Practical Issues in Developing and Maintaining a Computerized Adaptive Testing Program

Steven L. Wise* and G. Gage Kingsbury**

James Madison University (USA)*
Northwest Evaluation Association (USA) **

The basic principles of computerized adaptive testing are relatively straightforward. The practice of implementing and maintaining an adaptive testing program, however, is far more complex. A number of practical challenges await measurement professionals; we discuss four general types: establishing and maintaining item pools, choosing test administration procedures, protecting test security, and responding to examinee issues. The success of an adaptive testing program will largely depend on how well the measurement practitioner deals with these challenges.

Key words: computerized adaptive testing, adaptive testing program

The computerized adaptive test (CAT) has become increasingly common in large-scale testing programs. The primary advantage of a CAT to test developers and administrators is its promise of efficient testing. In theory, examinee testing times can be dramatically reduced while maintaining the quality of measurement provided by conventional (i.e., fixed-item) tests. This advantage is particularly attractive to testing programs that have traditionally required lengthy tests. In such testing contexts, the potential problem of examinee fatigue and, consequently, diminished effort can be alleviated by use of a CAT.

Virtually all operational CATs use measurement methods based on item response theory (IRT; Lord & Novick, 1968) to select test items to administer and to estimate examinee proficiency. The invariance principle of IRT allows one to administer different sets of items drawn from an item pool to different examinees, yet estimate their relative levels of proficiency on a common scale of measurement. A CAT's efficiency is realized through the targeting of item difficulty to examinee proficiency. IRT principles suggest that items targeted in this manner provide maximal information in the estimation of examinee proficiency.

A CAT administration is essentially the repetition of a two-step process. At step one, an item is administered whose difficulty is matched to the examinee's current (or initial) proficiency estimate. At the second step, the examinee's response to the administered item is scored, and the examinee's proficiency estimate is updated. These two steps are then repeated until some stopping criterion is met, which is usually either a predetermined number of items or a desired level of measurement precision. Through this process, the CAT algorithm converges on a final proficiency estimate for the examinee.

Although, in theory, adaptive testing is a relatively simple idea, the reality of planning, implementing, and maintaining a CAT program is substantially more complex. In this paper we discuss a number of challenging practical issues that must be addressed in establishing CAT programs. We address four general areas: item pools, test administration, test security, and examinee issues. In each area, we present the practical issues likely to be faced by administrators of CAT programs, discuss empirical research relevant to each issue, and provide recommendations for measurement practice. Although most of the research studies we report were conducted in the United States, we believe that the issues they address are relevant in Spain and throughout Europe as well.

Developing and Maintaining an Item Pool for Use in an Adaptive Test

In developing an item pool for adaptive testing, we need to address a number of practical issues that may affect measurement. The decisions we make about these issues will determine what our item pool looks like when we administer an adaptive test, and how the results of that test relate to the underlying trait being measured. These issues include the following:

1) Item pool size and control

2) Verifying the dimensionality of an item pool

3) The response models

4) Item removal and revision

5) Adding items to the item pool

6) Maintaining scale consistency

7) Using multiple item pools

Many of these issues come about because we almost never want to measure just one thing. If you look at the test development blueprint for a test in fifth grade mathematics it will contain actions like create, recognize,

and <u>apply</u>. At the same time, the blueprint will deal with content areas as diverse as geometry, problem solving, and computation with whole numbers.

As measurement professionals, the job of identifying and measuring the underlying traits for this type of test blueprint is not an easy one. We complicate the issue even more when we move from a traditional paper-and-pencil test to an adaptive test, because instead of dealing with a measurement scale for a single test to be administered in a group setting, we need to establish a measurement scale for a host of possible tests that examinees might see. To deal with this very complicated situation, it is useful to see it from the point of view of one of our fifth grade teachers, who suggested that it is easy to add apples and oranges. You just end up with mixed fruit. Therefore, in developing item banks and measurement scales for our adaptive tests, it is helpful to remember that we are normally measuring along the mixed fruit metric.

Item Pool Size for an Adaptive Test

At one time it was suggested that an item pool as small as 100 items could allow a CAT to improve the accuracy obtained from a wide-range paper-and-pencil test (Urry, 1977). While this is still true, no one currently developing an adaptive test would try to begin with as few as 100 items. Three factors have caused us to drastically increase the size of item pools that are viewed as appropriate. First, conventional tests have become considerably better in the past few decades. Functional level testing (NWEA, 1997a) provides educational tests which are approximately as accurate as a two-stage adaptive test. Second, constraints that we impose on the adaptive testing item selection procedures (such as content constraints and longitudinal testing constraints) require considerably larger item pools than adaptive tests with no such constraints, to provide the same amount of information. Third, for fairly high-stakes adaptive tests that are intended to be in use for more than a short time, the security of a test may be compromised if the item pool is too small. It is a relatively simple task for a group of examinees to lower the validity of a test with a small item pool by memorizing a few items each and telling a few friends.

