Georgiadou, E., Triantafillou, E. & Economides, A.A.: **Evaluation parameters for computer adaptive testing.** *British Journal of Educational Technology*, Vol. 37, No 2, pp. 261–278, March 2006.

Evaluation parameters for computer-adaptive testing

Elisabeth Georgiadou, Evangelos Triantafillou, and Anastasios A. Economides

Dr. Georgiadou Elisabeth is a teacher at Vocational Education and researcher at University of Macedonia, Thessaloniki, Greece. She is working on the field of educational hypermedia systems and in particular the design and evaluation of these systems. Her research has been presented at a number of international conferences. Dr. Evangelos Triantafillou is a teacher at Vocational Education and researcher in the Department of Informatics, Aristotle University of Thessaloniki. His research interests include Educational Technology, Multimedia Educational Technology, and Adaptive Hypermedia Systems. He has published several papers in international scientific journals. Dr. Anastasios Economides is an Assistant Professor and Vice-Chairman in the Information Systems Postgraduate Program at the University of Macedonia, Thessaloniki, Greece. His current research interests include mobile, collaborative and adaptive learning and networks. Address for correspondence: Dr. Anastasios A. Economides, Department of Informatics, University of Macedonia, Egnatia 156, Thessaloniki 54006, Greece. Email: economid@uom.gr

Abstract

With the proliferation of computers in test delivery today, adaptive testing has become quite popular, especially when examinees must be classified into two categories (pass/fail, master/nonmaster). Several well-established organisations have provided standards and guidelines for the design and evaluation of educational and psychological testing. The purpose of this paper was not to repeat the guidelines and standards that exist in the literature but to identify and discuss the main evaluation parameters for a computer-adaptive test (CAT). A number of parameters should be taken into account when evaluating CAT. Key parameters include utility, validity, reliability, satisfaction, usability, reporting, administration, security, and thoseassociated with adaptivity, item pool, and psychometric theory. These parameters are presented and discussed below and form a proposed evaluation model, Evaluation Model of Computer-Adaptive Testing.

Introduction

Testing is directly related to education and training as a way to measure the performance levels of students. Since education was established as an institution, different methods of assessment have been used in different contexts, such as class presentations, essays, projects, practicum, etc. However, the most common tools of measuring performance are the oral test and the paper-and-pencil test. Given that the computer has been an educational tool in the last few decades and that its use has spread rapidly in all levels of education and training, the use of computer-based tests (CBTs), as a result, has increased significantly over the last few years. Still, its use is limited, as there is a tendency to associate CBT with automated multiple-choice questions and not as a tool that can enrich students' learning experiences.

However, there are a number of perceived benefits in using computers for assessing performance, which Harvey and Mogey (1999) record as:

- Large numbers can be marked quickly and accurately.
- Students' response can be monitored.
- Assessment can be offered in an open-access environment.
- Assessments can be stored and reused.
- Immediate feedback can be given.
- Assessment items can be randomly selected to provide a different paper to each student.

Another benefit of CBTs would be to bring the assessment environment closer to the learning environment. Software tools and web-based sources are frequently used to support the learning process, so it seems reasonable to use similar computer-based technologies in the assessment process (Baklavas, Economides & Roumeliotis, 1999; Lilley & Barker, 2002).

The most common type of a CBTs is the linear, fixed-length computerised assessment that presents the same number of questions to each examinee in a specified order, and the score usually depends on the number of questions answered correctly. A linear test consists of a full range of easy and difficult test questions that are either randomly selected from a large pool of questions or are the same for all examinees. In some cases, the examinees can select a set of questions to be answered from a larger set, for example, 10 out of 15 questions.

Evidently, the type of CBTs described here imitates a paper-and-pencil test that is presented in a digital format and gives little or no attention to the ability of each individual examinee. Consequently, the examinees can be presented with questions that are either too easy or too difficult. Correctly answering the easy questions (or items) or incorrectly answering the difficult ones can provide information regarding a classification decision such as pass/fail, good/very good/excellent. However, it cannot give enough information about the level of performance for each individual. By contrast, in a computer-adaptive test (CAT), a special case of computer-based testing, the items that are administered to the examinees are tailored to the individual examinee's ability on the construct being measured.

CAT: an overview

CAT tailors itself to the ability of the examinee, as there is a different way on the selection and presentation of the items. The items presented are selected taking into account the examinee's individual performance during the test, or in other words, how each examinee answered previous items. If the examinee correctly answers the item presented, then a more difficult one is presented next. On the other hand, if the examinee incorrectly answers the item, an easier item is presented next. This way, low-ability examinees will be presented with relatively easy items, while high-ability ones will be presented with more difficult items. The score will be determined from the level of the difficulty, and as a result, while all examinees may answer the same percentage of questions correctly, the high-ability ones will get a better score as they answer more difficult items correctly.

During the assessment, the examinees might feel discouraged if the items are too difficult or, on the other hand, might lose interest if the items are too easy. Because the items in a CAT are selected in an interactive way, it works like a good quality oral assessment where the examiner has the opportunity to change the item's difficulty level constantly according to the examinee's performance.

First, the CAT presents an item of moderate difficulty to assess initially each individual's level. During the test, each answer is scored immediately, and if the examinees answer correctly, then the test statistically estimates their ability as higher and then presents an item that matches this higher ability. If the next item is again answered correctly, it reestimates the ability as still higher and presents the next item to match the new ability estimate. The opposite occurs if the item is answered incorrectly. The computer continuously reevaluates the ability of the examinee until the accuracy of the estimate reaches a statistically acceptable level or when some limit is reached such as when a maximum number of test items are presented. The items presented to the examinee are included in the item pool (or item bank), a collection of test items with a full range of proficiency levels.

