Economides, A. A. & Roupas, C. (2009). Evaluating computerized adaptive testing systems. In: Solutions and Innovations in Web-Based Technologies for Augmented Learning: Improved Platforms, Tools and Application, N. Karacapilidis (ed.), IGI-Global. ISBN: 978-1-60566-238-1

Chapter XI
Evaluating Computerized Adaptive Testing Systems

Anastasios A. Economides
University of Macedonia, Greece

Chrysostomos Roupas
University of Macedonia, Greece

ABSTRACT

Many educational organizations are trying to reduce the cost of exams, the workload, delay of scoring, and the human errors. Also, organizations try to increase the accuracy and efficiency of the testing. Recently, most examination organizations use Computerized Adaptive Testing (CAT) as the method for large scale testing. This chapter investigates the current state of CAT systems and identifies their strengths and weaknesses. It evaluates 10 CAT systems using an evaluation framework of 15 domains categorized into 3 dimensions: Educational, Technical and Economical. The results show that the majority of the CAT systems give priority to security, reliability, and maintainability. However, they do not offer to the examinee any advanced support and functionalities. Also, the feedback to the examinee is limited and the presentation of the items is poor. Recommendations are made in order to enhance the overall quality of a CAT system. For example, alternative multimedia items should be available so that the examinee would choose his preferred media type. Feedback could be improved by providing more information to the examinee or providing information anytime the examinee wished.

INTRODUCTION

The increasing number of students, the need for effective and fast student testing, multimedia-based testing, self-paced testing, immediate feedback, and accurate, objective and fast scoring push many organizations to use Computer-Based Testing (CBT) or Computer Assisted Assessment (CAA) tools (Brown, 1997). But this is not enough. Current learning theories lead towards student-centred and personalized learning. There is also increased interest for reducing the cheating, reducing the examinee's anxiety, challenging but not frustrating the examinees, as well as for

immediate and continuous examinee's guidance based on his knowledge, proficiency, ability and performance. Thus, many organizations are further driving towards computerized adaptive testing (CAT) tools (e.g. GMAT, GRE, MCSE, TOEFL). CAT is a special case of CBT. It is a computer-based interactive method for assessing the level of a student's knowledge, proficiency, ability or performance using questions tailored to the specific student. The CAT system selects questions from a pool of pre-calibrated items appropriate for the level of the specific student. Wainer (1990) indicated that two of the benefits of CATs over CBTs are higher efficiency and increased student motivation due to higher levels of interaction provided. CAT can estimate the student's level in a shorter time than any other testing method. CAT is based on either Item Response Theory (IRT) or Decision Theory (Welch & Frick, 1993; Wainer, 1990; Rudner, 2002). It is a valid and reliable testing method.

A CAT system tailors the test to the proficiency of the individual examinee. The CAT system adjusts the test by presenting easy questions to a low-proficiency examinee and difficult questions to a high-proficiency examinee. However, the score of each examinee depends not only on the percentage of questions answered correctly but also on the difficulty level of these questions. Even if both examinees answer the same percentage of questions correctly, the high-proficiency examinee gets a higher score because he answers correctly more difficult questions. Because each test is tailored to the individual examinee, far more information is gained from the examinee's response to each item than in conventional test (Young et al., 1996). The main advantage of a CAT is efficiency (Straetmans & Eggen, 1998). IRT-based CAT has been shown to significantly reduce testing time without sacrificing reliability of measurement (Weiss & Kingsbury, 1984). It has been shown that CAT needs fewer questions and less time than paper-and pencil tests to accurately estimate the examinee's level (Jacobson, 1993; Carlson, 1994; Wainer, 1990; Wainer et al., 2000). However, Lilley, Barker & Britton (2004) argued that the stop condition of a CAT can create a negative atmosphere amongst examinees, which could result in the rejection of the CAT altogether. Examinees might consider that the fairness of the assessment is jeopardised if the set of questions is not the same for all participants. Furthermore, examinees expressed their concern about not being able to return to review and modify previous responses. Olea et al. (2000) showed that allowing answer review decreases the examinee's anxiety, and increases the number of correct responses and the estimated ability level of the examinee. Similarly, Wise and Kingsbury (2000) pointed out that when examinees are allowed to change answers, they are more likely to decrease their anxiety improve their scores and score gains. Lilley & Barker (2003) showed that learners with different cognitive styles are not disadvantaged. Also, CAT has the potential to offer a more consistent and accurate measurement of examinee's abilities than that offered by traditional CBTs. Georgouli (2004) proposed an intelligent agent for self-assessment which adapts its material to reflect the needs of the individual learner, whether it is for studying or for testing. In addition to the examinee's achievement in the test, the system would also consider his personality characteristics (Triantafillou, 2007a). Taking into consideration the examinee's knowledge on the domain, background experience, preferences, personal data and mental model, efficient CATs would be produced (Triantafillou, 2007b).

Although major organizations develop and use CAT systems, there is little work on evaluating these systems in a comprehensive way. Most organizations performed a self-evaluation of their systems aiming at proving the validity and reliability of their CAT and their items. However, there are more parameters to consider when designing, developing or using a CAT system. Boyle & O' Hare (2003) addressed this need to evaluate educational software. As Wise and Kingsbury

(2000) stated, although CAT is a relatively simple idea, the reality of planning, implementing and maintaining a CAT program is substantially more complex. Zahorian et al. (2001) remarked that the usual online computer-based questioning systems have no built-in help, no guidance if questions are answered incorrectly, no method for selecting questions based on the students' needs, and no comprehensive monitoring of a student's progress through a knowledge map. Recently, Triantafillou et al. (2008) developed and evaluated a CAT application on mobile devices.