As a result, many of the adaptive tests in use today have item pools with more than 1000 items. At least one licensure test uses rotating item pools of more than 2000 items each. Obviously, the stakes involved in a test are going to be a primary factor in determining how many items are needed. It is clearly helpful to make a distinction between the bank of items that is available for the design of an adaptive test, and the pool of items that is actually made available in the adaptive test. Using a fraction of the items in

the item bank to form the item pool for any particular adaptive test is probably advised.

Verifying the Dimensionality of Item Responses

Before starting a discussion of dimensionality, it is important to clean up our language. Item pools do not have a dimensionality. Item responses do. Therefore, in investigating dimensionality, we are investigating the interaction between examinees and test questions. Thus, we need to be clear in identifying the population of interest. If the dimensionality analysis is done with uninterested volunteers or with students prior to instruction, it is unlikely that the dimensionality of their responses will be the same that we will observe when we actually administer the test to highly motivated individuals or students after instruction.

We would suggest two points at which it is imperative to check response dimensionality. The first is during the initial development of the item bank. At this time, we need to identify the dimensionality in order to determine whether a unidimensional or a multidimensional item response model is better suited to describing the responses of examinees. In general, we would search for the minimum number of dimensions that adequately explains the item responses.

The second point at which we need to check the dimensionality of the response space is when we systematically change the test content or the population of individuals taking the test. This is often neglected by test developers, to the detriment of the measurement scale.

Checking the fit of the item responses to the IRT model is helpful, but certainly not sufficient in identifying response dimensionality. Unfortunately, identifying dimensionality is closely tied to the response model that we plan to use in modeling multidimensional item responses. There are many procedures that help in identifying the dimensionality of the response space including exploratory factor analysis, confirmatory factor analysis, full-information factor analysis, multidimensional scaling, essential dimensionality analysis, and various techniques related more directly to IRT.

While there are many ways of modeling student performance in *n*-space, few have empirical evidence suggesting that their use might be justified. We need more evidence indicating how multiple traits interact to form an item response before we can be justified in deciding between compensatory and non-compensatory models. As a result, one approach to dealing with response dimensionality is to use the most parsimonious model that adequately explains examinee responses. Whenever possible, it seems

that a unidimensional model would be preferable, since it makes far fewer assumptions about responses.

A simple approach to identifying whether a set of responses can be represented adequately by a single dimension is to conduct an exploratory factor analysis, with a parallel analysis to identify whether there is more than one factor in the data set that explains more variance than the factors from a random data set. If this indicates a single factor, one can calibrate the items within content areas and within the total test to perform the principal-axis calibration analysis suggested by Bejar (1980) and couple the results of this analysis with the essential dimensionality analysis described by Stout (1987).

If these analyses indicate that a unidimensional approach is adequate, then it seems most parsimonious to apply a unidimensional item response model. If the results of the analyses indicate a multidimensional structure is necessary to explain the examinee responses, then it might be prudent to look carefully at the structure of the test and try to simplify it rather than using one of the multidimensional item response models.

Item Response Models

Many researchers have made the choice of a particular item response model into the focus of a debate that can only be termed religious. However, there is no theoretical or practical reason that we should confine ourselves to a single item response model, and a number of reasons why we should not. For instance, if one wishes to construct a measurement scale with a limited sample size, the 1PL (one-parameter logistic) model may be the most appropriate model to use (Lord, 1983) even if we don't expect that the items in our pool fit the 1PL model very well.

Later, when larger sample sizes are available, one might want to add to the existing item bank without disrupting the measurement properties of the scale. Is there anything preventing the use of the 3PL model with these new items? Of course not, as long as the procedures to add the items to the scale are appropriate. The result is an item pool in which the items are calibrated to different response models, but are linked to a common measurement scale. Religion shouldn't prevent good measurement. This type of approach can be used to create hybrid banks that include unidimensional and multidimensional models, and dichotomous and polychotomous models.

Another issue concerning the use of a particular item response model has to do with adaptive testing itself. Since adaptive tests choose items based on item parameter estimates, these parameter estimates need to be accurate. At the same time, the need for accuracy is greatest near the point of inflection of the item response function, since this is the vicinity in which

most items are administered. Since this is the range of the item response curve in which our different models differ the least, it may imply that differences in the various unidimensional models may not be as important as they would initially appear.

Item Removal, Revision, and Retesting

In building an item pool and measurement scale for use in an adaptive test, it is critical to determine procedures to use to help throw out items that don't perform well. This is the case regardless of the item response model(s) you choose to use. Poor items should be removed from the item pool as soon as they are identified, as they can cause errors in proficiency estimation and in decisions about examinees.

Some have suggested that with a flexible item response model, poor items will be seen rarely in an adaptive test. Therefore, removing oddly performing items should not be an important consideration. This idea may have had some merit in the early days of adaptive testing, because items that don't fit the response model tend to have lower values for the discriminatory power parameter. As a result, these items tended to be selected for use infrequently. However, the current use of exposure control and content balancing has made it more likely that these poor items will appear on some individuals' tests. As a result, it is quite important to remove poor performing items from the item pool.