There are many advantages recorded in the literature with regard to CAT (Linacre, 2000; Rudner, 1998) such as flexibility of test management, immediate availability of scores, increased test security, increased motivation, etc. However, the main advantage of CAT over a linear CBTs is efficiency. Because fewer questions are needed to achieve a statistically acceptable level of accuracy, significantly less time is needed to administer a CAT compared with a linear CBTs. CAT can reduce testing time by more than 50% while maintaining the same level of reliability.

The calculation of the score for CAT is mainly based on the principles of the Item Response Theory (IRT) (Hambleton, Swamination & Rogers, 1991; Lord, 1980; Wainer, 1990). The central elements of IRT are mathematical functions that calculate the probability of a specific examinee answering a particular item correctly. In IRT, examinees can be described by a set of one or more ability scores that are predictive, through mathematical models, linking actual performance on test items, item statistics, and examinee abilities (Lilley & Barker, 2002).

Brief history of CAT

The original idea of adaptive testing belongs to French psychologist Alfred Binet (1859–1911) who wanted to find a way to measure the ability to think and reason apart from education in any particular field. In 1905, he developed a test in which he made children do tasks such as follow commands, copy patterns, name objects, and put things in order or arrange them properly. The test began with questions suitable for each child's age, continued with questions that advanced up the age scale, and ended when the child failed frequently. From Binet's work, the term 'intelligence quotient', or 'IQ', entered the vocabulary.

Several methods were based on Binet's work, such as Lord's (1980) Flexilevel testing procedure, Lewis and Sheehan's (1990) Testlets, etc. In general, the procedures on the above methods determine the general ability level of the student within the first few test questions. Then, based on item-response statistics, the computer or the tutor selects items that are suitable for the student's particular level and administers those items to get a more precise estimate of the student's ability level. This strategy eliminates the need for students to answer numerous questions that are too difficult or too easy for them.

With the proliferation of computers in test delivery today, adaptive testing has become quite popular, especially when examinees must be classified into two categories (pass/fail, master/nonmaster). Several recognised testing programmes such as Graduate Record Exam, Graduate Management Admission Test, Scholastic Aptitude Test, Microsoft's qualifications, etc have adopted adaptive testing as their current method for testing (Giouroglou & Economides, 2004).

Evaluation parameters

Several organisations have provided standards and guidelines for the design and evaluation of educational and psychological testing. The American Council on Education (ACE) published in 1995 the *Guidelines for Computerized Adaptive Test Development and Use in Education* (ACE, 1995). In 1999, the *Standards for Educational and Psychological Testing* was published and adopted by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The *Standards* were published to provide criteria for the evaluation of tests, testing practices, and the effects of test use.

> Although the evaluation of the appropriateness of a test or testing application should depend heavily on professional judgment, the Standards provide a frame of reference to assure that relevant issues are addressed. It is hoped that all professional test publishers will adopt the Standards and encourage others to do so. (AERA, APA, NCME, 1999).

Moreover, in 2000, the Association of Test Publishers (ATP), in the face of the rapid growth of computer-based testing, published the *Guidelines for Computer-Based Testing* (ATP, 2000) with the intention to supplement, extend, and elaborate the *Standards for Educational and Psychological Testing* as they apply to computer-based and Internet-based testing and assessment. In 2004, the International Test Commission (ITC) published the *International Guidelines on Computer-Based and Internet Delivered Testing* (ITC, 2004) addressed to test developers, publishers, and users.

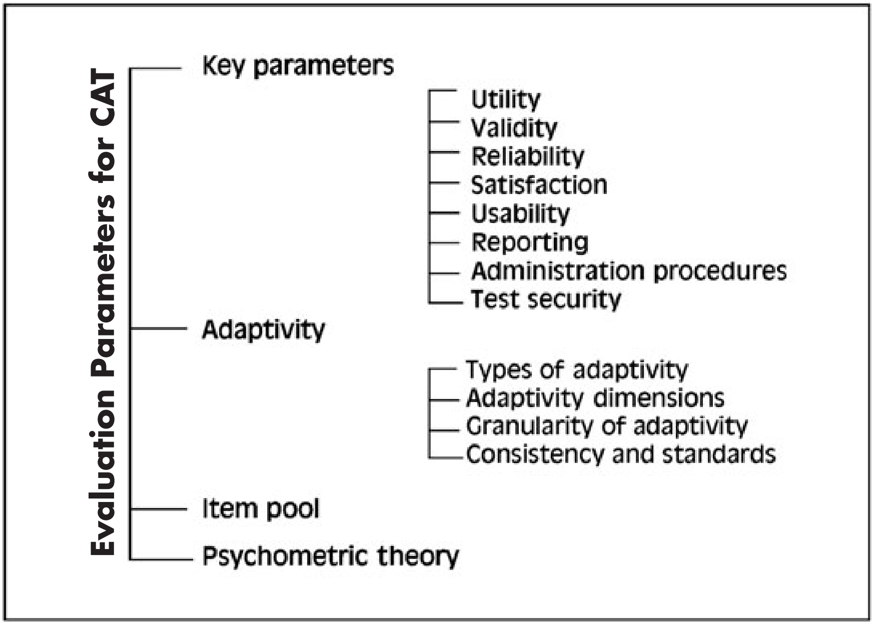

Figure 1: Evaluation model of computer-adaptive testing

The purpose of this paper was not to repeat the guidelines and standards that exist in the literature but to identify and discuss the main evaluation parameters for a CAT. In his article *Questions to Ask When Evaluating Tests*, Rudner (1994) made an effort to report the basic standards applicable to most test evaluation situations. This paper will extend Rudner's work, as there are many differences between a CAT and any other type of test, with adaptivity as the most important.

