The Development and Application of Computer-Adaptive Testing in a Higher Education Environment

Mariana Lilley

A thesis submitted in partial fulfilment of the requirements of the University of Hertfordshire for the degree of Doctor of Philosophy

The programme of research was carried out in the
School of Computer Science,
Faculty of Engineering and Information Sciences,
University of Hertfordshire

2007

Abstract

The research reported in this thesis investigated issues relating to the use of computer-assisted assessment in Higher Education through the design, implementation and evaluation of a computer-adaptive test (CAT) for the assessment of and provision of feedback to Computer Science undergraduates. The CAT developed for this research unobtrusively monitors the performance of students during a test, and then employs this information to adapt the sequence and level of difficulty of the questions to individual students. The information about each student performance obtained through the CAT is subsequently employed for the automated generation of feedback that is tailored to each individual student.

In the first phase of the research, a total of twelve empirical studies were carried out in order to investigate issues related to the adaptive algorithm, stakeholders' attitude, and validity and reliability of the approach. The CAT approach was found to be valid and reliable, and also effective at tailoring the level of difficulty of the test to the ability of individual students. The two main groups of stakeholders, students and academic staff, both exhibited a positive attitude towards the CAT approach and the user interface.

The second phase of the research was concerned with the design, implementation and evaluation of an automated feedback prototype based on the CAT approach. Five empirical studies were conducted in order to assess stakeholders' attitude towards the automated feedback, and its effectiveness at providing feedback on performance. It was found that both groups of stakeholders exhibited a positive attitude towards the feedback approach. Furthermore, it was found that the approach was effective at identifying the strengths and weaknesses of individual students, and at supporting the adaptive selection of learning resources that meet their educational needs.

This work discusses the implications of the use of the CAT approach in Higher Education assessment. In addition, it demonstrates the ways in which the adaptive test generated by the CAT approach can be used to provide students with tailored feedback that is timely and useful.

Acknowledgments

The work in this thesis has been carried out over a period of five and half years and could not have been completed without the support and patience of the supervisory team and my family.

I am extremely grateful to my principal supervisor, Trevor Barker, for his continuous guidance, constructive comments and enthusiasm towards this work. His expertise in the fields of statistics and evaluation of educational software applications were crucial to the success of the research.

I am also very grateful to Carol Britton, who was my second supervisor for most of the research, for all her encouragement and sound advice. In spite of retiring from the University in April this year, Carol continued to support this work; many thanks are due to Carol for her patience and useful discussions during the writing up period.

Thanks are also due to Jill Hewitt for taking over as second supervisor after Carol's retirement, and for supporting the allocation of research hours to this work.

I am indebted to a number of people with whom I have discussed many of the ideas presented in this work for their comments and suggestions, in particular Andrew Pyper, Jonathan Meere, Liz Wray and Marta Maia.

I should also like to acknowledge and thank all academics and students who participated in the studies conducted as part of this research for their time and valuable feedback. Thanks are also due to colleagues from the wider educational community for being willing to share their views about the use of adaptive testing in Higher Education at the conferences where this work has been presented.

Special thanks to my parents, Ana and Luis, for their unfailing love and support over the years. Finally, a big thank you to my husband Matthew, for providing a happy and stable home environment that made it all possible.

Contents

List of Figures

List of Tables

List of Equations

Abbreviations and Acronyms used

1-PL	One-parameter logistic model
2-PL	Two-parameter logistic model
3-PL	Three-parameter logistic model
ADO	ActiveX Data Objects
ASP	Active Server Pages
CAA	Computer-assisted assessment
CAT	Computer-adaptive test
CAT-ASVAB	Computerised adaptive testing version of the Armed Services Vocational Aptitude Battery
CBT	Computer-based test
CML	Conditional maximum likelihood
CRUD	Create, read, update, delete
CTT	Classic Test Theory
GMAT	Graduate Management Admission Test
GRE	Graduate Record Examination
HCI	Human-computer interaction
ICC	Item characteristic curve
IRT	Item Response Theory
JML	Joint maximum likelihood
MML	Marginal maximum-likelihood
OLE-DB	Object Linking and Embedding-Database
TOEFL	Test of English as a Foreign Language
VB	Visual Basic

1. Introduction

The past two decades have seen an increased use of computer-assisted assessment (CAA) applications in Higher Education, to the extent that the use of computer technology in student assessment is quickly becoming a common feature across the sector. In spite of its increased use, it appears that the full potential of the use CAA technology in student assessment has not materialised:

> "The computer has not been significantly exploited as an enabler of new assessment methods, rather it has been used to implement traditional assessment. Systems that use the computer's interactive nature, such as in adaptive testing and other types of guided learning, peer review systems, and so on, are few." (Joy et al., 2002: p. 3)
>
> Although the capability for adaptive testing exists, it "has yet to be exploited within higher education as a viable approach to assessment and as a contributor to quality learning." (Challis, 2005: p. 519)

As the statements above suggest, in spite of being underused, there is a growing awareness of the value of interactive software applications that dynamically adapt to their users, such as adaptive testing (or computer-adaptive test).

In this thesis, the term computer-adaptive test (CAT) is used to refer to a CAA application that unobtrusively monitors the performance of students during a test, and then employs this information to adapt the sequence and level of difficulty of the questions (or tasks) to individual students.

The research reported in this thesis attempts to exploit the potential of CAA through the design, implementation and evaluation of the CAT approach in the assessment of and provision of feedback to Computer Science undergraduates.

1.1 Aims of the research

The main aim of this PhD thesis is to answer the following two research questions:

- What are the potential applications of the CAT approach in the assessment of Computer Science undergraduates?
- In which ways can the CAT approach be used to provide automated feedback to students that is timely and useful?

In addressing the research aims above, the following list of objectives was generated:

(a) to identify the main issues in designing and implementing a CAT software application to be used in the assessment of Computer Science undergraduates;

(b) to design and implement a CAT software application;

(c) to identify the key issues in evaluating a computer-assisted assessment (CAA) application;

(d) to evaluate the CAT software application;

(e) to identify the key components of the CAT approach that are useful in the provision of feedback to students;

(f) to design and implement an automated feedback software application based on the CAT approach;

(g) to evaluate the automated feedback software application.

Outcomes from the research addressing these objectives are reported in the Conclusion chapter, see section 10.2.

The contribution to knowledge of the work described in this thesis is therefore to demonstrate:

- how the CAT approach can be applied to the assessment of Computer Science undergraduates;
- the ways in which the individually tailored test generated by the CAT approach can be employed to identify the strengths and weaknesses of individual students, and to support the adaptive selection of learning resources that meet their educational needs.

1.2 Methodology

This section describes the methodology employed to address the aims of the research. Section 1.2.1 presents the software development approach used in order to design and implement the applications created as part of this work. In section 1.2.2, the approach to evaluation is described.

1.2.1 Approach to software development

The approach to software development employed in this research was iterative prototyping. The prototypes built are what Preece et al. (2002) call "high-fidelity", given that they are fully functional and interactive.

The iterative prototyping method is particularly suitable for projects of an explorative nature, such as this research, because:

- a full, definite set of requirements was not available from the outset (Boyle, 1997; Preece et al., 2002);

- high-fidelity prototypes are useful in the exploration and testing of ideas with stakeholders (Boyle, 1997; Preece et al., 2002).

In this work, the CAT high-fidelity prototype was built based on ideas drawn from the literature, in particular Lord (1980), Wainer (2000a), Wainer (2000b), Wainer & Mislevy (2000), Wolfe et al. (2001a) and Guzmán et al. (2005). The software development cycle employed in this work can be summarised as:

- build (or revise) high-fidelity prototype;
- evaluate high-fidelity prototype, using a combination of quantitative and qualitative methods (see section 1.2.2);
- use evaluation data to refine high-fidelity prototype.

As can be seen from the list above, the iterative nature of the approach means that each iteration of the high-fidelity prototype is evaluated with the stakeholders (in the case of this thesis, academic staff, and students or test-takers), and their feedback used to evolve and improve the software application. Two high-fidelity prototypes were built and refined using the cycle described above:

- a CAT software prototype to be used as a tool for the assessment of Computer Science undergraduates (i.e. test-takers);
- an automated feedback prototype to deliver individual feedback to test-takers.

The evaluation of both prototypes comprised a series of empirical studies. The methodology applied in these studies is presented next.

1.2.2 Approach to evaluation

The main aims of the evaluation phase were to identify the extent to which the high-fidelity prototypes were fit for the purpose for which they were designed. Prototypes were built or revised based on the findings from the evaluation.

A number of writers including Laurillard (1993), Boyle (1997), Barker & Barker (2002), and Bull & McKenna (2004) have warned that the evaluation of educational software is complex and, in order to be effective, it should:

- involve the participation of the main groups of stakeholders;
- take place in a real educational setting;
- consist of both quantitative and qualitative evaluation methods.

Redmond-Pyle & Moore (1995) identify three types of evaluation that were considered useful in this research:

- expert review;
- end-user testing;
- survey of end-user attitudes.

Various techniques can be employed for data gathering in each of these three phases, and the methods used in this work are outlined below.

1.2.2.1 Expert review

Members of academic staff were recruited as experts in this research. Experts were employed in the evaluation of the CAT approach in two different ways: heuristic evaluation, and expert advice.

It should be noted that, as academic staff, the experts involved in the evaluation are also stakeholders in the student assessment process; their participation in the evaluation was therefore crucial in order examine factors and issues that are important to academic staff that could otherwise have been overlooked by the research team.

Heuristic evaluation. Experts conducted a structured inspection of the CAT software prototype in the form of a heuristic evaluation (Molich & Nielsen, 1990; Redmond-Pyle & Moore, 1995; Preece et al., 2002). Experts were requested to examine the CAT software prototype, and then rate different aspects of the interface based on a checklist provided by the research team.

Expert advice. McAteer & Shaw (1994) suggest that it is useful to elicit views of academic staff in order to explore ideas relating to the development of educational software. Semi-structured discussions and questionnaires were employed in order to gather information about expert views on the research.

Semi-structured discussions can be seen as a special case of focus groups (Kontio et al., 2004). All semi-structured sessions conducted as part of this research adhered to the same format. First, a member of the research team provided a presentation of the main concepts of the CAT approach (or automated feedback approach). The presentation was then followed by a semi-structured discussion, where the participants were asked to share their views on the approach. The sessions were video recorded, with the permission of the participants.

Questionnaires were also were employed in this work in order to elicit views of academic staff. The questionnaires used in this research comprised closed questions, where experts were required to rate statements using a five point Likert scale. Boyle (1997) states that Likert scales are particularly useful in obtaining quantitative data on subjective reactions to a system. The five point Likert scale was chosen for this work for two reasons: (1) it contains a neutral midpoint, which poses the least constraint on the participants (Boyle, 1997) and (2) scales of more than five points can be unnecessarily difficult to use (Preece et al., 2002). In addition to closed questions, the questionnaires contained text boxes, so experts could add comments if they wished to do so.

1.2.2.2 End-user testing

Students (or test-takers) are the end-users of the CAT and the automated feedback high-fidelity prototypes developed for this research. The aim of end-user testing was to gather data about user satisfaction, as well as to identify any usability issues regarding the user interface that could hinder end-users' performance. End-user testing data was gathered in three different ways: observation, focus group, and questionnaire. In addition, data gathered during

the use of the CAT software prototype by test-takers was subjected to statistical analysis.

Observation. This method involved trained observers watching end-users performing representative tasks using the application being evaluated (Redmond-Pyle & Moore, 1995; Boyle, 1997; Dunn et al., 2003; Bull & McKenna, 2004). This method was employed in order to identify any potential usability problems with the user interface of the CAT software prototype that could hinder test-takers' performance.

Focus group. The focus group was used in conjunction with the observation method (Litosseliti, 2003), in order to obtain supplementary data relating to the overall end-user satisfaction with the CAT software prototype.

Questionnaire. End-users were asked to rate statements that are commonly used in human-computer interaction studies (Jettmar & Nass, 2002), such as "I found the application easy to use", in order to gather information about their perceived ease of use of the system. Participants were presented with the questionnaire following their use of the CAT or automated feedback software prototype. In this work, each questionnaire statement contained a five point Likert scale, and a text box for entering free text comments.

Statistical analysis. In order to ensure user satisfaction with the system, it was important to examine whether the CAT approach was both valid and reliable. To this end, data analysis consisted of descriptive statistics (i.e. mean and standard deviation), one-way ANOVA procedures, and the Pearson's Product Moment correlations to determine the magnitude and the significance of the relationship between test and retest scores (Brown, 1988). Furthermore, t-test procedures were carried out to assess whether the means of two groups of scores (for example, formative and summative scores) were statistically different from each other.

1.2.2.3 End-user attitude

Three different methods were employed in order to gather information about end-users' attitude: focus group, questionnaire, and interview.

Focus group. Preece et al. (2002) and Litosseliti (2003) recommend the use of focus groups for exploring complex and sensitive topics. The focus group method was useful in the initial stages of this research in order to obtain information on participants' views and attitudes on the CAT approach prior to its implementation in a real assessment setting. One of the limitations of the focus group method is the issue of representativeness; the number of participants in a focus group session is relatively low. Moreover, it is possible that even in a focus group facilitated by an experienced moderator the views of less articulate or confident participants are not expressed (Litosseliti, 2003). For this reason, the focus group method in this research was used in conjunction with questionnaire and interview methods in order to obtain a more comprehensive picture of test-takers' attitude towards the CAT approach.

Questionnaire. McAteer & Shaw (1994), Boyle (1997) and Bull & McKenna (2004) suggest that questionnaires are useful in an educational context in order to elicit reactions from stakeholders to a software application. Redmond-Pyle & Moore (1995) add that questionnaires can be particularly useful when gathering information from large groups of users. In this work, end-users were presented with questionnaires following their interaction with the CAT or automated feedback software prototype. The participants were requested to rate questionnaire statements relating to the perceived level of difficulty of a CAT, and usefulness of the automated feedback using a five point Likert scale.

