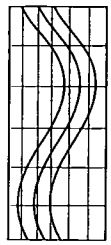
Sireci, S. G. (2004). Computerized-adaptive testing: An introduction. In J. Wall and G. Walz (Eds.), Measuring up: Assessment issues for teachers, counselors, and administrators (pp. 685-694), Greensboro, NC: CAPS Press.



Chapter 48
Computerized Adaptive Testing
An Introduction

Stephen G. Sireci

Computers are revolutionizing almost every aspect of our society and testing is no exception. Delivering tests on a computer often improves exam security, testing efficiency, and scoring, and it often allows for measurement of knowledge, skills, and abilities that cannot be measured using traditional assessment formats (Zenisky & Sireci, in press). One of the most widely cited benefits of computer-based testing is the ability to use the computer to tailor the test to specific characteristics of an examinee. In this chapter, I provide a brief overview of this type of adaptive testing, focusing on the issues most relevant to teachers, counselors, and administrators. Readers interested in more comprehensive or technical writings in this area are referred to Drasgow and Olson-Buchanan (1999); Hambleton, Zaal, and Pieters (1991); Mills, Pontenza, and Fremer (2002); Parshall, Spray, & Kalohn (2001); Sands, Waters, & McBride (1997); van der Linden and Glas (2000); and Wainer (2000).

The notion of adaptive testing dates back to the original academic screening tests developed by Binet in 1908. The Binet scales were designed to identify schoolchildren who were not likely to benefit from the typical educational system. Knowing that students who were unable to answer an easy question were unlikely to be able to answer a difficult one, Binet arranged the administration of test items in ascending order of difficulty and used different stopping rules for ending the test session based on a student's patterns of responses. This notion of adapting the test administration to the proficiency[1] level of a student carried over into contemporary intelligence tests that are individually administered (e.g., Stanford-Binet tests, Wechsler scales). Adaptive testing was not logistically feasible in large-scale assessment until the advent of the computer, however.

Computerized adaptive testing is a test administration system that uses the computer to select and deliver test items to examinees. These tests are called *adaptive* because the computer selects the items to be

administered to a specific examinee based, in part, on the examinee's proficiency on previous items. Unlike many traditional tests where all examinees take the same form, the computer adapts or tailors the exam to each examinee. This tailoring is done by keeping track of an examinee's performance on each test question and using this information to select the next item to be administered. The criteria for selecting the next item to be administered are complex, but the primary criterion is a desire to match the difficulty of the item to the examinee's current estimated proficiency. Presently, there are numerous examples of computerized adaptive testing programs, including the ACCUPLACER postsecondary placement exams, the Graduate Record Exam, and several licensure and certification exams.

The idea of using the computer to match the difficulty of an item to the proficiency of an examinee was initially proposed by Lord (e.g., Lord, 1980). Lord's idea was to begin a test administration by presenting an item of moderate difficulty. If the examinee answered the question correctly, a slightly more difficult item was administered. If the examinee answered the question incorrectly, a slightly easier question was administered. This iterative process continued until a sufficient number of items had been administered for confident estimation of the examinee's score.

How Computerized Adaptive Testing Works

The adaptive nature of a computerized adaptive test (CAT) stems from the procedure used to select the items to be administered to an examinee. This procedure is often referred to as the *item selection algorithm.* As described previously, a key goal of the algorithm is to match item difficulty to examinee proficiency. Obviously, the proficiency level of an examinee is not known at the time of testing. Therefore, estimates of examinee's proficiency must be used throughout the test session. At the beginning of the test, the proficiency estimate is typically set just below the average of the population of all test takers. (This estimate is usually selected based on extensive pretesting of the examinee population.) A value slightly below the average is used to reduce the chance that the first item on the test will be particularly difficult for an examinee. After each response to an item, the proficiency estimate for the examinee is updated.

