The Evaluation of a Computer-Adaptive Test

Autoria: Mariana Lilley, Trevor Barker, Marta de Campos Maia

Abstract

In a traditional computer-based test (CBT), the questions presented during a given assessment session are not tailored for the specific ability of an individual student. In contrast, in a computer-adaptive test (CAT), the questions are selected dynamically based on the student's individual performance during the assessment. A typical CAT is based on Item Response Theory (IRT), and the some of the characteristics of IRT and its Three-Parameter Logistic Model (3-PL) are outlined here. Furthermore, this paper presents a report on the development of research recently completed by the University of Hertfordshire in the United Kingdom and Fundação Getúlio Vargas in Brazil, in which both the increased use of computer-assisted assessment in Higher Education and the use of CATs within Business Administration distance learning were discussed. In this study, several evaluation methods were employed, including heuristic evaluation, online questionnaires and focus groups. These methods are explained here and their usefulness is discussed in the final part of this paper. It is hoped that the research described here will be of interest to practitioners and researchers in a wide range of educational contexts.

1. Introduction

The use of computer-assisted assessments has been growing [3,5,11] and one of the most popular types of computer-assisted assessment currently in use is the computer-based test (CBT). CBTs in many respects mimic a paper-and-pencil test, in which students are presented with a predefined set of questions that they must answer by entering responses into the computer. In contrast, in a computer-adaptive test (CAT), the questions are selected by the computer according to the student's performance during the test. The presentation of questions for each individual student is adapted dynamically as the test progresses according to his or her performance. If a student answers a question correctly, a more challenging question is next presented. Conversely, if the response provided by the student is incorrect, an easier question is picked next, until an equilibrium level is reached or the test ends.

An underlying principle within a CAT is to mimic aspects of an oral interview [7], in which an interviewer would start a session by asking a question of medium difficulty. If the answer given by the student is correct, a more difficult question follows. Conversely, if the answer provided by the student is incorrect, a simpler question is asked next. The idea behind this assessment method is that questions that are either too difficult or too easy do not provide substantial information regarding the level of ability of a particular student in a subject, nor do they challenge or motivate a student [13].

The use of computer-adaptive testing has been increasing, and indeed replacing the traditional CBTs in some areas. The implementation of the former in examinations such as the Graduate Management Admission Test (GMAT) [4], Test of English as a Foreign Language (TOEFL) [20] and Microsoft Certified Professional (MCP) [6] are evidence of this trend.

In order to evaluate the appropriateness of a CAT in a real Higher Education context, a prototype of CAT based on the 3-PL Model was designed and implemented. This prototype was then evaluated by both lecturers and students. The CAT prototype presented here is based on Item Response Theory (IRT) [15], which its central element is a family of mathematical functions that calculates the probability of a specific student answering a particular question correctly. The field of IRT is vast, and this paper only introduces the aspects of IRT that were applied to the prototype introduced here. Hence the reader interested in investigating this topic in more depth is referred to the writings of Lord [15], Wainer [21], Hambleton [9] and Linden [14]. A brief introduction to IRT is presented in the next section of this paper.

2. Item Response Theory (IRT)

The prototype of a CAT discussed here was based on the Three-Parameter Logistic Model (3-PL) within IRT. In this model, in order to evaluate the probability P of an examinee with an unknown ability θ answering an item of difficulty b correctly, the mathematical function shown in Equation 1 [15] is used.

Equation 1: The Three-Parameter Logistic Model

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

In Equation 1, e represents the natural logarithmic base (i.e. 2.71828…). The parameter b represents the item's difficulty, and within the prototype described here $-2 \leq b \leq 2$. The parameter a represents the item's discrimination, which facilitates the separation among examinees with abilities $\leq \theta$ from examinees with abilities $> \theta$ [9]. Finally, the values for the pseudo-chance, also known as "guessing parameter", vary from 0 to 1 or, in other words, $0 \leq c \leq 1$. For example, in a well-designed multiple-choice item with 5 options, an examinee with no knowledge has 1 in 5 chances of answering the item correctly by guessing, therefore c=0.2.

In order to demonstrate how the 3-PL Model is applied within the prototype introduced here, consider the information regarding a hypothetical item's database presented in Table 1.

Table 1: Hypothetical values of parameters from Equation 1

Item ID	*b*	*a*	*c*
1	-1.09	1.25	0.01
2	1.7	1.48	0.25
3	-1.09	0.95	0.10
4	0	1.5	0.10
5	-0.77	0.75	0.25
6	0.38	1.32	0.20
7	1.04	0.79	0.05
8	0.22	0.66	0.20
9	1.25	0.64	0.10
10	-1.29	1.59	0.25

The items represented in Table 1 are all objective items (e.g. multiple-choice or multiple-response questions) and therefore can be dichotomously scored or, in other words, scored as being either correct or incorrect.