Responses to these statements were treated as ordinal data, and analysed using the Kruskal-Wallis test, which is a suitable statistical method for analysing non-parametric data (Brown, 1988). The correlation between questionnaire responses and test-takers' performance on the CAT tests was analysed using Spearman rank-order correlation coefficients. This is because this coefficient is a non-parametric measure of correlation, commonly applied to ordinal data (Brown, 1988).

Interview. Some participants who answered the questionnaire were selected for interviews. Boyle (1997: p. 202) suggests that "interviews are a useful way to gain a rich understanding of users' reactions to a system". In this work, end-users (i.e. test-takers) were asked to rate the level of difficulty of a test dynamically generated by the CAT software prototype using a questionnaire, and interviews were then employed in order to gain an insight on the reasons for their ratings.

Each type of evaluation method described in section 1.2.2 provided different types of information to the research team, which taken together influenced the direction of the work. The impact of the evaluation findings on this research is evident in subsequent chapters of this thesis.

1.3 Structure of the thesis

Section 1.3.1 provides an overview of the thesis. In section 1.3.2, an overview of the empirical studies conducted as part of this research is presented.

1.3.1 Overview of chapters

The thesis contains ten chapters, including the introductory one.

Chapter 2 provides the background to the research, including the main issues surrounding the use of computers in student assessment in a Higher Education setting, and computer-adaptive tests (CATs) in particular. Different approaches to the development of adaptive testing are examined with emphasis on Item Response Theory (IRT) (Lord, 1980). Three-Parameter Logistic (3-PL) model (Lord, 1980) from IRT was chosen as the basis for the CAT adaptive algorithm, and the reasons for this choice as well as an overview of the 3-PL model are also provided in this chapter. In addition, Chapter 2 contains an overview of the key components of a CAT based on IRT, and potential advantages of and barriers to the implementation of the CAT approach.

Using as a starting point the key components of the CAT approach identified in Chapter 2, Chapter 3 examines the main issues surrounding the design and implementation of the CAT software prototype developed for this research. The chapter provides an overview of the approach used for the calibration of the item (i.e. question) database, and the CAT testing algorithm. Whilst some assumptions relating to how to select the items to be administered first and next were made based on ideas drawn from the literature, other issues were so central to the research that needed to be investigated directly. Thus, Chapter 3 describes three empirical studies concerned with database calibration, the effect of different stopping conditions, and the effect of question review in a CAT.

Once the CAT software prototype had been designed and implemented, the next stage of the research was concerned with the evaluation of the CAT approach. The evaluation involved the two main groups of users, test-takers and academic staff, and Chapter 4 focuses on the evaluation of the approach by the former. In Chapter 4, empirical studies are used to investigate test-taker attitude towards the CAT approach, test-taker perceived level of difficulty of a CAT, as well as to identify any usability issues relating to the CAT software prototype that could affect test-takers' performance in an adverse way. Chapter 4 also includes a section regarding the changes made to the CAT software prototype, in the light of the information gathered from the studies reported earlier in the chapter.

Chapter 5 is concerned with the evaluation of the CAT approach by academic staff, and reports on the findings from two empirical studies. In the first study, the CAT software prototype was subjected to a heuristic evaluation (Molich & Nielsen, 1990) in which the user interface was inspected by a group of experts. The second study was undertaken to investigate academic staff views on the pedagogical usefulness of the CAT approach.

Following the evaluation of the CAT software prototype by test-takers and academic staff, the research focused on examining whether the CAT approach is both valid and reliable. Issues of face, content and construct validities are

discussed in Chapter 6. The reliability of the CAT approach, including test-retest reliability, is also discussed in this chapter.

The need for enhanced feedback to CAT test-takers was an important outcome of the pedagogical evaluation described in Chapter 5. The issue of feedback on performance was investigated as part of this research, and a software prototype was designed and developed to provide CAT test-takers with individual feedback on performance. The automated feedback prototype, and the ideas that underpinned its design, are described in Chapter 7.

Similarly to the evaluation of the CAT software prototype, the automated feedback prototype was evaluated by the two main groups of stakeholders, namely test-takers and academic staff. Chapter 8 focuses on the evaluation of the feedback prototype by test-takers. Three empirical studies were conducted to ascertain whether the automated feedback provided by the prototype was useful, and the revision tasks recommended were within each individual test-taker's grasp.

In order to get a complete picture of stakeholders' reactions to the automated feedback prototype, three empirical studies involving academic staff were conducted. These studies examined academic staff attitude towards the feedback approach, and are reported in Chapter 9.

Chapter 10 presents a summary of the research, with emphasis on the conclusions drawn from the empirical studies conducted as part of this work. Chapter 10 also discusses the significance of the thesis in the context of an increased use of computer technology in student assessment and learning. Suggestions for future work are also included.

Chapter 10 is followed by the list of references and appendices. Appendix A contains a glossary of the terms frequently used in the thesis. The following appendices contain guidelines and research instruments used as part of empirical studies involving test-takers: Appendix B, Appendix C, Appendix D, Appendix E, Appendix F, Appendix G, and Appendix H. Guidelines and research instruments employed in empirical studies involving academic staff are included in: Appendix I, Appendix J, Appendix K, and Appendix L.

Appendix M provides a list of papers published as part of the research reported here.

As can be seen from the overview above, a series of empirical studies were conducted as part of this work. These are presented next.

1.3.2 Overview of empirical studies

The empirical studies conducted as part of this work can be divided into two main groups. The first group, which is presented in section 1.3.2.1, is concerned with studies relating to the CAT software prototype. The second group is related to studies concerning the automated feedback prototype, and a summary of these studies is provided in section 0.

It should be noted that the empirical studies reported in this thesis were conducted with the approval of the Ethics Committee from the Faculty of Engineering & Information Sciences. It was essential to the research to ensure that test-takers participating in studies were not disadvantaged, especially in those cases where the studies took place in a summative assessment context. To this end, test-takers always took an adaptive test using the CAT software prototype plus a traditional CBT test. In all cases, the highest score obtained by each test-taker (i.e. either CAT or CBT score) was employed to compute their final grade. Following each study, debriefing sessions that included a comprehensive description of the nature of the research were carried out.

1.3.2.1 CAT prototype

This section aims to provide an overview of the empirical studies regarding the CAT approach conducted as part of this research. For ease of reading, this information is summarised in the form of a table, see Table 1-1 (p. 29). The research reported here was carried out over a period of five and a half years, and Table 1-2 (p. 30) below shows how the sequence in which the studies were conducted relates to the thesis.

All the studies listed in this section are discussed in greater detail in subsequent chapters. Further details about the methodology employed in these studies can be found in section 1.2.

As part of this research a CAT prototype was designed, implemented and evaluated. In Table 1-1, studies (1), (2), (3) and (4), are concerned with practical design and implementation issues.

Study (1) is concerned with the calibration of items (i.e. questions); one of the goals of this process is to determine the difficulty of each question. There are various approaches to the calibration of items, and the approach used in this research was a combination of expert calibration and calibration based on actual responses from test-takers. The expert calibration was based on Bloom's taxonomy of cognitive skills (Bloom, 1956), which is a commonly used method for classifying objective questions (Ward, 1981; Bull & McKenna, 2004). Study (1) is concerned with assessing the usefulness of experts and Bloom's taxonomy of cognitive skills in the calibration of questions.

A number of factors need to be taken into account in determining the stopping condition for a CAT. Study (2) focuses on the use of the standard error for the ability estimated as a stopping condition, and whether or not this would be valid in the context of the CAT prototype developed for this research. Such a stopping condition would lead to increased testing efficiency, one of the major benefits of the CAT approach reported in the literature (Jacobson, 1993; Carlson, 1994; Ward, 1988; Wainer, 2000a; Wainer, 2000b). The implementation of standard error as a stopping condition, however, can result in test-takers having different test lengths. Study (3) examines test-taker attitude towards standard error as a stopping condition, as well as other alternatives such as test length.

A further practical design and implementation issue examined as part of this research was concerned with whether or not to include functionality that would allow test-takers to return to previous items. There are mixed views on this issue (see for example Vispoel et al., 2000; Olea et al., 2000; Revuelta et al., 2000; Thissen & Mislevy, 2000; Wainer, 2000b; Vicino & Moreno, 2001;

Guzmán & Conejo, 2004), with the majority of the work in the CAT area tending towards CATs where item review is not permitted. Those who argue in favour of item review, cite a reduction in student anxiety and greater resemblance with other assessment methods such as paper-and-pencil tests as motivating factors for implementing CATs where item review is allowed. In order to identify whether or not item review functionality should be added to the CAT prototype developed for this research, study (4) examines the effect of item review on proficiency level estimates.

Several authors including Lord, 1980; Hambleton & Cook, 1983; Hambleton & Swaminathan, 1990; Veerkamp & Berger, 1999; Guzmán & Conejo, 2005, have highlighted the benefits of the use of IRT psychometric models and their application in CAT. However, there are few examples in the literature of test-taker attitude towards key aspects of the CAT approach, such as the fact that test-takers can be presented with different sets of questions during the same assessment session and the scoring method employed within CATs. This is somewhat surprising, given the importance of stakeholder acceptance of approach (see for example Jacobson, 1993). Test-taker attitude towards the CAT approach was examined in study (6).

It was also important to investigate whether test-takers found the CAT software prototype developed for this research easy to use, as it was essential to ensure that the application would not have an adverse effect on test-taker performance. This aspect was the focus of studies (5) and (6).

A number of authors including Carlson (1994), Ward (1988) and Wainer (2000a) suggest that one of the benefits of the CAT approach is the possibility to match the difficulty of the questions to a test-taker's ability. Studies (7) and (8) investigate what was the test-takers' perceived the level of difficulty of the test when using the CAT software prototype developed for this research. It should be noted that this issue is also briefly explored in study (5).

An important aspect of this research was to examine the attitude of all major stakeholders towards the CAT approach. Whilst studies (5), (6), (7) and (8) are concerned with the views of test-takers as stakeholders in the assessment

process, studies (9) and (10) are concerned with a second group of stakeholders, namely academic staff. A heuristic evaluation was carried out as part of study (9) in order to investigate whether or not the CAT software prototype would disadvantage students. Study (10) focuses on the pedagogical evaluation of the CAT approach by a group of experts.

One of the main concerns of stakeholders in the assessment process would be as to whether or not the approach is valid and reliable. In order to investigate the validity and reliability of the CAT approach as implemented in this research, the empirical studies (11) and (12) are concerned with these issues.

Study number	Study title	Method	Section	Brief description	Year
(1)	Database calibration	Statistical analysis	3.2.2	The study aimed to investigate the effect of employing experts to calibrate a database.	2006
(2)	Stopping condition	Statistical analysis	3.5.2	The aim of this study was to identify whether or not the standard error for the proficiency level estimate would be a valid stopping condition.	2004
(3)	Stopping condition	Focus group	3.5.3	The aim of this study was to investigate test-taker attitude towards different stopping conditions.	2002
(4)	Reviewing previously entered responses	Statistical analysis	3.6	In a CAT, test-takers are not normally permitted to return to previous questions. The purpose of this study was to examine the effect of item review on proficiency level estimates.	2005
(5)	First user study	Questionnaire; Onsite observation	4.1	The aim of this study was twofold. First, to uncover any usability issues regarding the CAT prototype. Second, to examine test-taker perceived level of difficulty of an adaptive test using the CAT prototype.	2002
(6)	Test-taker attitude	Focus group session	4.2	The purpose of the focus group session was to examine test-taker attitude towards the CAT approach.	2002
(7)	Perceived level of difficulty: summative assessment	Statistical analysis; Interview	4.3.1	The study described in this section is concerned with the perceived level of difficulty of the CAT in a summative assessment context.	2005
(8)	Perceived level of difficulty: formative assessment	Statistical analysis	4.3.2	The study described in this section is concerned with the perceived level of difficulty of the CAT in a formative assessment setting.	2006
(9)	Usability evaluation	Heuristic evaluation	5.1	The CAT prototype was examined by a group of experts, in order to uncover any usability issues that might affect performance.	2002
(10)	Pedagogical evaluation	Questionnaire	5.2	This study is concerned with the pedagogical evaluation of the CAT approach by a group of academic staff.	2002
(11)	Construct validity	Statistical analysis	6.1.3	The aim of this study was to examine if the CAT approach has construct validity.	2006
(12)	Test-retest reliability study	Statistical analysis	6.2.2	The aim of this study was to investigate issues concerned with the reliability of the CAT approach.	2003

Table 1-1: Summary of empirical studies relating to the CAT approach reported in the thesis

Year	Study number	Study title	Method	Section	Brief description
2002	(3)	Stopping condition	Focus group	3.5.3	The aim of this study was to investigate test-taker attitude towards different stopping conditions.
2002	(5)	First user study	Questionnaire; Onsite observation	4.1	The aim of this study was twofold. First, to uncover any usability issues regarding the CAT prototype. Second, to examine test-taker perceived level of difficulty of an adaptive test using the CAT prototype.
2002	(6)	Test-taker attitude	Focus group session	4.2	The purpose of the focus group session was to examine test-taker attitude towards the CAT approach.
2002	(9)	Usability evaluation	Heuristic evaluation	5.1	The CAT prototype was examined by a group of experts, in order to uncover any usability issues that might affect performance.
2002	(10)	Pedagogical evaluation	Questionnaire	5.2	This study is concerned with the pedagogical evaluation of the CAT approach by a group of academic staff.
2003	(12)	Test-retest reliability study	Statistical analysis	6.2.2	The aim of this study was to investigate issues concerned with the reliability of the CAT approach.
2004	(2)	Stopping condition	Statistical analysis	3.5.2	The aim of this study was to identify whether or not the standard error for the proficiency level estimate would be a valid stopping condition.
2005	(4)	Reviewing previously entered responses	Statistical analysis	3.6	In a CAT, test-takers are not normally permitted to return to previous questions. The purpose of this study was to examine the effect of item review on proficiency level estimates.
2005	(7)	Perceived level of difficulty: summative assessment	Statistical analysis; Interview	4.3.1	The study described in this section is concerned with the perceived level of difficulty of the CAT in a summative assessment context.
2006	(1)	Database calibration	Statistical analysis	3.2.2	The study aimed to investigate the effect of employing experts to calibrate a database.
2006	(8)	Perceived level of difficulty: formative assessment	Statistical analysis	4.3.2	The study described in this section is concerned with the perceived level of difficulty of the CAT in a formative assessment setting.
2006	(11)	Construct validity	Statistical analysis	6.1.3	The aim of this study was to examine if the CAT approach has construct validity.