The statistical model underlying computerized adaptive testing is item response theory (IRT). IRT posits several mathematical models that characterize items and examinees on a common scale. In IRT, the

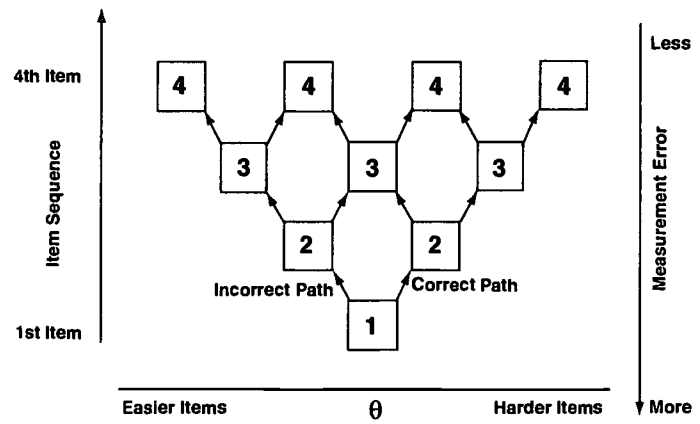
scale that indicates the difficulty of an item is the same scale that is used to assign scores to examinees. Thus, an item of average difficulty would have the same value on the scale as the value assigned to an examinee of average proficiency. There are several attractive features of IRT, including the ability to provide scores on a common scale for examinees who take different items, which is par for the course in computerized adaptive testing. The details of IRT are beyond the scope of this chapter, but several excellent textbooks on IRT are available (e.g., Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). Suffice it to say that several different types of IRT models are available and all have strengths and weaknesses in particular testing applications.

Using IRT in adaptive testing, an examinee's proficiency estimate is updated each time he or she answers a test item, and a new item is selected based on the updated estimate. When the proficiency estimate is calculated, an estimate of the amount of uncertainty in the estimate (i.e., an estimate of the error of measurement) is also calculated. As progressively more items are administered, the degree of uncertainty diminishes. Figure 1 presents a simplified example of how a traditional CAT works. The horizontal axis in this figure represents the item difficulty/examinee proficiency scale, which is typically denoted using the Greek letter theta (θ).

The vertical axis represents the sequence of test items administered. As one moves from left to right, the items become more difficult. As is evident from the figure, answering an item correctly results in the administration of a more difficult item, and answering an item incorrectly results in the administration of an easier item.

Figure 1. Illustration of a traditional computerized-adaptive test

Arrows pointing to the left indicate the item administered after an incorrect answer and arrows pointing to the right indicate the item administered after a correct answer.



There are several different methods for ending a computerized adaptive testing session. In some situations, fixed-length CATs are used, where all examinees are administered the same number of items, regardless of the measurement error associated with their score. However, many CATs use a variable-length procedure in which the test session ends when some pre-specified level of measurement precision is reached. Test stopping rules for variable-length CATs typically use one of two methods, depending on the testing context. In a norm-referenced context, where no performance standards are set on the test, a minimum standard error criterion is typically used. In this situation, an examinee's test ends when the measurement error associated with her or his score dips below a pre-specified level (Lord, 1980). This criterion ensures that the scores for all examinees meet a minimum standard of reliability. In criterion-referenced testing situations, such as in licensure or certification testing, a test session ends when it is clear that an examinee's proficiency is above or below a specific threshold, such as a passing score (Lewis & Sheehan, 1990). This criterion minimizes measurement error at specific cut scores, which increases the reliability of classification decisions made on the basis of test scores.

In addition to matching item difficulty to examinee proficiency and determining when a test ends, a CAT item selection algorithm also may control several other factors including content representation and item exposure. *Content representation* refers to the ability of the algorithm to ensure that the content specifications of the test are adhered to for each examinee. For example, if the content specifications for a ninth-grade social studies test require that 30 percent of the test items measure history, 25 percent measure geography, 25 percent measure economics, and 20 percent measure sociology, the algorithm can keep track of the content designations of each item to ensure these content specifications are met for all examinees. The algorithm can also keep track of how often an item is administered to make sure that item exposure levels do not get too high. If the same items were administered too often to examinees, knowledge of specific items may be relayed to future test takers, which would inflate their scores. Thus, the item selection algorithm is critically important for ensuring testing efficiency, content validity, and item security.

Benefits of Computerized Adaptive Testing

Many of the benefits of computerized adaptive testing stem from the fact that the administration of the test is computerized. The benefits of computer-administered tests include more flexible test administration schedules, improved test security, instantaneous scoring and score reporting, and inclusion of multimedia in the assessment (e.g., audio, video, and three-dimensional graphics). Given appropriate computerized infrastructures such as secure local area networks for storing items and examinee responses electronically, test security is increased because there are no test booklets that can be lost or stolen before, during, or after test administration. In addition, the computer can keep track of how often an item is administered so that coaching courses and others that try to "beat the test" will not be able to reproduce test questions.