The test starts with a randomly selected item of medium difficulty. Suppose that a given examinee is presented with Item 4, an item of medium difficulty (*b=0*), high discrimination (*a=1.5*) and a pseudo-chance *c* of *10%.* Given that in this example the examinee answered the first item correctly, Figure 1 represents the Item Characteristic Curve (ICC) for this item, which was calculated using Equation 1.

Figure 1: ICC curve for Item 4 answered correctly

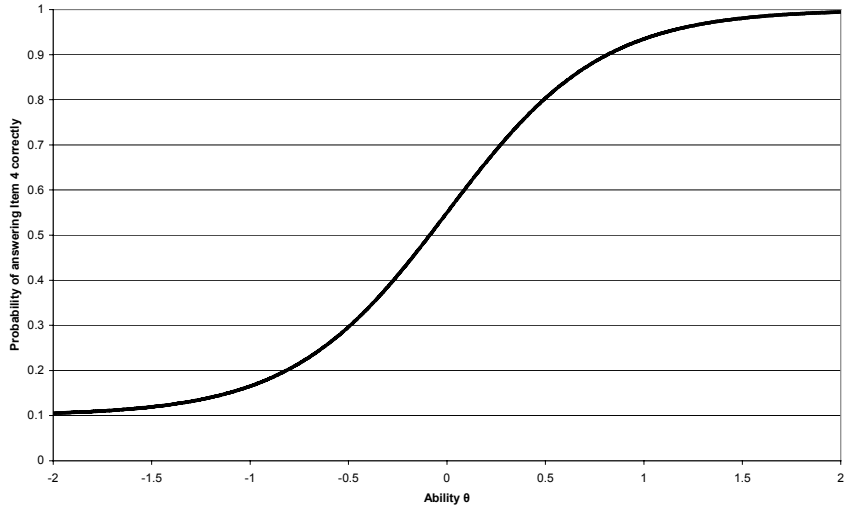
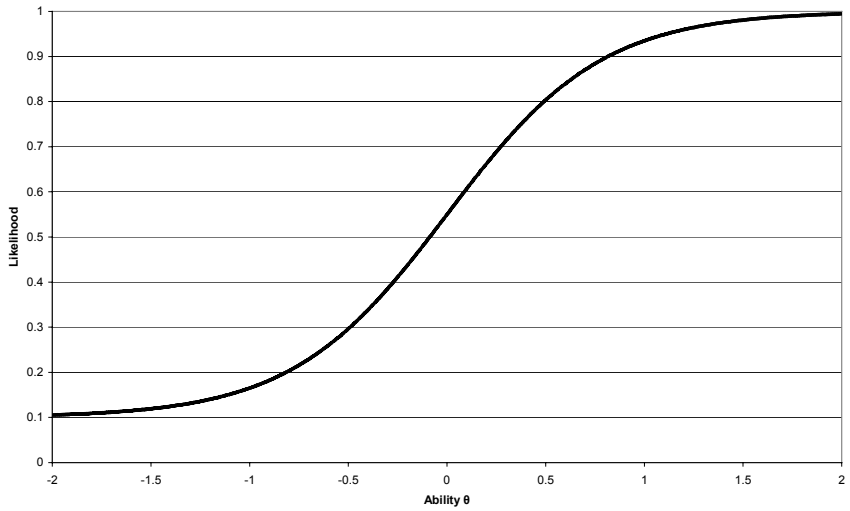


Figure 2 represents the *response likelihood curve* for this individual examinee. The response likelihood curve is the likelihood of an examinee answering a sequence of items, which is plotted by multiplying the ICCs for the relevant items. Since only one item has been answered so far, the ICC curve for Item 4 and the *response likelihood curve* are identical.

Figure 2: Response likelihood curve after Item 4 has been answered



In the event of the examinee answering the previous item correctly, a more difficult item follows. Item 7 has higher level of difficulty (*b=1.04*) than Item 4. The discrimination *a* is *0.79* and the pseudo chance *c* of this item is 5*%.* Suppose that our examinee has also answered Item 7 correctly, Figure 3 represents the ICC curve for Item 7.