Table 1-2: Summary of empirical studies relating to the CAT approach reported in the thesis, in chronological order

In addition to the CAT software prototype, an automated feedback prototype was designed, implemented and evaluated as part of this research. The evaluation of the automated feedback prototype involved a series of empirical studies, and these are summarised in the following section.

1.3.2.2 Automated feedback prototype

This section aims to provide an overview of the empirical studies regarding the automated feedback prototype carried out as part of this research. For ease of reading, Table 1-3 (p. 32) summarises the five empirical studies conducted as part of this research that are related to the automated feedback prototype, and further details about the methodology employed in these studies can be found in section 1.2.

Study number	Study title	Method	Section	Brief description	Year
(1)	Test-taker attitude: summative assessment	Statistical analysis	8.1	This study is concerned with test-taker attitude towards the automated feedback approach, in a summative assessment setting.	2005
(2)	Test-taker attitude: according to performance	Statistical analysis	8.1.2	This study aims to investigate if test-taker performance on the test had an effect on the perceived usefulness of the feedback.	2005
(3)	Test-taker attitude: formative assessment	Statistical analysis; Questionnaire	8.1.3	This study is concerned with test-taker attitude towards the automated feedback approach, in a formative assessment setting. This study also investigates the perceived ease of use of the automated feedback application.	2006
(4)	Academic staff attitude	Semi-structured discussion	9.2	This study aims to examine academic staff attitude towards the automated feedback prototype.	2006
(5)	Academic staff perceived usefulness of the feedback	Questionnaire	9.3	This study aims to examine academic staff perceived usefulness of the automated feedback prototype in summative and formative assessment settings. It also aims to investigate the academic staff perceived speed, quality and appropriateness of the feedback.	2006

Table 1-3: Summary of empirical studies relating to the automated feedback prototype reported in the thesis

As can be seen from Table 1-3, studies (1), (2) and (3) were concerned with test-taker attitude towards the automated feedback approach. In order to get a more comprehensive picture of stakeholders' attitude towards the approach, studies (4) and (5) examined the reactions of academic staff to the automated feedback prototype. All studies listed in Table 1-3 are discussed in greater detail in Chapters 8 and 9.

The following chapter introduces the theoretical background to the research.

2. Background to the research

This chapter aims to provide an introduction to the main issues associated with the use of computers in student assessment, and the use of computer-adaptive tests (CATs) in particular. To this end, the chapter is organised into two main sections. The first section provides an overview of the use of computer applications in student assessment. The second section provides a theoretical and practical context for the CAT approach.

The following section provides an introduction to the computer-assisted assessment (CAA) field, including an outline of early research in the area and examples of typical CAA applications.

2.1 Introduction to Computer-Assisted Assessment

The field of computer-assisted assessment (CAA) is concerned with the use of digital technologies in student assessment. This term is often used interchangeably with computer-aided assessment and computer-based assessment. The term e-Assessment has been increasingly used to refer to CAA software applications that are delivered over the Internet or an intranet. As an introduction to CAA, this section first looks at early research in the area.

Early research. Much research has been conducted on the use of CAA applications by Higher Education institutions. Some of the early research in

the field was concerned with the effect of the use of computers in student assessment and its equivalence with their paper-and-pencil counterparts (see for example Brosnan, 1999). Results from these studies were mixed, with some studies showing that computer anxiety had an effect on student performance (see for example Lee et al., 1986) and others showing no statistically significant difference in scores between computer-based and paper-and-pencil formats (see for example Chin et al., 1991; Vogel, 1994; Baydoun & Neuman, 1998). Recent literature in the CAA field supports the view that the use of computers in student assessment does not have an adverse material effect on student performance. Bull & McKenna (2004: p. 65), for instance, indicate that "the increasing use of computers in all sectors of education and society would suggest that computer familiarity and anxiety are diminishing factors". Moreover, Cann & Pawley (1999) and Bull & McKenna (2004) report on positive student reactions to the adoption of CAA as part of their assessment.

In addition to issues surrounding student anxiety and reactions to the adoption of CAA, early research focused on the identification of motivating factors for the adoption of CAA by Higher Education institutions. These motivating factors have remained stable in recent years and include the perceived need to:

- make use of the available computing infrastructure (Bull & McKenna, 2004);
- store and re-use assessment (Harvey & Mogey, 1999; Dunn et al., 2003; Bull & McKenna, 2004);
- produce reports on students' performance and progress in a fast and automated way (Brown, 1997; Brown et al., 1998; Miller et al., 1998; Conole & Bull, 2002; Dunn et al., 2003; Bull & McKenna, 2004);
- achieve speed and consistency of marking (Harvey & Mogey, 1999; Conole & Bull, 2002; Bull & McKenna, 2004);

- reduce the marking workload for academic staff (Brown, 1997; Conole & Bull, 2002; Dunn et al., 2003; Bull & McKenna, 2004) ;
- provide students with timely feedback (Harvey & Mogey, 1999; Conole & Bull, 2002; Dunn et al., 2003; Bull & McKenna, 2004);
- increase the frequency of student assessment (Conole & Bull, 2002; Bull & McKenna, 2004);
- broaden the range of skills being assessed (Conole & Bull, 2002; Bull & McKenna, 2004);
- broaden the range of assessment methods being used (Conole & Bull, 2002; Bull & McKenna, 2004).

As can be inferred from the two last items in the list above, CAA software applications can be used to deliver a wide range of assessments, and some examples of such applications are provided next.

Examples of CAA applications. There is a wealth of examples of CAA applications, such as the work of Callear et al. (2001), where a CAA application for marking short free-text responses is described. Foxley et al. (2001) describe a CAA software application capable of assisting in marking computer programs, diagrams and essays. Other examples of the use of CAA include electronic assessment of computer programming skills (Brown et al, 1998; Bull & McKenna, 2004), self-assessment of undergraduate projects (Bull & McKenna, 2004), peer review of essays (Robinson, 1999) and simulations (Bull & McKenna, 2004).

Although CAA software applications have been shown to be appropriate for a variety of assessment methods, Bull (1999), Joy et al. (2002), Warburton & Conole (2003) and Bull & McKenna (2004) suggest that most of the uptake of CAA focuses on the use of objective testing. As its name implies, an objective test is based on the use of objective questions. Objective questions are characterised by:

- a predefined set of possible answers;

- the possibility of marking without any subjective judgement on the part of the marker.

Objective questions are “admirably suitable for machine marking” (Miller et al., 1998: p. 153), and one can speculate that this characteristic has stimulated the uptake identified by Bull (1999), Warburton & Conole (2003) and Bull & McKenna (2004). Although marking objective questions using a CAA application is straightforward, the design and construction of good objective questions by academic staff is often a laborious and time-consuming task (Miller et al., 1998; Pritchett, 1999; Dunn et al., 2003).

Bull & McKenna (2004) identify four ways in which objective tests can be used in the assessment of students: summative, formative, diagnostic and self-assessment. It should be noted that in this work, diagnostic testing was considered a special case of summative testing (as it can be used to make a pass/fail decision), and self-assessment was considered a special case of formative assessment (as it has no effect on a student’s final grade). There is also the issue of what types of skills can be assessed using objective tests, and this is discussed next.

Objective tests: skills assessed. Much of the literature uses Bloom’s six levels of cognitive skills (Bloom, 1956; Anderson & Krathwohl, 2001) – namely knowledge, comprehension, application, analysis, synthesis and evaluation – as a tool to classify objective questions according to the skills being assessed. McBeath (1992), for instance, suggests that all levels of cognitive skills can be assessed using objective questions. Ward (1981), Pritchett (1999), Davies (2001), Biggs (2002) and Joy et al. (2002) however, have reported on the unsuitability of objective testing to assess higher level cognitive skills, such as synthesis and evaluation. Indeed Bull & McKenna (2004) indicate that a common assumption amongst academic staff and practitioners is that only the three or four lowest levels of cognitive skills defined by Bloom can be assessed using objective tests. In the work reported in this thesis, it was assumed that objective questions can be effectively employed to assess the three first levels of cognitive skills: knowledge, comprehension and application.

The implications of this assumption to this research are discussed later in sections 3.2 and 7.3.

In spite of the potential limitation with regards to skills being assessed, objective testing is considered convenient and useful (Brown et al., 1998; Biggs, 2002) as well as valuable (Cann & Pawley, 1999), when it supplements other forms of assessment. In addition to skills being assessed, an issue to be considered when using objective tests is the approach employed to select the questions to be administered during the test. This issue is discussed next.

Objective tests: selection of questions. Much of the objective testing carried out in the United Kingdom is based on Classic Test Theory (CTT) principles (Bull & McKenna, 2004).

Weiss & Yoes (1991: p. 69) indicate that CTT is also known as "true score" or "number-correct score" theory. Lord (1980) and Weiss & Yoes (1991) indicate that CTT is based on the assumption that each test-taker has a true score. The true score is an unobservable quantity, which represents the hypothetical perfect measurement of a test-taker's ability. In order to estimate a test-taker's ability, CTT employs the following two concepts:

- observed score, which is the number-correct score as measured in a test;
- error, which represents the amount of error of the observed score as a measure of the true score.

In summary, CTT is based on the assumption that:

$$\text{Observed Score} = \text{True Score} + \text{Error}$$

Equation 2-1: Classical Test Theory Model (Weiss & Yoes, 1991)

In the work reported here, the term computer-based test (CBT) is used to refer to CAA software applications that are based on CTT principles. In a typical CBT, the same set of questions is administered to all test-takers (Romero et al., 2006).

In a CBT, questions are normally selected by academic staff prior to the test, in such a way that a broad range of ability levels, from low to advanced, is catered for (Ward, 1981; Pritchett, 1999). Reports in the literature suggest that this technique is the most commonly employed (see for example Brown, 2003; Dunn et al., 2003; Race et al., 2004). There are, however, other less frequently used techniques such as:

- randomly selecting questions from a pool (see for example Thelwall, 2000);
- automatically generating questions during the test (see for example Williams et al., 1999).

Although the two techniques listed above could result in students being administered different sets of questions, it is expected that such different tests would be of similar difficulty and duration. This often means that the final CBT score is determined by the number of questions answered correctly out of the total number of questions. In some cases, negative marking (Ward, 1981; Bull & McKenna, 2004), guess correction (Ward, 1981; Bull & McKenna, 2004) and confidence rating (Davies, 2001) techniques are applied in order to minimise the potential occurrence of inflated scores due to guessing. It should be noted that there is some debate amongst educationalists as to whether these techniques should be applied at all (Ward, 1981; Bull & McKenna, 2004).

Bull & McKenna (2004) recognise the value as well as the extensive use of the CTT approach (and, consequently, of the CBT approach) in student assessment. However, Bull & McKenna (2004), Lord (1980) and Hambleton & Swaminathan (1990) highlight that a limitation of this approach is that the ability of a given test-taker is determined by the difficulty of the test. Bull & McKenna (2004: p. 77), for instance, point out that "if a test is difficult, students appear to have a lower ability than when a test is easy".

There is also the issue of assessing groups of test-takers with mixed abilities. Given that in a conventional CBT all students are presented with the same set of questions, it is possible that high-performing students are presented with one or more questions that are below their level of ability. Similarly, low-

performing students can be presented with questions that are above their level of ability. Inappropriate levels of question difficulty might lead those less proficient students to experience frustration when overly difficult questions are presented. In a similar way, more proficient students might feel bored if the questions administered during a given session of assessment were unchallenging. In both cases, there is a risk of student de-motivation.

One potential solution to address the problem of de-motivation is adaptive testing. The underlying idea of an adaptive test is to present each test-taker with a set of questions that is appropriate to their level of ability. In the next section, different approaches to adaptive testing are introduced.

2.2 Introduction to Computer-Adaptive Tests

Computer-adaptive tests differ from the conventional CBTs primarily in the approach used to select the set of questions to be administered during a given assessment session. A computer-adaptive test (CAT) is, as its name implies, a CAA software application where the content and/or sequencing of the test items is adapted to each individual test-taker. The following section provides an overview of different approaches that support the implementation of adaptive testing.

2.2.1 Approaches to Adaptive Testing

The term computer-adaptive test (CAT) is commonly used to describe a CAA software application where Item Response Theory (IRT) is employed to estimate a test-taker's ability and, based upon this estimate, select the item (i.e. question) to be administered next. However, not all approaches to CAT use IRT.

Trentin (1997) proposes a system based on a hierarchical representation of the content domain, where the test starts with questions of high difficulty. Trentin's (1997) system aims to spare high achieving students from being administered low level questions. Rudner (2001) proposes a CAT based on

Measurement Decision Theory (MDT), which is a measurement model for classifying test-takers based on statistical decision theory. Rudner's (2001) CAT is mostly concerned with classifying test-takers into one of a finite number of discrete categories, such as pass/fail. Lütticke (2004) describes an adaptive test where students are presented with questions from different domains of Computer Science. Student responses to the questions are automatically analysed by the system. An incorrect response will cause a tutoring component to provide some feedback and then the question is re-administered. This process is repeated until the student provides a correct response. Steven & Hesketh (1999) and Tzanavari et al. (2004) depict an adaptive test where a set of If-Then rules created by the tutor is used to select the question to be administered next. Kaburlasos et al. (2004) and Cristea & Tuduce (2005) describe an adaptive test where the adaptive algorithm is based on a tree structure.

