Triantafillou, E., Georgiadou, E. & Economides, A.A.:**The design and evaluation of a computerized adaptive test on mobile devices**. *Computers & Education*, Vol. 49, 2007

The Design and Evaluation of a Computerized Adaptive Test on Mobile Devices

Evangelos Triantafillou [*], Elissavet Georgiadou, Anastasios A. Economides
University of Macedonia
Egnatia 156
Thessaloniki 54006, GREECE
vtrianta@csd.auth.gr, elisag@otenet.gr, economid@uom.gr

Abstract

The use of computerized-adaptive testing (CAT) has expanded rapidly the last years mainly due to the advancements in communication and information technology. Availability of advanced mobile technologies provides several benefits to e-learning by creating an additional channel of access with mobile devices such as PDAs and mobile phones. This paper describes the design issues that were considered for the development and the implementation of a CAT on mobile devices, the CAT-MD (Computerized Adaptive Test on Mobile Devices). Throughout the development of the system, formative evaluation was an integral part of the design methodology. The recommendations, suggestions and the results of the formative evaluation were used to improve the system in order to make the assessment procedure more effective and efficient. These results of the formative evaluation are also presented here.

Keywords: Computerized-adaptive testing, Mobile Learning, Formative Evaluation

1. Introduction

Recent advancements in computer technology and psychometric theories have accelerated the change of test format from conventional paper-and-pencil tests to computerized adaptive testing (CAT). A CAT is a method for administering tests that adapts to the examinee's ability level. Moreover, in the recent years, much attention has been paid to mobile computing. Availability of advanced mobile technologies have started to extend e-learning by adding an extra channel of access for mobile users with mobile devices such as hand phones, Personal Digital Assistants (PDAs) or pocket PCs. Although, mobile computing has become an important and interesting research issue, little research has been done regarding the implementation of CAT using mobile devices (and this is the focus of our research.

Our research is an attempt to examine the design and development issues, which may be important in the implementation of a CAT on different mobile products such as mobile phones and PDA. As a case study an educational assessment prototype was developed, called CAT-MD (Computerized Adaptive Test on Mobile Devices). A formative evaluation of the CAT-MD was carried out in order to investigate the effectiveness and efficiency of the system and also to assess its usability and appeal.

This article describes the design, the development and the formative evaluation of the CAT-MD. It first discusses the design issues that were considered for the development of the computerized adaptive test, next, it describes the architecture and the implementation of the computerized adaptive test on a PDA and finally, it presents the description and the results of the formative evaluation.

2. Computerized Adaptive Test

Testing is one of the most widely used tools in higher education. The main goal of testing is to measure student knowledge level in one or more concepts or subjects, i.e. in pieces of knowledge that can be assessed. Since education was established as an institution different methods of assessment were used in different contexts, such as class presentations, essays, projects, practicum, etc. However,

[*] Corresponding Author.
E-mail address: vtrianta@csd.auth.gr (E.Triantafillou)

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 the last few decades and its use is spread rapidly in all levels of education and training, the use of computer-based tests (CBTs) has increased significantly over the last few years. CBTs became feasible for licensure, certification and admission.

The most common type of a CBT is the linear one that is a fixed-length computerized assessment that presents the same number of items to each examinee in a specified order and the score usually depends on the number of items answered correctly. A linear test consists of a full range of easy and difficult test items that are either randomly selected from a larger pool or are the same for all examinees. Evidently the type of CBT 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.

By contrast, in computerized adaptive test (CAT), a special case of computer-based testing, each examinee takes a unique test that is tailored to his/her ability level. As an alternative of giving each examinee the same fixed test, CAT item selection adapts to the ability level of individual examinees and after each response the ability estimate is updated and the next item is selected to have optimal properties at the new estimate (van der Linden & Glas, 2003). The CAT presents first an item of moderate difficulty in order to initially assess each individual’s level. During the test, each answer is scored immediately and if the examinee answers correctly then the test statistically estimates her/his ability as higher and then presents an item that matches this higher ability. If the next item is answered again correctly it re-estimates 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 re-evaluates the ability of the examinee until the accuracy of the estimate reaches a statistically acceptable level or when some limit is reached; such as a maximum number of test items. The score is 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 correctly more difficult items.