Figure 3: ICC curve for Item 7 answered correctly

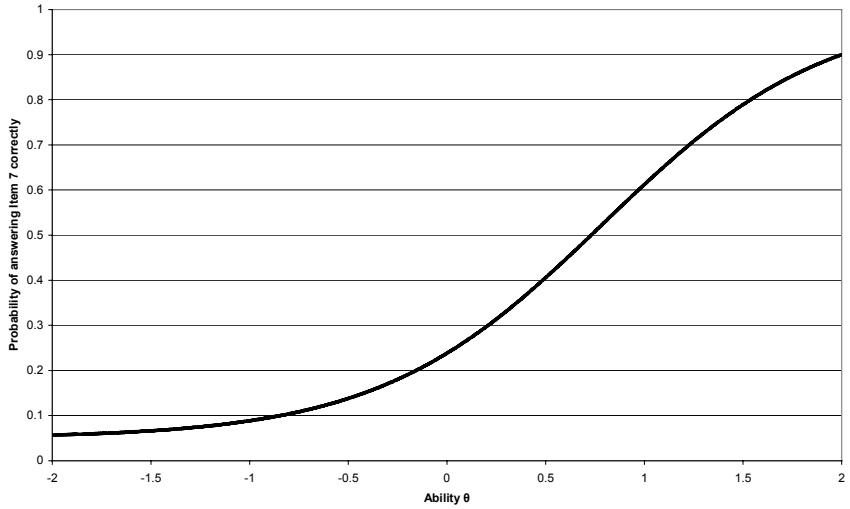
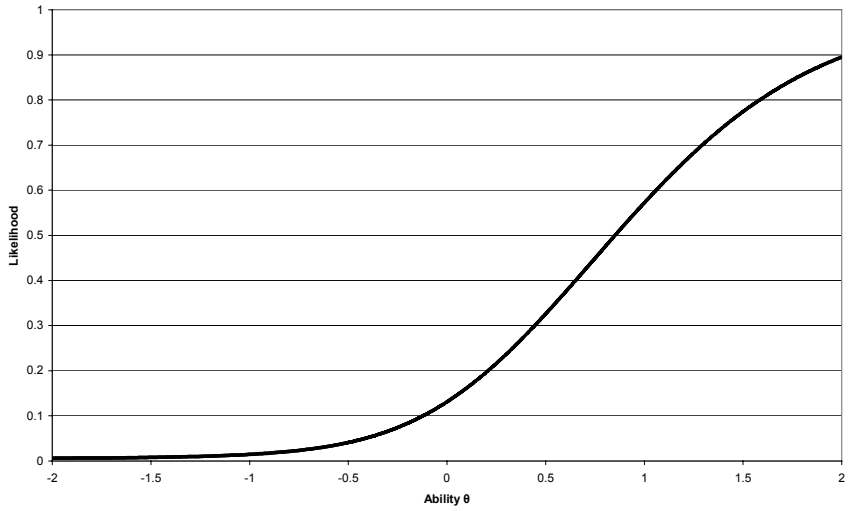


Figure 4 illustrates the currently response likelihood curve, which is the product of the ICC curves shown in Figures1 and 3.

Figure 4: Response likelihood curve after two items have been answered



In this example, the examinee has answered all the items presented correctly. The examinee's response likelihood curve is composed of the product of two S-shaped curves of type $P(\theta)$ and, therefore, the curve does not have a peak value. The same characteristic (i.e. no peak value) would have occurred if the examinee has answered all the items presented incorrectly, since the response likelihood curve would be calculated as being the product of various $(1-P(\theta))$ and, consequently, the curve would also not have a peak value within the range $-2 \leq \theta \leq 2$.

The examinee's response is evaluated as either being correct or incorrect, and a relevant ICC is generated for each response. If the response has been evaluated as correct, a more difficult item is presented next; otherwise an easier item is presented. This process would be repeated until at least one item has been answered correctly and one item has been answered incorrectly. The selection of which more difficult or easier item would follow is fairly random.

Suppose that our examinee is now presented with a more difficult item, which is Item 2. This item has difficulty *b=1.7*, discrimination *a=1.48* and pseudo-chance *c=0.25*. Given that the examinee's response for this answer has been evaluated as incorrect, Figure 5 illustrates the ICC curve for this item.