A number of parameters should be taken into account when evaluating CAT. Key parameters include utility, validity, reliability, satisfaction, usability, reporting, administration, security, and those associated with adaptivity, item pool, and psychometric theory. These parameters are presented and discussed below and form a proposed evaluation model, Evaluation Model of Computer-Adaptive Testing, that is illustrated in Figure 1.

Key parameters

Utility is an important parameter to consider when evaluating a CAT and is related mainly to the test purpose. Computerised systems are not always the best option for testing; for example, when the examinations involve essay writing, then any computerised test is not applicable.

CAT is an important testing instrument in:

- identifying whether the examinee has met the specific objectives of a course;
- indicating the examinee's level of achievement in a skill domain;
- identifying specific areas in which a student needs additional educational experiences;
- diagnosing the examinee's skill area strengths and weaknesses; and
- detecting whether candidates have met minimum course requirements as demonstrated in a mastery test (Dunkel, 1999).

CAT developers need to state clearly the assessment purpose and to ensure that the test is able to measure the examinee's true professional level. To achieve this goal, a CAT must provide examinees with a broad range of content areas and skill tasks—depending on the subject matter—to ensure that their proficiency level is measured properly. Because examinees may be of high- or low-proficiency levels, the CAT must be designed in such a way that it provides adequate assessment for the entire range of ability represented in the examinee population (Green, Bock, Humphreys, Linn & Reckase, 1984). In other words, the item pool must include items that correspond to the entire range of ability, which is not identified only by difficulty level but also from a variety of designated tasks depending always on the subject matter.

Validity refers to whether a test actually measures what it is supposed to measure. It relates to the aptness of the inferences made on the basis of the test scores. The three major conceptions of validity are content validity, criterion validity, and construct validity (American Psychological Association, 1985).

Content validity involves inspecting the test to see whether the items included are valid for testing purposes. For example, a valid mathematical test for the notion of addition would probably contain addition exercises. If the test contains more multiplication exercises than addition ones, then its validity is questionable. Moreover, content validity is concerned with sample-population representativeness; for example, the knowledge and skills covered by the test items should be representative of the larger domain of knowledge and skills.

Criterion validity refers to how the test compares with another criterion measure that can be, for example, some other aspects of achievement such as grade-point average. This kind of validity is determined by examining the statistical correlation between the test and the criterion measure (high positive, +1.0, to low, −1.0). For an acceptable criterion validity, the correlation should be positive and high.

Construct validity is the degree to which a test measures the theoretical construct that it intends to measure. A variety of statistical techniques may be used to see it the test behaves in ways predicted by the given construct. For example, a new test of computer programming skills would be expected to correlate highly with other valid tests of computer skills. Conversely, there would also be the expectation that this new test would have little correlation with a different type of test, such as a test of social intelligence.

Reliability refers to the precision and consistency of scores derived from a test instrument. Therefore, if a test is reliable, the person will achieve the same score on repeated testing plus or minus acceptable measurement error. Also, reliability is a function of general, situational, and individual factors that can be used to frame evaluative questions for the developers of the test.

General factors influencing reliability include clear and explicit instructions for the examinees and familiarity of the examinees with the CAT format before taking it. Situational factors are related to the testing environment such as noise level. Finally, individual factors include transient and stable factors such as the physical and psychological health of the examinees and the examinees' experience with similar tests. Consequently, only a test of high reliability is useful for making decisions about an examinee (Dunkel, 1999).

There are three types of reliability: test–retest, internal consistency, and interrater reliability. Test–retest reliability is measured by looking at whether an examinee receives an equivalent score on the same CAT at two different points in time. In other words, the examinee takes the test twice, and the scores are correlated to determine whether each examinee performs about the same in both administrations of the test.

Internal consistency reliability is measured by looking at the statistical relationship among items from a single CAT. If all the items are supposed to measure the same construct, then there should be a strong correlation among the items. For example, for a given examinee and item-difficulty level, getting one item correct means that it is likely that the other related items also will be answered correctly.

Interrater reliability is the degree to which the measuring instrument generates similar results at the same time with more than one evaluator.

The satisfaction parameter concerns with the examinee's behaviour towards the CAT in terms of its interface and its overall function. With regard to the interface, the issues associated with the examinee's satisfaction are the terminology used, the feedback given by the system, the help and documentation provided, and screen design issues.

Screen design is very important, as different screen elements should be used to present stimulating information that will motivate and assist the users in retaining and recalling information. The psychological limitations to consider when designing hypermedia learning systems—CAT can be considered as such—include: (1) memory load (ie, how many different control icons are reasonable for learners to remember at any one time?), (2) perception (ie, what colours and fonts provide the best readability?), and (3) attention (ie, how can the users' attention be drawn to information that is relevant when there is a lot of different information on the screen?) (Preece, 1993). A large number of screen design guidelines produced from several researchers on educational technology exist in the literature (Clarke, 1992; Cox & Walker, 1993; Mcatter

& Shaw, 1995; Morris, Owen & Fraser, 1994) that a CAT designer should take into consideration.

However, the most important is that screens should be designed as clear and as self-explanatory as possible. The overall satisfaction with the software has to do with this, as it concerns whether the CAT is simple, clear, easy, and pleasant to use so that the examinees can focus on answering the items and not on trying to understand how the system works.