Regardless of some disadvantages reported in the literature –for example, high cost of development, item calibration, item exposure (Eggen, 2001; Boyd, 2003), the effect of a flawed item (Abdullah, 2003), or the use of CAT for summative assessment (Lilley & Barker, 2002, 2003)– CAT has several advantages. Testing on demand can be facilitated so as an examinee can take the test whenever and wherever s/he is ready. Multiple media can be used to create innovative item formats and more realistic testing environments. Other possible advantages are flexibility of test management; immediate availability of scores; increased test security; increased motivation etc. However, the main advantage of CAT over any other computerized based test is efficiency. Since fewer questions are needed to achieve a statistically acceptable level of accuracy, significantly less time is needed to administer a CAT compared to a fixed length Computerized Based Test (Rudner, 1998; Linacre, 2000).

Since the mid 80’s when the first CAT systems became operational i.e. Armed Services Vocational Aptitude Battery for US Department of Defence account (van der Linden & Glas, 2003) using adaptive techniques to administer multiple-choice items, much research and many technical challenges make possible new assessment tools. Current research in CAT is not limited to educational admissions, yet, focus on applications, in small and large scale, that address self-assessment, training, employment, teacher professional development for schools, industry, military, assessment of non-cognitive skills etc.

3. Mobile Learning

In the recent years the use of different mobile products such as mobile phones and Personal Digital Assistant (PDA) devices has increased rapidly. Moreover, much attention has been paid to mobile computing within Information Technology industry. Availability of advanced mobile technologies, such as high bandwidth infrastructure, wireless technologies, and handheld devices, has started to extend e-learning towards mobile learning (Sharples, 2000). Mobile learning (m-learning) intersects mobile computing with e-learning; it combines individualized (or personal) learning with anytime and anywhere learning. The advantages of m-learning include: flexibility, low cost, small size, ease of use and timely application (Jones & Jo, 2004).

M-learning is often thought of as a form of e-learning, but Georgiev, Georgieva, and Smrikarov (2004) believe it would be more correctly defined as a part, or sub-level, of e-learning. They believe m-learning is a new stage in the progress of e-learning and that it resides within its boundaries. M-learning is not only wireless or Internet based e-learning but should include the anytime/any place concept without permanent connection to physical networks (Jones & Jo, 2004). The place independence of mobile devices provides several benefits for m-learning environment like allowing students and instructors to utilize their spare time while travelling in a train or bus to finish their homework or lesson preparation. Furthermore, m-learning have the potential to change the way students behave and interact with each other (Luvai, 2005).

The introduction of mobiles devices into the learning pedagogy can compliment e-learning by creating an additional channel of assessment with mobile devices such as PDAs, mobile phones, portable computers. Due to their convenient size and reasonable computing power, mobile devices have emerged as a potential platform for computer-based testing. Although, there is previous research evaluating the impact of computer-based testing (using desktop systems), there is little research on the impact of using mobile devices to administer tests since handheld and wireless technologies are relatively new and continually evolving. Most of the literature is focused on practical topics concerning with the development and implementation of new educational applications that can be delivered with these devices (Chen, Myers, & Yaron 2000; Cook, 2000). However, Segall, Doolen and Porter (2005) conducted a research in order to compare the usability effectiveness, efficiency, and satisfaction of a PDA-based quiz application to that of standard paper-and-pencil quizzes in a university course. Their study showed that the PDA-based quiz is more efficient since students completed it in less time than the paper-and-pencil quiz. Furthermore, in their study no differences in effectiveness and satisfaction were found between the two quiz types. However, the authors concluded that PDAs are an attractive test administration option for schools and universities.

Research and development on CAT is mostly targeted towards PC based users. Although, mobile computing has become an important and interesting research issue, little research has been done on the implementation of CAT using mobile devices and this is the focus of the research presented in this paper. The current study is an attempt to examine the design and development issues, which may be important in the implementation of a CAT using mobile devises such as mobile phones and PDAs. As a case study an educational assessment prototype was developed, called CAT-MD (Computerized Adaptive Test on Mobile Devices), to support the assessment procedure of the subject "Physics" which is typically offered to second grade students in senior high school in Greece.