Figure 5: ICC for Item 2 answered incorrectly

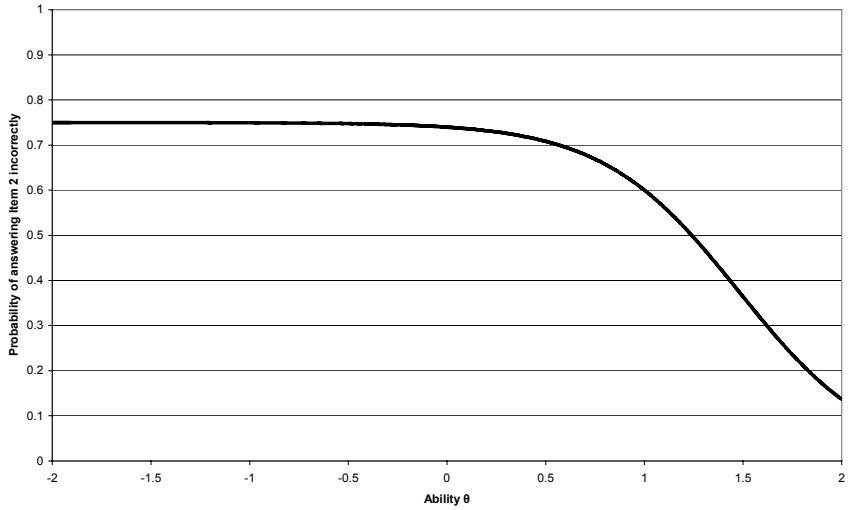
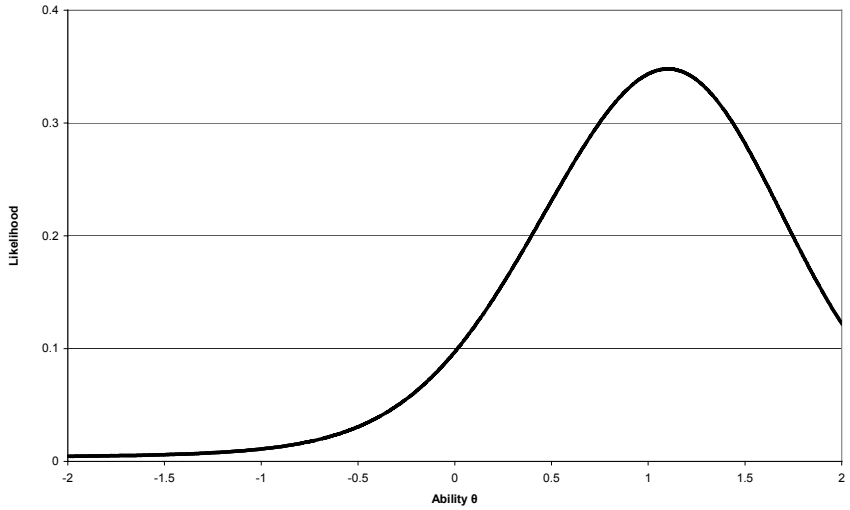


Figure 6 shows the response likelihood curve after three items have been answered.

Figure 6: Response likelihood curve after three items have been answered



When the examinee's response likelihood curve is formed by the product of at least one $P(\theta)$ and one $(1-P(\theta))$, the curve would typically have a peak. The value of the X-axis at the curve's peak, which in this example is *1.25*, is taken to be the new provisional ability θ. Wainer [21] indicates that at $\theta=b$, the mathematical functions provided by IRT can provide maximum information about the examinee's ability. Thus once a provisional ability has been established, the examinee is then supplied with an item from the item's bank for which the difficulty *b* is the closest value to the provisional ability θ. In other words, the items to be administered are not randomly selected anymore. In this specific example, the item to be administered next would be item 9, since it has *b=1.25*. This is one of the fundamental points of an *adaptive* test, to adapt the items

according to the responses and then provide the most appropriate items according to each examinee's individual responses. Typically the responses from many questions are necessary in order to estimate a candidate's ability.

In the prototype introduced here, the process of presenting items, evaluating the responses and dynamically selecting the next item to be administered is repeated until a time limit has been reached or a certain number of questions has been administered, whichever happens first.

In the following section of this paper, we describe how the prototype was evaluated by a group of experts, in addition to a summary of the main findings from this evaluation.

3. The Heuristic Evaluation

The prototype introduced here was designed to estimate the level of proficiency in English for those students whose first language is not English. This prototype was initially evaluated by a group of eleven lecturers formed by both lecturers in Computer Science and in English for Academic Purposes within the University of Hertfordshire [12].

This first evaluation took the form of an heuristic evaluation [16] and prior to this evaluation, the experts attended a session in which the main characteristics of a CAT were explained. It was considered important that the experts were clear about how the computer-assisted assessment tool was intended to work as well as the objectives of the prototype.

After the briefing session, each expert independently evaluated different elements of the prototype's interface. These interface elements were then rated according to ten usability principles, using a Likert scale from 1 (poor) to 5 (excellent). These usability principles (i.e. heuristics) have been widely used in the software evaluation area and the interested reader is referred to the work of Molich and Nielsen [16], Preece [18] and Preece *et al.* [19]. Table 2 illustrates these guidelines and the scores obtained.

