Practical Questions in Introducing Computerized Adaptive Testing for K-12 Assessments

Walter D. Way

April 2006

PEARSON
Educational Measurement
rr0503

Using testing and assessment to promote learning

Pearson Educational Measurement is the largest commercial processor of student assessments. We provide services and products in support of assessment programs to local and state education agencies as well as other assessment organizations and publishers. We understand how assessment activities can promote learning and benefit students, teachers, parents and schools.

PEM Research Reports provide dissemination of PEM research and assessment-related articles prior to publication. PEM reports in .pdf format may be obtained at:

http://www.pearsonassessments.com/research

Abstract

In this paper, a number of practical questions related to introducing CAT for K-12 assessments are discussed. These questions relate to four main topic areas: 1) the CAT algorithm; 2) CAT item development; 3) the CAT test delivery system; and 4) psychometrics to support CAT. In considering these questions, we describe some of the successful practices that have been used in operational high-stakes CAT programs, as well as the challenges these programs face. This discussion is aimed to assist state departments of education in considering the use of CAT as they move to transition testing programs to online delivery in the future.

Practical Questions in Introducing Computerized Adaptive Testing for K-12 Assessments

Introduction

In many states, using the computer to deliver standards-based assessments is becoming increasingly attractive to departments of education, legislators, and other policy makers. The potential advantages of online assessment include immediate score reporting, decreased administrative burdens on school district personnel, increased security of testing materials, and more flexible test scheduling. In many states, policy makers are particularly excited about the potential for efficient measurement of students' abilities through innovative test delivery models. The most popular of these is computerized adaptive testing (CAT).

The basic idea of CAT is that test items are selected by the computer to individually match the ability level of each student. In this manner, the test is tailored to each student. In fact, some of the earliest applications of CAT in the testing literature were referred to as "tailored testing" (Lord, 1970). With CAT, the computer begins by choosing an initial item, usually one at an intermediate level of difficulty, to administer to the student. If the student answers the item correctly, a more difficult item is selected next for administration. If the student's answer is incorrect, an easier item is selected next for administration. This process of selection and evaluation is carried out by the computer throughout the test. By "adapting" the difficulty level of the items selected, the computer is able to accurately measure student ability using far fewer items than a traditional paper assessment.

The selection, administration, and scoring of items in CAT depends upon an item selection algorithm. Virtually all CAT algorithms are based on item response theory (IRT), which is used to calibrate items to a common difficulty scale and to score students on a common ability scale. In most CAT applications, the algorithm selects items from an item pool consisting of several hundred items spanning a range of content and difficulty levels. The basis of CAT has been well-researched and its theoretical foundations are well established in the testing literature.

Conceptually, because CAT capitalizes on the power of the computer to deliver a more efficient test, it is a compelling alternative to consider with the introduction of online assessments. From a policy

standpoint, testing in the schools using CAT seems like an obvious improvement over traditional paper-and-pencil testing methods. However, along with the conceptual advantages, there are also a number of challenges that arise in practice when a state attempts to implement CAT as part of a high-stakes standards-based assessment program.

There is, for example, some controversy related to the use of CAT in the context of the No Child Left Behind (NCLB) Act of 2001. Specifically, the U.S. Department of Education's final regulations stipulate that all of the items administered in a CAT used for NCLB accountability purposes must be developed to measure grade-level standards. This stipulation derailed efforts in several states to introduce CAT as part of their assessment programs because item development efforts for the CAT tests they intended to use were not targeted within each grade to the specific grade-level standards (Trotter, 2003). However, this stipulation does not prohibit the use of CAT in state assessment programs for NCLB purposes, as long as items in the CAT pools developed for each grade and content are purposefully written to measure the relevant grade-level standards.