It can be seen from the work outlined above that adaptive testing is not dependent on IRT. However, IRT has been shown to be useful in the efficient implementation of adaptive testing (Lord, 1980; Weiss, 1983; Hambleton & Swaminathan, 1990; Wainer & Mislevy, 2000). This is because IRT can be used to select questions that provide the most information about each test-taker, regardless of their ability.

The research introduced here focuses on the practical application of adaptive testing rather than on a comparison of different underpinning theories. IRT was chosen over the other theories presented in this section, as it has the largest body of research supporting its use and therefore it was considered the most appropriate choice. The following section provides a brief introduction to IRT.

2.2.2 Introduction to IRT

Item Response Theory (IRT) is a family of mathematical functions that attempts to predict the probability of an individual answering an item (i.e.

question) correctly (Lord, 1980; Hambleton & Swaminathan, 1990; Wainer & Mislevy, 2000).

Weiss (1983), Hambleton & Swaminathan (1990) and Baker & Kim (2004) ascribe the origins of IRT to the 1911 Binet-Simon Intelligence Scale. In this work, Binot and Simon described, in the form of tables, the relation between the proportion of correct responses to an item and children's chronological age. In 1916, Terman used this same type of tabular information to plot curves relating the probability of a correct response to an item and age. This curve is now known as an item characteristic curve (ICC). Examples of ICCs will be provided later in this chapter.

Weiss (1983), Hambleton & Swaminathan (1990), Van der Linden & Hambleton (2000) and Baker & Kim (2004) provide a full historical account on the development of IRT that dates back to the development of parameter estimation procedures by Richardson in 1936 and Lawley in 1943. This historical perspective also highlights the development of item response models by Lord in 1952, Birnbaum in 1957 and Rasch in 1960.

Prior to the 1960s, IRT parameter estimation was a very laborious job. Weiss (1983), Hambleton & Swaminathan (1990) and Van der Linden & Hambleton (2000) suggest that much of the IRT research in the 1960s, 1970s and 1980s was stimulated by the availability of computer resources. Important milestones that exemplify this are the release of computer programs for parameter estimation such as BICAL in 1969 and LOGIST in 1974.

Despite the pioneering work of Lord (1971a, 1971b) amongst others, McBride (2001a) indicates that IRT research prior to 1977 focused on theoretical analyses and computer simulation studies. Thus, it lacked of compelling empirical evidence involving real test conditions and test-takers.

The publication of "Applications of Item Response Theory to Practical Testing" (Lord, 1980) coincides with a greater interest in practical applications of IRT and, in particular, with IRT at the core of computer-adaptive tests (CATs). As evidence of this trend, one could refer to the report by McBride & Martin (1983)

on an evaluation study involving real test-takers rather than computer simulations.

At the time of writing, Lord's (1980) “Applications of Item Response Theory” is over 25 years old and there has been a significant amount of research in the IRT field since it was first published. Nonetheless, the item response functions defined in Lord’s text have remained stable and are still in use.

The Armed Services Vocational Aptitude Battery (ASVAB) (McBride, 2001b), (Graduate Management Admission Test (GMAT) (Guo et al., 2006), Test of English as a Foreign Language (TOEFL) (Glas et al., 2003), Graduate Records Examination (GRE) (Wainer & Eignor, 2000) and Microsoft Certified Professional (Microsoft Corporation, 2006) are examples of the application of IRT in large, high stake admission and certification tests.

Recent practical applications of IRT include its use for summative and formative assessments in various educational contexts, such as computing (Yong & Higgins, 2004; Pérez & Alfonseca, 2004; Guzmán & Conejo, 2004; Alfonseca et al., 2005; Guzmán et al., 2005), languages (Chalhoub-Deville et al., 2000; Gonçalves et al., 2004; Ho & Yen, 2005) and mathematics (Fernandez, 2003; He & Tymms, 2004; He & Tymms, 2005).

In addition to work concerned with practical applications of IRT for admission tests, certification tests and educational purposes, a significant amount of research has been dedicated to item calibration (Guzmán & Conejo, 2005); item selection procedures (Veerkamp & Berger, 1999; Vos, 2000); item exposure control (Hetter & Sympson, 1997; Revuelta & Posanda, 1998) and issues related to response time in ability estimate (Thissen, 1983; Hornke, 2000; Wheadon & He, 2006).

The historical account provided here does not consider parallel developments in IRT such as the work published by Samejima (1969). The reason for this is that Samejima’s work focused on polychotomous items (i.e. item where options are ordered along a continuum, as in Likert scales), and this research centers on the use of IRT models for dichotomously scored items.

2.2.3 IRT Models for Dichotomously Scored Items

This work focuses on the use of IRT for scoring dichotomous items or, in other words, items where the test-takers' responses can be considered to be either being 'correct' or 'incorrect'. The research is concerned with the use of objective questions and focuses on IRT models for dichotomously scored items.

As defined by Weiss (1983: p. 9), "IRT models specify the probabilistic relationship between the observed responses of an individual to a test item and the individual's level on the latent trait". A variety of IRT models have been developed for dichotomously scored items (Lord, 1980; Hambleton & Swaminathan, 1990; Wainer & Mislevy, 2000). The simplest IRT model for dichotomously scored items is the One-Parameter Logistic Model (1-PL), often referred as to the Rasch Model in the honour of the Danish mathematician George Rasch (1901-1980).

Within The 1-PL model it is assumed that items vary only in their difficulty. In the Two-Parameter Logistic Model (2-PL), items vary in both difficulty and discrimination. In the Three-Parameter Logistic Model (3-PL) items vary in difficulty, discrimination and guessing (also known as pseudo-chance parameter).

An important aspect of this work was to identify an appropriate IRT model. An extensive discussion on the merits of the different psychometric models is beyond the scope of this work and the interested reader is referred to Hambleton & Murray (1983), Lord (1983), Divgi (1986) and Heaning (1989).

To summarise, the Rasch model is a special case of the 3-PL model, where discrimination is equal to 1 and pseudo-chance is equal to 0. Lord (1980) and Heaning (1989) support the selection of the Rasch model when the number of test-takers available is less than 100 or 200, regardless of the Rasch model's limitations with respect to guessing. In contrast, research reported by Divgi (1986) supports the view that "the Rasch model should not be used with multiple-choice tests" (p. 296). In a similar vein, Hambleton & Murray (1983) established that the 2-PL and 3-PL models are more suitable than the Rasch

model in many situations, as it cannot be always assumed that all items have the same discrimination and pseudo-chance. Ward (1988: p. 272) highlights that the 3-PL model "generally provides a more accurate representation of the characteristics of real test questions". Wainer & Mislevy (2000: p. 68) add that "3-PL is the IRT model that is most commonly applied in large scale testing applications". However, Ward (1988: p. 272) also warns that the application of the 3-PL model "is more demanding computationally".

An assumption of the work reported in this thesis was that it was important to take into account the effect of guessing and item discrimination when estimating a test-taker's ability and, for this reason, the 3-PL model was chosen.

The following section provides a brief introduction to Item Response Theory concepts that are helpful for an understanding of subsequent chapters.

2.2.4 Three-Parameter Logistic Model Overview

The CAT software prototype described here was based on the Three-Parameter Logistic Model (3-PL) within IRT. In this model, in order to evaluate the probability P of a test-taker with an unknown ability θ answering an item correctly, the mathematical function shown in Equation 2-2 (Lord, 1980: p. 12) is used.

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

Equation 2-2: Three-Parameter Logistic Model (Lord, 1980)

The scaling of ability θ is arbitrary, and ability scores are typically placed on a scale with mean of zero and a standard deviation of one (Lord, 1980). Two different scales were used in the work reported here. In earlier stages of the research, the scale used varied from -2 to +2. In later stages of the research, the scale varied from -3 to +3 in order to make it possible to use the CAT software prototype developed for this research in combination with the

commercial software application XCalibre (Assessment Systems Corporation, 2007; Gierl & Ackerman, 1996). The use of XCalibre in the research is later described in section 3.2.

In Equation 2-2, e represents the natural logarithmic base (i.e. 2.71828…). The parameter b represents the item's difficulty, and within the prototype described here $-3 \leq b \leq 3$. The parameter a represents the item's discrimination, which facilitates the separation among test-takers with abilities $\leq \theta$ from test-takers with abilities $> \theta$ (Hambleton & Swaminathan, 1990). 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, it can be assumed that in a well-designed multiple-choice item with 5 options, a test-taker 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 this work, consider the information regarding a hypothetical item's database presented in Table 2-1. The database contains only ten items. Although this would not be feasible in a scenario involving real test-takers, a pool of ten calibrated items is sufficient for illustrative purposes.

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	2.38	1.32	0.20
7	1.04	0.79	0.05
8	0.22	0.66	0.20
9	1.26	0.64	0.10
10	-1.29	1.59	0.25

Table 2-1: Hypothetical item pool containing 10 items

The items represented in Table 2-1 are all objective items – such as multiple-choice or multiple-response questions – and therefore can be dichotomously scored.

The test starts with a randomly selected item of medium difficulty. Let us assume that a given test-taker is presented with item 4, an item of medium difficulty (b=0), high discrimination (a=1.5) and pseudo-chance c = 0.10. Given that in this example the test-taker answered the first item correctly, Figure 2-1 represents the Item Characteristic Curve (ICC) for this item, which was calculated using Equation 2-2.

The response likelihood curve is the likelihood of a test-taker 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 (see Figure 2-1) and the *response likelihood curve* are identical. The response likelihood function (Lord, 1980) is shown in Equation 2-3 below.

$$L(u_1, u_2, ..., u_n \mid \theta) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j}$$

Equation 2-3: Response Likelihood Function (Lord, 1980)

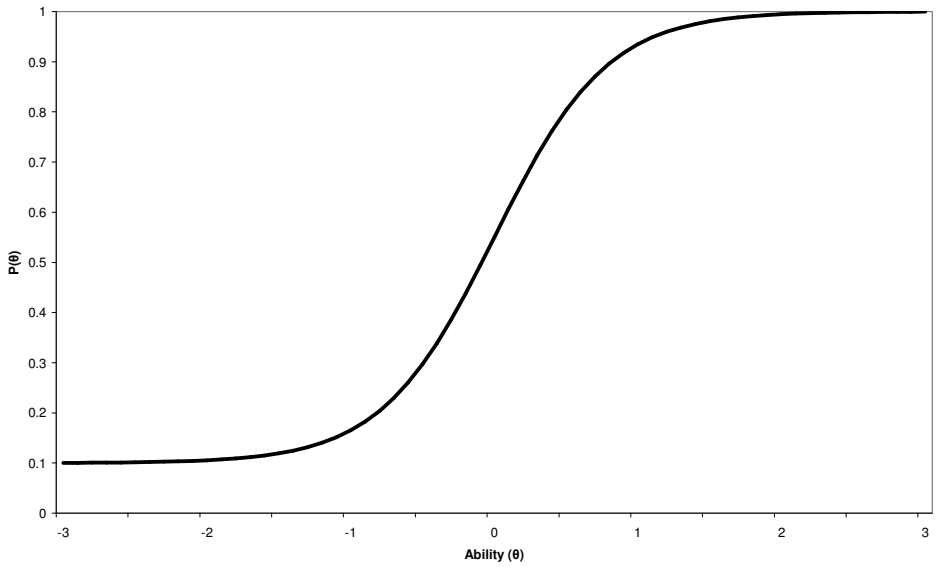

Figure 2-1: ICC curve for item 4 answered correctly

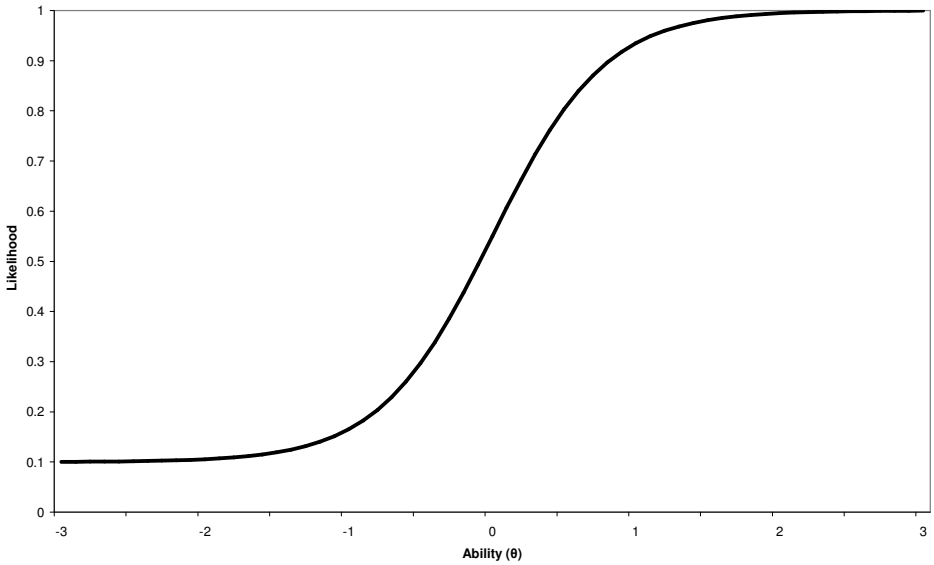

Figure 2-2: Response likelihood curve after item 4 has been answered

In the event of the test-taker 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 the test-taker has also answered item 7 correctly; Figure 2-3 represents the ICC curve for item 7 and Figure 2-4 illustrates the current response likelihood curve, which is the product of the ICC curves shown in Figure 2-1 and Figure 2-3.

