Learners' Perceived Level of Difficulty of a Computer-Adaptive Test: A Case Study

Mariana Lilley, Trevor Barker, and Carol Britton

University of Hertfordshire, School of Computer Science,
College Lane, Hatfield, Hertfordshire AL10 9AB, United Kingdom
{M.Lilley, T.1.Barker, C.Britton}@herts.ac.uk

Abstract. A computer-adaptive test (CAT) is a software application that makes use of Item Response Theory (IRT) to create a test that is tailored to individual learners. The CAT prototype introduced here comprised a graphical user interface, a question database and an adaptive algorithm based on the Three-Parameter Logistic Model from IRT. A sample of 113 Computer Science undergraduate students participated in a session of assessment within the Human-Computer Interaction subject domain using our CAT prototype. At the end of the assessment session, participants were asked to rate the level of difficulty of the overall test from 1 (very easy) to 5 (very difficult). The perceived level of difficulty of the test and the CAT scores obtained by this group of learners were subjected to a Spearman's rank order correlation. Findings from this statistical analysis suggest that the CAT prototype was effective in tailoring the assessment to each individual learner's proficiency level.

1 Introduction

Computer-adaptive tests (CATs) are computer-assisted assessment applications in which the level of difficulty of the questions is dynamically tailored to the proficiency level of individual learners. Wainer [8] suggests that CATs mimic aspects of an oral interview in which the tutor would adapt the interview by choosing questions appropriate to the proficiency level of individual learners. Jettmar & Nass [5] describe CATs as a special case of intelligent user interfaces, in which user performance is unobtrusively monitored and the level of difficulty of the questions adapted accordingly. Brusilovsky [2] cites CATs as one of the elements of a paradigm shift within educational software development, from "one size fits all" to one capable of offering higher levels of interaction and personalisation. Conejo et al. [3], Fernandez [4], Lilley et al. [6] amongst others have reported on the benefits of the CAT approach in a range of educational contexts.

The main aim of a CAT software application is to provide learners with questions that are sufficiently challenging, and yet not so difficult that could lead to frustration or bewilderment on the part of the learners.

CATs are based on Item Response Theory (IRT). IRT is a family of mathematical functions that attempts to predict the probability of a user successfully completing a task or, more specifically, answering a question correctly. An overview of our CAT prototype is provided in the next section of this paper.

M.F. Costabile and F. Paternò (Eds.): INTERACT 2005, LNCS 3585, pp. 1026 – 1029, 2005.

2 Prototype Overview

The CAT prototype described here comprised a graphical user interface, a question bank and an adaptive algorithm based on the Three-Parameter Logistic (3-PL) Model from IRT [7, 8].

Equation 1 [7] shows the 3-PL Model function used to predict the probability of a test-taker with an unknown proficiency level θ correctly answering a question of difficulty *b*, discrimination *a* and pseudo-chance *c*. In Equation 1, questions with greater values for the difficulty *b* parameter require greater proficiency on the part of the test-taker to answer the question correctly than those questions with lower values. The discrimination *a* parameter describes the question's usefulness when distinguishing amongst test-takers near a proficiency level θ [7]. The pseudo-chance *c* parameter indicates the probability of a test-taker answering a question correctly by chance.

$$P(\theta) = c + \frac{1-c}{1+e^{-1.7a(\theta-b)}} \tag{1}$$

A typical CAT starts with a question of medium difficulty. In general terms, a correct response will cause a more difficult question to be administered next. Conversely, an incorrect response will cause a less difficult question to follow. As the test proceeds, the mathematical function shown in Equation 1 is employed to estimate the test-taker's proficiency level. The proficiency level estimate is then used to select the question to be administered next. A detailed description of IRT is beyond the scope of this paper and the interested reader is referred to Lord [7] and Wainer [8].

3 The Study

A sample of 113 Computer Science undergraduate students participated in a summative assessment session using the CAT application. The assessment session took place in computer laboratories, under supervised conditions. Participants had 35 minutes to answer 24 objective questions organised into 4 topics within the Human-Computer Interaction (HCI) subject domain. Participants' performance on this assessment is summarised in Table 1.

In Table 1, the value for the proficiency level ranged from -3 (lowest) to +3 (highest). In a CAT, we are not concerned with the number of correct responses. Indeed most participants are expected to answer approximately 50% of the questions correctly, as it is anticipated that the questions administered to each individual test-taker would be tailored to that individual's proficiency level within the subject domain. The focus is therefore on the level of difficulty of the questions answered correctly by each individual test-taker.