One way of trying to investigate item misfit is to examine the empirical item response curve, and compare it to the theoretical curve. The comparison of the theoretical curve to the empirical curve (based on the observed proportion of correct answers to an item from groups of examinees with nearly identical proficiency estimates) is an often overlooked, extremely straightforward approach to the visual identification of poorly performing items.

Only through careful rejection and revision procedures can we maintain an item pool that measures consistently. This is extremely important in an adaptive test, which tends to be a shorter test, and therefore tends to be more influenced by oddly performing items. Items are frightfully expensive to develop, and so we have a tendency to want to keep items that seem to be functioning, but not functioning very well. In the long run, it is better to revise and retest these items, even though it is a more expensive process.

One final point with regard to item rejection and revision. If the process that you set up seems to be consistently rejecting a large percentage of the items that you field test, you may have a deeper problem. If these

rejected items don't have some readily identifiable flaws, you should reconsider the dimensionality and model specification questions.

Adding Items to the Item Pool

Once you have a pool of items calibrated to a particular measurement scale, you can add additional items to the pool through any number of linking designs. Normally a group of examinees will take a set of old, calibrated items and a set of new, uncalibrated items. Then a linking procedure is used to calibrate the new items to the existing scale. This is a very reasonable procedure to use in an adaptive test, in which one can seed new items throughout the test at will. The new items that are administered in this fashion can then be brought onto the measurement scale using one of two linking procedures. (You will notice that we didn't say "equating". The processes of linking and equating are often confused, and in this case, score equating is exactly what you do not want to do.)

In one common linking procedure, all items administered to a person are calibrated (the old and the new), and then the difference in calibrations for the old items is used to transform the calibrations of the new items onto the original scale. This procedure is less than optimal for use in an adaptive test for several reasons. First, since different individuals take different items in the body of the adaptive test, you have to collect much more data than normal in order to calibrate the old items (or, alternatively, you need to seed old items into the test along with the new items, lengthening the test considerably.) Second, the transformation to the measurement scale is a group process which is extremely sensitive to the items used. Unless some double checking process is available, this procedure is not suggested.

In a second linking procedure, the new items are calibrated using the test taker trait level estimates obtained from the old items as if they were the actual trait levels. This fixes the trait level parameter for each person, and reduces the estimation of item parameters to a single step. Item parameters obtained using this fixed-parameter design (Ingebo, 1997) are by definition on the desired measurement scale. The accuracy of this procedure is directly related to the accuracy of the trait level estimate, and the number of items in the adaptive test (which affects the granularity of the trait level data, and therefore affects the values that we can obtain for the item parameter estimates for our new items.) While this procedure has weaknesses, it is preferable to the first linking procedure for adaptive testing, because the adaptive test should give us very stable trait level estimates, and because it doesn't depend on the administration of certain items during the adaptive portion of the test.

Maintaining Scale Consistency

If a scale is in use for several years, the population being tested is likely to change in its characteristics. As a result, means and percentiles and other sample specifics will change across years. In this changing environment, the fixed-parameter linking design described above will enable you to fix the measurement scale so that item parameter estimates shouldn't drift very much. However, it is rational to conduct a drift study by recalibrating some previously calibrated items occasionally, to verify that drift is non-directional and within the bounds that we would expect due to sampling error.

In addition to scale drift, scale consistency can be threatened by the conditions under which field testing is done. Issues such as administrative modifications, time limits, sample suitability, and other environmental conditions can cause instability in the measurement scale which a drift study can identify, but not correct. We need to control or account for these factors if we intend to have a scale that stays stable over long periods of time. This is particularly important in larger testing efforts in which field testing is being done in various sites or by multiple organizations. Since almost all IRT models assume that items are given under power conditions, the most important of these factors may be speededness.

If we assume that field test items are scattered throughout an operational test, we have a strong field testing paradigm. The examinee is in the same state while taking the field test items as he or she is when taking the operational items, because they are given at the same time. However, we can still do damage to the measurement scale simply by changing the time constraints of the test.

If the time limits of the test were changed so that it was slightly speeded, field test items that are administered toward the end of the test would appear slightly more difficult than they would if they had been administered at the beginning. This is a problem that will continue into future years, as the items with this slightly-too-difficult calibration are used as part of the operational test, where they become the items that help calibrate the new items for next year—resulting in the new items having slightly-too-difficult calibrations. This cyclical process can cause the measurement scale to drift farther and farther from the original measurement scale. As a result, longitudinal growth and change estimates become suspect. The worst thing about this phenomenon is that it isn't correctable by the handling of omitted questions, because we still have the effects that occur due to hurried responses.