In this document, a number of practical questions related to introducing CAT for K-12 assessments are discussed. These questions relate to four main topic areas: 1) the CAT algorithm; 2) CAT item development; 3) the CAT test delivery system; and 4) psychometrics to support CAT. Table 1 lists the questions that we believe are most important for those considering the implementation of a CAT testing system. In the sections that follow, we address these questions and describe some of the successful practices that have been used in operational high-stakes CAT programs. The discussion in this document has been kept at a non-technical level, though some of the concepts under discussion are relatively complex. For those readers seeking more detailed discussions of these issues, we have provided a number of references to the CAT literature. For more extensive introductions to CAT and computer-based testing in general, the reader is referred to Wainer (2000) and Parshall et al. (2002).

Table 1. *Practical Questions in Introducing CAT for K-12 Assessments*

Questions about the CAT Algorithm

- What IRT model is being used to administer the CAT?
- How does the CAT algorithm choose items?
    - to maximize score precision
    - to meet test specifications
    - to handle items associated with common stimuli
    - to keep certain items from being over-used
- How does the CAT algorithm end the test?
- How does the CAT algorithm deal with constructed response items?
- Should the CAT algorithm allow for item review?
- How is the CAT algorithm certified as working properly prior to the test?

Questions about CAT Item Development

- How many items should be developed to support CAT?
- What is the best way to write questions to support CAT?
- How are the items developed for CAT field-tested?
- How are the items developed for CAT reviewed?
- For how long should items developed for CAT be used?

Questions about the CAT Test Delivery System

- Does the test delivery system support innovative questions?
- How secure is the test delivery system?
- How flexible is the delivery system with respect to differences in school-level computer infrastructure?
- Can the delivery system tolerate network problems and/or power outages?
- Does the delivery system provide sufficient testing accommodations?

Questions about Psychometrics in support of CAT

- Are scores on CAT comparable to scores based on paper versions of the test?
- How are CAT field-test items calibrated and linked to a common scale?
- How are the technical characteristics of CAT documented?
- What alternatives to CAT exist?

Questions about the CAT Algorithm

What IRT model is being used to administer the CAT?

There are two IRT models commonly used in high-stakes testing programs, the Rasch or one-parameter logistic model, and the three-parameter logistic (3PL) model. The Rasch model considers performance on a test item to be a function of the student's ability and the item's difficulty level. The 3PL model considers performance on a test item to be a function of the student's ability and three

characteristics of the item: difficulty, discriminating power, and the likelihood that a student of extremely low ability will answer the item correctly.[1]

If a particular IRT model has been used in the scaling and equating work for an existing paper-based testing program, it most likely makes sense to continue using that model to administer a CAT. This is especially true if an item bank has been built up over several years using a particular IRT model. However, there are differences between the Rasch and 3PL models that could affect the characteristics of CAT, and explorations of CAT should include a more detailed consideration of these models and their relative advantages and disadvantages.

How does the CAT algorithm choose items?

The CAT algorithm evaluates characteristics of questions to determine which ones are chosen and administered to students. IRT provides standard statistics by which the computer can evaluate questions based on their contributions to score accuracy. However, a CAT algorithm is also programmed to take additional considerations into account. Perhaps the most important consideration is the test content specifications. When transitioning to CAT, it is usually required that the CAT algorithm keep track of content measured by each item and choose items so that the test blueprint is met by the end of each CAT administration. Another consideration concerns the selection of items associated with stimuli. To deliver a reading test using CAT, the algorithm must choose appropriate reading passages and then select some number of questions associated with each passage. This introduces a number of complexities that the algorithm must be programmed to address. Finally, it is necessary to program controls into the CAT algorithm to ensure that items in a pool are not overexposed. There are a number of different approaches to CAT item exposure control that have been documented in the measurement literature. These vary in complexity and in how the limits to exposure are implemented (Way, 1998).

How does the CAT algorithm end the test?

CAT can be administered by giving each student the same number of items (fixed-length) or by giving each student items until a pre-specified level of score precision has been met (variable-length). Either of these approaches may make sense for a particular application, although it is easier with fixed-

[1] A detailed discussion of IRT models is beyond the scope of this document. For an introduction to IRT models that includes some interesting graphical illustrations, see Parchev (2004).

length CAT to balance the test blueprint and to establish controls over item exposure. In high-stakes K-12 assessments, there may be policy reasons to prefer fixed-length CAT as this ensures that each student receives the same number of items and that the test blueprint can be accurately reflected in each student's test.