Table 1. Summary of test-takers' performance (N=113)

	Mean	Standard Deviation
Proficiency Level	0.08	1.07
% Correct responses	47.64	10.37

At the end of the assessment session, all participants were asked to rate the difficulty of the test that they had just taken from 1 (very easy) to 5 (very difficult). The mean test difficulty, as perceived by the participants, was 3.37 (SD=0.60, N=113). Their ratings are illustrated in Table 2.

Table 2. Level of difficulty of the test as perceived by the participants (N=113)

1 (Very easy)	2	3 (Just right)	4	5 (Very difficult)	Total
0	2	72	34	5	113

It was important to investigate whether or not the correlation between participants' performance and their perceptions on the level of difficulty of the overall test was statistically significant. Such statistical analysis is the focus of the next section of this paper.

4 Perceived Test Difficulty According to Learner's Performance

Participants' results and their perception of the test's difficulty were subjected to a Spearman's rank order correlation.

No statistically significant correlation was found between participants' proficiency levels and the test's difficulty rating, such as rs = -0.092, Sig. (2-tailed) = 0.333, N=113. The data gathered in this study was also subjected to a Kruskal-Wallis Test, where Chi-Square = 0.736, df = 2, Asymp. Sig. = 0.692. The results from this statistical analysis are summarised in Table 3.

Table 3. Level of difficulty of the test as perceived by the participants, according to test performance (N=113)

	Group	N	Mean Rank
Rating	Low-performing participants	38	58.96
	Intermediate-performing participants	36	58.24
	High-performing participants	39	53.95

The results shown in Table 3 were taken to indicate that the participants' performance on the test had no effect on the perceived difficulty of test. This is of particular importance, since one of the goals of our CAT prototype was that test-takers would be presented with tasks that are challenging and motivating, rather than tasks that are either too difficult and therefore frustrating, or too easy and thus uninteresting.

5 Summary and Concluding Remarks

Interactive software applications that adapt to their users have been rapidly gaining in importance within the HCI field. CATs are an example of such interactive applications, as the level of difficulty of the tasks is adapted to the proficiency level of indi-

vidual users. Despite the substantial amount of work that has been conducted in this area, it can be argued that users' perceptions of the CAT approach have been under represented in the literature.

In previous studies we have shown that our CAT prototype supports accurate measurement of learners' proficiency levels [1, 6]. This paper is concerned with users' perceptions of the level of difficulty of an assessment session that was interactively created using our CAT prototype. Findings from this empirical study were taken to indicate that the CAT approach was effective in providing individual users with a sufficient challenge whilst using the application. An important assumption of our work was that an appropriate degree of challenge would enhance test-takers' motivation. This could, in turn, contribute towards an enhancement of their learning experience. The use of computers alone is not sufficient to motivate users of educational software applications. Interaction is a valuable tool to maintain learners' motivation and therefore whenever possible, educational software should adapt to the learner's proficiency levels and skills.

References

1. Barker, T. & Lilley, M. Are Individual Learners Disadvantaged by the Use of Computer-Adaptive Testing? In *Proceedings of the 8th Learning Styles Conference*. University of Hull, European Learning Styles Information Network (ELSIN), pp. 30-39, 2003.
2. Brusilovsky, P. Knowledge Tree: A Distributed Architecture for Adaptive E-Learning In *Proceedings of the 13th World Wide Web Conference*, May 17-22, New York, New York, USA, pp. 104-113, 2004.
3. Conejo, R., Millán, E., Pérez-de-la-Cruz, J. L. & Trella, M. An Empirical Approach to On-Line Learning in SIETTE In Proceedings of the 2000 Intelligent Tutoring Systems Conference, LNCS **1839**, pp. 605-614, 2000.
4. Fernandez, G. Cognitive Scaffolding for a Web-Based Adaptive Learning Environment In *Proceedings of 2003 International Conference on Web-based Learning*, LNCS **2783**, pp. 12–20, 2003.
5. Jettmar, E. & Nass, C. Adaptive testing: effects on user performance In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*. Minneapolis, Minnesota, USA, pp. 129-134, 2002.
6. Lilley, M., Barker, T. & Britton, C. The development and evaluation of a software prototype for computer adaptive testing. *Computers & Education Journal* **43**(1-2), pp. 109-122, 2004.
7. Lord, F. M. *Applications of Item Response Theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates, 1980.
8. Wainer, H. *Computerized Adaptive Testing (A Primer)*. 2nd Edition. New Jersey: Lawrence Erlbaum Associates, 2000.