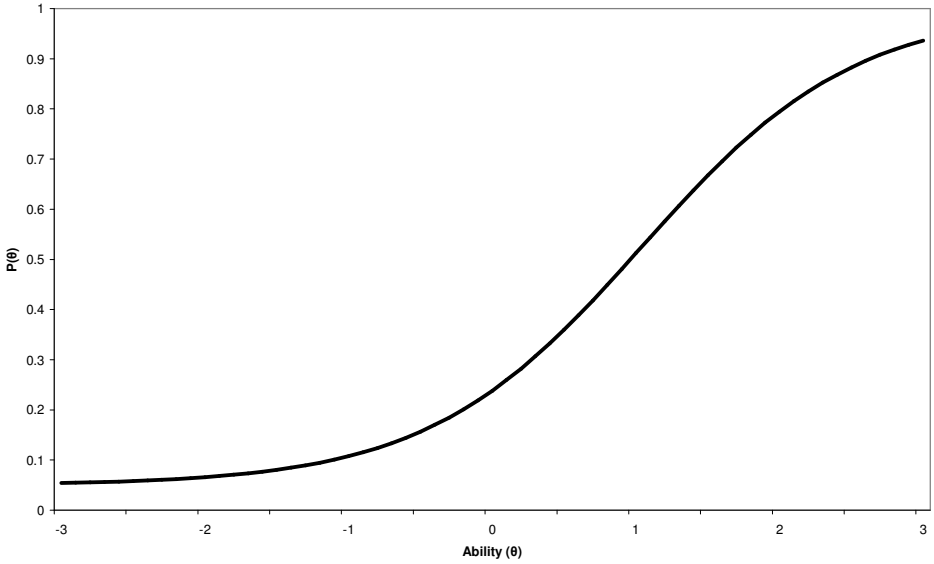

Figure 2-3: ICC curve for item 7 answered correctly

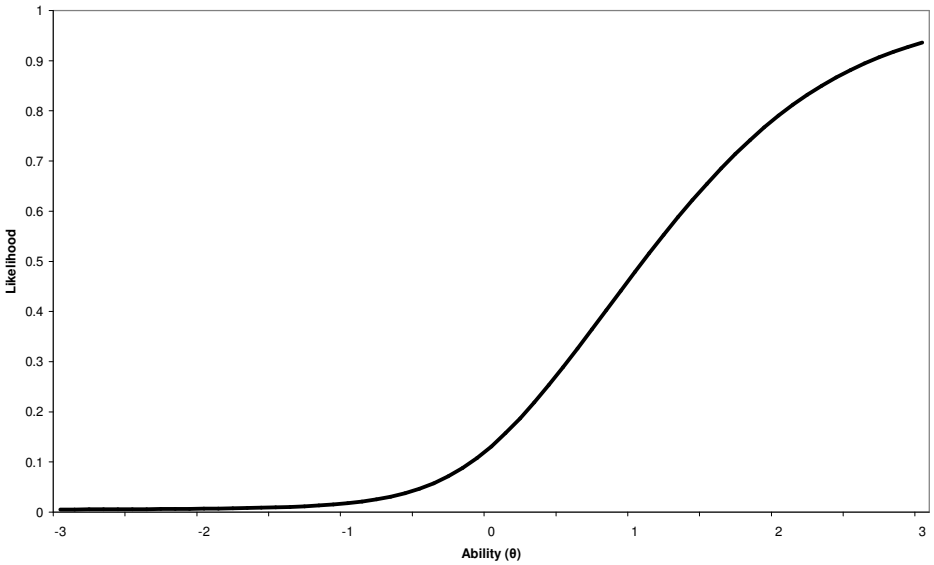

Figure 2-4: Response likelihood curve after two items have been answered

In this example, the test-taker has answered all the items presented correctly. The test-taker's response likelihood curve is composed of the product of two S-shaped curves of type *P(θ)* and, therefore, the curve does not have a peak value. The same characteristic (i.e. no peak value) would have occurred if the test-taker has answered all the items presented incorrectly, since the response likelihood curve would be calculated as being the product of various *(1-P(θ))* and, consequently, the curve would also not have a peak value within the range $-3 \leq \theta \leq 3$.

The test-taker'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 is 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. It is important to note that CAT test-takers are not normally permitted to return to previous items (Vicino & Moreno, 2001) or to omit responses (Lord, 1980; Wainer et al., 2000). The issue of returning to previous item is discussed later in section 3.6. It should be noted that omitting responses is not permitted in the CAT software prototype developed for this research.

Assume that the test-taker 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 test-taker's response for this answer has been evaluated as incorrect, Figure 2-5 illustrates the ICC curve for this item and Figure 6 shows the response likelihood curve after three items have been answered.

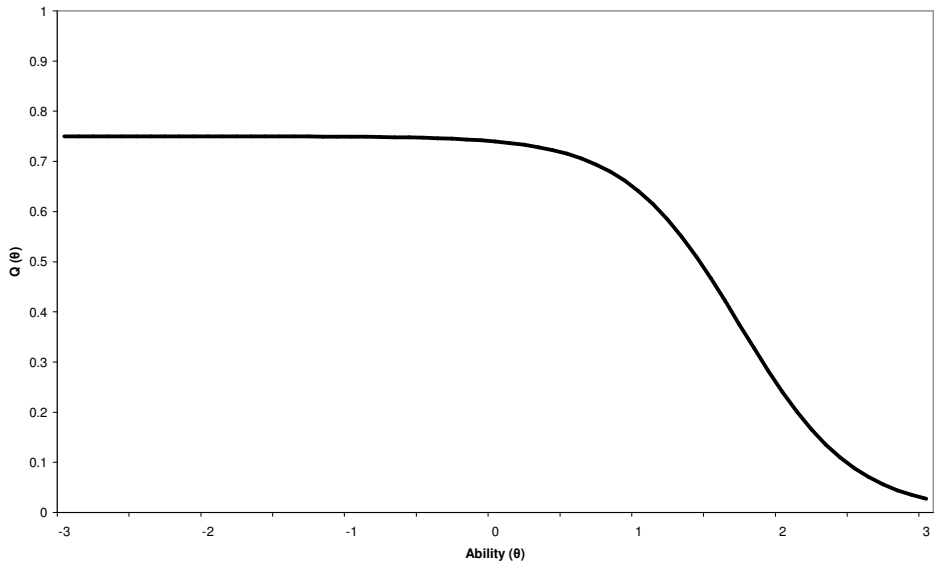

Figure 2-5: ICC for item 2 answered incorrectly

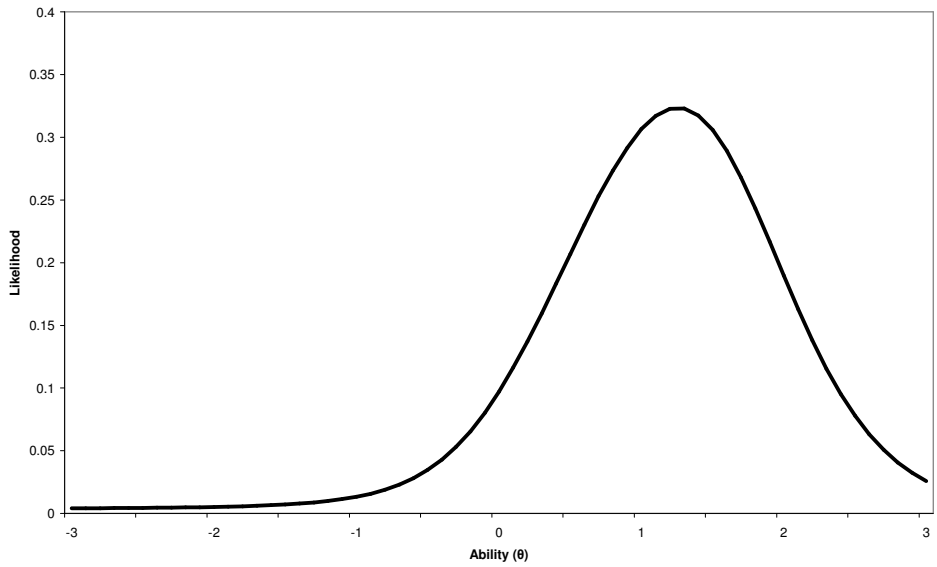

Figure 2-6: Response likelihood curve after three items have been answered

When the test-taker's response likelihood curve is formed by the product of at least one *P(θ)* and one *(1-P(θ))*, 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.

Thus once a provisional ability has been established, the test-taker 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. This item selection criterion is known as difficulty-based criterion (Guzmán et al., 2005).

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=1.26. 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 test-taker's individual responses.

Typically the responses from many questions are necessary in order to estimate a test-taker's ability. The process of presenting items, evaluating the responses using the 3-PL Model and dynamically selecting the next item to be administered is repeated until a stopping condition is met. Examples of stopping conditions include: error of estimation of the test-taker's ability, a fixed number of items has been administered and a certain time has elapsed. Stopping conditions are discussed further in section 3.5.

The following section introduces the six main components of a CAT, as identified as part of the research reported here.

2.2.5 Key components of the CAT approach

The identification of the six key components of the CAT approach was based on the work of Carlson (1994), Linacre (2000), Flaugher (2000), Wainer & Mislevy (2000) and Thissen & Mislevy (2000), and these components are listed below.

A calibrated item pool. In a CAT, the items (i.e. questions) in the pool must be calibrated. In the case of the 3-PL model calibrated item statistics are used to describe the item's difficulty b, discrimination a and pseudo-chance c. It is also possible that other parameters, such as item content, are considered and these issues are later discussed in section 3.4. With regards to its size, it is recommended that the item pool is as large as possible and that the difficulty of the items is widely spread out, in order to cover the entire range of test-taker ability. Wainer & Egnor (2000) suggest that the pool should contain thousands of items. McBride (2001c) suggests that the number of items in the pool should exceed the number of items administered to a test-taker by a ratio of 5 or 10 to 1. Carlson (1994: p. 219) reports on a body of research that suggests that "satisfactory results can be obtained with pools of approximately 100 items" provided that the items "span the entire difficulty range". Issues related to the calibration of items relevant to this research are later discussed in section 3.2.

An item response model. As mentioned earlier, examples of item response models include the 1-PL, 2-PL and 3-PL models (Lord, 1980). The chosen model should form the basis for the calibration of items, and algorithms for item selection and ability estimate.

A method for selecting the item to be administered first. Generally, very little information (if any) about the test-taker's ability is available at the start of the test. The selection of the item to be administered first can be totally random, or based on an educated guess about the test-taker's ability. How to start a CAT test is discussed later in section 3.3.

A method for computing the test-taker's ability (and provisional test-taker's ability). Computing the test-taker's ability can be achieved through methods such as maximum likelihood. This is discussed in section 2.2.4 and section 3.4.

A method for selecting the item to be administered next. This involves searching through the calibrated item pool in order to identify a non-administered item that best matches the item selection criteria. Often this can

be translated into administering the item from the calibrated pool for which the difficulty b is the nearest value to the most recent ability estimate. Section 3.4 provides a description of the item selection method employed in this research.

A stopping condition. As it name implies, the stopping condition in a CAT is employed to terminate the test. For example, the test can be terminated when a predefined number of questions has been administered, or the ability estimate is considered to be sufficiently accurate. Different approaches to terminating a CAT are discussed later in section 3.5.

In addition to the key components of a CAT, the main advantages and barriers to the implementation of the approach were identified as part of this work. These are reported next.

2.2.6 CAT approach: advantages and barriers

Rather than looking at the advantages of the CAT approach that are generic to CAA – such as increased speed of marking – this section focuses on the advantages that are characteristic of the CAT approach. The key advantages of the CAT approach, as reported in the literature, are presented next.

Measurement precision. Sands & Waters (2001) argue that the measurement precision of a conventional (classic) test where all test-takers answer the same set of questions is peaked around the middle ability level of the target population. As a result, high measurement precisions are obtained for the average test-taker, and less so for those test-takers at the low and high end of the ability scale. In contrast, in a CAT the measurement precision in a CAT is improved overall given that questions are tailored to individual test-takers.

Test efficiency. Much of the CAT literature focuses on the benefits of the approach in terms of efficiency. Jacobson (1993), Carlson (1994) and Wainer (2000a), for example, cite increased efficiency of testing as one of the main benefits of the CAT approach. Items that are too easy or too difficult for a given test-taker provide very little useful measurement information regarding

this test-taker's ability. By tailoring the difficulty of the question to each test-taker, the test length can be reduced with no loss in measurement precision (Jacobson, 1993; Carlson, 1994; Wainer, 2000a). Jacobson (1993) and Carlson (1994), for instance, suggest that it is possible to reduce test length by up to 50% without jeopardising test validity and reliability.

Test security. Ward (1988) suggests that the use of CATs can lead to improved security. This is due to the fact that the questions administered to each test-taker are dynamically selected according to ability. This can result in different test-takers being administered different sets of questions. This would, in turn, make it more difficult for test-takers to share detailed information about the test that could improve their scores in the future.

Test-taker motivation. In a CAT, test-takers are challenged and motivated by test items at an appropriate level, rather than discouraged by items that are far above or below their ability level (Wainer, 2000a).

In addition to the main advantages of the CAT approach, some barriers to its implementation were identified. These barriers are listed next.

The model itself. The CAT approach as proposed in this thesis relies on the 3-PL model from IRT and, naturally, on the assumptions of this theory. One of the assumptions of IRT (and Classical Test Theory, for that matter) is the existence of a single dimension of knowledge or trait (for example, mathematical facility) that accounts for an individual's performance when answering an item. As Wainer et al. (2000) point out, this premise is rather limited as it does not represent the complexity of, for instance, how individuals solve problems. In practical terms, however, Lord (1980), Wainer & Mislevy (2000), Van der Linden & Hambleton (2000) amongst others have shown IRT models for dichotomously scored items to be useful in real world applications.

Stakeholder attitude towards the approach is under-represented in the relevant literature. A number of authors including Lord, (1980), Hambleton & Swaminathan (1990); Weiss & Yoes (1991), Wainer & Mislevy (2000), Segall & Moreno (2001), Wolfe et al., (2001b), Krimpen-Stoop & Meijer (2003) and Eggen (2004) have reported on research concerned with psychometric

aspects of IRT models, such as efficiency and measurement precision. However, relatively little attention has been paid to practical issues for small-medium implementations of the CAT approach, in particular stakeholder attitude towards the approach.