4. System Architecture

The prototype CAT-MD uses the Item Response Theory (IRT) as an underlying psychometric theory, which is the base for many adaptive assessment systems and depicts the relationship between examinees and items through mathematical models (Lord, 1980; Hambleton 1991; Wainer, 1990). Psychometric theory is the psychological theory or technique of mental measurement, which is the base for understanding general testing theory and methods. The central element of IRT is mathematical functions that calculate the probability of a specific examinee answering a particular item correctly. IRT is used to estimate the student's knowledge level, in order to determine the next item to be posed, and to decide when to finish the test. There are four main components needed for developing IRT-based CAT: the item pool, the item selection procedure, the ability estimation and the stopping rule (Dodd, De Ayala & Koch, 1955). The following sections describe these components of the CAT-MD system.

4.1 Item pool

The most important element of a CAT is the item pool that 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. The success of any CAT program is largely dependent on the quality of the item pool that can be conceptualized according to two basic criteria: a) the total number of items in the pool must be sufficient to supply informative items throughout a testing session, and b) 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 suggests 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).

The item parameters included in the pool are dependent upon the Item Response Theory (IRT) model selected to model the data and to measure the examinees' ability levels. In IRT-based CATs, the difficulty of an item describes where the item functions along the ability scale. For example, an easy item functions among the low-ability examinees and a hard item functions among the high-ability examinees; thus, difficulty is a location index.

An ideal item pool needs many items, best spread evenly over the possible range of difficulty. In our approach CAT-MD includes a database that contains 80 items related to the chapter "Electricity" from the 'Physics" subject. For every item, the item pool includes the item's text, details on the correct answer and the difficulty level. The difficulty level varies form “very easy” to “very hard” and the values used cover the range between -2 and +2.

4.2 Item selection

In IRT theory, the item selection procedure is the process of selecting an item from the item pool to be administered to the examinee. A reasonable assumption is that each examinee responding to a test item possesses some amount of the underlying ability. Thus, one can consider each examinee to have a numerical value, a score that places him or her somewhere on the ability scale. This ability score will be denoted by the Greek letter theta, θ. At each ability level, there will be a certain probability that an examinee with that ability will give a correct answer to the item. This probability will be denoted by $P(\theta)$. In the case of a typical test item, this probability will be small for examinees of low ability and large for examinees of high ability (Baker, 2001).

Two common classes of IRT models are determined by the way items responses are scored. Items with only two response options (correct or incorrect) are modelled with the dichotomous IRT models. Items with more than two response options can be modelled with polytomous IRT models (Boyd, 2003). Our prototype CAT-MD includes multiple choice items and true false items. Since, these are examples of items that can be scored dichotomously, CAT-MD is based on dichotomous IRT model.

The main aspect of IRT is the Item \Characteristic Curve (ICC) (Baker, 2001). ICC is an exponential function, which gives the probability of answering a question of certain difficulty level correctly by a learner with certain skill level. ICC is a cumulative distribution function with a discrete probability. The models most commonly used as ICC functions are the family of logistics models of one (1PL), two (2PL) and three parameters (3PL).

The 1-parameter logistic (1PL), or Rasch model is the simplest IRT model. The Danish mathematician Georg Rasch first published the 1-parameter logistic model in 1960s and as its name implies, it assumes that only a single item parameter is required to represent the item response process. This item parameter is termed difficulty and the equation for this model is given by:

$$P(\theta)=\frac{1}{1+e^{-1(\theta-b)}} \qquad \textbf{(1)}$$

where, e is the constant 2.718, b is the difficulty parameter and θ is an ability level.

In CAT-MD, as each student answers a question, his or her response is evaluated as being either correct or incorrect. In the event of a correct response, the probability $P(\theta)$ is estimated applying the formula shown in Eq. (1). Otherwise, the function $Q(\theta)=1-P(\theta)$ is used.

The Item Information Function (IIF) is also considered as an important value in the item response theory’s item selection process. It gives information about the item to be presented to the leaner in an adaptive assessment. For selecting a question appropriate to the learner, IIF for all the questions in the assessment should be calculated and the question with highest value of IIF is presented to the learner. This provides more information about the learner’s ability and is given by the equation:

$$I_i(\theta)=P_i(\theta)(1-P_i(\theta)) \qquad \textbf{(2)}$$

where $Pi(\theta)$ is the probability of a correct response to item i conditioned on ability θ (Baker, 2001; Lord, 1980).