Satisfaction is closely related to the usability of the system. The usability of a CAT can be measured (1) by having a small set of evaluators examine the interface and by judging its compliance with recognised usability principles and (2) by observing and measuring the end-users' attitudes. It is important for most of the evaluation to occur before the administration of the test (formative evaluation) to judge the strengths and weaknesses of the test in its developing stages, for the purpose of revising, to improve its effectiveness and appeal. Therefore, the observation and measurement of the end-users' attitudes can form the summative evaluation to determine the value or worth of the test, usually compared with another form of test (eg, linear CBT, paper and pencil).

One of the most important evaluation studies for measuring usability is the heuristic evaluation suggested by Nielsen (Molich & Nielsen, 1990; Nielsen, 1994). In this study, these are 10 general principles for user-interface design. They are called 'heuristics' because they are more of the nature of rules of thumb than of specific usability guidelines. These are as follows:

- Visibility of the system status: The system should always keep users informed about what is going on through appropriate feedback within reasonable time.
- Match between the system and the real world: The system should speak the users' language with words, phrases, and concepts familiar to the user rather than with system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
- User control and freedom: Users often choose system functions by mistake and will need a clearly marked 'emergency exit' to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
- Consistency and standards: Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
- Error prevention: Even better than good error messages is a careful design, which prevents a problem from occurring in the first place.
- Recognition rather than recall: Make objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for using the system should be visible or easily retrievable whenever appropriate.
- Flexibility and efficiency of use: Accelerators—unseen by the novice user—may often speed up the interaction for the expert user, such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

- Aesthetic and minimalist design: Dialogues should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
- Help users recognise, diagnose, and recover from errors: Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
- Help and documentation: Although it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

The above heuristics apply to the design and evaluation of CATs except maybe for the 'user control and freedom' and the 'flexibility and efficiency of use' heuristic. A CAT allows limited user control and freedom, as each item is scored after it is answered, and the score determines the next question in the adaptive sequence. It is not possible for the examinee to review and change the answers and as a result, undoes and redoes are not allowed. Moreover, a CAT test is flexible and efficient as a testing procedure compared with paper-and-pencil tests, but it is not designed to provide unlimited flexibility and efficiency for the examinee, as testing is adapted interactively from the programme to match the ability level of the examinee by means of a statistical method. Therefore, the examinee cannot speed up the interaction, as this is controlled by the system. Also, despite the fact that the items presented are adapted to the level of the ability of the user, the way in which the information is presented in most CATs is the same for all items regardless of their difficulty level. As a result, the interface cannot be adjusted according to individual user characteristics.

Yet, learners' motivation is increased when they control the navigation of a hypermedia environment. Therefore, a CAT could be designed in a way that will provide more learner control. For example, the examinee could have the option to choose between two or more items of the same difficulty level at any time. This feature, however, requires a very large item bank, as more than one item is exposed at the same time. Moreover, when the system adapts to dimensions other than knowledge level, such as cognitive style, learning style, etc, learner control is essential so that the examinees can select the appropriate type of items (eg, text or visual items).

Reporting refers to the methods used to report test results. Currently, most of the CATs return test results accurately, quickly, and by using different methods for the same test such as scaled scores, subtest results, combined test results, etc. However, a CAT should provide resources that would help examiners interpret the results and most importantly, throughout this interpretation, to detect the students' educational needs. Therefore, for each examinee and/or for all statistical purposes, a test report should include: (1) the results of any particular item; (2) the results of the items included in a particular unit; (3) the results of similar difficulty level items; (4) the wrong answers; and (5) the time consumed, etc.

Administration procedures

Most systems are case based: they have been developed for a particular cognitive subject. It is difficult to change the cognitive subject without knowing the system and without having programming skills. An ideal adaptive test should be independent from the cognitive subject and should allow teachers to organise tests with few steps through relatively easy screens. It is therefore important for examiners to have an authoring system with a simple and easily navigable interface that would help them to add or alter questions to the question pool. In addition, test authors should be able to edit or delete assessments, sections, and items.

Furthermore, basic administration procedures would permit the teacher to determine the entry level (the test-entry procedure), the selection of the items from the item pool (item-selection procedure), the stopping rules (test-termination procedure), as well as the way of scoring (test-scoring procedure). The test-entry procedure refers to the selection of the first item in CAT. To be as accurate as possible, this selection should be based on knowledge level, learning style, and psychometric theory. All subsequent item selection is based on student performance (Wise & Kingsbury, 2000).

Test security

Security is a basic factor that should not be ignored. If a test is not secure, it loses reliability and validity regardless of how well it is designed and developed. Test security starts during the development of test questions, continues through the distribution of the test, and culminates in the administration and scoring of each test.

Courseware packages now offer the features of timed quizzes and quizzes where the instructor only allows the students the opportunity to answer each question once. Although these features certainly improve test security, they are not perfect. To further enhance test security, some courseware packages allow tracking via the Internet protocol address of the student. However, it has to be mentioned that in cases when the CAT is administered via the Internet, it is not possible to know for certain the identity of the examinee. Some support the view that in traditional paper-and-pencil tests also, examiners are not sure of the identity of the examinees unless they ask for a picture ID, and not many do so.

Currently, most systems are web based. In this case, a client-server architecture is essential, and communication between client and server can be achieved basically by three different ways: (1) Hyper Text Mark-up Language (HTML) forms based on the client's side and a Common Gateway Interface (CGI) on the server side; (2) JAVA client or HTML on the client's side and JAVA Servlets on the server side; and (3) dynamic communication by using Active Server Page (ASP) or Hypertext Preprocessor (PHP) technology.