Table 2: Summary of the heuristic evaluation

	Poor				**Excellent**	
Usability Principle	**1**	**2**	**3**	**4**	**5**	**Mean**
Visibility of the system status	0	0	1	6	4	4.3
Match between system and the real world	0	0	1	4	6	4.5
User control and freedom	0	0	3	5	3	4.0
Consistency	0	0	0	5	6	4.5
Error Prevention	0	0	1	6	4	4.3
Recognition rather than recall	0	0	1	3	7	4.5
Flexibility and efficiency of use	0	0	5	2	4	3.9
Aesthetic	0	1	1	6	3	4.0
Feedback and errors	0	0	1	6	4	4.3
Help and documentation	0	2	0	6	3	3.9

It has been suggested that within an heuristic evaluation, five evaluators are able to detect 75% of the usability problems within a software application [18]. The scores obtained from the eleven experts involved in the evaluation process would therefore suggest that there were no major usability problems within the prototype.

After carrying out the heuristic evaluation, the experts were asked to rate ten statements from 1 (Unlikely) to 5 (Likely) in order to gather data on the prototype's usefulness as an educational tool. Table 3 summarises the results obtained in this section of the evaluation.

Table 3: Summary of the pedagogical evaluation

	Unlikely				Likely	
Pedagogical Measure	**1**	**2**	**3**	**4**	**5**	**Mean**
CAT would enable lecturers to mark summative assessments more quickly.	1	1	1	2	6	4.0
CAT would enable lecturers to mark summative assessments more accurately.	1	1	1	4	4	3.8
CAT as summative assessment tool would enable lecturers to detect students' educational needs.	1	0	7	1	2	3.3
Students would be receptive to using CAT in a summative assessment environment.	0	1	3	4	3	3.8
CAT as summative assessment tool would enable students to detect their educational needs.	4	0	4	2	1	2.6
CAT as formative assessment tool would enable lecturers to detect students' educational needs.	1	1	1	5	3	3.7
Students would be receptive to using CAT in a formative assessment environment.	0	0	2	5	4	4.2
CAT as formative assessment tool would enable students to detect their educational needs.	2	3	3	2	1	2.7
Students' interaction with the system would be simple and clear.	0	0	1	4	6	4.5
Students would find the system easy to use.	0	0	0	1	10	4.9

The results obtained indicate that the lecturers considered that the prototype would be valuable in terms of both speed and accuracy. However, the experts suggested that the use of objective items to assess the examinee's abilities of synthesis and evaluation is restricted, and this opinion is shared by the authors [13]. The evaluators also emphasised that the accuracy of the score given to an examinee relies on the correctness of the item parameters used in order to estimate the examinee's ability and therefore without an adequately large and calibrated items' bank the use of a CAT is limited. The experts also reported that the prototype would give greater assistance in a formative rather that in a summative assessment environment. The next stage in the process of evaluating the CAT prototype was to involve students in the evaluation procedure.

4. User Evaluation

In this evaluation, the software was used to test students' ability in English grammar and language. The students received no prior training on how to use the application and all students managed to complete the test without having difficulties related to the software's interface.

Twenty-seven students took part in the first student evaluation, during which each student was presented with a set of twenty items on the use of English grammar and language. These twenty items were divided into two groups of ten items each, in which the items were either dynamically selected (i.e. CAT) or predefined (i.e. CBT). Fifteen students answered the CAT items followed

by the CBT ones, and the remaining twelve students answered the items in the opposite group order (i.e. CBT followed by CAT). In both cases, the students were unaware of the group order.

For each item answered, starting with the second, the students were asked to rate the level of difficulty of both the item they had just answered and the test up to that point, from 1 (more difficult) to 5 (easier), as illustrated in Figure 7.