4.3 Ability estimation

After each item is administered and scored, an interim estimate of examinees’ ability (θ) is calculated and used by the item selection procedure to select the next item. The most commonly used estimation procedure is maximum likelihood estimation (MLE) (Lord, 1980). Similar to the item parameter estimation, this procedure is an iterative process. It begins with some a priori value for the ability of the examinee. In CAT-MD, it begins with $\theta=1$. The estimation calculation approach is the modification of the Newton-Raphson iterative method for solving equations method outlined by Lord. The estimation equation used is shown below:

$$\theta_{n+1} = \theta_n + \frac{\sum_{i=1}^{n} S_i(\theta_n)}{\sum_{i=1}^{n} I_i(\theta_n)} \qquad \textbf{(3)}$$

where

$$S_i(\theta) = [u_i - P_i(\theta)] \frac{P_i'(\theta)}{P_i(\theta)[1 - P_i(\theta)]} \qquad \textbf{(4)}$$

where

is the skill level after n questions, and
ui = 1 if the response is correct and ui = 0 for the incorrect response.

4.4 Stopping Rule

One important characteristic of CAT is the test termination criterion. The termination criterion is generally based on the accuracy with which the examinees' ability has been assessed. In most CATs, the termination of the test may be based on the number of items administered, the precision of measurement or a combination of both (Boyd, 2003). Measurement precision is usually assessed based on error associated with a given ability. The standard error associated with a given ability is calculated by summing the values of the item information functions (IIF) at the candidate's ability level to obtain the test information. Test information, TI( ), is given by the equation:

$$TI(\theta) = \sum_{i=1}^{N} I_i(\theta) \qquad \textbf{(5)}$$

Next, the standard error is calculated by using the equation:

$$SE(\theta) = \frac{1}{\sqrt{\mathrm{TI}(\theta)}} \qquad \textbf{(6)}$$

After each administration of an item, the standard error associated with a given ability is calculated to determine whether a new item must be selected or whether the administration of the test can be terminated. It is common in practice, to design CATs so that the standard errors are about .33 or smaller (Rudner, 1998). In CAT-MD the test terminates for each examinee when the standard error associated with a given ability ( ) is less than 0.30 or when the maximum number (that is 20) of items has been administered.

5. System Implementation

Currently, the basic architecture of the system has been implemented. The prototype software has been developed using Macromedia Flash as it offers competitive advantages. It is a lightweight, cross-platform runtime that can be used not just for enterprise applications, but also for communications, and mobile applications. According to Macromedia Company the 98 percent of all Internet enabled computers and 30 million mobile devices use the Flash technology (www.macromedia.com). To date, many manufacturers license Macromedia Flash on their branded consumer electronics devices, such as mobile phones, portable media players, PDAs, and other devices. These licensees include leading mobile device manufacturers such as Nokia, Samsung, Motorola, and Sony Ericsson.

Figure 1 presents two screenshots of the implementation of CAT-MD on a mobile phone and on a PDA. Moreover, the CAT-MD is portable to any device that has installed the Macromedia Standalone-Flash Player. In addition, if a Macromedia plug-in for the web browser (Internet Explorer, Mozilla, etc.) is installed, the CAT-MD can be also accesses as flash shockwave film.

6. Formative Evaluation

Throughout the development of CAT-MD, formative evaluation was an integral part of the design methodology. Formative evaluation is the judgment of the strengths and weaknesses of the system in its developing stages, for the purpose of revising the system. The goals of the evaluation were to test the satisfaction of the learners with the use of CAT-MD, and to examine the effectiveness and efficiency of the assessment procedure. In CAT-MD case, three types of formative evaluation were used: expert review, one-to-one evaluation and small group evaluation. Tessmer (1993) defines these evaluations types as follows:

- Expert review – experts review the system with or without the evaluator present. The experts can be content experts, designers or instructors.
- One-to-one evaluation – one learner at a time reviews the system with the evaluator and comments upon it.
- Small group evaluation – the evaluator tries out the system with a group of learners and records their performance and comments.