How does the CAT algorithm deal with constructed response items?

Although CAT is typically carried out with items that are scored as either right or wrong, it is possible to include items that can be scored in multiple categories, such as essays and other constructed-response item types. There are extensions of both the Rasch and the 3PL models that can account for items scored with partial credit; these extensions can be incorporated into the CAT algorithm for item selection and scoring. However, to obtain the benefits of CAT, it must be possible for the computer to score the constructed-response items. Recently, technologies for computer scoring of essays and other item types involving text entry have become relatively commonplace (Shermis & Burstein, 2003).

Should the CAT algorithm allow for item review?

One controversial question related to the administration of CAT has to do with whether students should be permitted to review and change items that they answered previously. Early researchers working in adaptive testing argued that allowing students to go back and change previous answers would negatively affect the measurement efficiency of CAT: if an answer was changed, the items following in the sequence selected by the computer would no longer be the best ones. Some researchers also warned that students could use trickery to improve their scores on CAT by purposefully answering questions incorrectly so the computer would give the easiest questions possible, and then changing their answers to correct at the very end of the test (Wainer, 1983). For these reasons, a number of operational CAT programs do not allow students to review previous responses. However, recent research has found that allowing students to review and revise answers in CAT does not appreciably affect measurement efficiency, and strategies to obtain higher scores by trickery are extremely unlikely to benefit students (Stone & Lunz, 1994; Vispoel, Hendrickson & Bleiler, 2000).

Based on research to date, there seems to be no valid psychometric reason not to allow students the benefit of reviewing and revising their answers in CAT. Nevertheless, permitting review introduces added complexity into the CAT algorithm, and the manner in which review is allowed must be

programmed thoughtfully. For example, in some CAT applications students are allowed to review and/or change the previous five questions that they answered, and the set of five questions that are eligible for review changes as each new item is administered.

How is the CAT algorithm certified as working properly prior to the test?

CAT is a dynamic process and the expected results of the CAT algorithm should be thoroughly tested each time a CAT item pool is prepared for use. There are two complementary steps to certifying that a CAT algorithm is working properly prior to using it in an actual administration: 1) generating thousands of computer simulated CAT exams and evaluating the results; and 2) using a quality assurance team to "take" CAT exams by following specific scripts that dictate item-by-item responses, and verifying that the results match expectations. The computer simulations allow psychometricians to verify over thousands of simulated cases that the test blueprint is satisfied, that no items are used too frequently, and that the expected CAT scores are sufficiently reliable. The quality assurance team confirms that all aspects of the CAT administration software and item-selection algorithm are working properly at the level of selected individual tests.

A critical component of this certification process is to use the same executable software component for the CAT algorithm in both the computer simulations and the delivery system. This permits psychometricians to prepare item-by-item scripts that show each item that should be administered by the CAT algorithm when the responses generated in the computer simulations are entered by a person taking the real CAT. Note that CAT algorithms often involve a random component. For example, in some CAT algorithms a random number is generated before each item is selected and is used to determine which of several eligible items is actually administered. In this case, the random number generator used in the computer simulations must work identically to the random number generator used by the CAT algorithm in the operational test.

Questions about CAT Item Development

How many items should be developed to support CAT?

Based on research on testing programs such as the GRE® and SAT®, Stocking (1994) concluded an item pool equal to about 12 times the length of a fixed-length CAT was adequate for a variety of content areas and test structures. For example, a fixed-length CAT of 30 items would be supported by a pool of about 360 items, a fixed-length CAT of 40 items would be supported by a pool of about 480 items, and so forth. Stocking's research did not address variable-length CAT, although an average pool size equal to 12 times the average number of items administered in a variable-length CAT would likely be sufficient.