The objective of this paper is to evaluate contemporary CAT systems. We do not aim at comparing the CAT systems among themselves in order to find the best one. After all, each one of these has been developed for a different subject and a different purpose. Rather, we want to identify the current state of the art in this area, and discover the best characteristics and major drawbacks. Based on the results of the evaluation, we propose directions for enhancement of these CAT systems. Also, we determine best practices for designing and developing future CAT systems. In the next Section 2, we present the framework for evaluating the CAT systems. In Section 3, we present the evaluation results for the educational dimension. In Section 4, we present the evaluation results for the technical dimension. In Section 5, we present the evaluation results for the economical dimension. In Section 6, we conclude and suggest directions for improvements.

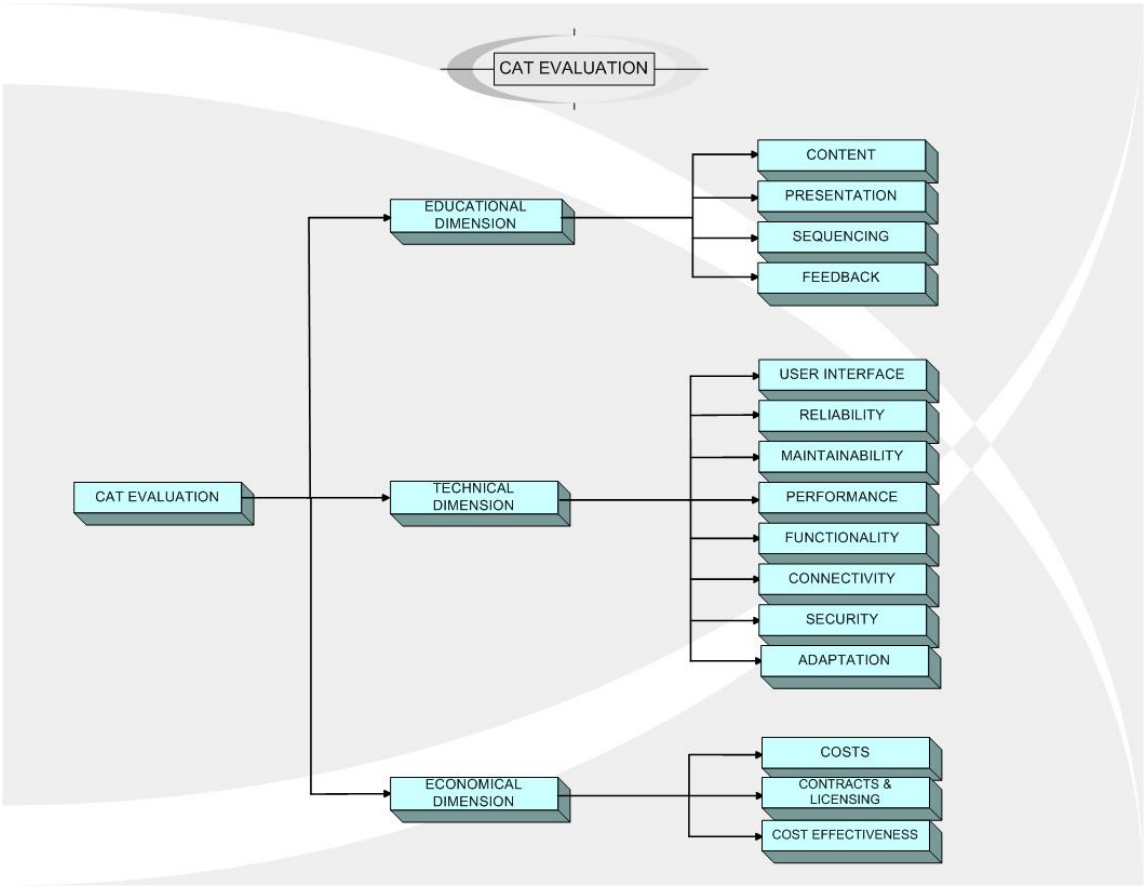
EVALUATION OF CAT SYSTEMS

Based on our previous work and experience with CAT (Baklavas et al. 1999, Giouroglou & Economides, 2004 and 2005; Economides, 2005a) we resulted to a number of CAT systems. We contacted the corresponding organizations and repeatedly asked the full version of their CAT systems. It was extremely difficult to even get an answer from some organizations. Finally, we were able to gather the following 10 CAT demos: Graduate Management Admission Test (GMAT), Graduate Record Examination (GRE), Test Of English as a Foreign Language (TOEFL), Microsoft Certified Systems Engineer (MCSE), Cisco, the Computing Technology Industry Association (CompTIA), Cito Group NT2-CAT Lezen computer adaptive test for reading Dutch as a second Language, FastTEST Pro, Maryland State Department of Education (MSDE), "An On-line Interactive Computer Adaptive Testing Tutorial by Lawrence M. Rudner". The last two systems belong to non profit organizations, while the rest to commercial ones.

We based our evaluation on CATE (Economides, 2005a). CATE (Computer Adaptive Testing Evaluation) is a framework for evaluating CAT systems across three dimensions: educational, economical and technical (Figure). The educational dimension includes the following domains: Content, Presentation, Sequencing and Feedback. The technical dimension includes the following domains: User Interface, Reliability, Maintainability, Performance, Functionality, Adaptation, Connectivity and Security. The economical dimension includes the following domains: Costs, Contract and Cost-effectiveness.

Previous studies on evaluating testing tools using specific criteria include the following. Baklavas et al. (1999) evaluated Web-based testing tools with respect to the variety of question types that support, the capabilities for multimedia use, the security, the easiness of development, maintenance and delivery of tests, the automatic grading and the statistical analysis of the results. Dunkel (1999) pointed out the importance of the appropriateness, reliability, validity and utility of CAT. Valenti et al. (2001) considered criteria for the interface, the question management, as well as the test management and implementation issues. Valenti et al. (2002) suggested the use of suitability, security, interoperability, operability, understandability, learnability and reliability in order to evaluate a computer based assessment

Figure 1. CATE domains (Economides, 2005a)



system. Sclater and Howie (2003) considered various types of users (system administrator, question author, test author, learner, marker, etc.) and propose requirements for each user type. Georgiadou et al. (2006) identified the following important parameters for a CAT: utility, validity, reliability, satisfaction, usability, reporting, administration, security; as well as parameters associated with adaptivity, item pool, and psychometric theory. Finally, Triantafillou et al. (2008) evaluated a mobile CAT application on mobile devices. The students commented among other on the clarity of the test, adaptive test procedure, results' accuracy, feedback, mobility.