The easiest way to fix this problem is to a) vary the position of any particular item in the field test and b) use only untimed tests. In the name of economics, timed tests may look appealing, but the long term costs that come from slippage in the measurement scale are substantial, and hardly ever considered when trying to establish time limits for adaptive tests. If you must use a timed test, placing field test items in the first portion of the test seems advisable, and one of the recently developed hybrid IRT models to help control for speededness may be useful. However, if examinees are allowed to continue to work on the test as long as they are working constructively, the measurement scale will be much stronger in the long run.

Using Multiple Item Pools

In a high-stakes adaptive test, which must maintain a high degree of item security, one might use multiple item pools that periodically rotate in and out of use. The primary purpose for use of more than one pool is to give a testing organization a way to respond if an item pool is stolen or compromised. This approach is almost never appropriate for low-stakes tests. In general, it is a more effective measurement procedure to include more items in a single item pool for the adaptive test, rather than splitting the items into smaller pools. Ten years ago, developers of adaptive tests sometimes created small item pools because of storage considerations, but that is no longer an important issue.

It is unclear whether the use of multiple item pools helps or hinders test security in high-stakes tests. Using multiple item pools reduces the size of each item pool, making memorization easier if the pool is stolen, and making it more probable that examinees will see individual items that have been exposed. An interesting research question is whether or not we need secure item pools for high stakes tests, if our item pools are extremely large and our item selection algorithm limits or balances item exposure. It may be that if it would take a concerted effort to memorize enough items to influence their scores, examinees might choose instead to learn the content being tested.

Test Administration

A substantial body of research has been conducted concerning the procedures to be used in administering an adaptive test. It is the single most investigated aspect of adaptive testing. The early theoretical work by Lord on procedures for adaptive testing (1970, 1976) set the foundation for virtually all later work. The work by Weiss and his colleagues (Betz & Weiss, 1974; McBride & Weiss, 1976) investigated practical procedures for

adaptive testing, and provided much of the early work in actually implementing adaptive tests.

This led to work by a host of researchers dealing with practical applications of adaptive testing for specific admissions tests, licensure tests, certification tests, and general educational tests. Each of these application areas has its own particular needs, but in every case the developers have to make decisions in the areas of test entry, item selection, scoring, and test termination.

The Test Entry Procedure

A procedure for selecting the first item in an adaptive test should almost always use all of the information that is available about a examinee. This information may differ from one testing situation to another, or even from one examinee to another.

An example of this may be seen in the NWEA adaptive tests (NWEA, 1997b) which are used to assess student achievement in mathematics, reading, and language usage. In these tests a triage procedure is used to identify the first item to be given to a student. First, the system checks to see if the student has taken a previous test. If so, the previous achievement level estimate is used to start the current test. If no previous test score is available, the system looks to see what grade the student is in, and starts the test at the grade level mean (based on a large norming sample). If no grade level mean is available the last procedure to be used is to start the students test at a predefined achievement level.

It should be noted that the entry level only identifies the characteristics of the first item. All subsequent item selection is based on student performance. Procedures that allow an examinee's final score to be affected by anything but the performance on the current test are unlikely to be acceptable and are likely to be challenged by those adversely affected.

The Item Selection Procedure

Once the entry point for a test is established, the item selection procedure for the body of the test needs to be delineated. In early adaptive tests, it was common for test developers to choose the most informative item that hadn't been administered, and administer it. Information is almost never the only consideration in current adaptive tests. Virtually every current adaptive test chooses items using some variety of constrained CAT (e.g., C-CAT; Kingsbury and Zara, 1989) item selection procedure. The varieties of constraints put on item selection commonly include content constraints, item exposure constraints, conflicting item constraints.

One approach to item selection uses a Bayesian estimate of proficiency. This estimate can start with a mean equal to the performance level that is used as the entry point, and a very diffuse prior. Each question selected in the test is the one that is expected to reduce the variance of the prior distribution the most. This approach allows the item difficulty to vary quite a bit at the beginning of the test. The prior distribution is updated after each item is taken, and is used for item selection throughout the test. As the test progresses, we become more confident about the examinee's proficiency level, and so the changes in item difficulty become smaller and smaller.

This individualized Bayesian item selection results in less radical difficulty changes than the use of the maximum-likelihood estimate of proficiency. It can be made even more efficient by putting constraints on the way in which the posterior variance changes from item to item.

The Scoring Procedure

While a Bayesian achievement level estimate can be used for item selection, as described above, a maximum-likelihood proficiency estimate is probably more appropriate as a reported score for virtually all circumstances. This procedure of scoring is asymptotically unbiased, and should result in the most informative proficiency estimates across all examinees. In addition, it avoids an awkward characteristic of the Bayesian score by not being affected by anything other than the current test performance.