Although the number of items that are psychometrically sufficient for a CAT pool is an important consideration, an equally important question is this: how many CAT pools are necessary? For a large state testing program, perhaps 200,000 or more students in a particular grade might take a CAT in a certain content area during a spring testing cycle. Most likely, these CAT exams would have to be administered over a period of time, perhaps over several weeks or even several months. With sufficient item exposure controls, assume that most items in an appropriately sized CAT pool would be seen by 10 percent or less of the students testing during that time. This sounds reasonably good, except that 10 percent of 200,000 is still 20,000 students! Because items in a CAT item pool are used again and again over time, there is an unavoidable vulnerability to organized efforts to memorize and share items. For admissions testing programs that introduced CAT in the 1990s, such as the GRE and GMAT, maintaining item pools to support continuous CAT programs quickly became extremely burdensome and difficult (Davey & Nering, 2002).

Although the stakes associated with standards-based state assessment programs are undoubtedly high, it does not seem likely that the excessive item pool maintenance that goes on in some CAT environments is necessary in K-12 testing. Nevertheless, for a spring testing window, it would seem prudent for a state to have at least two CAT pools prepared for a given grade and content area: one primary pool that could be used during the testing window, and a second pool that could be held in reserve in case of a security breach and/or to be used for retesting students who initially fail tests used for graduation or promotion.

What is the best way to write items to support CAT?

To best support CAT, test developers must think differently about item development. For a traditional test, most item writing is targeted at middle difficulty items, that is, items that can be answered correctly by 60 to 70 percent of the student population. In general, few extremely difficult or extremely easy items are written, and often these items are rejected for statistical reasons. With CAT, items should be written to uniformly cover a wide range of difficulties so that high and low ability students can be measured accurately. Depending upon the specificity of the grade-level standards to which the items are being written, this may be quite challenging. In addition, approaches to writing items that are associated with reading passages need to be reconsidered. For example, a traditional approach to writing items associated with reading passages is to write a set of items that span a range of difficulties. However, with CAT, it may be preferable to write all of the items associated with one passage to be relatively difficult, and all of the items associated with another passage to be relatively easy. This way, a high ability student will be given the passage with the difficult items, the low ability student will be given the passage with the easy items, and neither student wastes time answering questions associated with a passage that are not appropriate for their ability.

How are the items developed for CAT field tested?

One of the advantages of CAT is that it is easy to embed field test items. Thus, any CAT algorithm used with a K-12 testing program should be able to routinely administer field test items that do not count toward the student's score. Often, field testing within CAT is done using a pool of field-test items. If the items in the test are discrete, the field test items can be randomly selected from the pool and administered in a systematic fashion, for example, as every Nth item in the test. However, for content areas such as reading, in which items are associated with passages, field test item administration becomes more complicated. Because the numbers of items associated with the operational passages may differ from student to student, the appropriate points in the adaptive test where the field-test passages and field-test questions are to be administered may also differ. In this case, the computer algorithm must be sophisticated enough to recognize these transition points in each student's test.

How are the items developed for CAT reviewed?

An extensive process of item and test review is a critical component of any high quality testing program. With CAT, test development quality control procedures are nearly the same as those followed in a traditional testing program. However, one critical difference in test development for CAT versus traditional test development occurs at the final phase of test assembly. With a traditional test, a content expert reviews a draft test as a whole and looks for potential flaws such as items that clue each other or items that contain redundancies in settings, problem features, or vocabulary. For example, in a math test the content expert may notice that all of the items in a draft test measuring geometric relationships happen to involve triangles. In a reading test, it may be found during this review phase that there are two passages pertaining to civil rights. The content experts would routinely flag these redundancies and make sure that replacement items or passages were found that resolved them. Content experts also tend to find more subtle reasons to replace items in the final phase of a test review that are based on their professional experience and pedagogical expertise.

With CAT, test forms are replaced by item pools that may include 300 to 400 questions. Because of the sheer number of questions involved and the dynamic nature of CAT item selection, the craft of test form review must be re-thought. In testing organizations experienced with CAT, several strategies have emerged to streamline the development and review of CAT pools. First, these organizations have developed more extensive item coding taxonomies that incorporate, in addition to the existing content standards, features of test items that content experts consider important. Similarly, coding has evolved for potential issues beyond those related strictly to content, such as the gender and ethnicity of subjects referred to in items and passages. When combined with information about item difficulty, these taxonomies facilitate analyses of the existing item bank and provide prescriptive information to guide the item writing process. These taxonomies can be used to establish specifications for forming item pools, and in some cases, become part of the specifications that determine how the CAT algorithm selects items.