The CGI-based approach does not offer persistent communication channels between a server and a client. Matching between a client and a corresponding persistent application process would have to be managed by cookies, which cannot be regarded as an ideal solution with respect to security. Furthermore, an expert user can obtain the answers by viewing the code source of the HTML page. Therefore, it is preferable to use JAVA/ASP/PHP technology.

Adaptivity

CAT is closely related to adaptive hypermedia. Adaptive hypermedia is an alternative to the traditional 'one-size-fits-all' approach in the development of hypermedia systems. Currently, there are a large number of hypermedia environments covering a wide range of topics. A hypermedia environment is considered to be a flexible instructional environment in which all learning needs can be addressed (Ayersman & Minden, 1995). Many researchers have been working to construct sophisticated hypermedia systems, which can identify the user's interests, preferences, and needs and give some appropriate advice to the user throughout the learning process. Adaptive Hypermedia was introduced as one possible solution. Adaptive Hypermedia Systems (AHS) combine hypermedia systems with Intelligent Tutoring Systems to adapt web-based educational material to particular users. AHS build a model of the goals, preferences, and knowledge of each individual user and use this model throughout the interaction with the user to adapt to the needs of that user (Brusilovsky, 1996).

Adaptive hypermedia techniques can be useful to solve a number of problems associated with the use of CAT. As discussed, the most important feature in CAT's function is the knowledge level of the examinee on a particular subject. The knowledge of different users can vary greatly, and the knowledge of a particular user can grow quite fast. The same question can be unclear for a novice and, at the same time, trivial and boring for an advanced learner. In an adaptive test, questions are selected from a repository that is based on the learner's present skill level and that gets updated when the learner attempts the question. The difficulty level of the question also gets updated in accordance to the learner's response. The test continues until some termination criterion is reached. The test result is based on the level of difficulty of the questions correctly answered by the learner. Summarising and authoring CAT requires additional considerations because of the adaptive question selection and skill level estimation and of adaptively updating the question's difficulty level. This paper attempts to highlight a number of parameters that should be taken into account when developing or evaluating CAT, such as types of adaptivity, dimensions, consistency, and standards.

Types of adaptivity

Different authors in different areas have used the term adaptivity. Brusilovsky (2001) defines an AHS as one that builds a model of goals, preferences, and knowledge for every user. It uses this model, by means of interactions, to adapt to the needs of the user. Oppermann, Rashev, and Kinshuk (1997) takes the characteristics of the user into account and distinguishes between Adaptable Systems and Adaptive Systems. De Bra (1999) takes into account user preferences as a variable that decides the adaptation and classifies the hypermedia environments or websites built according to their capacity to carry out some type of personalisation in Adaptable Hypermedia, Adaptive Hypermedia, and Dynamic Hypermedia.

After examining the development of AHS, the authors classify adaptive systems in three types with regard to adaptivity: (1) an adaptive system; (2) an adaptable system; and (3) a combination of the adaptive and adaptable system. Adaptive hypermedia refers to hypermedia systems that rely on a user model to support the delivery of user-specific content. On the other hand, in adaptable systems, users select from a variety of parameters to adapt hypermedia to their needs. A problem with adaptive systems in general is that they might make wrong adaptations based on guesses that they make about the user (Kay, 1994). It is therefore of crucial importance to allow the user to control the adaptivity and to alter the assumptions made by the system. Furthermore, the combination of the adaptive and the adaptable system is suitable in situations where users cannot customise effectively on their own and as a result, need assistance from the system. This approach looks the most promising, because it can provide support for novices, help users to customise from the beginning of the test and to be selective about what they customise, and finally help users maintain over time the interfaces that result from the customisation.

Adaptivity dimensions

AHS make it possible to deliver 'personalised' views or versions of a hypermedia document without requiring any kind of programming by the author(s). Also, although it is possible to offer users a way to initialise the user model through a questionnaire, an AHS can do all the adaptation automatically simply by observing the browsing behaviour of the user.

Traditionally, the adaptation decision in adaptive systems was based on taking into account various user characteristics represented in the user model such as user goals/tasks, knowledge, background, hyperspace experience, and preferences. That was true for pre-1996 AHS. Currently, the situation is different. A number of adaptive web-based systems are able to adapt to something else other than user characteristics such as adaptation to the user's environment.

Adaptation to the user's environment is a new kind of adaptation that was brought by web-based systems. Because users of the same server-side web application can reside virtually everywhere and can use different equipment, the adaptation to the user's environment has become an important issue. A number of current AHS suggested some techniques to adapt to both the user location and the user platform. Simple adaptation to the platform (hardware, software, network bandwidth) usually involves selecting the type of material and media (ie, still picture vs. movie) to present the content. More advanced technologies can provide considerably different interfaces to the users with different platforms and can even use platform limitation to the benefits of user modeling (Brusilovsky, 2001).

Furthermore, currently adaptive systems can be developed to accommodate various learner needs and is the ideal way to accommodate a variety of individual differences including learning style and cognitive style (Ayersman & Minden, 1995; Triantafillou, Pomportsis & Georgiadou, 2002). In the ideal educational environment, a tutor with instructional experience on a learning domain can identify the students' individual differences with regard to cognitive styles and acquired knowledge and thus, can provide them with a learning material that is individually selected and structured. In order to simulate, in a sense, an ideal educational environment, an AHS should provide learners with the ability to use different instructional modes to accommodate their individual needs and to improve their performance. Therefore, it has to include in its design both issues of cognitive style and teaching strategy. Teaching strategy refers to the instructional material and the instructional strategy.