CATE includes not only technical criteria for software quality, but also educational and economical. Regarding the technical quality, CAT is based on the ISO 9126 quality standard which defines six software quality characteristics: Functionality, Reliability, Usability, Efficiency, Maintainability, and Portability. However, CATE gives special attention also to Adaptation and Security, since they are extremely important in CAT systems.

Our objective was to identify the current state of CAT systems, their strengths and weaknesses. We did not aim at comparing them among themselves, since each one of them has been developed for different purpose. We (the authors) have qualitatively evaluated the 10 CAT demos taking into consideration comments from graduate students who had experienced them. For every CAT demo,

we have evaluated each one of the 15 domains (Figure). Our evaluation of each domain was qualitative based on the CATE framework.

EDUCATIONAL DIMENSION

The educational dimension consists of the following domains: 1) Content, 2) Presentation, 3) Sequencing, and 4) Feedback.

Content

First, we examined the various CAT systems regarding their Content. The Content refers to the quantity and quality of the items in the item bank. It is very important since the test is based on these items. It determines not only the test topic but also the test difficulty levels.

The content of CAT should be based and supported by currently acceptable didactic and pedagogical theories, such as: creative, explorative, active, constructive, problem solving, critical thinking learning. It should be personalized. The items should be of high quality, i.e. valid, trustworthy, correct and accurate without any errors. The item authors should possess credentials and reputation. The items should be useful, up-to-date, and will be valid for long time. They should be relevant, suitable and appropriate for the indented tests, ages and educational level of the examinees. They should objectively present a variety of "points of view" without discriminating with respect to age, gender, race, religious, political ideas etc. They should be acceptable and compatible to the examinee's language, social, cultural, racial, political, and religious values and ideas. They should adjust and support the values of the examinees and the value of learning.

The quantity of the items should be comprehensive and complete covering all main ideas and key points at the right quantity. It should also be sufficient and balanced to cover the intended topics, difficulty levels, skills and abilities to be tested. It should support various social interaction types (e.g. formal, informal), cognitive and conational types. Finally, it should be easy, time and cost efficient to develop, calibrate, manage, validate and update the items.

Regarding the 10 CAT demos, their Content is based and supported by currently acceptable didactic and pedagogical theories. The items have been validated and are accurate without any errors. Most of the systems have high quality items, which are useful, up-to-date and valid for long time. Some of the tests have technological Content so the items need to be up-to-date. Other tests examine the language skills of the examinee so the items need to be valid for a long time. In both cases the test providers stood up well to these challenges. The items are also relevant and appropriate for the indented tests. They do not discriminate with respect to age, gender, race, culture, religious, political ideas etc. In most systems, the quantity of the items is sufficient and according to the amount of the topic that the test must cover. Most of the tests are covering the main ideas and the key points of the topic. Most of the tests took extra consideration to find and wording the deceitful answers, which must have the same attractiveness, convenience and plausibility to the right answer. Many of the tests use items in which the examinee needs to solve a problem in order to answer the question. Finally, most of the systems use content balancing in order to utilize efficiently the item bank and prevent item over-exposure and under-exposure. Item exposure control strategies have been discussed in Georgiadou (2007).

The majority of the CAT systems score higher or equal to "Fair" with respect to the Content (Table). Two of them distinguish and score "Excellent": i) Graduate Management Admission Test (GMAT), and ii) Test Of English as a Foreign Language (TOEFL). GMAT covers two different topics: mathematics and language. TOEFL covers

Table 1. Distributions and average scores of contemporary CAT systems

Domain	NE	Very Poor	Poor	Fair	Good	Excellent	Average score
Content	-	0%	10%	40%	30%	20%	3,6
Presentation	-	30%	20%	40%	10%	0%	2,3
Sequencing	-	0%	10%	30%	30%	30%	3,8
Feedback	-	10%	20%	60%	10%	0%	2,7
User Interface	-	10%	0%	60%	30%	0%	3,1
Reliability	-	10%	0%	0%	40%	50%	4,2
Maintainability	-	0%	0%	30%	40%	30%	4
Performance	-	0%	10%	40%	40%	10%	3,5
Functionality	40%	40%	20%	0%	0%	0%	0,8
Connectivity	-	10%	0%	50%	30%	10%	3,1
Security	-	0%	10%	10%	30%	50%	4,2
Adaptation	-	0%	0%	40%	30%	30%	3,9
Costs	-	-	-	-	-	-	-
Contracts and Licensing	20%	0%	10%	60%	10%	0%	2,4
Cost-Effectiveness	-	0%	0%	80%	0%	20%	3,4

reading, grammar and listening comprehension. Three of the systems score "Good", four of them score "Fair", and one scores "Poor".

Presentation

Presentation refers to the presentation, media and format of the items in the CAT. The presentation, media and format of the items should be personalized. It should be clear, simple, and of low overhead. It should be rich, be based on a variety of media (e.g. text, picture, image, graphs, diagrams, audio, video, immersion) of high quality (e.g. resolution, number of colors, sound fidelity). There should be the right mix of media objects at the appropriate positions with low distraction. The result should be enjoyable.

Regarding the 10 CAT demos, their items are simple and of low media overhead. However, the Presentation with respect to multimedia is poor. The luck of pictures, images, graphs and diagrams is obvious, especially at the first splash screen where a form rich in multimedia is usually expected. The media quality is low along with the resolution. Some audio exists but only in listening comprehension.