Figure 7: Screen dump of online questionnaire

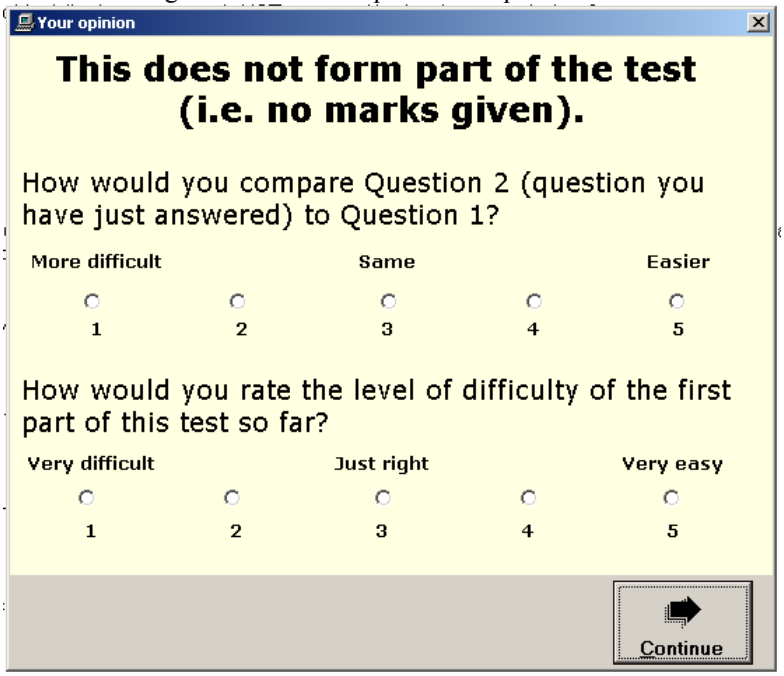


Table 4 summarises some of the data collected using the online questionnaire. Although the differences in the mean values obtained for both types of tests (i.e. CAT and CBT) presented in this table are not substantial, the slight difference in these values seems to indicate that the students perceived that within the CAT the items were more appropriate to their level of ability. For the CAT test, the mean values for level of difficulty of item (i.e. question) and test were respectively 2.9 and 3.1, while for the CBT these values were 2.5 and 2.7 within a 1 (more difficult) to 5 (easier) scale.

Table 4: Summary of the answers obtained for the research questions

		More difficult		Just right		Easier	
		1	**2**	**3**	**4**	**5**	**Mean**
CAT	Level of difficulty of item	21	70	88	32	32	**2.9**
	Level of difficulty of test	10	45	124	33	31	**3.1**
CBT	Level of difficulty of item	53	70	81	29	10	**2.5**
	Level of difficulty of test	24	84	100	23	12	**2.7**

One of the main objectives of an adaptive test is to dynamically select the items presented to each individual student in order to match his or her estimated ability, and this characteristic is illustrated in Table 5.

Table 5: Number of correct responses

Amount of Correct Responses	CAT Number of students	CBT Number of students
Less than 50% of correct responses	2	10
Greater than or equal to 50% of correct responses	25	17

The fact that the items are dynamically selected, and therefore not identical to all students, presents potential limitations. This issue was further explored by the focus group session reported later.

In order to find any severe usability problems with the software, the students' use of the software was observed during the test by a trained observer in Human-Computer Interaction (HCI). In addition to the observation of the students' behaviour during the test and analysis of the data collected using the online questionnaire, a Pearson's Product Moment correlation was performed on the scores obtained by the students on each part of the test, and the results of this correlation are shown in Table 6.

The results obtained from the statistical analysis shown in Table 6 show that there is an important correlation between the CAT score, CAT level obtained and CBT score ($p<0.001$). Furthermore, we interpret the data shown in Table 6 as a corroboration of the experts' opinion in that the prototype's interface does not negatively affect students' performance during the test.

Table 6: Pearson's Product Moment correlation between the scores and levels for participants in CBT and CAT sections of two assessments N = 27

Variable	Correct responses CBT	Correct responses CAT	Level CAT
Correct responses CBT			
5. Pearsons R	*	0.486	0.398
6. Significance		$p < 0.001$	$p < 0.001$
Correct responses CAT			
Pearsons R	*	*	0.516
Significance			$p < 0.001$

In order to understand at the individual level what was happening as students used the CAT, a second user evaluation was undertaken, in which seven randomly selected students participated. This evaluation involved a CAT in which students were presented with forty items on the use of English grammar and language. Figure 8 illustrates the pattern of responses obtained for two randomly selected students. The curves for the five remaining students have similar characteristics and therefore were not included in the graph for the purpose of clarity.