As discussed above the CAT-MD is portable to any device that has installed the Macromedia Standalone-Flash Player. However, we evaluated the PDA version (HP iPAQ 1950) as it is a more useful device to an educational setting. The current interface of CAT-MD and its functionalities are the result of revisions based on the analysis of the data collected during the formative evaluation. Next, we will discuss the process of the formative evaluation i.e. subjects, procedure, and results.

6.1 Expert review and one-to-one evaluation

Four experts acted as evaluators in the expert review phase: a teaching/training expert, an instructional design expert, a subject-matter expert, and an educational technologist. In this phase, a semi-structured interview aimed at determining the reactions of experts and a debriefing session were used. During the semi-structured interview, although the subjects were prompted with questions, the main aim was to get their subjective reactions on the clarity, completeness and ease of use of the prototype CAT-MD. The debriefing session was used to conclude the evaluation with general questions about the assessment procedure and the design of the prototype CAT-MD and to prompt subject's suggestions for the improvement. In addition, the experts were asked to evaluate the prototype's effectiveness as a computer-assisted assessment tool in an academic context.

All the experts stated satisfied with the use of the PDA as an additional tool of the assessment procedure and they agreed with its user friendliness. They indicated that PDAs are attractive devices because they are lightweight, extremely portable, and feature touch-sensitive displays. Furthermore, they felt that the introduction of mobiles devices into the learning pedagogy adds extra value to the use of computers in educational settings. They pointed out that the mobility eliminates the need for a specialized computer lab. PDA can be used anywhere, including a traditional classroom.

Furthermore, they stated that computerized assessment procedure is a promising alternative to the traditional paper-and-pencil assessment procedure. Experts stated that they were satisfied with the use of CAT-MD and they indicated that such assessment can be also implemented within other subjects and especially those that use labs such as physics, chemistry etc. Furthermore, they agreed that CAT-MD interface was clear and straightforward. In addition, they observed that the adaptation based on examinee's knowledge estimate was easy to understand.

The expert review phase was followed by the one-to-one evaluation. Ten subjects participated in the one-to-one evaluation. They were students from senior high school. They were at the second grade and studying the chapter of "Electricity" from the 'Physics' subject. In this phase, a semi-structured interview and a debriefing session were used as well. The interview covered the following topics: 1) the use of the PDA, 2) the review of the computerized adaptive testing procedure, and 3) the evaluation of the CAT-MD system.

Most of the students agreed on the user friendliness of the PDA. They did not have any previous experience with a similar tool and they felt challenged by using it. Furthermore, they stated that they wanted to use the PDA again. With regard to the adaptive testing algorithm, they agreed that the structure of the CAT was clear and easy to understand. They also felt that the CAT was accurate, exact and reliable. Concerning the CAT-MD system, their comments were also satisfactory. They agreed on its effectiveness and user friendliness. Furthermore, most of the students felt that the CAT-MD was enjoyable mainly due to the use of the PDA.

Although most comments were positive, the students pointed out some weaknesses of the software. These comments from one-to-one evaluation were compared and processed together with experts' suggestions. Two of the more significant recommendations that were implemented during the revision phase are as follows:

- At the initial implementation of the CAT-MD, the examinee selects his/her response and then the test didn't forward automatically to the next item in order to give the student the possibility to rethink and alter his/her response. To move to the next item, the examinee must press the button "next". Although, the students received a short introduction on how to use the CAT-MD system, most of them expected that the test will administer automatically the next item after their response. Most of the students and experts suggested that this would be improved by altering the state of button "next" in order to indicate that the selection has been finished and that the examinee can proceed to the next phase (Figure 2).
- The process of displaying items, evaluating responses and selecting the next item to be administered is repeated until the accuracy of the estimate reaches a statistically acceptable level or twenty (20) items has been administered. In the prototype CAT-MD, the system didn't display any message about the correctness of student's response. Most of the students indicated that they prefer to know if their answer was correct or not immediately after the processing of each item. Furthermore, the pointed out as a weakness that they didn't know the total number of the items that they had already answered. After the revision, these suggestions were implemented (Figure 3).