On the other hand, this is quite expected because adding multimedia in a test will dramatically increase the size of the test in the disk and the downloading time. Moreover a new and inexperienced user prefers ease of learning rather than ease of use. This means that the examinee would prefer an interface easy to understand rather than an interface easy to use (e.g. shortcuts) (Dennis, Wixom & Tegarden, 2005). Therefore most of the tests include enough blank space and use only the necessary information in order to keep the test functional. Furthermore, the main concern of the CAT provider is to create error free software with accurate scoring that would increase their

reliability to the public, rather to focus on an attractive Presentation of the items in the CAT.

From the aesthetic point of view, most of the tests use readable fonts and never use capital letters except if they serve a purpose such as for titles. Moreover, they use colour and patterns carefully and sparingly. The tests try to provide pleasant readability and not art. So, the colour is used either to separate and categorize the items or to highlight important information (Dennis, Wixom & Tegarden, 2005). Another weakness is that the user does not have the possibility to personalize the test. In other words, he cannot change the Presentation parameters according to his personal taste.

The majority of the CAT systems score lower or equal to "Fair" with respect to the Presentation (Table). The Test Of English as a Foreign Language (TOEFL) achieves a "Good" score. It is the only CAT system that includes multimedia not only in the splash screen but also in the listening comprehension items. Three systems score "Very Poor", two systems score "Poor", and four systems score "Fair".

Sequencing

Sequencing refers to the sequencing of the items presented to the examinee. In CAT, the Sequencing of the items depends on the examinee's answers. An adaptive algorithm is employed to select the next item to be presented to the examinee. This algorithm should be based on a valid and accredited pedagogical and psychometric theory. The duration and the number of items in the CAT should be enough to produce valid results. The selected items should accurately represent the content, skills and abilities that are intended to be measured. The exposure of the items should be kept low (Georgiadou, 2007) and the test-overlap minimum. The algorithm should be easy, time and cost efficient to initiate, manage and terminate. It should be fair, non-discriminating, and consistent. It should be intuitive, logical and appropriate for the examinee. There prioritization of important items. It should enhance student's motivation and enjoyment. It should support a variety of item types, sequencing methods and scoring methods. It should support a large number of concurrent tests and examinees. It should avoid guessing and cheating. It should result to valid, reliable and error-free scores. The scores should be stable, reproducibility, and consistent.

Different allocation control levels among the examinee, the teacher and the system should be possible. For example, the examinee may have the option to overtake control over the CAT ignoring any suggestions of the system. The examinee could select the next item, skip an item, go back and alter an answer, retry an item.

Usually, the test starts with a question of average difficulty, and then proceeds to an easier or a more difficult one depending on the examinee's answer. So, a test with five levels of difficulty will have one concrete item for the first question, two concrete items for the second (an average item is not an option, because depending on the examinee's answer to the first question the second question must have an easier item or a more difficult one), three concrete items for the third question, four concrete items for the fourth question and five concrete items for the rest of the questions.

The previous algorithm predetermines the Sequencing of the items. However, some tests don't share this logic. The Sequencing is not predetermined. Each question will acquire an item from an item bank according to the question's difficulty. The items are divided into multiple levels of difficulty. For example, if the next question should be an easy one then the test will search the item bank and find all easy questions that have not been presented previously. Then, it will select randomly or according to an algorithm (e.g. information maximization) one of them.

Both algorithms are easy to initiate and fair to the examinee, because in both cases the next item is presented according to the examinee's last

answer. The second algorithm though, creates more unique tests then the first. The motivation of the examinee is high as the questions are not too difficult or too easy. The scores are stable, consistent and have fine distinctions because answering a difficult question provides a higher score than answering an easier one. Cheating is excluded but guessing is impossible to avoid for multiple choice, true/ false, etc. (A. Economides, 2005a). However, there are two serious limitations: i) the examinee cannot skip an item, and ii) the examinee cannot go back, review and change his answer to a previous item.

Regarding the 10 CAT demos, most of them score higher or equal to "Fair" with respect to the sequencing (Table). FastTEST Pro, "An On-line Interactive Computer Adaptive Testing Tutorial by Lawrence M. Rudner", and Microsoft Certified Systems Engineer (MCSE) score "Excellent". From the rest, three systems score "Good", three systems score "Fair", and one scores "Poor".

Feedback

Feedback refers to the response of the CAT system to the examinee's actions. It may aim to control, guide and regulate the examinee, or instruct and teach him, or help and support him. It may inform him about his progress, his strengths and weaknesses. It may also try to develop, enhance and improve his strengths as well as reduce and correct his weaknesses (Economides, 2005b, 2006). It is a powerful educational tool which would substantially improve the learning. Most educators and psychologists agree that instantiation and accuracy in scoring of a test helps the examinee to improve him-self and discover his weaknesses (Kapsalis, 2004). Feedback may be useful if an examinee's performance is hampered because of testing situation and not because of limited proficiency (Noijons, 1994).

The feedback to the items should be personalized. It should be timely, quality, accurate, relevant, clear and easy to understand. It should be of proper quantity, media and format. It should inform the examinee about the content, the skills and abilities to be tested, the required prerequisites, the options, the available tools and resources, the CAT method and the score. It should advise the examinee on test strategies and the use of time. It should notify the examinee on deadlines. It should provide hints on the items as well explanations on the answers. It should encourage, inspire, motivate, and stimulate the examinee. Finally, it should praise and congratulate the examinee.

There should exist a variety of support facilities (e.g. searching, communication, collaboration, sharing, glossary, dictionary, FAQ, bibliography, references, links, help, documentation). Also, various educational tools should be provided to the examinee and the teacher (e.g. designing, creating, and organizing the items, as well as monitoring, helping, evaluating, and recording the examinee) with no programming need. Finally, there should be a variety of communication and collaboration tools (e.g. e-mail, chat, videoconferencing, etc.).

Most of the 10 CAT demos satisfy some of these criteria. They may provide the examinee's final score immediately. However, they do not provide any extra information. Furthermore, they do not praise or congratulate the examinee for his effort. This cause low motivation and discourages the examinee to try harder. Without the appropriate feedback there is no improvement or progress.