Figure 8: Estimation of ability

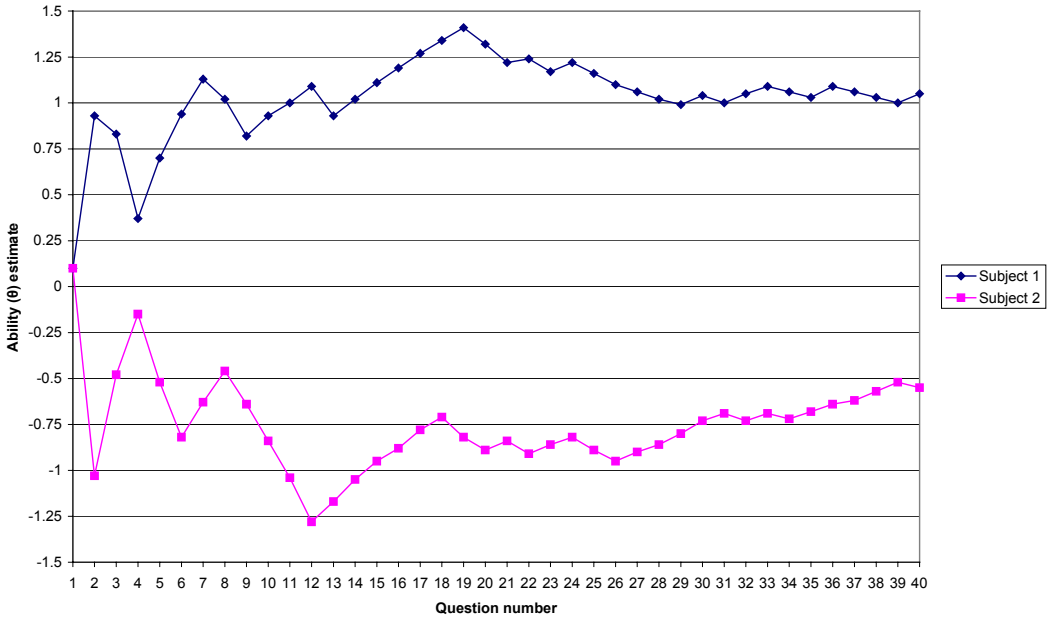


Figure 8 illustrates how the adaptive process worked. The CAT prototype first presents an item of medium difficulty. After a correct answer is given, the adaptive algorithm statistically estimates the student's ability as higher than previously estimated, and then presents an item that matches that new higher estimate of the student's ability (i.e. $\theta=b$). If the answer for the following item is once more correct, the algorithm estimates the student's ability as higher still. On the other hand, if the answer provided is incorrect, the adaptive algorithm estimates the student's ability as being lower. In the prototype introduced here, this process of estimating the student's ability and presenting a new item continues, with the algorithm gradually locating the student's level of ability, until forty items have been administered.

By observing the estimated level of ability after each item makes it possible to identify not only that this estimation becomes more accurate with each item answered, but also that the estimation of the students' abilities tends to be restricted to a definite range of values in the longer term. In addition, it can be seen from Figure 8 that both students were not presented with the same set of items as they performed differently during the test.

The data gathered within the heuristic and pedagogical evaluations provided the authors with relevant information regarding the interface's usability and the prototype's potential in pedagogical terms. The first user evaluation made it available information regarding the prototype's usability through observation and the students' perceived level of difficulty through online questionnaire. In order to investigate the attitude of students to the CAT approach, twelve randomly selected students from the original group were invited to participate in a focus group session and the findings from this session are described next.

5. Focus Group

One of the main advantages of a focus group is the possibility of gathering information about diverse or sensitive issues that were likely to be overlooked when using quantitative methods Preece *et al.* [19], such as the online questionnaire described in the previous session of this paper.

The focus group was guided by an experienced facilitator in HCI and lasted forty minutes. The session was recorded on video to facilitate later analysis. The main purpose of this focus group session was to investigate usability issues related to the user interface and students' attitude towards the use of CATs in summative and formative assessments.

During the focus group session, vital information related to the use of computer-adaptive testing was gathered. For example, the participant students indicated potential limitations of the CAT approach, such as the students' perception of fairness within the adaptive test. In a computer-adaptive test, the set of items answered by one student is very unlikely to be the same set presented to another. Moreover, the final score depends not only on the number of items answered correctly but also on their level of difficulty. Thus one student may provide the same amount of correct responses as any other and yet achieve a lower grade.

Despite the potential limitations described earlier, students were mainly positive about the use of such a computer-assisted assessment tool. The participant students indicated that tests that are too easy are meaningless, and tests that are too difficult are usually frustrating and incite them to "guess the answer" since they cannot answer based on their knowledge. This fact would suggest a benefit of computer-adaptive tests over traditional computer-based tests, since the items presented are interactively selected according to each student's previous responses and therefore are more likely to be fit for each student's individual ability.

6. Discussion

At the time of writing there is an increased demand for the use of computers within the Brazilian educational sector [2,8,10]. The reasons for this growth range from the relationship between computer skills and employability [8], the potential to make a more efficient use of the computational resources already available within some Higher Education (HE) institutions to the popularisation of distance learning [16].

For some, the use of CAT in HE offers the potential to assess students in a more efficient and interactive way, especially when compared to those levels of efficiency and interaction provided by traditional CBTs. Furthermore, the CAT approach would be particularly interesting for Business Administration programmes, since this approach would bring their assessment practices in line with the practices of leading organisations, such as the Graduate Management Admission Test (GMAT), which is already using CAT.

Notwithstanding the benefits listed earlier, our study has identified several classes of potential limitations associated with this approach to assessment. These problems may be classified under the following headings:

1 Interface design

2 Application performance

3 Tutor attitude

4 Student attitude

5 Pedagogical issues

6 Ethical issues

If CAT is to become an important and useful tool in assessing students in Business Administration programmes and HE in general, then the solutions to at least some of these problems will be necessary. We argue that the evaluation approach adopted in this research was useful both in identifying the range of problems involved in the use of CAT and also in suggesting potential solutions. The use of expert evaluation methods helped clarify interface design and performance issues. It also provided insight into tutor attitude that will be one focus of future research. The use of online data collection methods was able to provide information on the performance and attitude of students, as they were actually involved in taking the assessments, using the software. Issues that arose from these two lines of research were further explored in the focus group study, where pedagogical and ethical issues could be explored more fully.