6.2 Small group evaluation

The final stage of formative evaluation consisted of the small group evaluation in a "real world" environment. Twelve subjects took part in the small group evaluation. Students at the second grade of the senior high school were asked to volunteer for the small group evaluation. The subjects were selected carefully from the volunteers in order to represent the target population. There were 7 males and 5 females. While some of the students had little experience with computers, none had received any test on a computer and none had ever use PDA. The evaluation was conducted by collecting data about the assessment procedure from a variety of sources, using a variety of data gathering methods and tools.

In the beginning of the evaluation, all students were administered a paper-and-pencil test in order to compare testing time between CAT-MD and paper-and-pencil test. Twenty (20) multiple choice items and true false items were included in the paper-and-pencil test. These items were not included in the item pool of the CAT-MD to avoid the items' exposure. However, their difficulty level was similar to those included in the item pool in order to have comparable results. After this procedure the students received a short introduction on how to use the CAT-MD system. During the assessment procedure, the students used comment logs in order to note specific strengths or weaknesses of the system. When all the students have finished using the CAT-MD, an attitude questionnaire was used to determine their experience in using the system. Finally, a debriefing session was used to assess the subjective satisfaction of subjects on the instructional and interface design of CAT-MD.

One of the major purposes of the small group evaluation was to investigate the efficiency and the effectiveness of the CAT-MD. Effectiveness and efficiency are interrelated evaluation goals. The efficiency of the assessment is related to the time required for learners to conclude the test. The majority of the students completed the assessment procedure with the CAT-MD in significantly less time than the paper-and-pencil test since in general fewer items were needed with CAT-MD to estimate their ability. According to student's comments they didn't waste time attempting items that are too hard or trivially easy. So, regarding the efficiency, the instruction was considered successful.

Furthermore, small group evaluation can offer useful implementation information concerning the usability and the appeal of the system. Usability is a measure of the ease with which a system can be learned and used. The usability and the appeal of the system were investigated in this study through the attitude questionnaires and the debriefing session. An analysis of the data collected showed that overall the subjects were satisfied with the CAT-MD. Most of the students indicated that they were pleased with the design of the interface. They pointed out that it didn't contain heavy components. However, some students noticed that the images included in some items were not so clear due to small resolution and size.

Moreover, the majority of the students felt that the test was clear and secure. They indicated that it was easy to understand the adaptive assessment procedure. They were satisfied with the test adaptation based on their knowledge estimate. Furthermore, the majority of students felt that the CAT-MD offers more accurate results as they were graded not just on the basis of correct answers, but also on the difficulty level of the items answered. In addition, they stated pleased with the fact that the score was available immediately, providing instantaneous feedback. Summarizing, positive comments referred to the ease of taking the mobile test, the speed of the mobile test, and the relative absence of stress during the exam.

Most of the students felt challenged by the system's mobility. They indicated that the use of the PDA was very interesting and attractive. The majority of the students were enthusiastic with the implementation of the CAT on the PDA and they stated that they wanted to use the CAT-MD again. Most of them answered affirmatively to the question if they would like to use this tool for other subjects. In addition, they pointed out that CAT-MD can be used to any subject and especially to those subjects that use labs such as physics, chemistry etc.

7. Conclusion

This article describes the design and development of the CAT-MD (Computerized Adaptive Test on Mobile Devices), a prototype CAT on mobile devices such as PDAs. The design process was driven by continuous formative evaluation. Recommendations from experts and suggestions that resulted from one-to-one and small group evaluation are summarized below:

- CAT-MD is an effective and efficient assessment tool. The mobility eliminates the need for a specialized computer lab and it can be used anywhere, including a traditional classroom.
- The system needs to clearly indicate when the next item is ready to be administered.
- The system should provide immediate feedback on the correctness of examinee's answer.
- Students should be able to access information at any time with regards to the total numbers of items administered.
- Images on PDAs should be carefully selected taking into considerations the small resolution and size of the device.

These recommendations resulting from the formative evaluation of the CAT-MD, although they do not have a universal value as the design and development of a system always depends on the target population and the subject matter, could be seen as some points worth considering from designers of computerized adaptive test on mobile devices. Summative evaluation is the next step of our research in order to assess the effectiveness of the system with reference to validity.