The Test Termination Procedure

As adaptive tests are developed, it is reasonable to tailor the termination procedure to the test purpose. For instance, a test that is to be used for initial screening of students entering a school might benefit from a fixed test length, to provide for fairly consistent testing times across students. On the other hand, a licensure test designed to make a high stakes decision about a candidate might benefit from a variable length test that provides more information for candidates near the decision point. This type of test might also benefit from the use of a variation of the sequential probability ratio test (Wald, 1947) to provide a consistent level of confidence in decisions made (Reckase, 1983). Between these extremes, a test designed to provide high-quality proficiency estimates for a broad spectrum of individuals might benefit from the use of a stopping rule that allows test length to vary, but terminates the test when a predefined amount of information is obtained.

Test Security

Security is a concern of any testing program. No matter how strong the psychometric characteristics of a test, if test security is compromised then the validity of score-based inferences is undermined. In this section, we briefly discuss security issues that are particularly relevant to CATs. An extensive discussion of these issues is provided by Way (1998).

As was discussed earlier, the success of a CAT is dependent on the integrity of its item pool. The higher the consequences associated with a CAT, the more likely that persons or organizations will try to acquire information regarding the particular items in the pool. To the extent that an item is known in advance by examinees, its item parameters (estimated from a calibration sample) no longer apply. As proportionately more examinees know the content of the item, its difficulty parameter shifts toward the easier end of the proficiency scale, the discrimination parameter shifts toward zero, and the guessing parameter becomes increasingly irrelevant. It is therefore essential to the CAT item selection and scoring processes that the items remain secure. There are two major issues concerning the security of items in a CAT environment: item disclosure, and item theft.

Item Disclosure

In paper-and-pencil testing, all examinees can usually be tested simultaneously. Sufficient test booklets to test each examinee can be inexpensively produced. In contrast, computer-based testing is typically asynchronous. It is likely that there will be fewer computers available than examinees to be tested, which implies that some examinees will be tested before others. Moreover, an attractive advantage of computer-based testing is its capability for on-demand testing, in which examinees are tested at various times, and not in a large group administration.

It is a common practice for examinees to talk among themselves about test items, particularly when the consequences for test performance are high. Students who study the particular items they have heard about from others and then take the test later during the testing period would be potentially at an advantage, which would tend to positively bias their proficiency estimates. A solution to this problem is to use large item pools, which would diminish—but not eliminate—the impact of such item disclosures.

In school settings, a related problem occurs when teachers find out about specific test items from earlier-tested students and quickly "teach to the test," which could increase the test performances of students who had not yet been tested. Although being taught the material from specific items would probably incrementally increase the proficiency levels of students,

such a practice would invalidate inferences from performance on the specific test items to proficiency on the content domain(s) from which the items were sampled.

Item Theft

All testing programs must be concerned regarding theft of its items. Some proponents of CATs have contended that such tests are inherently more secure, because there are no paper copies of test forms that can be stolen or photocopied. CATs, however, can be quite vulnerable to covert item theft. Colton (1998) discussed numerous electronic devices that can be used to steal test items, including pagers, miniature cameras, video transmitters, and micro video recorders. In addition, there is the threat of electromagnetic spectrum interception, in which relatively inexpensive equipment can be used to covertly intercept the electromagnetic signals from a computer—essentially creating an exact replica of what is being displayed on a computer monitor. Such types of theft can occur without the test giver being aware of its taking place. Colton discussed various observation and electronic countermeasures that can be employed; he notes, however, that such countermeasures can be costly.

It is probably realistic for us to acknowledge that we could not prevent sufficiently determined individuals from gaining access to test items from the pool. Presuming, then, that total security is infeasible, we should instead focus on deterrence measures, such as the use of multiple large item pools that are frequently rotated, close monitoring of test examinees, and limited use of on-demand testing.

Examinee Issues

The CAT administrator must find adequate solutions to a number of practical technical problems. A CAT is used to assess the proficiency of people, however, and test givers would be prudent to not overlook potential problems that a CAT administration might cause for examinees. Although it is important to consider the perspective of the examinee in any achievement testing, it is particularly important for us to understand how the unique—and relatively new—testing methods used in a CAT may affect examinees.

At first glance, the experience of taking a CAT may not appear to be very different from a conventional test. An item appears on the computer screen, the examinee enters his or her answer, and the next item appears. This process continues until the test is completed. There are, however, a number of unique aspects to the CAT experience that might influence an examinee's test performance. First, computer-based testing is unfamiliar to many examinees. Items presented on a computer screen may be more

difficult or fatiguing for examinees to read. Longer items, whose size exceeds the dimensions of the computer screen, require examinees to scroll through the item content. The entry of examinee responses using a keyboard or mouse is different from circling answers in a test booklet or filling in bubbles on a machine-scorable answer sheet.

Second, in a conventional test, examinees are usually given all of their test items at once. This provides examinees a great deal of freedom to browse through the items, skip some items to be answered later in the test, and review—and possibly change—answers. In contrast, CAT examinees have far less control, because items are typically administered one at a time, without an opportunity for review.