There are some test strategies and instructions, mainly in the first page of the test, but there is no notification of deadlines and only few provide support facilities (e.g. frequently answered questions, dictionary, etc.) or explanations for the answers, though they inform the examinee which questions were incorrectly answered. However, this is the only information they provide. Taking everything into consideration, the quantity and quality of the feedback information is average and the lack of media is more than obvious.

Regarding the feedback, six of the CAT demos score "Fair" (Table). Two systems score "Poor", and one system scores "Very Poor". "An On-line

Interactive Computer Adaptive Testing Tutorial by Lawrence M. Rudner" gets the highest score of "Good". It presents the probability for a correct response to each item according to previous answers to this question by other users the same time that the item is presented. At the end, it provides information about the response of the user (correct or incorrect), the true score of each item, the item difficulty and the estimated ability.

We assigned the following scores: "Very Poor"= 1, "Poor"= 2, "Fair"= 3, "Good"= 4, and "Excellent"= 5. Then, the average scores are presented at the last column of the Table. In the educational dimension domains, the CAT demos score above average regarding the Content and the Sequencing. However, they fail regarding the Presentation and the Feedback. Designers and developers of CAT systems should not overlook Presentation and Feedback. Rather, they should put effort to improve these domains.

TECHNICAL DIMENSION

The technical dimension consists of the following domains: 1) User Interface, 2) Reliability, 3) Maintainability, 4) Performance, 5) Functionality, 6) Connectivity, 7) Security, and 8) Adaptation.

User Interface

The User Interface is the aggregate of input and output means by which the examinees interact with the CAT system. It includes the graphical, textual and auditory information the CAT system presents to the examinee, and the control sequences (e.g. keystrokes with the computer keyboard, movements of the computer mouse, and selections with the touch screen) the examinee employs to interact with the CAT system. The design of a User Interface affects the amount of effort the examinee must expend to provide input for the system and to interpret the output of the system, and how much effort it takes to learn how to do this. Usability is the degree to which the design of a particular User Interface takes into account the human psychology and physiology of the examinees, and makes the process of using the system effective, efficient and satisfying. Usability is the capability of the CAT system to be understood, learned, used and attractive to the examinee. The less effort the examinee needs to understand and learn the CAT system's operation, as well to use it, the better. Also, the more the CAT system catches the examinee's attention the better.

Most of the CAT systems have a friendly User Interface. It is important not to overload an examinee under pressure. As it has already been mentioned in the Presentation domain, the examinee prefers a simple, easy to learn and use Interface. Thus, most of the CAT systems tried to create an interface that helps the user to be always aware of where he is in the test and what information is being displayed. All areas are clear and well defined. So, the user is not confused in any area.

Furthermore, most of the User Interfaces are consistent. Consistency in the navigation controls conveys how action in the system should be performed. The same icon or command has the same operation throughout the test. Moreover, the icon for a specific operation in all tests is always in the same area in the test (Dennis, Wixom & Tegarden, 2005).

The operation is correct and precise. Most of the CAT systems present a confirmation button so that the examinee confirms his answer before he is allowed to press the next button to proceed to the next question. This confirmation button prevents the user to go by mistake to the next question before he is sure for his answer to the current question. In CAT, the examinee cannot return to a question and change his answer. The structure is simple and effective, as most tests don't have more than six buttons on each form. Many tests provide feedback, help documentation and high quality of interactivity. The responses to examinee's actions are immediate and error

free. However the design, as it has already been mentioned in the Presentation, is very poor.

Regarding the User Interface, the majority (six out of the 10) of the CAT demos score "Fair" (Table). Three systems score "Good": i) Graduate Management Admission Test (GMAT), ii) Graduate Record Examination (GRE), and iii) the Test Of English as a Foreign Language (TOEFL). All three systems follow the usability rules. The systems are easy to understand, learn and use even for a beginner user.

Reliability

Reliability refers to the capability of the CAT system to maintain a specified level of operation during the assessment. The CAT system should achieve the following capabilities with minimum effort at minimum time: i) avoid failures and faults, ii) maintain consistent operation even in case of failures, iii) recover from failures re-establishing its previous state of operation, and iv) be available to the examinee at any moment during the assessment. Roever (2001) points out that the most severe technical problem is the failure of the server which houses the CAT system. A simple way around this problem is to have "mirror sites" on alternate servers. Additionally, keeping on alternative communication paths between the examinee and the CAT system increases the reliability.

Many of the test providers are large organizations or institutions with years of experience. Most of them provide official diplomas. So, the Reliability is very important for their reputation and they took extra consideration to achieve a sufficient degree of Reliability. Most of the systems are error free and handle efficiently an unexpected situation. The algorithms are designed in such a way that saves all users' actions and can load the test from the last action of the user. So if for example, the power goes off at the seventh item, the user can continue his test from the seventh item and on. The six previous responses are stored. An unexpected situation by mistake of the user is limited because most of the tests guide the user to take a specific action and block all other undesirable actions. For example, the user cannot press the "next" button before the "confirm" button.

The operation of the tests is stable, consistent, correct and accurate. The tests treat similar states in a similar way. They also keep on back up of the data, items, scores, statistics, etc. No data or other useful resources are lost in case of error. For example, in a situation of hardware fault (e.g. power off), the CAT systems not only maintain data by saving the test but also detect the previous save operation and allow the user to continue.

Regarding the Reliability, almost all CAT systems achieve high scores (Table). Four systems score "Good". Five systems score "Excellent": i) Graduate Management Admission Test (GMAT), ii) Graduate Record Examination (GRE), iii) the Test Of English as a Foreign Language (TOEFL), iv) Microsoft Certified Systems Engineer (MCSE), and v) the Computing Technology Industry Association (CompTIA).

Maintainability

Maintainability refers to the effort and time needed for installation, fault removal, update, upgrade, expansion and other modifications of the CAT system. Also, it is related to the risk taken from unexpected effects of modifications.