Practical implementation issues. The effort required to implement a CAT is much greater than that required to implement a CBT. The item selection procedure in a CBT is less critical than in a CAT, as in the former all test-takers are presented with the same set of items. In a CAT, the questions presented to test-takers are selected in such a way to maximize the level of information about the test-taker at a particular ability level. Another salient difference between CBTs and CATs is related to the database of questions. In a CBT, a database containing only the questions to be administered during the test is required. In a CAT, a large and calibrated database of questions, spanning the entire difficulty range is required. The calibration of the item pool can be an arduous process, as discussed later in section 3.2.

2.3 Summary

There is a large body of research to support the view that computer-assisted assessment (CAA) is a regular component of student assessment in Higher Education (Joy et al., 2002; Conole & Bull, 2002; Warburton & Conole, 2003; Bull & McKenna, 2004; Warburton & Conole, 2004). Much of the use of CAA in Higher Education focuses on the use of objective testing, in particular computer-based tests (CBTs) (Bull & McKenna, 2004).

The prime difference between computer-adaptive tests (CATs) and conventional CBTs is the way in which the questions are selected. In a CBT, the same set of fixed questions is administered to all test-takers. However, this static approach often poses problems for individual test-takers, as a typical CBT contains items that are intended to cover a broad range of abilities. As a result, low performing test-takers must answer numerous questions that cause frustration, as they are above their level of ability. Similarly, high performing

test-takers are required to answer a number of questions below their ability level before reaching a level where they are challenged. In both cases, such questions provide little useful information about the ability of test-takers.

In contrast, in a CAT the proficiency level of individual test-takers is estimated during the test so the questions can be tailored to match each test-taker's ability within the subject domain. CATs are typically based on Item Response Theory (IRT) (Lord, 1980), which is a well-known family of mathematical functions that aim to predict the probability of a test-taker answering an item correctly. Earlier in this chapter, a brief historical account of the development of IRT and some examples of IRT applications were provided.

There is more than one IRT model for dichotomously scored items, but the Three-Parameter Logistic (3-PL) model was chosen as the underlying model for this research because this model:

- takes into account not only the question's difficulty but also: the discrimination provided by the question and the probability of a test-taker answering a question correctly by chance (Lord, 1980);
- is a widely used IRT model for objective testing (Wainer & Mislevy, 2000).

CATs are more difficult to construct than conventional CBTs, due to the need for an adaptive algorithm, and a large and calibrated item pool. The CAT approach, however, presents various benefits over its CBT counterpart such as improved measurement precision, increased efficiency, enhanced security and increased test-taker motivation. Barriers to the adoption of the CAT approach include: limitations of the 3-PL model, lack of compelling evidence from stakeholders reporting their acceptance of the approach, and practical implementation issues.

The following chapter focuses on practical design and implementation issues, specifically:

- database calibration;
- how to start a CAT;

- how to select the item to be administered next;
- how to terminate a CAT;
- item review in a CAT.

3. The design and implementation of the CAT prototype

The previous chapter introduced the underlying concepts of the computer-adaptive test (CAT) approach. As part of the research described here, a high-fidelity software prototype of a CAT was designed, implemented and evaluated. In this chapter, an overview of how the concepts introduced in the previous chapter were applied to the design and implementation of the testing algorithm for a CAT based on the Three-Parameter Logistic model (Lord, 1980) is provided. The evaluation of the prototype is presented later in Chapters 4 and 5.

This chapter is organised into six main sections. The first section provides a brief overview of the CAT software prototype implementation. The second section is concerned with the calibration of database questions; as mentioned in section 2.2.5, a calibrated question database is a key element of a CAT. The following three sections focus on the testing algorithm of the CAT software prototype developed for this research. A testing algorithm can be defined as a collection of steps that describe how a test is performed. The testing algorithm of a CAT can be generically described as “start”, “select the item to be administered next” and “stop”. The “start” step is mostly concerned with the level of difficulty of the question to be administered first. The “select the item to be administered next” step relates to the factors that might be taken into

account when dynamically selecting the question to be presented next. Such factors might include one or more of the following: difficulty, discrimination, pseudo-chance, content and exposure. Finally, the "stop" step is concerned with the specification of stopping conditions for a CAT test. The establishment of stopping conditions might depend on a wide range of factors, such as the need for test efficiency. These aspects are discussed later in this chapter.

The sixth section introduces the issue of whether or not test-takers should be permitted to return to previously answered questions.

The following section focuses on practical implementation issues relating to the CAT software prototype.

3.1 Implementation overview

The CAT software prototype is an application developed for the Microsoft Windows platform, and was implemented in Visual Basic (VB) version 6. VB is an event-driven programming language that is suitable for the software development method chosen for the research, i.e. iterative prototyping (Preece et al., 2002).

VB provides means of accessing databases using ActiveX Data Objects (ADO) through an OLE-DB (Object Linking and Embedding-Database) provider (Microsoft Corporation, 2007a). In this work, a Microsoft Access database was employed to store information regarding test-takers, test-taker performance during the test (including responses to individual test items) and item characteristics.

In the early stages of the search, it was planned that all test-takers would access the same instance of the CAT database over a computer network. Microsoft Access back-end databases can support up to 255 simultaneous users; however, better performance is typically achieved with 25 to 50 users (Microsoft Corporation, 2007b). Performance issues merited special consideration because one read and one write operation was carried out for each item (i.e. question) answered by each test-taker during the test: one read

operation to retrieve the data relating to the item (i.e. question) to be administered next, and one write operation to store information regarding the test-taker's response to the current item.

In addition to potential performance problems when assessing over 50 test-takers simultaneously, sporadic network connection and performance problems at the University's laboratories have led the research team to adopt a decentralised approach. Interruptions during assessment sessions were likely to lead to increased test-taker anxiety and, for this reason, it was important to the research to adopt such an approach in order to eliminate (or, at the very least, minimise) network related problems. In this decentralised approach, each workstation contained its own local copy of the CAT database. A batch VB program was employed at the end of the test to collate all data from each workstation into a master CAT database. The collation of data at the end of the test is necessary in order to: (1) allow academic staff to examine information regarding test-taker performance and (2) release results to test-takers.

The following section focuses on different approaches to the calibration of items (i.e. questions).

3.2 Database calibration

In any practical implementation of the CAT approach, a calibrated item (i.e. question) database is required. In the case of the Three-Parameter Logistic (3-PL) model (Lord, 1980), the calibration of items is concerned with assigning values to each of the IRT parameters: difficulty b, discrimination a and pseudo-chance c (Ward, 1988); a brief introduction to the 3-PL model can be found in section 2.2.4. The calibration of item parameters is central to the CAT approach, as these parameters are employed in selecting the question to be administered next (i.e. selecting the question in the database that best matches a test-taker's proficiency level) and calculation of a test-taker's proficiency level (see Equation 2-2, p. 44).

There is more than one approach to the calibration of items, and the next section provides an overview of the main issues that were considered as part of this work.

3.2.1 Overview

In addition to the conventional approach to item calibration, this section describes two other approaches: expert and online calibrations.

Conventional calibration. This approach involves using methods such as the joint maximum likelihood (JML), the conditional maximum likelihood (CML) and the marginal maximum likelihood (MML) estimation to analyse actual test-taker responses, and compute item parameter estimates. The MML method is considered particularly suitable for settings with fewer test-takers (Gierl & Ackerman, 1996). The number of test-takers required in order to estimate item parameters varies, with some recommending actual responses from at least 1,000 suitably selected test-takers (Wainer & Mislevy, 2000; McBride, 2001c), and others recommending between 200 and 1,000 test-takers (Huang, 1996).

The most common ways of obtaining response data from test-takers are:

- recruiting suitable test-takers for the sole purpose of item calibration;
- using data already available; in some cases, this involves analysing items that were previously administered as part of paper-and-pencil tests or CBTs.

Expert calibration. A further approach to item calibration would be the use of subject domain experts to define IRT parameters, in particular the level of difficulty b, of non-calibrated questions. Yao (1991), for example, describes an application of CAT where language experts rated the difficulty of 69 Chinese newly-written items according to 9 levels of language proficiency. In a similar vein, Linacre (2000) depicts a CAT application where experts rated the difficulty of reading comprehension items based on Lexile difficulty. The Lexile text difficulty takes into account factors such as word and sentence length

(Linacre, 2000). It should be noted that both examples above refer to CAT applications that employ the Rasch model (1-PL model, see section 2.2.3).

Fernandez (2003) describes the implementation of a 3-PL CAT where five experts classified a set of 30 questions into five categories, from 1 (very easy) to 5 (very difficult). Gonçalves et al. (2004) depict a 3-PL CAT where experts were required to assign values to a question's difficulty b as follows: difficult (-2.5), medium (-1.0), easy (1.0) and very easy (2.5). Conejo et al. (2000) also propose a CAT 3-PL application where expert calibration is used for the initial calibration of items.

Online calibration. This calibration method entails using test-taker responses to previously calibrated items to estimate parameters of new items during the course of a test (Wainer & Mislevy, 2000). It is also possible to employ online calibration to refine existing IRT parameter estimates. Conejo et al. (2000), for instance, employ expert calibration for setting the initial IRT parameters and then online calibration for refining the parameter values.

As part of this research, the three item calibration approaches described above were considered. In the following section the approach to question calibration employed in this work is described.

3.2.2 Approach employed in the research

The approach to calibration employed in this research was organised into two different stages. The first refers to newly-written questions or, in other words, questions with no historical data. The second stage refers to questions with historical data.

Calibration of newly-written items. The conventional approach for newly-written questions was not employed, as it was considered that smaller applications of the CAT approach, such as the application introduced here, would benefit from a calibration procedure that did not depend on actual responses from test-takers. The expert calibration was chosen over the online one, as its implementation was faster and simpler.

The expert item calibration, as implemented in this work, was based on Bloom's taxonomy of cognitive skills (Bloom, 1956; Anderson & Krathwohl, 2001) and difficulty within the subject domain. The CAT prototype developed for this research was centred on the use of objective questions, and an important assumption of this work was that such questions are capable of effectively assessing the first three cognitive skills: knowledge, comprehension and application. Based on this assumption, questions could then be categorised as assessing of one of these three cognitive skills.

In this work, subject domain experts were employed to classify the items (i.e. questions) according to skills assessed (see section 2.1). After this initial classification, subject experts were then required to rank questions according to their difficulty. Their ratings were then translated into a value between -3 and +3, in order to serve as the difficulty b for the item. Table 2-1 illustrates the range of values and corresponding cognitive skills.

Difficulty b range	**Skill assessed**	**Brief description**
$-3 \le b < -1$	Knowledge	Ability to remember and/or recall previously taught material
$-1 \le b < +1$	Comprehension	Ability to interpret and/or translate previously taught material
$+1 \le b \le +3$	Application	Ability to apply taught material to novel situations

Table 3-1: Difficulty b range and corresponding cognitive skills

After the classification according to difficulty b, experts were then required to classify items according to their usefulness at differentiating between examinees within each cognitive skill range. This was then employed to assign a value to the item discrimination a parameter as follows:

- a set to 0 for items with lesser usefulness;
- a set to 1 for items with greater usefulness.

Finally, the parameter c was estimated depending on the number of item options and key answers. For example, a multiple-choice item (i.e. one option is the key answer) with four options would have its c value set to 0.25 (i.e. 1 / 4). This estimate was based on the assumption that all options are equally plausible.

Once questions have been answered by test-takers, their responses are used to refine IRT parameters or, in other words, to recalibrate the questions. In the next section, the approach to recalibration employed in this research is described.

Recalibrating existing questions. As mentioned above, expert calibrations were employed only for newly-written items or, in other words, items with no historical data. Expert calibrations were refined (i.e. recalibrated) after each CAT assessment session. The recalibration was carried out using actual responses from test-takers, and performed by importing test-taker actual responses to the commercial software application XCalibre (Assessment Systems Corporation, 2007; Gierl & Ackerman, 1996). The XCalibre software employs the MML estimation method; this method, as mentioned earlier, requires fewer test-takers than other methods such as CML and JML in order to perform the item parameter estimation.

To summarise, a combination of expert calibration and MML item parameter estimation method was employed in this work in order to calibrate the question database as follows:

- expert calibration, based on Bloom's taxonomy of cognitive skills and difficulty within the subject domain, was used for newly-written items (i.e. items with no historical data);
- test-taker responses and the MML parameter estimation method were employed to recalibrate existing items.

An empirical study was carried out in order examine the usefulness of expert calibration, as proposed in this work, when setting initial values for the difficulty b.

Method. The database employed in this study comprised 150 items within the Visual Basic.NET subject domain. The database was initially calibrated by experts, and then re-calibrated three times after CAT assessment sessions using the XCalibre software package (Assessment Systems Corporation, 2007).

Table 3-2 shows the difficulty b means, after each calibration. As can be seen from Table 3-2, the difficulty b mean value for the expert calibration was the lowest (mean=0.855, SD=1.203, N=150). This can be taken to indicate that, on average, experts perceived the questions to be easier than test-takers.

	Difficulty b Mean	**Std. Dev.**
Expert calibration	0.855	1.203
MML Recalibration 1	1.189	0.943
MML Recalibration 2	1.151	1.155
MML Recalibration 3	1.113	1.146

Table 3-2: Mean values for the difficulty b value (N=150)

The data in Table 3-2 were subjected to statistical analysis, and the main findings are reported next.

Findings. In order to test for significant differences between the difficulty b means, a repeated measures analysis of variance (ANOVA) was performed on the data shown in Table 3-2. For the purpose of clarity, this ANOVA will be referred to as ANOVA_01. The findings of ANOVA_01 showed that there were significant differences between the difficulty b means (see Table 3-2), such as df=3, F=19.935, p<0.001.

