Chapter 1.21
Evaluation of Computer 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 the exams, the workload and delay of scoring, and the human errors. Also, they try to increase the accuracy and efficiency of the testing. Recently, most examination organizations use computer 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 three 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 a 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 cheating, reducing the examinee's anxiety, challenging but not frustrating the examinees, as well as for immediate and continuous examinee guidance based on knowledge, proficiency, ability, and performance. Thus, many organizations are further driving towards computer 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 precalibrated items appropriate for the level of the specific student. Wainer (1990) indicates 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 (Rudner, 2002; Wainer, 1990; Welch & Frick, 1993). 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 the examinee 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, Shermis, Brutten, & Perkins, 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 (Carlson, 1994; Jacobson, 1993; Wainer, 1990; Wainer, Dorans, Eignor, Flaugher, Green, Mislevy, Steinberg, & Thissen, 2000). However, Lilley, Barker, and Britton (2004) argue 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 jeopardized 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, Revuelta, Ximenez, and Abad (2000) show 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) point out that when examinees are allowed to change answers, they are more likely to decrease their anxiety and improve their scores and score gains. Lilley and Barker (2003) show 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) proposes 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.

Although major organizations develop and use CAT systems, there is no work to evaluate 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 and O'Hare (2003) address this need to evaluate educational software. As Wise and Kingsbury (2000) state, although CAT is a relatively simple idea, the real-