The installation of all tests is very easy and needs very small disc space (due to the luck of multimedia). Some tests do not need installation at all and are compatible with the most common operating systems. All organizations gave effort to create a software easy to maintain and easy to reconfigure in case of changes that could be required. Usually the only thing that needs to be changed is the item bank according to the topic that needs to be examined. The guarantees are for long time and cover almost any possible case, as

most of the test providers are large and respectable organizations.

Some tests provide to the user the right to change the software or to add and delete items in the item bank. This is very useful because it keeps the items up to date and produces new tests according to the topic that must be covered. So, an institution could create a new test for private use without asking the CAT system provider for a new item bank.

Regarding the Maintainability, all CAT systems score higher or equal to "Fair" (Table). Three systems score "Excellent": i) Graduate Management Admission Test (GMAT), ii) Graduate Record Examination (GRE), and iii) FastTEST Pro. FastTEST Pro also gives the user the right to add, alter and delete items from the item bank. From the rest, four systems score "Good", and three systems score "Fair".

Performance

The Performance domain examines the achieved performance and efficiency of the CAT system. If the test is delivered via the Web, download times can be negligible or considerable, depending on server traffic, complexity of the page, client computer speed, etc. It is therefore important for timed tests to stop the timer during downloads and restart it when the page is fully displayed (Roever, 2001).

In all CAT systems, the processing is immediate so that the examinee won't worry of loosing precious time. The response of the systems is also immediate. All CAT systems took extra consideration to have high processing speed, even if the adaptive test is online. The delay of storing and receiving data is almost zero. This efficiency is achieved because most of the systems do not use a database separate from the main program. So, they do not waste time to connect to a remote database in order to retrieve and store data. Also, the memory capacity is high since each item is very small (due to the lack of multimedia). The effectiveness and efficiency of the systems are very high.

On the other hand the user produces the input data by checking the correct answer, which is very easy to store. The CAT systems avoid to employing advanced input devices such as camera, handwritten recognizer or speech recognizer.

Regarding the Performance, almost all CAT systems score higher or equal to "Fair" (Table). The Test Of English as a Foreign Language (TOEFL) scores "Excellent". TOEFL manages to keep the delay small even if the retrieved item is large (e.g. sound in the listening comprehension). From the rest CAT systems, three systems score "Good", and three systems score "Fair".

Functionality

Functionality refers to available functions, features (e.g. alerting and reminding), tools (e.g. calculator, editor, scratch-work space, drawing, ruler, protractor, audio recorder, photo camera, etc.), and applications in the CAT systems. It examines the quantity, quality, appropriateness and the properties of these functions to support the examinee during the assessment.

Unfortunately, most of the CAT systems tend to avoid using these tools or not to use them at all. The main consideration of the test providers is to concentrate on producing an error-free "multiple question" test. A possible reason may be that many examinees are not familiar with computers, or even if they are, they may not be familiar with the CAT system capabilities. So during the test, they might get confused and as consequence loose precious time.

Regarding the Functionality, all CAT systems score low (Table). Four systems do not have any extra functions and features. Four systems score "Very Poor", and two systems score "Poor".

Connectivity

Connectivity refers to the ability of the CAT system to interact and communicate with other software and hardware systems. It examines the capability of writing/reading to/from various systems via various networks in various formats using various protocols. For example, items from various item banks would be used by the system. The test results would be reported to statistical analysis and administration software at the school or state. The portability of the system and the capability to execute the CAT on different types of computers are also important issues.

Most of the tests comply with international standards and are compatible with many software and hardware devices. As it has already been mentioned, some tests do not need installation and are compatible with many operating systems. On the other hand, the CAT systems use very few extra tools.

The importation and exportation of data, items, scores and statistics is quite easy without the need of additional plug-ins. The integration of the parts of the test is transparent to the examinee. All parts are successfully combined to produce a correct and autonomous test.

Regarding the Connectivity, almost all CAT systems score higher or equal to "Fair" (Table). Graduate Record Examination (GRE) scores "Excellent". Three systems score "Good", and five systems score "Fair".

Security

Security refers to the protection of the CAT system against unauthorized access to or modification of information, whether in storage, processing or transit, and against the denial of service to authorized users or the provision of service to unauthorized users, including those measures necessary to detect, document, and counter such threats. It ensures a state of inviolability from hostile acts or influences. It prevents unauthorized persons from having access to restricted information. It also ensures confidentiality so that information is accessible only to those authorized to have access.

Most of the CAT providers are large organizations or institutions. Security is a very important issue for them. A Security error would harm the organization's reputation. The organizations usually certify and guarantee their Security. So, the items are well protected. Especially in a predetermined algorithm the items are not stored in an item bank but they are part of the test, so no one can separate and process or store them.

The examinee's confidentially, anonymity and privacy is protected. Cheating, plagiarism, unauthorized notes taking, reproduction and copying are prevented. This is to be expected because the user's actions are restricted. The items are rarely in text format, even if they are composed only from text so they cannot be copied during the test and a user cannot add or alter an item or write any notes to the examiner. All data activities, decisions and applications are visible and available to the examinee whenever he requests them. Furthermore, every examinee answers a unique test tailored around his proficiency level. So, no two examinees answer the same items. In addition, the possible answers in an item are scrambled. This improves the security.

It is obvious that Security is a crucial issue in tests. Almost all CAT systems score high in Security (Table). Five systems score "Excellent": i) Graduate Management Admission Test (GMAT), ii) Graduate Record Examination (GRE), iii) the Test Of English as a Foreign Language (TOEFL), iv) Microsoft Certified Systems Engineer (MCSE), and v) the Computing Technology Industry Association (CompTIA).

Adaptation

The CAT systems select the next item according to the last answer of the examinee. If the examinee answers an item correctly then the next item is more difficult than the current item. On the contrary, if the examinee answers incorrectly then the next item is an easier one. The possibility of two examinees to view exactly the same questions is very small. So, the CAT systems adapt the Content to the level of knowledge of the examinee. However, the systems do not adapt the Presentation to the personal taste of the user and the Sequencing algorithm is hidden. The examinee sees only the questions and the possible answers. Usually, the examinee does not know that the next item is presented according to his last answer. The Feedback is adapted in some tests but most of the tests provide standard information.