Finally, in many CATs the length of the test (in number of items) can vary markedly across examinees. In norm-referenced measurement, different test lengths across examinees result whenever a common standard error of proficiency estimation is used as the criterion for terminating a CAT. In criterion-referenced measurement, for which the goal of measurement is to identify examinees whose proficiency levels exceed some standard, testing for a given examinee will continue only until a confident pass/fail decision can be made. During these types of testing situations, examinees may have little idea how close they are to completion of their tests. This is quite different from conventional tests, in which examinees can continually tell how many items they have yet to answer, and can allocate their efforts accordingly.

The purpose of this section is to discuss the examinee's perspective during a CAT administration. Three examinee issues are discussed. The issues are interrelated, as decisions made by the test giver concerning each issue may affect other issues as well.

Item Review

In developing a CAT, we must make a decision regarding item review. Currently, virtually no operational CATs provide an opportunity for examinees to go back and review their answers to previously administered items. Item review in a CAT has generally been viewed by test administrators as a threat to the increased efficiency of adaptive testing. Item review requires additional testing time, changed answers may increase the standard error of an examinee's proficiency estimate, and examinees may strategically use item review to artificially increase their scores. Although the reactions of examinees to CATs have been generally positive, however, they have consistently reported dissatisfaction with the absence of item

review (Baghi, Ferrara & Gabrys, 1992; Legg & Buhr, 1992; Vispoel, Rocklin & Wang, 1994; Vispoel, Wang, de la Torre, Bleiler & Dings, 1992).

Should we be concerned about the strong examinee preference for item review? It is worth noting that the availability of item review to examinees during paper-and-pencil tests was an unplanned, uncontrollable aspect of group-administered achievement and ability tests. With computer-based tests, however, test givers can effectively prevent examinees from reviewing their answers. Moreover, if no one is allowed item review on a CAT, then everyone is treated equally. This test giver-imposed control over item review is therefore consistent with the idea of increased test standardization.

Eliminating item review, however, may negatively impact examinee test performance. Over sixty years of research on answer changing has consistently shown that (a) when examinees are allowed to change answers, they are more likely to improve their scores (albeit typically slightly), and (b) score gains due to answer changes are overwhelmingly due to reasons such as re-thinking or re-reading the item, or making a clerical error. It follows, therefore, that denying item review denies an opportunity for answer changing, which would tend to improve test performance.

There is also the possibility that denying item review results in increased levels of anxiety—and possibly impaired test performance—for some examinees. While denying item review represents increased control for the test giver, it also means decreased control for the examinee. And it has been found, in many contexts, that individuals better tolerate stressful situations (such as tests) when they feel that they have some control over their environment. Increased perceived control has been associated with decreased anxiety and improved task performance (Glass & Singer, 1972; Blechman & Dannemiller, 1976; Perlmuter & Monty, 1977).

There are important arguments on either side of the item review issue, and test givers should weigh these arguments in deciding whether item review should be provided. Several useful discussions of the advantages and disadvantages of item review with CATs are available (Lunz, Bergstrom, & Wright, 1992; Stone & Lunz, 1994; Vispoel et al., 1992; Vispoel, 1998; Wainer, 1993; Wang & Wingersky, 1992; Wise, 1996).

Time Limits

Establishing a reasonable time limit for conventional standardized tests is challenging. If the testing time is too long, then time needed to administer a test is needlessly lengthened, with a consequent loss of testing efficiency. If the testing time is too short, then some examinees will not be able to

complete all of the test items in the allotted time. For these examinees, the resultant test scores will underestimate their true levels of proficiency—which means that the test validity has been compromised.

For a CAT, however, establishing a time limit is more complicated. One reason is that CATs using score precision as a stopping criterion will administer tests of different lengths. And if one does not know in advance how long a given examinee's test will be, how does one know how much time to allow? Even when fixed-length CATs are used, the time limits issue is complex. Imagine two CAT examinees: a more able examinee who receives 40 harder math items, and a less able examinee who receives 40 easier math items. Should we use the same time limit? What if we knew that the harder items generally require more time for an examinee to answer, because they involve more time-consuming computations? Because examinees each receive a unique set of items, it is more difficult to choose a single time limit that would be equally appropriate for each of their tests.

The decision regarding appropriate time limits to provide on a CAT is an important issue. One might argue that the imposition of any time limit is antithetical to a goal of a testing program that promotes students exhibiting their optimal levels of performance. Time limits can also cause difficulties in item calibration, as discussed earlier. The practical goal is to identify a time limit that does not meaningfully limit student performance, while keeping the testing session reasonably short. This issue is complicated by findings that some ethnic minority groups take more time to complete CATs (Baghi et al., 1992; Legg & Buhr, 1992; O'Neill & Powers, 1993; Zara, 1992), though some research has indicated that allowing minority students more time on conventional tests has not enhanced their performance relative to majority students (Evans & Reilly, 1972; Wild, Durso & Rubin, 1982).