A repeated measures analysis of variance (ANOVA) was also performed on the difficulty b means for Recalibration 1, Recalibration 2 and Recalibration 3 (i.e. excluding the expert calibration from Table 3-2). For the purpose of clarity, this ANOVA will be referred to as ANOVA_02. The findings of

ANOVA_02 showed that there were no statistically significant differences, such as df=2, F=1.198, p=0.303. Thus, the expert calibration appeared to account for the statistically significant differences in means reported in ANOVA_01.

One of the assumptions of this work was that items could be classified according to Bloom's taxonomy of cognitive skills, as shown in Table 3-1. Table 3-3 shows the number of items per cognitive skill assessed, after the MML recalibration (observed) and as estimated by the experts (expected).

Cognitive skill	MML Recalibration 3 (Observed)	Expert calibration (Expected)	Residual
Knowledge $-3 \le b < -1$	10	14	-4
Comprehension $-1 \le b < +1$	97	92	5
Application $+1 \le b < +3$	43	44	-1
Total	150	150	

Table 3-3: Total number of items per cognitive skill

As can be seen from Table 3-3, the number of items per cognitive skill in the expert calibration was different from that observed in 'MML Recalibration 3'. In order to test if the difference was statistically significant, a Chi-Square test was performed. No statistically significant difference was found (df=1, Chi-Square=0.482, p=0.487). This is an interesting finding, as although there was a significant difference in the difficulty level calibration (see Table 3-2), this was not reflected in the cognitive skill calibration where the difference was not significant (p=0.487).

For example, there is the case of one question that had its difficulty b set to -1 by experts, but this value was refined to 0.18 after recalibration. Although expert and MML estimates are different, both estimates for the difficulty b (-1 and 0.18) are within the comprehension range (see Table 3-1).

There is also the issue of questions that do not fit the expert calibration model as proposed in this work. One of the questions, for example, was calibrated as having its difficulty b equals to -2 by experts. This would denote the cognitive skill knowledge. After recalibration, however, this question had its difficulty b recalibrated to 0.79 (comprehension). One can speculate that the reason for this is that, although the question assessed the cognitive skill knowledge, the question assessed what Ward (1980: p. 55) calls "abstruse facts" within the subject domain, and therefore its difficulty b was increased. In such a scenario, recalibration can be employed to identify questions that do not fit the model for later removal from the database.

Finally, there are other aspects that can affect a question's difficulty b estimate, such as question exposure. Assume that there is a question that assesses the cognitive skill application. If test-takers have been exposed to the question and its correct response before, it is possible that a test-taker would be able to answer the question correctly based on the cognitive skill knowledge (recall) rather than application.

The calibration of questions is a very complex topic (Wainer & Mislevy, 2000; Conejo et al., 2000; McBride 2001c; Guzmán & Conejo, 2005). With hindsight, the use of experts for the initial calibration of items may be seen as an over-simplification of item parameter estimation; however, in the context of the research this approach proved to be adequate and useful. Barker et al. (2006b) support the findings reported in this section.

The following section discusses how to start a CAT assessment session.

3.3 Starting the test

In a CAT, the question to be selected next depends on the set of previous responses as described in section 2.2.4. There remains the issue of how to decide the first question to be administered, although Lord (1980: p. 153)

states that “unless the test is very short, a poor choice of the first item will have little effect on the final result”.

In the early stages of the research, three different approaches to starting the CAT test were considered. The first approach was to start the test with a random question from the question bank. A potential limitation of this approach is that the test could start with a question from either end of the difficulty scale, i.e. very difficult or very easy. In the context of the research reported here, questions from either end of the difficulty scale were considered less useful than questions from the middle of the scale.

The second approach considered was to utilise information about test-takers obtained prior to the test, such as education history, previous CAT scores or performance in similar subjects (Lord, 1980; Thissen & Mislevy, 2000). For instance, test-takers who performed well in a similar subject area would start the test with a question of greater difficulty than those who performed less well.

Given that historical information about test-takers is not always available, the third approach considered was to start the test with a question of middle difficulty. An important assumption of the work reported here was that proficiency levels ranged from -3 to +3 with a mean of 0. Therefore, it was considered practical to start the test with a question for which the difficulty parameter b was near 0. This is a similar approach to that reported by Wolfe et al. (2001a), where the ability levels ranged from -2.250 to +2.125 with a mean of 0 and it was also assumed at the start of the test that the test-taker’s ability was 0.

3.4 Selecting the item to be administered next

In the work reported here, three issues – difficulty, content, exposure – were taken into account when selecting the item to be administered next. It is important to note that the item selection relies on the existence of a calibrated pool of items. The item selection based on item difficulty is introduced next.

Difficulty. The first issue was the level of difficulty of the item (i.e. question) to be administered next. The underlying idea was to present each test-taker with different items based on estimated proficiency level and set of previous responses. To this end, the Three-Parameter Logistic (3-PL) model from Item Response Theory (IRT) was employed to estimate the proficiency level of each test-taker. For the purposes of clarity, the mathematical function from the 3-PL model used to model used to evaluate the probability *P* of a test-taker with an unknown ability θ correctly answering a question of difficulty b, discrimination a and pseudo-chance c is shown in Equation 3-1 (Lord, 1980):

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

Equation 3-1: Three-Parameter Logistic Model (Lord, 1980)

Once a proficiency level has been estimated, methods such as "maximum information" (Thissen & Mislevy, 2000), "maximum expected precision" (Thissen & Mislevy, 2000) and "difficulty-based" (Guzmán et al., 2005) can be employed to select the most informative item for a test-taker's estimated proficiency level. The work reported here employed the "difficulty-based" method (Guzmán et al., 2005), as this was less computationally demanding and therefore potentially easier to implement than its item selection counterparts. In summary, in the work reported here, once a provisional proficiency level has been estimated, the test-taker is then supplied with:

- an item from the item's bank for which the difficulty b is the nearest value to the most recent proficiency level estimate;
- if there is more than one item with the same difficulty b, then the item with the highest value for the discrimination a is administered next;
- if there is more than one item with the same difficulty b and discrimination a, then the item with the lowest pseudo-chance parameter c is administered next;

- if there is more than one item with the same difficulty b, discrimination parameter a and pseudo-chance c, then one of the selected items will be randomly administered next.

The proficiency level estimate is calculated using the response likelihood function (Lord, 1980) shown in Equation 3-2.

$$L(u_1, u_2, ..., u_n \mid \theta) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j}$$

Equation 3-2: Response Likelihood Function (Lord, 1980)

As shown in section 2.2.4, a proficiency level estimate can only be computed once the test-taker has answered at least one item correctly and one item incorrectly; this is because "there will be no finite maximum likelihood estimate of the examinee's ability as long as his answers are all correct or incorrect" (Lord, 1980: p. 153). In the event of a test-taker answering all items correctly, the item to be administered next will be an item of difficulty b higher than the previous item. Conversely, in the event of a test-taker answering all items incorrectly, the question to be administered next will be an item of difficulty b lower than the previous item.

There is also the potential for unusual response patterns, commonly known as aberrant responses (Thissen & Mislevy, 2000). This would occur when test-takers provide correct responses for difficult items and incorrect responses for easy items. Aberrant response patterns are atypical (Lord, 1980; Hambleton & Swaminathan, 1990; Thissen & Mislevy, 2000), and were not observed as part of the research reported here.

Content. Content balancing was the second issue taken into account when selecting the item to be administered next. In early stages of the research, the item selection mechanism focused solely on 3-PL parameters. Although this was an effective approach, it was observed that some subject areas have a diverse content. Given that such diverse content can be divided into topic areas, it was considered necessary to modify the item selection algorithm so that it would rotate through the different topics areas within the subject domain

being tested. This is in line with the work of Thissen & Mislevy (2000), where it was found that content balancing might be an important factor in some applications of the CAT approach. In practical terms, to ensure content balancing, items are first selected according to topic area and then according to difficulty.

Exposure. Similarly to topic area, item exposure control is a factor that is not incorporated into the 3-PL model but received special consideration in the work described here. Interesting work in the area of item exposure has been carried out by Stocking & Lewis (2003), who implemented an algorithm capable of controlling item exposure conditional on ability. The focus of Stocking & Lewis's (2003) work is on practical applications of CAT, where it is possible that an item has an overall low exposure but a high exposure amongst test-takers of similar ability. There is also the work of Hetter & Sympson (2001), who implemented a randomization scheme in order to reduce the exposure of certain items in the pool. Unlike Stocking & Lewis's (2003) work, Hetter & Sympson's (2001) method is not conditional on ability levels.

The method of controlling item exposure as employed in the research described here is simple and straightforward, and is based on the work of Hetter & Sympson (2001). The CAT prototype introduced here keeps track of the number of times each item from the bank is administered to test-takers. In the event of more than one item from the pool presenting the same values for topic area, difficulty b, discrimination parameter a and pseudo-chance c, the item with the least exposure is administered next.

Figure 3-1 provides an overview of how the CAT software prototype developed for this research works. Figure 3-2 illustrates the method employed in this research for dynamically selecting the item to be administered next. It can be seen from Figure 3-2 that items are selected based on the following criteria in descending order of priority: content, difficulty, discrimination, pseudo-chance and exposure.

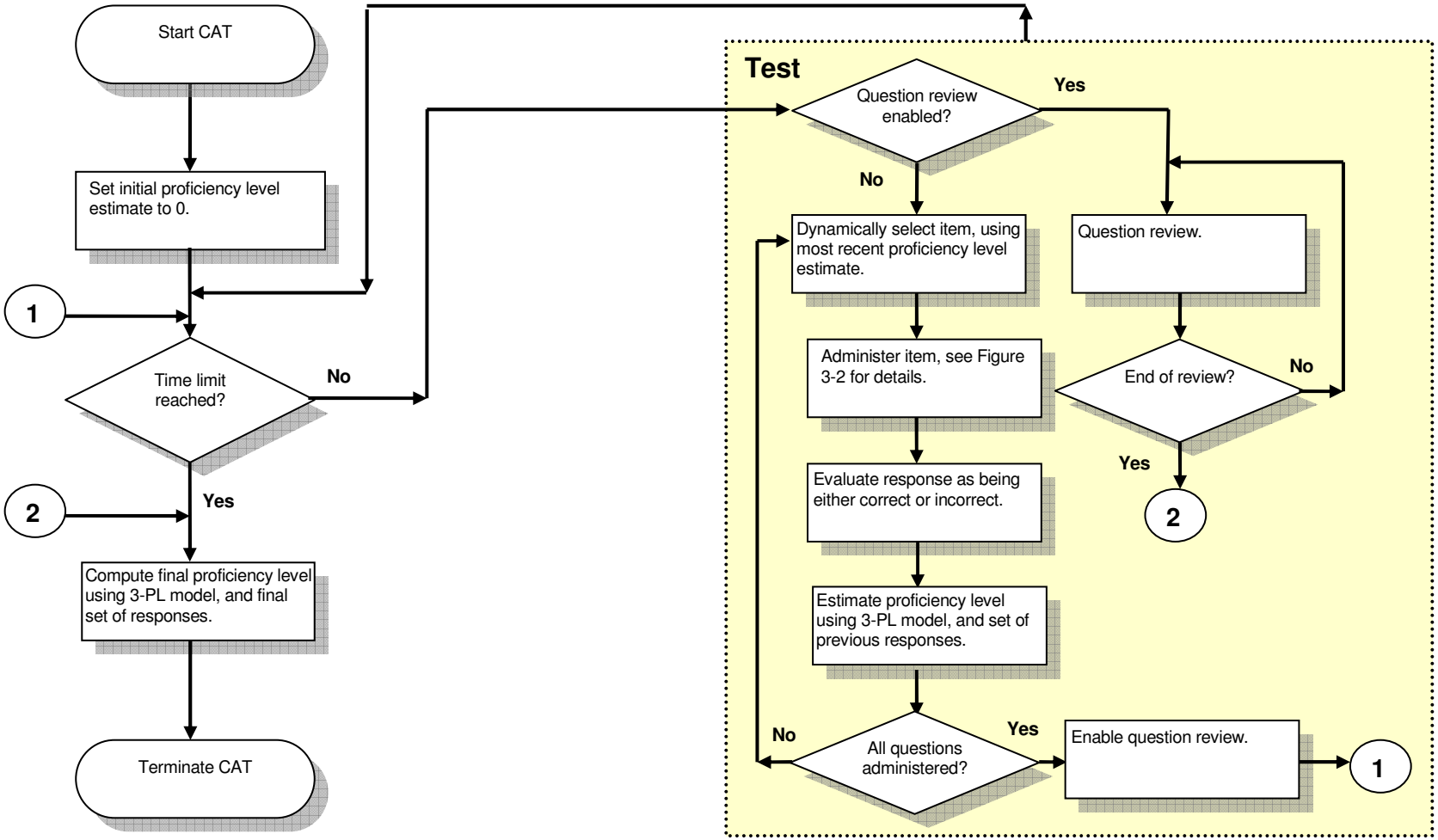

Figure 3-1: Overview of how the CAT prototype works
It should be noted that test-takers can terminate the test at any given time, by selecting the "Exit" option.