In addition to the practical benefits that arise from computerization of a test, computerized adaptive testing offers improved testing efficiency, which means we can obtain confident estimates of examinees' performance using fewer items than are typically required on nonadaptive tests. This gain in efficiency stems directly from the CAT item selection algorithm, which avoids administering items that are too easy or too difficult for an examinee. Therefore, CATs are often significantly shorter than their paper-and-pencil counterparts—typically about half as long as a parallel nonadaptive test (Wainer, 1993). This reduction in testing time is appreciated by examinees, as well as by teachers and counselors who hate to lose valuable instructional or counseling time.

Another widely cited benefit of computerized adaptive testing is a reduction in test anxiety for many examinees (Gershon & Bergstrom, 1991). In traditional testing, some examinees may freeze when presented with an item that is much too difficult for them to answer. Such examinees may find taking an adaptive test less anxiety-provoking. Recent research suggests, however, that a reduction in test anxiety due to the adaptive nature of the test may apply only to examinees of relatively low proficiency (Wise, 1996).

The benefits of computerized adaptive testing explain its growing prevalence in educational and psychological assessment. Test administrators and examinees like it because it reduces testing time and allows for instantaneous score reporting. Psychometricians and test developers like it because it provides precise scores for examinees using far fewer items than are required using traditional testing formats,

which is important in terms of minimizing item exposure and potentially lowering the costs associated with developing new items. Given these benefits, we can expect to see its prevalence increase in the future. There are some problems and limitations with computerized adaptive testing, however, which may restrict its applicability in some situations.

Limitations of Computerized Adaptive Testing

Although there are many positive features of CATs, there are some disadvantages and limitations as well. A disadvantage for many testing agencies is the increased cost of developing and administering a test on a computer. Computer programs must be written to select, administer, and score items; large banks of items must be created to have many items available at all proficiency levels; and computerized testing centers must be leased or acquired to administer the tests. Each of these activities involves substantial investment of money and personnel, which can be daunting in many testing situations.

Another limitation of computerized adaptive testing is the inability to review test forms in advance of test administration. In paper-and-pencil testing, committees of content experts and sensitivity reviewers can evaluate test forms for their appropriateness for all examinees. Such reviews are more difficult in computerized adaptive testing because there is no single form of the exam.

Perhaps the most serious criticism of computerized adaptive testing is that examinees are typically not allowed to skip test questions or go back and review items answered previously. These actions are common in paper-based testing, but because the item selection algorithm in a CAT needs an examinee response to a previous question to select the next question, these behaviors can affect accurate proficiency estimation. In fact, Wainer (1993) pointed out that if examinees are allowed to skip and change answers to questions, they may be able to "trick" the algorithm into administering them the easiest possible set of test questions and subsequently bias their scores upward.

Other limitations of CATs pertain to their reliance on the computer. If schools and other organizations are unable to secure adequate numbers of appropriate computers for test administrations, CATs and other computer-based tests are not an option. In addition, in some situations, examinees' computer proficiency may interact with the construct being measured, such that examinees who are more familiar with computers do better on the test compared with examinees who have equal competence in the subject matter tested but are less familiar with

computers (Huff & Sireci, 2001).

Although CATs have their weaknesses, many testing agencies weigh the pros and cons of computerized adaptive and nonadaptive tests and conclude that the strengths of CATs outweigh their limitations. Others seek a compromise between a traditional CAT and a nonadaptive test. These compromises, such as testlet-based testing and computerized multistage testing, are discussed in the next section.

The Future of Computerized Adaptive Testing

Presently, there is increased interest and activity in testing, with most states administering high-stakes tests to students in grades K–12 (Linn, 2000). Recent federal mandates such as the No Child Left Behind legislation and the evaluation requirements for federally funded programs suggest that testing activities will increase substantially over the foreseeable future. Given this increase in testing and a desire to reduce testing time, computerized adaptive testing is likely to become more popular in our schools.

A relatively recent development in the computerized adaptive testing world is the idea of using the computer to administer pre-assembled sets of items, rather than a single item, to an examinee. Wainer and Kiley (1987) introduced the concept of a *testlet* to describe a subset of items, or a "mini-test," that could be used in an adaptive testing environment (Wainer & Lewis, 1990). Examples of testlets include sets of items that are associated with a common reading passage or graphic, or a carefully constructed subset of items that mirrors the overall content specifications for a test. After the examinee completes the testlet, the computer scores the items within it and chooses the next testlet to be administered. Thus, this type of test is adaptive at the testlet level rather than at the item level. This approach allows for better control over exam content and can allow examinees to skip, review, and change answers within a block of test items.