The systems adapt the Content to the screen size. However, the image resolution is not adapted to the available transmission bandwidth. The Adaptation is consistent; similar reasons cause similar Adaptation results. The tests were observed several times, either with exactly the same actions or with different actions. The third item for exampled was answered two times correctly and one time incorrectly. The correct answers led to the same (in a predetermined algorithm) more difficult question, while an easier one followed an incorrect answer.

Regarding the Adaptation, almost all CAT systems score higher or equal to "Fair" (Table). Three systems score "Excellent": i) FastTEST Pro, ii) "An On-line Interactive Computer Adaptive Testing Tutorial by Lawrence M. Rudner", and iii) Microsoft Certified Systems Engineer (MCSE). FastTEST Pro tries to adapt even the type of question as it gives to the user the possibility to select among "Multiple choice", "Check all that apply", and "True/False" questions. Three systems score "Good", four systems score "Fair", and one scores "Poor".

The average scores are presented at the last column of the Table. The CAT systems score above average in all technical domains except the Functionality. Designers and developers of CAT systems should not overlook Functionality. Rather, they should provide extra features and tools to support the examinee.

ECONOMICAL DIMENSION

The economical dimension consists of the following domains: 1) Costs, 2) Contracts and Licensing, and 3) Cost Effectiveness.

Costs

This domain includes the Costs for developing, validating, operating, administering, maintaining, upgrading, etc. the item bank and the CAT system. It has already been pointed out that the cost of developing a CAT system can be significant (Meijer & Nerling, 1999; Hableton et al., 2000). For example, developing and validating an item bank of 1000 items for a specific topic is not an easy task. For obvious reasons, the CAT systems providers did not provide any information on these costs. So, it was not possible to evaluate the various Costs.

Contracts and Licensing

This criterion applies only to the for-profit organizations since the CAT systems by the non-profit organizations are free. All CAT systems provide information about the examination fees. However, there are not alternative types of contracts with respect to the number of subjects, number of examinees, number of items, etc. For example, a class of 100 students cannot negotiate for lower fees. Regarding the Contracts and Licensing, the majority of the CAT systems score "Fair" (Table). Graduate Record Examination (GRE) scores "Excellent".

Cost Effectiveness

The Cost Effectiveness domain is related to the overall examinee's satisfaction of using the CAT system versus the fees he pays. Almost all CAT systems score "Fair" (Table). The two systems by non-profit organizations score "Excellent": i) the Maryland State Department of Education (MSDE), and ii) "An On-line Interactive Computer Adaptive Testing Tutorial by Lawrence M. Rudner", since they are free.

Then, the average scores are presented at the last column of the Table. The CAT systems score above average in Cost Effectiveness and below average for the Contracts and Licensing.

CONCLUSION

The aim of this paper was to investigate the current state of CAT systems, to identify their strengths and weaknesses and suggest directions for improvements. First, it should be mentioned that the results regarding the evaluation of the CAT systems are subjective. A large scale evaluation, let say, by hundreds of students is not possible due to the complexity of these systems and the CATE framework. The authors evaluated these systems taking into consideration comments by graduate students who had experienced them. While most CAT systems met most of the CATE requirements, there are some domains which have not yet been fully developed.

It is obvious that the contemporary CAT systems give priority to Security, Reliability and Maintainability. However, they almost ignore issues related to the Presentation, Functionality, Feedback, Contracts and Licensing. They target to provide error-free and easy to understand tests at the expense of reducing the availability of multimedia, supporting tools and applications.

The evaluation's purpose was to comprehend the existing situation in order to proceed to the development of new advanced CAT systems. The evaluation tries to find the strengths and weaknesses of contemporary CAT systems in order to enhance the strengths and reduce the weaknesses. For example, the Feedback could be improved by providing more information to the examinee, or providing information anytime the examinee wishes. The Presentation and Adaptation could be improved by personalizing the test to the examinee's personal taste. For example, the examinee would select his favourable ways of Presentation, Feedback, User Interface, etc. in a pre-test screen. So, the examinee would select how the items would be presented (e.g. using sound, video or text), what orientation information to see (e.g. time alerts), colours and fonts, the types of the feedback (e.g. instant feedback to know if he answered correctly the same time that he confirms his answer). This way the examinee will be more comfortable with the test, and improve his performance and his scoring. It is not difficult to employ these capabilities into the current CAT systems. However, there are other limitations inherent to IRT (Item Response Theory). These include the following restrictions for the examinee. He cannot review all items and then answer them. He cannot skip an item without answering it. He cannot go back and revise his answer to a previous item.

On the other hand, it might be difficult to enhance the Functionality since the examinees have different operating systems or use different devices. An improvement on the Functionality could affect the Maintainability and the Connectivity because these domains demand stability.

The Security and Reliability domains have the fewest weaknesses. It is important that the CAT developer provides capabilities such as anonymity, privacy and back up of all the examinees actions in case of unexpected situations. Finally, efficient control of the item exposure can protect the item security.

REFERENCES

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*, 81-86, AACE.

Boyle, A., & O Hare, D. (2003). Finding appropriate methods to assure quality computer-based development in UK Higher Education. In *Proceedings of the 7th computer-assisted assessment conference*, 67-82, Loughborough University, United Kingdom.

Brown, J. D. (1997). Computers in language testing: present research and some future

directions. *Language Learning & Technology*, *1*(1), 44-59.

Carlson, R. D. (1994). Computer-adaptive testing: A shift in the evaluation paradigm. *Journal of Educational Technology Systems*, *22*(3), 213–224.

Cisco http://www.cisco.com

http://www.topshareware.com/Cisco-Practice-Tests-from-Boson-download-10944.htm

Cito Group http://www.cito.nl/

CompTIA http://www.comptia.org/certification/

Dennis, A., Wixom, B.H., & Tegarden, D. (2005). *Systems analysis and design with UML version 2.0*, 2nd edition, John Wiley & Sons Inc.