The evaluation of educational software is a complex exercise that is often performed poorly. It has been argued [1] that unless evaluation is performed fully and in correct context, it is better not to do it at all. The use of a range of techniques in our study was able to cross the boundaries of interface design, pedagogical and ethical issues, and stakeholder attitude. An understanding of any one area is only useful in the context of the others in order to obtain a balanced assessment of CAT in Business Administration programmes and HE in general. This was provided in our study by such an integrated evaluation approach.

A major limitation to this approach is the need to design and conduct complex evaluation. Typically this involved an additional effort equivalent to the effort involved in designing and implementing the software itself. It is important that this effort is factored in during the early stages of the project.

7. References

1 Barker, T and Barker, J. (2002). ***The evaluation of complex, intelligent, interactive, individualised human-computer interfaces: What do we mean by reliability and validity?*** Proceedings of the European Learning Styles Information Network Conference, University of Ghent, June 2002.

2 ***Brasil tem quase 150 milhões de excluídos digitais.*** (10 Apr 2003) [online]. Available from http://br.news.yahoo.com/030410/16/bi3u.html [27 Apr 2003].

3 Brown, S. (1999). ***Institutional strategies for assessment in Brown, S. and Glasner, A. Assessment Matters in Higher Education: Choosing and Using Diverse Approaches.*** Suffolk: The Society for Research into Higher Education and Open University Press.

4 ***Computer-Adaptive Format (2002)*** [online]. Available from http://www.mba.com/mba/TaketheGMAT/TheEssentials/WhatIstheGMAT/ComputerAdaptiveFormat.htm [28 Apr 2003].

5 Conole, G. and Bull, J. (2002). ***Pebbles in the Pond: Evaluation of the CAA Centre***. Proceedings of the 6th Computer-Assisted Assessment Conference, Loughborough, p. 63-73.

6 ***Exam and Testing Procedures (2002) [online].*** Available from http://www.microsoft.com/traincert/mcpexams/faq/procedures.asp [23 Apr 2003].

7 Freedle, R.O. and Duran, R.P. (1987). ***Cognitive and Linguistic Analyses of test performance.*** New Jersey: Ablex Publishing Corporation.

8 ***FSM: A era dos "cérebros-de-obra"*** (27 Jan 2003) [online]. Available from http://br.news.yahoo.com/030127/7/abav.html [27 Apr 2003].

9 Hambleton, R.K. (1991). ***Fundamentals of Item Response Theory.*** California: Sage Publications Inc.

10 ***Intel investe US$ 700 milhões em educação.*** (9 Apr 2003) [online]. Available from http://br.news.yahoo.com/030409/13/bgjr.html [27 Apr 2003].

11 Jolliffe, A., Ritter, J. and Stevens, D. (2001). ***The online learning handbook: developing and using web-based learning.*** London: Kogan Page.

12 Lilley, M. and Barker, T. (2002). ***The Development and Evaluation of a Computer-Adaptive Testing Application for English Language.*** Proceedings of the 6th Computer-Assisted Assessment Conference, Loughborough, p. 169-184.

13 Lilley, M., Barker, T. and Maia, M. (2002). ***Web-based adaptive testing in distance learning: an overview*** [CD-ROM]. 5th Simpósio de Administração da Produção, Logística e Operações Internacionais, São Paulo (Brazil), October 1-4, 2002.

14 Linden, W.J. Van Der (1997). ***Handbook of Modern Item Response Theory.*** NewYork: Springer-Verlag.

15 Lord, F.M. (1980). ***Applications of Item Response Theory to practical testing problems.*** New Jersey: Lawrence Erlbaum Associates (Publishers).

16 Meirelles, F. And Maia, M. (2001). *Educação a Distância: o caso da Open University* [CD-ROM]. 4th Simpósio de Administração da Produção, Logística e Operações Internacionais, Guarujá (Brazil), August 12-14, 2001.

17 Molich, R. and Nielsen, J. (1990). ***Improving a human-computer dialogue.*** Communications of the ACM 33(3), 338-348.

18 Preece, J. (1994). ***Human-Computer Interaction.*** Harlow, England: Addison-Wesley.

19 Preece, J., Rogers, Y. and Sharp, H. (2002). ***Interaction design: beyond human-computer interaction.*** New Jersey: John Wiley & Sons, Inc.

20 ***TOEFL Testing on Computer… (2002)*** [online]. Available from http://www.toefl.org/educator/edcomptest.html [27 Apr 2003].

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