Triantafillou, E., Georgiadou, E., & Economides, A. A. (2008). CAT-MD: **Computerized adaptive testing on mobile devices**. *International Journal of Web-Based Learning and Teaching Technologies* (extended versions of the best papers presented at m-ICTE2006 Conference), Vol. 3, No. 1, pp. 13-20. IGI Global. ISSN: 1548-1093, E-ISSN: 1548-1107

CAT-MD: Computerized Adaptive Testing on Mobile Devices

Evangelos Triantafillou
Center of Educational Technology
Dodekanisou 21, Thessaloniki 55131, Greece
vtrianta@edutech.gr

Elissavet Georgiadou
Center of Educational Technology
Karaoli & Demetriou 46, Thessaloniki 57001, Greece
elisag@otenet.gr

Anastasios A. Economides
University of Macedonia, Department of Computer Networks
Egnatia 156,Thessaloniki 54006, Greece
economid@uom.gr

ABSTRACT

In the last decade the use of different mobile products such as mobile phones and Personal Digital Assistant (PDA) devices has increased rapidly. In parallel, the use of computerized-adaptive testing (CAT) has expanded mainly due to the advancements in communication and information technology. The introduction of mobiles devices into the learning pedagogy can compliment e-learning and e-testing by creating an additional channel of assessment with mobile devices. Although, mobile computing has become an important and interesting research issue, little research has been done on the implementation of CAT using mobile devices. The current study describes the design issues that were considered for the development and the implementation of a CAT on mobile devices, the CAT-MD (Computerized Adaptive Testing on Mobile Devices).

Keywords: *Computerized Adaptive Testing; Mobile Learning*

COMPUTERIZED ADAPTIVE TESTING

The recent years Computer Based Testing (CBT) is widely used in educational and training as there are a number of perceived benefits in using computers for assessing performance such as: (a) large numbers can be marked quickly and accurately, (b) students response can be monitored, (c) assessment can be offered in an open access environment, (d) assessments can be stored and reused, (e) immediate feedback can be given, (f) assessment items can be randomly selected to provide a different paper to each student Harvey and Mogey (1999). Moreover, another benefit of CBTs would be to bring the assessment environment closer to the learning environment. Software tools and web-based sources are frequently used to support the learning process, so it seems reasonable to use similar computer-based technologies in the assessment process (Baklavas et al., 1999, Lilley & Barker, 2002).

Most types of CBT are based on 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, giving little or no attention to the ability of each individual examinee. However, in Computerized Adaptive Testing (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. 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) – 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 items 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 Testing (Rudner, 1998; Linacre, 2000).

Since the mid-80s when the first CAT systems became operational, i.e. the Armed Services Vocational Aptitude Battery for the US Department of Defense account (van der Linden & Glas, 2003) using adaptive techniques to administer multiple-choice items, much research and many technical challenges have made new assessment tools possible. The availability of advanced mobile technologies have started to extend e-learning by creating an additional channel of assessment with mobile devices such as hand phones, Personal Digital Assistants (PDAs) or pocket PCs.

MOBILE LEARNING

In the last decade 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).

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, 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 our research. 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 Testing 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.

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, Swamination & Rogers, 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.

In IRT-based CAT as each student answers a question, his or her response is evaluated as being either correct or incorrect. The process of displaying questions, evaluating responses and selecting the next question to be administered based on the student's latest estimated ability is repeated until a stopping rule has been reached or a certain number of questions has been administered, whichever happens first. 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, 1995). The following sections describe these components of the CAT-MD system.

Item Pool

The most important element of a CAT is the item pool, which is a collection of test items that includes a full range of levels of proficiency, and 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 from "very easy" to "very hard" and the values used cover the range between -2 and +2.

Item Selection

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 a dichotomous IRT model.

The main aspect of IRT is the Item Characteristic Curve (ICC) (Baker, 2001). ICC is an exponential function, which expresses the probability of a learner with certain skill level correctly answering a question of a certain difficulty 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(θ) is estimated applying the formula shown in Eq. (1). Otherwise, the function Q(θ)=1-P(θ) is used.

The Item Information Function (IIF) is also considered as an important value in the IRT'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:

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

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

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 θ=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:

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

where

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

where θ is the skill level after n questions, and $u_i = 1$ if the response is correct and $u_i = 0$ for the incorrect response.

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(θ), 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 (θ) is less than 0.30 or when the maximum number (that is 20) of items has been administered.

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.

The CAT-MD on HP iPAQ (PDA) | *The CAT-MD on Motorola MPx220 (mobile)*

Fig. 1 Interface of CAT-MD

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.

Figure 2 illustrates how a test item is presented to an examinee within the prototype. Each question has four multiple choice answers and the user can select the correct one by clicking the corresponding button. The system responds immediately indicating whether the selected answer is correct or not. The user can not alter his/her selection as this is not permitted from the CAT's architecture. Every time the test statistically estimates the user's ability based on the answer given and then presents an item that matches this new ability. At the lower right corner of the screen, there is a button that becomes active whenever the user completes the selection in each item. As a result, the user cannot omit any item as this would conflict with the item selection algorithm.

Further, at the lower left corner of the screen, a number appears that corresponds to the total number of the items that the user has already answered. The user does not know when the test will terminate, however, it is considered useful to display the total number of the answered items.

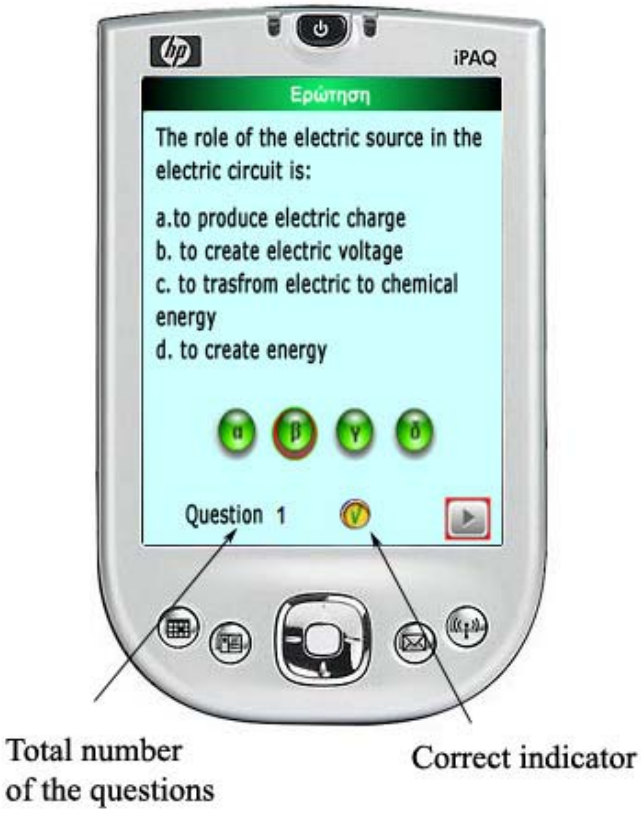

Fig. 2 Screenshot of CAT-MD

SUMMARY

This article describes the design and development of the CAT-MD (Computerized Adaptive Testing on Mobile Devices), a prototype CAT on mobile devices such as PDAs. Currently, the basic architecture of the system has been implemented. The prototype uses the Item Response Theory (IRT) as an underlying psychometric theory. Four main components are developed within the prototype: the item pool, the item selection procedure, the ability estimation and the stopping rule. Further research is on progress concerning the evaluation in order to investigate the effectiveness and efficiency of the system and also to assess its usability and appeal.

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.

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

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

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

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.

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

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

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

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

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

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

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.

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

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