Dunkel, P. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology*, *2*(2), 77-93.

Economides, A.A. (2005a). Computer adaptive testing quality requirements. In *Proceedings E-Learn 2005, World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education,* 288-295, AACE.

Economides, A.A. (2005b). Personalized feedback in CAT. *WSEAS Transactions on Advances in Engineering Education*, *3*(2), 174-181.

Eduventures http://www.eduventures.com

FastTest Pro http://www.assess.com/Software/FTP16Main.htm

Georgiadou, E., Triantafillou, E., & Economides, A.A. (2006). Evaluation parameters for computer adaptive testing. *British Journal of Educational Technology*, *37*(2), 261-278.

Georgiadou, E., Triantafillou, E., & Economides, A.A. (2007). A review of item exposure control strategies for computerised adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*(8).

Georgouli, K. (2004). WASA: An intelligent agent for Web-based self-assessment. In Kinhuk, Sampson, D. & Isaias, P. (Eds.), *Cognition and Exploratory Learning in Digital Age (CELDA 2004),* ISBN: 972-98947-7-9, 43-50. Assoc. Editors, L. Rodrigues and P. Barbosa, Lisbon, December.

Giouroglou, H., & Economides, A. (2004). State-of-the-art and adaptive open-closed items in adaptive foreign language assessment. In *Proceedings 4th Hellenic Conference with International Participation: Informational and Communication Technologies in Education*, Athens, 747-756.

Giouroglou, H., & Economides, A.A. (2005). The development of the adaptive item language assessment (AILA) for mixed-ability students. In *Proceedings E-Learn 2005 World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, 643-650, AACE.

GMAT http://www.gmat.org , http://www.mba.com, http://www.gmat-mba-prep.com/, http://www.800score.com/gmat-home.html

GRE http://www.ets.org , http://www.800score.com/gre-index.html

Hableton, R.K., Zaal, J.N., & Pieters, J.P. (2000).

Computerized adaptive testing : theory, applications, and standards. Reston, MA: Kluwer,

Jacobson, R. L. (1993). New computer technique seen producing a revolution in educational testing. *Chronicle of Higher Education, 40*(4), pp. 22–23.

Kapsalis, A.G. (2004). *Pedagogic psychology.* 3rd edition, Kiriakidis S.A.

Lilley, M., & Barker, T. (2003). An evaluation of a computer-adaptive test in a UK University context. In *Proceedings of the 7th computer-assisted assessment conference,* 171-182, United Kingdom: Loughborough University.

Lilley, M., Barker, T., & Britton, C. (2004). The development and evaluation of a software prototype for computer-adaptive testing. *Computers & Education 43*, 109-123.

MCSE http://www.microsoft.com/learning/mcp/mcse/http://www.sybex.com/sybexbooks.nsf/AdditionalContent/2946OnlineDemo?OpenDocument#

Meijer, R.R., & Nering, M.L. (1999). Computerized adaptive testing: Overview and introduction. *Applied psychological measurement. 23*(3), 187-194.

Olea, J., Revuelta, J., Ximenez, M.C., & Abad, F.J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. *Psicologica 21*,. 157-173.

Roever, C. (2001). Web-based language testing. *Language Learning & Technology*, *5*(2), 84-94.

Rudner, L.M. (2006). An on-line interactive computer adaptive testing tutorial. Retrieved on August 02, 2006, from, http://edres.org/scripts/cat/catdemo.htm

Rudner, L.M. (2002). An examination of decision-theory adaptive testing procedures, *Conference of American Educational Research Association*, New Orleans, LA April 1-5.

Sclater, N., & Howie, K. (2003). User requirements of the ultimate online assessment engine. *Computers & Education*, *40*, 285-306.

Straetmans, G.J.M., & Eggen T.J.H.M. (1998). Computerized adaptive testing: what it is and how it works. *Educational Technology*, 82-89, January-February.

TOEFL http://www.ets.org, http://www.toefl.org, http://toeflpractice.ets.org/

Triantafillou, E., Georgiadou, E., & Economides, A.A. (2007a). Applying adaptive variables in computerised adaptive testing. *Australasian Journal of Educational Technology, AJET, 23*(3).

Triantafillou, E., Georgiadou, E., & Economides, A. (2007b).The role of user model in CAT: Exploring adaptive variables. *Technology, Instruction, Cognition and Learning: An International, Interdisciplinary Journal of Structural Learning*, *5*(1), 69-89.

Triantafillou, E., Georgiadou, E., & Economides, A.A. (2008). The design and evaluation of a computerized adaptive test on mobile devices. C*omputers & Education*, *50*.

Valenti, S., Cucchiarelli, Al, & Panti, M. (2001). A framework for the evaluation of test management systems. *Current Issues in Education*, 4(6).

Valenti, S., Cucchiarelli, Al & Panti, M. (2002). Computer-based assessment systems evaluation via the ISO9126 quality model. *Journal of Information Technology Education*, *1*(3).

Wainer, H. (1990). *Computerized Adaptive Testing: A Prime*r. New Jersey: Lawrence Erlbaum Associates, Publishers.

Wainer, H., Dorans, D. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A Primer.* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Weiss, D.J., & Kingsbury,G.G. (1984). Application

of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4),361-375.

Welch, R.E., & Frick, T.W. (1993). Computerized adaptive testing in instructional settings. *Educational Technology Research & Development*, *41*(3), 47-62.

Wise, S.L., & Kingsbury, G.G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologica 21*, 135-155.

Young, R., Shermis, M.D., Brutten, S.R., & Perkins, K. (1996). From conventional to computer-adaptive testing of ESL reading comprehension. *System*, *24*(1), 23-40.

Zahorian S.A, Lakdawala, V.K., Gonzalez, O.R., Starsman, S., & J.F. Leathrum Jr. (2001). Question model for intelligent questioning systems in engineering education. *Proceedings 31st ASEE/ IEEE Frontiers in Education Conference*, pp. T2B7-12, IEEE.