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

Abdullah, S.C. (2003), *Student Modelling By Adaptive Testing - A Knowledge-Based Approach*. Unpublished PhD Thesis, University of Kent at Canterbury, June 2003.

Baker, Frank (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.

Boyd, A. M. (2003). *Strategies for Controlling Testlet Exposure Rates in Computerized Adaptive Testing Systems*. Unpublished PhD Thesis, The University of Texas at Austin, May 2003.

Chen, F., Myers, B., & Yaron, D. (2000). *Using handheld devices for tests in classes* (Tech. Report CMU-CS-00-152). Pittsburgh: Carnegie Mellon University, School of Computer Science.

Cook, R. P. (2000). The national classroom project an experience report. *In Proceedings of the 30th ASEE/ISEE frontiers in education conference* (pp. T1E/1-6). Champaign, IL: Stripes Publishing.

Dodd, B. G., De Ayala, R. J., & Koch W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 19 (1), 5-22.

Eggen, T.J.H.M. (2001). *Overexposure and underexposure of items in computerized adaptive testing*. Measurement and Research Department Reports 2001-1, Citogroep Arnhem.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Georgiev, T., Georgieva, E. and Smrikarov, A. (2004). M-Learning - a New Stage of e-Learning, International *Conference on Computer Systems and Technologies* - CompSysTech'2004, Rousse, Bulgaria, 17-18 June 2004. Available at http://ecet.ecs.ru.acad.bg/cst04/Docs/sIV/428.pdf

Hambleton, R. K., Swamination, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. California: Sage Publications Inc.

Jones V. & Jo H. J. (2004).Ubiquitous learning environment: An adaptive teaching system using ubiquitous technolog. *Proceedings of the 21st ASCILITE Conference*, Perth, Western Australia, 5-8 December 2004.

Lilley, M and Barker, T. (2002). The Development and Evaluation of a Computer-Adaptive Testing Application for English Language, *6th Computer Assisted Assessment Conference*, July 2002, Loughborough.

Lilley, M and Barker, T. (2003). An Evaluation of a Computer Adaptive Test in a UK University Context, *7th Computer Assisted Assessment Conference*, 8th and 9th July 2003, Loughborough.

Linacre, J. M. (2000). Computer-Adaptive Testing: A Methodology whose Time has Come. MESA Memorandum No. 69. Published in Sunhee Chae, Unson Kang, Eunhwa Jeon and 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*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Luvai F. M. (2005). Mobile learning: A framework and evaluation, *Computers & Education*, In Press, Available online 20 December 2005.

Rudner, L.M. (1998). "An online, interactive, Computer Adaptive Testing Tutorial". 11/98. Available at http://EdRes.org/scripts/cat

Segall N., Doolen T., Porter D. (2005). A usability comparison of PDA-based quizzes and paper-and-pencil quizzes. *Computer & Education*, 45 (2005), 417-432.

Sharples, M. (2000). The Design of Personal Mobile Technologies for Lifelong Learning, *Computers and Education*, 34 (2000), 177-193.

Tessmer M. (1993). *Panning and Conducting Formative Evaluations*. Kogan Page Limited.

van der Linden, W.J., Glas, C.A.W. (2003). Preface. In van der Linden, W.J., Glas, C.A.W (Eds). *Computerised Adaptive Testing: Theory and Practice*. Dordrecht, Boston, London: Kluwer Academic Publishers, vi-xii.

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

Wise, S. L. (1997, March). Overview of practical issues in a CAT program. *Paper presented at the annual meeting of the National Council on Measurement in Education*, Chicago IL.

Zurita, G., & Nussbaum, M. (2004). Computer Supported Collaborative Learning Using Wireless Interconnected Handheld Computers. *Computers & Education*, 42 (2004), 289–314.

Vitae

Dr. Evangelos Triantafillou is 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 and conferences.

Dr. Georgiadou Elisabeth is a lecturer in Technical Institute of Thessaloniki, Greece and she is working on the field of educational hypermedia systems and in particular the design and evaluation of these systems. Her research was presented at a number of international conferences.

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.

Figures

A. The CAT-MD on HP iPAQ (PDA) B. The CAT-MD on Motorola MPx220 (mobile)

Figure 1: Interface of CAT-MD

Figure 2: System screen before an after the revision

Figure 3: System screen after the revision