A variation of the testlet CAT model is computerized *multistage testing*. Multistage testing refers to the administration of several testlets in an adaptive, sequential fashion. At the first stage, examinees are administered a *routing test* that determines the difficulty level of the test they will take at the second stage. Their performance on the second stage of the test determines the test they will take at the third stage, and so on. The difference between a testlet CAT and a multistage test is that with the latter the mini-tests administered at each stage can be much larger than a typical testlet, and the number of stages is relatively small,

with two or three stages being most common. Both testlet CATs and multistage tests offer a compromise between the traditional nonadaptive format and computerized adaptive testing. Content experts and sensitivity reviewers can review the testlets to evaluate content quality; examinees can skip, review, and change answers to questions within a testlet or stage; and their responses are still used to tailor the remaining portions of the test to their specific proficiency level.

Another potential area in schools where computerized adaptive testing may become particularly beneficial is by tailoring the test to examinee characteristics other than proficiency. For example, information gained from a student's individualized education program could be used to select an appropriate starting point or sets of questions to be administered. The computer could also access different language versions of test directions or test questions for students with limited proficiency in the dominant language used in a school district. The computer could also be used to address test speededness issues by selecting for some students items that require less time to answer. Finally, one other way in which computerized adaptive testing may help teachers and counselors is by providing enhanced information about examinee performance that could be used for diagnostic and instructional purposes. For example, information regarding the amount of time taken to answer an item could be used to assess the strategies examinees used to answer the item.

Conclusion

In this chapter I attempted to provide a basic overview of computerized adaptive testing. This type of testing, or a variant of it, is gaining popularity at a rapid rate and is likely to become more prevalent in educational and psychological testing. I hope that reading this chapter gave you a general understanding of how computerized adaptive testing works and how to explain this type of test to students, parents, and those who make test-selection decisions. For those readers who want to gain a more complete understanding about the specifics of how such tests work, I highly recommend the references provided in the first paragraph of this chapter, most of which are textbooks. Computerized adaptive testing represents the most sophisticated test administration technology that psychometrics currently has to offer. It will remain an attractive testing model for the foreseeable future.

Notes

The author thanks Mary Pitoniak and April Zenisky for their helpful comments on an earlier version of this chapter.

1. In the context of assessment, the term *proficiency* refers to the knowledge, skills, and abilities a student possesses with respect to the construct being measured by the test.

References

Drasgow, F., & Olson-Buchanan, J. B. (Eds.). (1999). *Innovations in computerized assessment.* Mahwah, NJ: Erlbaum.

Gershon, R. C., & Bergstrom, B. (1991, April). *Individual differences in computer adaptive testing: Anxiety, computer literacy, and satisfaction.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Hambleton, R. K., & Swaminathan, H. R. (1985). *Item response theory: Principles and applications.* Hingham, MA: Kluwer.

Hambleton, R. K., Swaminathan, H. R., & Rogers, J. (1991). *Fundamentals of item response theory.* Thousand Oaks, CA: Sage.

Hambleton, R. K., Zaal, J. N, & Pieters, P. (1991). Computerized adaptive testing: Theory, applications, and standards. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 341–366). Norwell, MA: Kluwer.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16–25.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computer mastery test. *Applied Psychological Measurement, 14,* 367–386.

✦Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

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

Mills, C. N., Potenza, M. T., & Fremer, J. J. (2002). *Computer-based testing: Building the foundation for future assessments.* Mahwah, NJ: Erlbaum.

Parshall, C., Spray, J. A., & Kalohn, J. C. (2001). *Practical considerations in computer-based testing.* Springer Verlag.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.

van der Linden, W. J., & Glas, C. (Eds.). (2000). *Computer-adaptive testing: Theory and practice.* Boston, MA: Kluwer.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12*(1), 15–20.

Wainer, H. (2000). *Computerized-adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Erlbaum.

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

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1–14.

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 Council on Measurement in Education, New York.

Zenisky, A. L., & Sireci, S. G. (in press). Technological innovations in performance assessment for licensure and certification exams. *Applied Measurement in Education.*

✦ Document is included in the Anthology of Assessment Resources CD