The relationship between time limits and test performance appears to be moderated by examinee test anxiety. The differences in test performance between timed and untimed tests have been found to be greater for highly test anxious examinees (Hill, 1984; Onwuegbuzie & Seaman, 1995). This suggests that lengthening a time limit on a CAT may benefit some examinees more than others.

Given the differences among examinees, it appears that a single time limit is likely to be difficult to defend as equitable. Therefore, we should consider adopting very liberal time limits with CATs, or consider imposing no time limits at all. Keep in mind that a CAT is dramatically shorter than its conventional counterpart; we should consider giving some of that saved time back to examinees. Examination-related stress should thereby be reduced and test validity may be enhanced.

Equity

By *equity,* we refer to a set of factors that may compromise the fairness and comparability of scores from different types of examinees. An understanding of these factors should help guide the development of CAT testing programs that minimize their effects.

There is evidence that, in the United States, poor and minority children have had less access to computers at home and at school (Sutton, 1997). Because less access implies less experience, the relationship between computer experience and CAT performance becomes of increased importance. The limited research on this issue specifically related to CATs is mixed. One study found differences among racial/ethnic groups (Buhr & Legg, 1989) related to computer usage, while the other (Baghi et al., 1992) did not.

As discussed earlier, there are differences in racial/ethnic groups concerning testing time used on a CAT. Hence, any time limit that is imposed may have a differential effect on different groups—which may exacerbate test performance differences among these groups.

Research regarding subgroup differences in test performance between CATs and conventional tests is mixed. Zara (1992) found that the differences in performance between CAT and conventional versions of a national nursing licensure exam varied substantially across ethnic groups. White examinees showed a modest difference in favor of the conventional version, whereas Black examinees showed virtually no difference in performance between the test versions. In contrast, Buhr and Legg (1989) found that, although all ethnic groups scored higher on their CAT reading test, differences between scores for White examinees and those for Blacks and Hispanics were greater on the conventional test than on the CAT. Hence, the limited research regarding subgroup differences in test performance between CAT and conventional tests has not yielded consistent evidence that ethnic minority groups would be disadvantaged by a CAT.

At this point, it is too early to tell whether use of a CAT is likely to increase or decrease test score differences among subgroups. CAT developers should, however, be prepared to investigate this issue with their own CATs. Again, adopting liberal time limits is likely to minimize any subgroup score differences that are attributable to differences in the time needed to take a CAT. Sutton (1997) provides a good discussion of this equity issue.

CONCLUSIONS

One thing that should be clear from the nature of the comments above is that an adaptive test is much more than the test itself. In order to put a high quality adaptive test into place, the developer needs to think systemically. The test score that a single individual receives is only as accurate as the system allows it to be. Because the systems surrounding adaptive tests tend to be more complex than those surrounding paper tests, we should expect that issues of item pool maintenance, test administration, test security, and examinee issues should also be comparably more complex. This paper has tried to detail some of that complexity.

Because we have spent the bulk of this paper describing the many issues that must be addressed in the development and deployment of an adaptive test, it is probably useful to discuss what it is that makes this additional complexity worthwhile. The two features that set an adaptive test apart from a paper test for the examinee are its immediacy and its individualization. The feature that sets an adaptive test apart from a paper test for the test developer is the control of the immediacy and individualization.

That control gives the adaptive test developer the ability to mimic that which occurs in a paper test very closely, or to chose to keep or discard features of a paper test as they are desired. An example of this can be seen in the options for item review in an adaptive test. The only reason that item review is a topic of conversation within the context of adaptive testing is that it can be controlled. One can think about item review as a bad piece of baggage that couldn't be avoided in paper testing but can in an adaptive test. Alternatively, one can think of item review as a desirable feature that allows examinees to be more comfortable in the testing situation. The ability of the developer to have an opinion and act on it comes directly from the nature of adaptive testing. If a test developer views the characteristics of a paper test as part of a list of options for an adaptive test, the best adaptive test can be designed by keeping only those options that are appropriate for the situation at hand.

Research regarding adaptive testing has become progressively more applied over the past decade. With the increased emphasis on the implementation of adaptive tests, practical considerations have taken precedence over psychometric concerns for many researchers. There are areas in which this emphasis on the practical over the theoretical will soon slow practical development. Two of these areas are the development of psychometric models appropriate for use with simulation assessments and the development of calibration procedures that are designed for item pools

with an existing measurement scale. Substantial developments in each of these areas will be required to allow progress in adaptive test development within the next decade.

In this paper, we have discussed a number of practical challenges faced by measurement professionals involved in a CAT program. The success with which these challenges are met will largely determine the ultimate utility of the program.