A second strategy that is used in CAT item pool development to address issues such as clueing is to establish lists of items that should not be administered together. These are often referred to as "overlap" or "enemies" lists, and become part of the parameter information used by the CAT algorithm. As each item is selected for a given student, the computer checks the overlap list and excludes any enemy items from selection for the remainder of that CAT. Although overlap lists help to improve the content validity of individual CAT exams, the measurement efficiency of CAT can be severely compromised if

these lists are too extensive. Thus, it is important for content experts to use these lists only in cases where administering pairs of items together is clearly problematic.

A final strategy that is useful when CAT programs are under initial development is to use computer simulations to generate several CAT exams at different ability levels (e.g., high, medium and low), and to have content experts review them. Such reviews can help to confirm that the CAT exams are appropriately reflecting the test content, and in some cases, can identify content issues that need to be addressed in developing the CAT pools. It should be noted, however, that reviews of selected simulated CAT exams are extremely time-consuming and can become a bottleneck in the CAT development process. For this reason, they are not considered an efficient way to ensure test quality on an ongoing basis. If content reviews of simulated tests consistently find the same types of issues, there is likely a pervasive problem that should be addressed through item coding and the rules that govern how the CAT algorithm chooses items.

For how long should items developed for CAT be used?

With CAT, items are no longer thought of as being part of a particular test form, but rather are considered resources in a larger bank of items. Thus many testing agencies that sponsor CAT programs tend to use items over a long period of time, perhaps for years. This practice raises several issues. One issue is that some items tend to be selected more often than other items, and it is worth considering whether there is some point when the items used most often should be retired from use. The rationale for retiring items is the expectation that eventually, items given frequently will become known to test-takers. On the other hand, the CAT algorithm selects the most popular items for a reason, that is, these items tend to provide better information about many examinees' abilities than do the other items in the pool. If too many of these are abruptly removed from use, the efficiency of CAT may suffer.

A second issue with maintaining CAT pools over time is that items may become outdated, especially in science and social studies content areas. Because CAT pools can involve large quantities of items, it may be difficult to detect issues with item currency as pools are formed for a given CAT administration. Thus, if items remain eligible for use in a CAT over several years, formal procedures to review and confirm item currency should be introduced.

A final issue related to maintaining items for CAT is the need to disclose items to the public. Many state assessment programs require test forms to be disclosed to the public after they are administered. With CAT, it is no longer feasible to disclose test forms as they are given, since each student receives a different test form. To satisfy disclosure needs, one option would be to disclose the entire operational CAT pool once a testing window was completed. Of course, such a disclosure policy would be very expensive to maintain, as significant item-development efforts would be necessary to continuously replenish the disclosed pools. A less elaborate disclosure policy would be to release sample CAT exams chosen at different ability levels. This would serve to illustrate the content of the CAT exams, but it would not be possible to tie this content to individual or class level performance. To some extent, the value of disclosed tests as tools for teachers to use in diagnosing student weaknesses is diminished because the item analysis statistics that classroom teachers typically look at, such as percent correct values, are not interpretable in a CAT context. With CAT, if items are appropriately targeted to the ability level of students, high ability students should receive more difficult items and low ability students should receive relatively easier items. In theory, every item given with CAT should be correctly answered by the about one-half of the students to whom that item is administered, regardless of how difficult or easy the item would be if given to all of the students.[2]

Questions about the CAT Delivery System

Does the test delivery system support innovative questions?

A CAT delivery system should be capable of administering not only multiple-choice questions but also less common item types such as text formula input, clicking on "hot spots", graphing item types, and text entry and drawing solutions. In addition, various tools should be available to assist students in answering items as needed, such as rulers, protractors, calculators, highlighting capability, and drawing tools. Finally, the CAT delivery system should be capable of calling items or performance tasks that contain applets or compiled executables.

How secure is the CAT test delivery system?