Currently, only a few systems have been developed with regard to cognitive and learning styles. Many authors use the terms 'cognitive' and 'learning style' interchangeably. However, there is a difference between their uses. Cognitive style deals with the 'form' of cognitive activity (ie, thinking, perceiving, remembering), not its content. Learning style, on the other hand, is seen as a broader construct, which includes cognitive along with affective and psychological styles. Systems like Intelligent System for Personalized Instruction in a Remote Environment (Grigoriadou, Papanikolaou, Kornilakis & Magoulas, 2001), CS388 (Carver, Howard & Lavelle, 1996), and Incorporating Learning Strategies in Hypermedia (Bajraktarevic, Hall & Fullick, 2003) provide adaptation to various learning styles. A good example of adaptation in an educational system with regard to cognitive styles is Adaptive Education System based on Cognitive Style (AES-CS) (Triantafillou, Demetriadis, Pomportsis & Georgiadou, 2004). AEC-CS is based on the field dependent/field independent cognitive styles. The system uses navigational support tools (concept map, graphic path, advance organiser) and adaptive presentation techniques. Users are provided with instructional strategies that suit their cognitive preferred style with an option to switch it to a nonpreferred version.

Granularity of adaptivity

Current AHS provide adaptation based on a stereotypical user model with limited levels of user differentiation. However, the complexity of the learner's profile stresses that a different approach of the user model should be considered. AHS must incorporate multiple dimensions of the user including expertise, user goals, interests, preferred learning, and cognitive style. These dimensions may be declared by the user and/or measured by the adaptive system.

The proper type and number of dimensions remains an open research issue. Adding additional dimensions will not always increase the accuracy of the user model but will always increase the complexity of the user model and the requirements to collect additional user information. There is a balance between the number of dimensions, model complexity, and the accuracy of the model. Techniques for modifying the weights associated with different dimensions dynamically to better represent the user are open research issues.

In addition to using multiple dimensions of the user, AHS must incorporate multiple levels in each user-model dimension. Users are not just novice, intermediate, or expert users but are ranged in a scale of many intermediate values. Users are not simply analytical or global learners but instead are some combination of both characteristics. According to Carver, Hill and Pooch (1999), the AHS should not only model multiple dimensions of the user, but each dimension should have as much delineation as necessary to truly model the user.

Consistency and standards

Consistency and standards is one of the heuristic evaluations identified by Nielsen (1994) discussed in the usability section above. In terms of adaptivity, consistency and standards are very important. The adaptive behaviour should not alter the usage of the system, allowing learners to use the same interaction approach regardless of the adaptivity features and/or the learning model.

Item pool

The most important element of a CAT is the item pool (or item bank), which is a collection of test items that includes a full range of levels of proficiency from which varying sets of items are presented to the examinees. For every item, the item pool includes the question text, details on the correct answer(s), and the difficulty level. More detailed item pools also include information on the content area that each item belongs, instructional grade level, author details and also history of the item development, use, and recalibration. 'Although item pools are critical to the proper functioning of CATs no specific procedures have been identified for developing the specifications for a CAT item pool' (Reckase, 2003).

Each item in a pool must be assigned a range of difficulty that will dictate the score received by the examinee for a correct answer and also provide a cue to move on to the next question in the sequence or to end the test. There are several factors that affect the end of a CAT test, which are called the stopping rules. A test is finished when (Linacre, 2000) (1) the item bank is exhausted; (2) the maximum test length is reached; (3) the ability measure is estimated with sufficient precision or, in other words, a specified standard error of measurement is less than the specified value set by the administrator of the test; and (4) the test taker is exhibiting off-test behaviour such as responding too quickly or too slowly. Moreover, the CAT test cannot stop before (1) a minimum number of items have been given;, (2) every test topic area has been covered;, and (3) sufficient items have been administered to maintain test validity under challenge or review.

The success of any CAT programme is largely dependent on the quality of the item pool that can be conceptualised according to two basic criteria: (1) the total number of items in the pool must be sufficient to supply informative items throughout a testing session; and (2) the items in the pool must have characteristics that provide adequate information at the proficiency levels that are of greatest interest to the test developer. This criterion mainly means that at all important levels of proficiency, there are sufficient numbers of items whose difficulty parameters provide valuable information. Therefore, a high-quality item pool will include sufficient numbers of useful items that allow efficient, informative testing at important levels of proficiency (Wise, 1997). Moreover, the item pool must be sufficiently large to ensure that test items do not reappear with a frequency sufficient to allow the examinees to memorise them. The size and depth of the item pool from which individual items are drawn strongly affect the validity and utility of the resulting CAT scores.

With regard to the first criterion, the size of the item pool depends on the intended purpose and characteristics of the tests being constructed. Weiss (1985) argues that satisfactory implementations of CAT have been obtained with an item pool of 100 high-quality, well-distributed items. Moreover, properly constructed item pools with 150–200 items are preferred. However, in cases when the administrator of the test wants to minimise item exposure or when maximum content balance is required or administered a very high stakes examination, then the item pool should be larger.