RESUMEN

Aspectos básicos en el desarrollo y mantenimiento de programación de Test Adaptativos Informatizados. Los principios básicos de los tests adaptativos informatizados están relativamente bien establecidos. Sin embargo, la puesta en funcionamiento y el mantenimiento de un programa de tests adaptativos es bastante más complejo. Los profesionales de la medición habrán de enfrentarse a un conjunto de desafíos de tipo aplicado. En el trabajo se discuten desafíos de cuatro tipos: establecimiento y mantenimiento de los bancos de ítems, elección de los procedimientos de administración del test, protección de la seguridad del test, y respuesta a los asuntos relacionados con las personas que responden a los tests. El éxito de un programa de tests adaptativos dependerá mucho de cómo el elaborador de tests resuelva estos desafíos.

Key words: test adaptativos informatizados, programas para test adaptativos

REFERENCES

Baghi, H., Ferrara, S. F., & Gabrys, R. (1992, April). *Student attitudes toward computer-adaptive test administrations.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17,* 283-296.

Betz, N. E., & Weiss, D. J. (1976). *Psychological effects of immediate knowledge of results and adaptive testing* (Research Report 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Blechman, E. A., & Dannemiller, E. A. (1976). Effects on performance of perceived control over noxious noise. *Motivation and Emotion, 2,* 191-200.

Buhr, D. C., & Legg, S. M. (1989, March). *Investigating the validity of a computerized adaptive test for different examinee groups.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Colton, G. D. (1998). Exam security and high-tech cheating. *The Bar Examiner, 67(3),* 13-35.

Evans, F. R., & Reilly, R. R. (1972). *A study of test speededness as a potential source of bias in the admission test for graduate study in business quantitative score. .* Princeton, NJ: Educational Testing Service.

Glass, D. C., & Singer, J. E. (1972). Behavioral aftereffects of unpredictable and uncontrollable aversive events. *American Scientist, 60*, 457-465.

Hill, K. T. (1984). Debilitating motivation and testing: A major educational problem— Possible solutions and policy implications. In R. E. Ames & C. Ames (Eds.), *Research on motivation in education* (Vol. 1, pp. 245-274). New York: Academic Press.

Ingebo, G. S. (1997). *Probability in the measure of achievement.* Chicago, IL: MESA Press.

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

Legg, S. M., & Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Education Measurement: Issues and Practice, 11(2),* 23-27.

Lord, F. M. (1970). Some test theory for tailored testing. In Holtzman, W. H. (Ed.), *Computer-assisted instruction, testing, and guidance.* New York: Harper & Row.

Lord, F. M. (1976). A broad-range test of verbal ability. *Applied Psychological Measurement, 1*, 95-100.

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

Lord, F. M. (1983). Small n justifies the Rasch model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Lord, F. M. & Novick M. R. (1968). *Statistical theory of mental test scores.* Reading, MA: Addison-Wesley.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement, 16*, 33-40.

McBride, J. R., & Weiss, D. J. (1976). Some properties of a Bayesian adaptive ability testing strategy. *Applied Psychological Measurement, 1*, 121-140.

NWEA (1997a). *Achievement Level Tests: Technical manual.* Portland: NWEA.

NWEA (1997b). *Computerized adaptive testing user's manual.* Portland: NWEA.

O'Neill, K., & Powers, D. E. (1993, April). *The performance of examinee subgroups on a computer-administered test of basic academic skills.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Perlmuter, L. C., & Monty, R. A. (1977). The importance of perceived control: Fact or fantasy? *American Scientist,65*, 759-765.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Onwuegbuzie, A. J., & Seaman, M. A. (1995). The effect of time constraints and statistics test anxiety on test performance in a statistics course. *Journal of Experimental Education, 63*, 115-124.

Stone, G. E., & Lunz, M. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. *Applied Measurement in Education, 7*, 211-222.

Stout, W. F. (1987). A non-parametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Sutton, R. E. (1997). Equity and high stakes testing: Implications for computerized testing. *Equity and Excellence in Education, 30*(1), 5-15.

Urry, V. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

Vispoel, W. P. (1998). Review and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement, 35,* 328-347.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized adaptive, and self-adapted testing. *Applied Measurement in Education, 7,* 53-59.

Vispoel, W. P., Wang, T., de la Torre, R., Bleiler, T., & Dings, J. (1992, April). *How review options and administration modes influence scores on computerized vocabulary tests.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12,* 15-20.

Wald, A. (1947). *Sequential analysis.* New York: Wiley.

Wang, M., & Wingersky, M. (1992, April). *Incorporating post-administration item response revision into CAT.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*(4), 17-27.

Weiss, D. J. (Ed.) (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Wild, C. L., Durso, R., & Rubin, D. R. (1982). Effect of increased test-taking time on test scores by ethnic group, years out of school, and sex. *Journal of Educational Measurement, 19*, 19-28.

Wise, S. L. (1996, April). *A critical analysis of the arguments for and against item review in computerized adaptive testing.* Paper presented at the annual meeting of the National Conference on Measurement in Education, New York, NY.

Zara, A. R. (1992, April). An investigation of computerized adaptive testing for demographically-diverse candidates on the national registered nurse licensure examination. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.