Wang, M. M., Wingersky, M., & Zhu, R. (2001). Preliminary operational item monitoring method: Final Report. ETS internal communications.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, 2, 277-292.

Stocking, M. L., & Lewis, C. (1995). Controlling item exposure conditional on ability in computerized adaptive testing. RR-95-24, Educational Testing Service, Princeton, NJ 08541.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In Howard Wainer (Ed.), Computerized adaptive testing: A primer (2nd ed. Pp.

Wang, L., & Gawlick, L. (2001). The influence of item characteristics and administration position on CAT Scores. Paper presented at the 33rd annual meeting of the Northestern Educational Research Association, Hudson Valley, NY, October 26, 2001.

Wang, X., & Braswell, T. (2001, April). Summary of TOEFL CBT item pool creation, maintenance, and usage. Paper presented at the annual meeting of the National Conucil on Measurement in Education, Seattle, WA.

Way, W. D. (1997, March). Protecting the integrity of computerized testing item pools. Paper presented at the annual meeting of the National Conucil on Measurement in Education, Chicago, IL.

Way, W. D., Swanson, L., Steffen, M., & Stocking, M. L. (2001). Refining a system for computerized adaptive testing pool creation. RR-01-18, Educational Testing Service, Princeton, NJ 08541.