[2] This statement is not exactly accurate when the 3PL model is used for CAT administration, since the inclusion of a guessing parameter leads to the expectation that more than one-half of the students will correctly answer each item.

The CAT test delivery system should be secure on multiple levels. The security of test items should be maintained through encryption routines, and transmissions of data between a central server and computers in the schools should be monitored and logged. In this manner, full audit and acknowledgement tracking accompany data transmissions. Access to the system at the schools should be secure and password protected. In addition, sessions should be proctored and access to content should be possible only by authorized personnel and only during testing windows. As part of the administration, desktop access should be frozen so that it is not possible for students to access the internet or other desktop tools that might aid them.

How flexible is the CAT delivery system with respect to differences in school-level computer infrastructure?

The CAT delivery system should be flexible enough to support a variety of school platforms and infrastructures so that schools with older equipment and networking capabilities can still administer CAT efficiently. In addition, the response of the delivery system with respect to scoring a response, selecting the next item, and presenting the item on the screen should be within a specified time requirement so that students are not frustrated by having to wait for the computer to respond to their actions.

Can the delivery system tolerate network problems and/or power outages?

The CAT delivery system should be capable of restarting from network disruptions or power outages by returning to the same item that was displayed prior to the disruption. In addition, the system should be capable of recovering all responses made by students to previous items once power is restored.

Does the delivery system provide sufficient testing accommodations?

Many testing accommodations, such as large print and audio administration, may be easier to implement by computer. Other accommodations, such as screen readers, text-to-speech conversion programs, or special input devices, may be more difficult to provide. In general, a CAT delivery system should be capable of implementing a variety of testing accommodations, provided that the accommodated students are given the opportunity to familiarize themselves with the computer environment. If the assessment is given via CAT and by paper, it may be necessary to provide students with special accommodations in paper until online versions of the accommodations have been established.

Questions about Psychometrics in Support of CAT

Are scores on CAT comparable to scores based on paper versions of the test?

Although several state testing programs are moving their assessments online, they are finding that invariably some districts and/or schools are not prepared to test their students online. For this reason, states must be prepared to administer both online and paper versions of their assessments. Because of the high stakes nature of these testing programs, it is important that results based on the assessments be comparable, regardless of the mode under which testing occurs. Professional testing standards clearly state that evidence of comparability between CAT and paper versions of a test should be documented (AERA, APA, NCME, 1999, Standard 4.10).

Score comparability between CAT and paper test versions is challenging to establish, not only because of potential differences in performance when items are given online rather than on paper, but also because of differences arising from the adaptive capability of CAT tests versus the non-adaptive nature of paper-and-pencil administrations (Wang & Kolen, 2001). In some programs where CAT has been introduced, researchers have attempted to separate these two issues (Schaeffer, Steffen, Golub-Smith, Mills & Durso, 1995; Eignor, Way & Amoss, 1994). In other CAT programs, comparability between CAT and paper versions of a test has been established using score adjustments similar to those used in test equating (Segall, 1993; Eignor, 1993).

How are CAT field-test items calibrated and linked to a common scale?

Although field testing items as part of a CAT administration is relatively straightforward, the process of calibrating CAT field-test items and linking these items to a common scale can present a challenge because the operational CAT data are not evenly distributed across students but rather are tailored to each individual student based on that student's ability level. To overcome this challenge, one solution is to embed items that have been previously calibrated into field-test positions and to use these items to link the other field-test items to the existing IRT scale.

How are the technical characteristics of CAT tests documented?

In preparing to implement CAT, a number of design decisions must be made. These include, but may not be limited to, decisions about how the first item in the CAT is to be selected, how the CAT will end (e.g., after a certain number of items are administered or after a certain level of precision is reached), how content specifications will be satisfied, and how the exposure of items will be controlled. Professional standards state that the rationales and supporting evidence for these various decisions should be documented (AERA, APA, NCME, 1999, Standard 3.12). A method commonly used for making these design decisions is through CAT simulations (Eignor, Stocking, Way & Steffen, 1993). Using CAT simulations, different alternatives for various design elements can be evaluated and compared. Once design decisions are finalized, CAT simulations can document the expected characteristics of the CAT administration, and these expectations can be compared with the empirical CAT results.