Concerning the second criterion, unless the developed items have a distribution of difficulties that is reasonably matched to the important levels of proficiency, there will likely be areas of proficiency in which the test information provided by the CAT will accumulate at a too slow rate. In the sense of the pool analogy, the item pool will be too 'shallow' in some proficiency regions. In these regions, the CAT will be less efficient, resulting in either higher standard errors of proficiency estimation (for a fixed-length CAT) or a longer test being required (for a variable-length CAT) to reach a desired level of precision. An obvious solution to this problem is for additional pool items to be developed that provide additional information (depth) where it is most needed. The pool depth issue becomes more complicated when the pool is subdivided into a number of content domains, each of which must be represented to a prespecified degree in the CAT. Ideally, each content domain should exhibit a distribution of item difficulties that resembles that of the entire pool (Wise, 1997).

Moreover, another issue of major importance is the item exposure control that depends on the item selection method and on the testing algorithm of the CAT (see Eggen & Straetmans, 1996). The exposure control deals with the problem that in a CAT, pool items may be used either too often (overexposure) or too infrequent (underutilisation). Overexposure may jeopardise the confidentiality of items. If an item is presented too often, the examinees may become familiar with it and may prepare for it, particularly in a high-stakes test. This would result in a decrease in the item's actual difficulty, which in turn would positively bias proficiency estimation. On the other hand, underutilisation of an item is a waste of time and energy spent on the development of an item pool.

Psychometric theory

CAT should be based on a robust underlying psychometric theory. Psychometric theory is the psychological theory or technique of mental measurement, which is the base for understanding general testing theory and methods. A specific type of psychometric theory that assigns mathematical equations to individual test-taker proficiencies and test-item parameters, such as difficulty, and so that predictions can be made about 'the probability of that person getting the item correct' is called IRT (Wainer, 2000).

The calculation of the score for CAT is mainly based on the principles of the IRT (Hambleton *et al*, 1991; Lord, 1980; Wainer, 1990), yet item response theory is not applicable to all skills and item types, as it is fairly complex and relies on several fairly restrictive assumptions (Rudner, 1998, 2002). Decision theory is an alternative theory for sequential testing, which according to Rudner 'is very attractive due to its simplicity, wide acceptance in many fields, lack of assumptions, robustness and computational ease' (Rudner, 2002).

In general, IRT is a statistical framework in which examinees can be described by a set of one or more ability scores that are predictive, through mathematical models, linking actual performance on test items, item statistics, and examinee abilities. IRT begins with the proposition that an individual's response to a specific test item or question is determined by an unobserved mental attribute of the individual. Each of these underlying attributes, most often referred to as latent traits or abilities, is assumed to vary continuously along a single dimension usually denoted θ. Under IRT, both the test items and the individuals responding to them are arrayed on θ from lowest to highest. The position of person i on θ, denoted θ_i, is usually referred to as the person's ability or proficiency. The position of item j on θ, usually denoted b_j, is termed as the item's difficulty. Intuitively, we expect the probability of a correct response to the j–θ item to increase monotonically as θ_i–b_j increases (see Rudner, 1998).

Summary

CAT is a useful tool in testing performance and shows promise in becoming one of the basic testing procedures especially in large-scale examination for licensing and certification purposes. There are numerous potentials and advantages with efficiency as the most important. However, there are limitations such as (1) CATs are not applicable to all subjects and skills; (2) hardware limitations may restrict the types of items; (3) the test administration procedures are different, and this may cause problems to some examinees; (4) with each examinee receiving a different set of questions, there can be perceived inequities; and (5) the examinees are not usually permitted to review and to change their answers.

CAT is still a research subject, as several issues need to be addressed such as CAT's effect on user performance (Jettmar & Nass, 2002); use of psychometric theory (Rudner, 1998); different approaches for different subject matters, etc. However, in terms of CAT's evaluation, there are a number of parameters involved. This paper is an attempt to summarise all these parameters in an evaluation model that test developers and administrators need to know to implement a valid and reliable CAT.

Acknowledgments

The work presented in this paper is partially funded by the General Secretariat for Research and Technology, Hellenic Republic, through the E-Learning, EL-51, FlexLearn project.

References

American Council on Education (ACE). (1995). *Guidelines for computerized adaptive test development and use in education*. Washington DC: American Council on Education.

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

American Psychological Association (1985). *Standards for Educational and Psychological Tests*. Washington DC, APA.

Association of Test Publishers (ATP). (2000). *Guidelines for computer-based testing*. Washington DC: Association of Test Publishers.

Ayersman, D. J. & Minden, A. V. (1995). Individual differences, computers, and instruction. *Computers in Human Behavior*, *11*, 3–4, 371–390.

Bajraktarevic, N., Hall, W. & Fullick, P. (2003). ILASH: Incorporating learning styles in hypermedia. In *Proceedings of the AH2003 Workshop*, Budapest, Hungary, 20 May 2003 (pp. 145–154).

Baklavas, G., Economides, A. A. & Roumeliotis, M. (1999). Evaluation and comparison of web-based testing tools. In *Proceedings WebNet-99, World Conference on WWW and Internet* (pp. 81–86). Honolulu, Hawaii: AACE, 1999.

Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *User modelling and user-adapted interaction*, *6*, 2–3, 87–129.

Brusilovsky, P. (2001). Adaptive hypermedia. *User modeling and user-adapted interaction*, *11*, 87–110.

Carver, C., Hill, M. & Pooch, U. (1999). Third generation adaptive hypermedia systems. WebNet 99, Honolulu, Hawaii, 1999.

Carver, C. A., Howard, R. A. & Lavelle, E. (1996). Enhancing student learning by incorporating learning styles into adaptive hypermedia. *Proceedings of 1996 ED-MEDIA World Conference on Educational Multimedia and Hypermedia* (pp. 118–123). Boston, MA.

Clarke, A. (1992). *The principles of screen design for computer based learning materials*. London, UK: Department Of Employment.

Cox, K. & Walker, D. (1993). *User interface design* (2nd ed.) New York: Prentice Hall.

De Bra, P. (1999). Design issues in adaptive web-site development. *Proceedings of the 2nd Workshop on Adaptive Systems and User Modelling on the WWW*, 29–39.

Dunkel, P. A. (1999). Considerations in developing and using computer-adaptive tests to assess second language proficiency. *Language Learning & Technology*, *2*, 2, 77–93.

Eggen, T. & Straetmans, G. (1996). Computerized adaptive testing for classifying examinees into three categories. *Measurement and Research Department Reports*. 96–3. Cito, Netherlands: Arnhem.

Green, B., Bock, R. D., Humphreys, L., Linn, R. & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347–360.

Giouroglou, H. & Economides, A. (2004). State-of-the-Art and adaptive open-closed items in adaptive foreign language assessment. *Proceedings 4th Conference on Information and Communications Technologies in Education, Athens 2004* (pp. 747–756).

Grigoriadou, M., Papanikolaou, K., Kornilakis, H. & Magoulas, G. (2001). INSPIRE: an intelligent system for personalized instruction in a remote environment. *Proceedings of 3rd Workshop on Adaptive Hypertext and Hypermedia*, Sonthoven, Germany (pp. 13–24).

Hambleton, R. K., Swamination, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. California: Sage Publications Inc.

Harvey, J. & Mogey, N. (1999). Pragmatic issues when integrating technology into the assessment of students. In S. Brown, P. Race & J. Bull (Eds), *Computer-assisted assessment in higher education* (pp. 7–20). London: Kogan-Page.

International Test Commission (ITC). (2004). *International guidelines on computer-based and internet delivered testing*. Draft Version 0.6, July, 2004.

Jettmar, E. & Nass C. (2002). Adaptive Testing: effects on user performance. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 20–25, Minneapolis, Minnesota, USA (pp. 129–134).
Kay, J. (1994). Lies, damned lies, and stereotypes: pragmatic approximations of users, In A. Kobsa & D. Litman (Eds), *Proceedings of the 4th International Conference on User Modeling* (pp. 73–78), Hyannis, MA: Mitre Corp.
Lewis, C. & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, *14*, 367–386.
Lilley, M. & Barker, T. (2002). The development and evaluation of a computer-adaptive testing application for English language. In *Proceedings of the 6th Computer-assisted Assessment Conference*. Loughborough University, UK.
Linacre, J. M. (2000). Computer-adaptive testing: a methodology whose time has come. MESA Memorandum No. 69. Published in S. Chae, U. Kang, E. Jeon & J. M. Linacre. *Development of computerised middle school achievement test (in Korean)*. Seoul, South Korea: Komesa Press.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. NJ: Lawrence Erlbaum Associates, Publishers.
Mcatter, E. & Shaw, R. (1995). *The design of multimedia learning programs*. Glasgow: University of Glasgow, The EMASHE Group.
Molich, R. & Nielsen, J. (March 1990). Improving a human-computer dialogue, *Communications of the ACM*, *33*, 3, 338–348.
Morris, J. M., Owen, G. S. & Fraser, M. D. (1994). Practical issues in multimedia user interface design for computer-based instruction. In S. Resman (Ed.), *Multimedia computing: preparing for the 21st century* (pp. 225–284). Harrisburg, London: Idea Group Publishing.
Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R. L. Mack (Eds), *Usability inspection methods* (pp. 25–62). New York, NY: John Wiley & Sons.
Oppermann, R., Rashev, R., Kinshuk. (1997). Adaptability and adaptativity in learning systems. In A. Behrooz (Ed.), *Knowledge transfer* Vol. 2 (pp. 173–179). London, UK: Pace.
Preece, J. (1993). Hypermedia, multimedia and human factors. In C. Latchem, J. Williamson & L. Henderson-Lancett (Eds), *Interactive multimedia: practice and promise* (pp. 135–150). London: Kogan Page.
Reckase, M. D. (2003). Item pool design for computerized adaptive testing. Annual meeting of the national council of measurement in education, Chicago, IL, April 2003.
Rudner, L. M. (1994). Questions to ask when evaluating tests. *Practical assessment, research and evaluation*, *4*, 2.
Rudner, L. M. (1998). An online, interactive, computer adaptive testing tutorial. Retrieved 12 December 2004, from http://EdRes.org/scripts/cat
Rudner, L. M. (2002). An examination of decision theory adaptive testing procedures. Annual meeting of the American educational research association, New Orleans, LA, April 1–5.
Triantafillou, E., Demetriadis S., Pombortsis, A. & Georgiadou E. (2004). The value of adaptivity based on cognitive style: an empirical study. *British Journal of Educational Technology*, *35*, 1, 95–106.
Triantafillou, E., Pomportsis, A. & Georgiadou E. (2002). AES-CS: adaptive educational system based on cognitive styles. *Proceedings of the AH2002 Workshop*, Malaga, Spain, 2002 (pp. 10–20).
Wainer, H. (1990). *Computerized adaptive testing (a primer)*. NJ: Lawrence Erlbaum Associates.
Wainer, H. (Ed.) (2000). *Computerized adaptive testing: a primer* (2nd ed.) Mahwah, NJ: Lawrence Erlbaum.
Weiss, D. J. (1985). Adaptive Testing by Computer. *Journal of Consulting and Clinical Psychology*, *53*, 6, 774–789.
Wise, S. (1997). *Examine issues in CAT*. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago.
Wise S. L. & Kingsbury G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica*, *21*, 135–155.