What alternatives to CAT exist?

CAT represents one of many ways to design a test for computer administration. In contrast to CAT, *linear* computer-based testing involves simply delivering a traditional test form by computer. A variation on linear computer-based testing is to deliver computer-based tests by randomly choosing a set of questions from a larger pool of items. In some applications of this approach, a sophisticated item selection algorithm is used to balance characteristics similar to those considered by a CAT algorithm (e.g., content and item difficulty). Still other applications of computer-based testing use collections of items called "testlets" (Wainer & Kiely, 1987). In one application, called Computer Adaptive Sequential Testing (CAST), the computer starts the exam by assigning a testlet of middle difficulty and, based on the student's score on this testlet, assigning a second testlet that is either easier, more difficult, or of the same difficulty (Luecht & Nungester, 1998).

It is important for policy makers and other stakeholders in state assessment programs to realize that alternative online delivery models may be appropriate, either instead of or in addition to CAT in some cases. For example, English language arts tests often involve items associated with passages, essays, and short answer items. For such tests, CAT may not be the most appropriate choice for online administration.

In general, it makes sense for a state assessment program to consider both CAT and other online testing models as it explores options for online assessment. Ultimately, the administration of tests online

may differ by content area and grade level, and initial choices for online administration may be revisited as innovative items and performance tasks are developed that more fully exploit the power of the computer.

Conclusion

This paper highlights several of the practical questions that state testing programs need to consider in exploring the use of CAT for their assessments. These questions relate to the CAT algorithm, CAT item development, desirable features of a CAT delivery system, and the psychometric support of CAT. The past 15 years have seen a steady increase in the use of CAT and other online testing applications in a variety of testing settings. An enormous amount of research on CAT has been conducted and many lessons have been learned from operational experience. The evolution of computer technology in the schools has reached a point where effective online state assessments are achievable. CAT will play an important role in these online programs; strong psychometrics must also play a major role if professional testing standards are to be maintained.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, NCME) (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Davey, T., & Nering, M. (2002). Controlling item exposure & maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments*. Hillsdale, NJ: Lawrence Erlbaum.

Eignor, D. R. (1993). *Deriving comparable scores for computer adaptive and conventional tests: An example using the SAT* (Research Report 93-55). Princeton, NJ: Educational Testing Service.

Eignor, D. R., Way, W. D., & Amoss, K. E. (1994). *Establishing the comparability of the NCLEX using CAT with traditional NCLEX examinations*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation* (RR-93-56). Princeton, NJ: Educational Testing Service.

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

Luecht, R. M. & Nungester, R. J. (1998). Some practical applications of computer- adaptive sequential testing. *Journal of Educational Measurement, 35,* 229-240.

Parchev, I. (2004). *A visual guide to item response theory*. Retrieved November 9, 2004 from http://www2.uni-jena.de/svw/metheval/irt/VisualIRT.pdf.

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

Schaeffer, G. A., Steffen, M., Golub-Smith, M.L., Mills, C.N., & Durso, R. (1995). *The introduction and comparability of the computer adaptive GRE general test* (GRE Board Professional Report No. 88-08aP). Princeton, NJ: Educational Testing Service.

Segall, D. O. (1993). *Score equating verification analyses of the CAT-ASVAB.* Briefing presented to the Defense Advisory Committee on Military Personnel Testing. Williamsburg, VA.

Shermis, M. D., & Burstein, J. C., (Eds.). (2003). *Automated essay scoring: A cross- disciplinary perspective*. Hillsdale, NJ: Erlbaum.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Report RR-94-5). Princeton NJ: Educational Testing Service.

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.

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

Vispoel, W. P., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computer adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement*, *37*, 21-38.

Wainer, H., (Ed.) (2000). *Computerized adaptive testing: A primer* (2nd Edition). Hillsdale, NJ: Erlbaum.

Wainer, H. (1983). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, *12*, 15-20.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185-201.

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement*, *38* (1), 19-49.

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