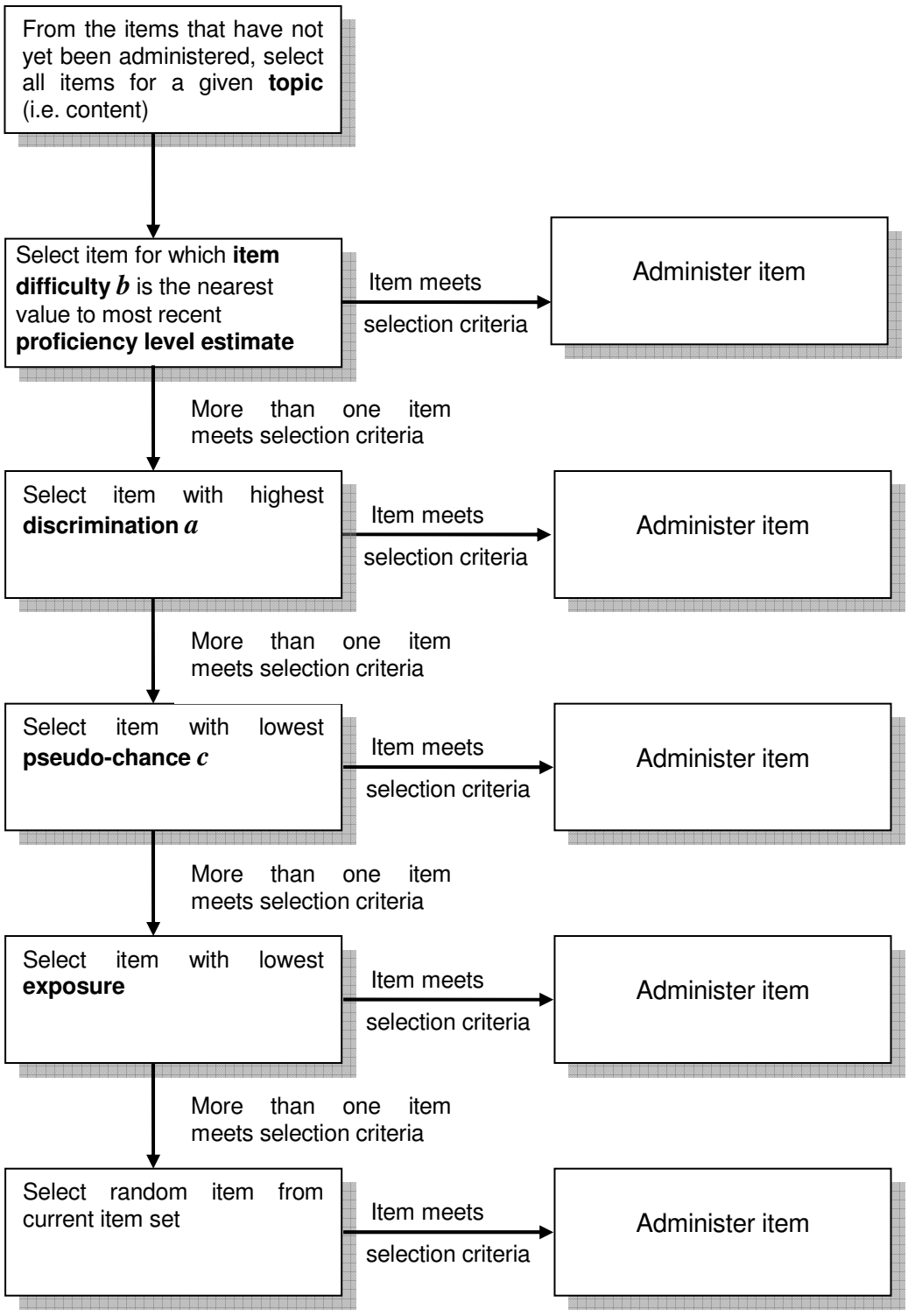

Figure 3-2: Item selection method used in this research

An important consequence of the item selection algorithm presented here is that, in practice, test-takers participating in the same CAT assessment session are likely to be presented with different sets of questions. The implications of this will be discussed in Chapters 4. Finally, it is important to note that in a CAT, the question to be administered next depends on the set of previous responses. For this reason, test-takers are not generally allowed to skip questions.

The next section describes different stopping conditions.

3.5 Stopping the test

A CAT may be stopped when a fixed number of items has been administered or a predetermined amount of time has elapsed. Such tests are commonly referred to as fixed-length CATs. It is also possible to stop a CAT once a satisfactory level of target measurement precision is achieved, for example the standard error for the test-taker's ability estimate reaches a predefined level. Such tests are commonly referred to as variable-length CATs. When defining the stop condition for a CAT, the tester can either adopt a single stopping condition or a combination of stopping conditions. For example, the tester can specify that the test stops when all items have been answered or a certain time has elapsed, whichever happens first. The next section presents the factors that are of greater relevance when establishing the stopping condition for a CAT.

3.5.1 Major factors regarding stopping conditions

This section is concerned with the major factors that need to be considered when choosing the stopping condition for a CAT. These factors include: efficiency, measurement precision, practical implementation issues and test-takers' attitude.

Efficiency. Variable-length CATs have the potential to achieve proficiency level estimates that are as accurate as these obtained from fixed-length CATs in a more efficient way, with less testing time and fewer test questions (Thissen & Mislevy, 2000; McBride, 2001b). Indeed Jacobson (1993) and Carlson (1994) suggest that in a variable-length CAT, test length can be reduced up to 50% without jeopardising test validity and reliability.

Measurement precision. In addition to greater efficiency, McBride (2001b: p. 56) reports that several researchers favour variable-length CATs in order to "achieve equal measurement precision for all examinees".

Despite the predicted benefits listed above, Thissen & Mislevy (2000) recommend that a combination of stopping rules – for example, standard error for the test-taker's ability estimate and maximum number of questions – is used in real-world applications of variable-length CATs. This is because there is the risk of not attaining the predefined degree of precision for the standard error of test-takers' ability estimate within reasonable testing time, which could lead to test-takers' fatigue or uncooperative behaviour. Furthermore, it is theoretically possible that all items in the pool are administered in a variable-length CAT without reaching the specified degree of precision.

Practical implementation issues. Fixed-length CATs present the advantage of being easier to implement than their variable-length CAT counterparts (Thissen & Mislevy, 2000). For instance, one can argue that the algorithm for a fixed-length CAT is less complex to design and implement than that required for a variable-length one, as the former would not involve performing calculations to determine whether or not a specified degree of precision has been achieved. Furthermore, fixed-length CATs that use the number of items to be administered as a stopping condition make it possible for the examiner to predict how many items are required in the item pool in order to support a CAT assessment session. Carlson (1994), for example, recommends that the question pool should contain at least three to four times the number of questions to be administered at all different levels of ability.

Test-taker attitude towards different stopping conditions. There is no substantial evidence on which stopping condition is the most suitable from the test-takers' perspective, as empirical studies of test-takers' views on different stopping conditions are under-represented in the CAT literature. Research reported by McBride (2001a) has shown that, in a variable-length CAT setting, "low ability examinees took much shorter tests than high ability examinees" (p. 56) and "this could lead to questions of equity" (p. 56). Furthermore, Hambleton et al. (1991) reported that "short tests are often viewed suspiciously by examinees" (p. 351). Sands et al. (2001) reported on previous research that "concluded that variable-length stopping rules based on the reliability of the ability estimate offered no advantage in precision over fixed-length tests, and that fixed length tests were probably more fair to lower ability examinees" (p. 75).

Two topics related to CAT stopping conditions did merit further investigation as part of the research reported here, namely the use of standard error for the proficiency level estimate as a stopping condition, and test-takers' attitude towards different stopping conditions. Both of these issues are discussed below.

3.5.2 Standard error as a stopping condition

The potential to achieve an accurate proficiency level estimate with less testing time and fewer questions is a theoretically appealing characteristic of variable-length CATs. Hence, it was important to investigate whether or not this was a valid stopping condition in a real educational setting. To this end, an empirical study in such a setting was designed and conducted. The method, summary of test-taker performance and findings regarding the study are reported below.

Method. In this study, 139 Level 2 Computer Science undergraduates took a test using the CAT software prototype developed for this research as part of their summative assessment for a second year programming module. The test took place in computer laboratories, under supervised conditions. The test comprised 30 questions, within a 40-minute time limit.

Summary of test-taker performance. The mean proficiency level was θ=0.066, SD = 1.081, N= 121.

Findings. Test-takers' proficiency level estimates were divided into three groups according to performance in the test, namely "low performing" (N=44), "average performing" (N=50) and "high performing" (N=45). Figure 3-3 below summarises the standard error for the estimate of proficiency level for the three different groups of participants.

The standard error of a random sample of test-takers – 15 low performing, 15 average performing and 15 high performing test-takers – was examined. It can be seen from Figure 3-3 that the standard error tends to decrease as the test progresses, and the estimate of the test-taker's proficiency level estimate becomes more accurate. This is an important finding, as it supports the view that the standard error for the proficiency level estimate is a valid stopping condition.

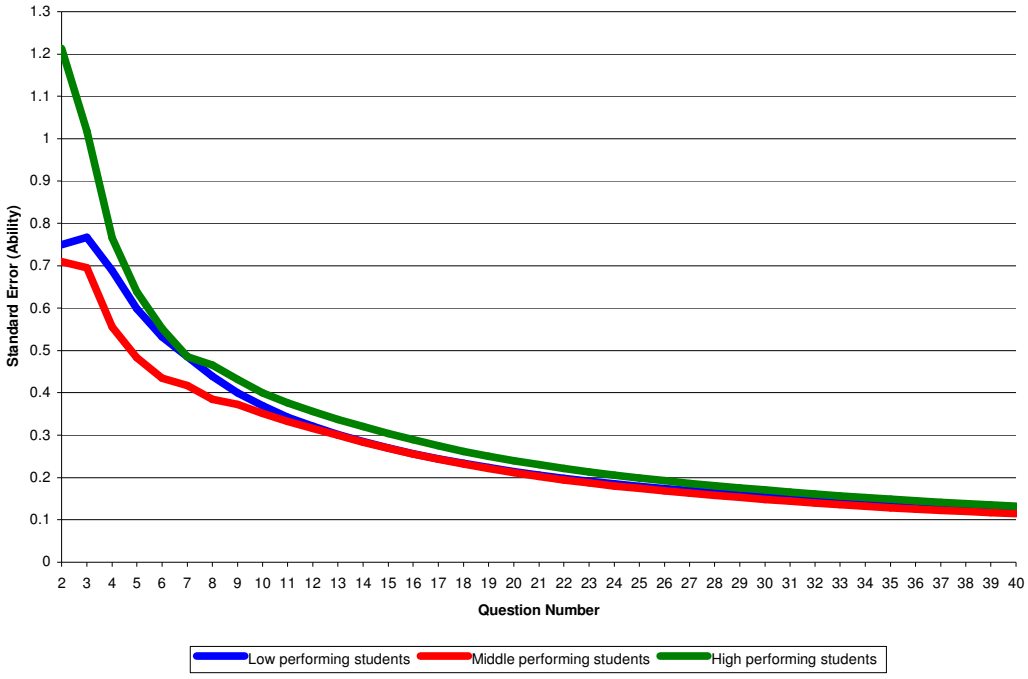

Figure 3-3: Standard Error for a random sample of test-takers

The results shown in Figure 3-3 also show that after 16 questions the standard error reached a value below 0.3, which could be an acceptable level for the

standard error for the estimate of proficiency level, for all three different groups. In practical terms, this means that the test could have stopped after 16 questions rather than continuing until question 40. In spite of the validity of the standard error for the proficiency level estimate as a stopping condition, there is the issue of syllabus coverage. This is due to the fact that stopping the test earlier might mean that fewer questions per topic are administered, and some of the required learning outcomes might not be assessed.

Similar findings regarding the use of standard error for the test-taker's proficiency level estimate as a CAT stopping condition were reported as part of this research, see Lilley et al. (2002c) and Lilley et al. (2004a).

3.5.3 Test-taker attitude

As it can be seen from Figure 3-3, a variable-length CAT might lead to different test lengths based on test-taker's performance during the test. This could, for example, result in some test-takers receiving lower scores than other test-takers who took longer tests (McBride, 2001a). One can speculate that, in a real educational setting, such a scenario would not be well received by those test-takers with lower scores. Hence, it was important to investigate test-takers' attitude towards different stopping conditions. A focus group session was carried out in order to gain a deeper understanding of test-takers' attitude towards different stopping conditions. The method, summary of test-taker performance and main findings from the session are presented below.

Method. In this study, 27 international students took a test on the use of English language and grammar, using the CAT software prototype developed for this research. The test comprised 20 questions, administered in 2 sessions, one of 10 dynamically selected CAT questions and the other of 10 static CBT questions. The total time limit was 40 minutes. The order in which the questions groupings were presented to each participant (i.e. CBT section followed by CAT or vice-versa) was randomly selected. The participants were unaware of the presentation order.

Summary of test-taker performance. The mean proficiency level was θ=0.020, SD = 1.020, N = 27.

Focus group information. Twelve volunteers took part in a focus group study immediately after undertaking the CAT test described above. The focus group guidelines can be found in Appendix B.

The focus group was guided by a moderator experienced in the area of Human-Computer Interaction (HCI) and lasted 40 minutes. The session was recorded on video to facilitate later analysis, with the agreement of the participants. The main purpose of this focus group session was to investigate usability issues related to the user interface and participants' attitude towards the use of CATs in summative and formative assessments. Immediately prior to the focus group session, the participants were provided with a paper copy of the questions that comprised the CBT element of the test and were briefed on how both the adaptive and non-adaptive elements of the test worked.

Participants were then given a standard introduction to the session, including establishment of guidelines for the session and issues of confidentiality.

After the introduction, each participant gave a brief summary of their attitude to the test they had just undertaken. This was followed by a general discussion led by the moderator.

Findings. The fairness of different stopping conditions for a CAT was a topic that generated a substantial amount of discussion within the group.

The stopping condition used in this study was a combination of number of questions administered or reaching the time limit, whichever happened first. A possible further stopping condition that was discussed was to stop the test after the standard error for the estimate of the test-taker's ability reached a predefined level. In so doing, the efficiency of the assessment process could be improved, as the time required to assess each test-taker could be reduced. For example, the stipulated standard error could be achieved after 15 questions for one test-taker and 19 questions for another.

Although the participants seemed to have understood the underlying principle of the standard error as a stopping condition, they did not fully approve the concept within a summative assessment scenario. The participants in the focus group were concerned that the standard error as stopping rule might prevent participants who started the test poorly from improving their performance and thereby achieving a better grade. Nevertheless, they seemed happy to have the test stopped if their performance was consistently high, as this would result in a good grade.

In a formative assessment scenario, however, participants indicated that they would be more likely to accept the standard error as a stopping condition for a CAT. The main reason for this was the fact that this would be a more efficient method of assessment than a traditional CBT.

In summary, participants considered a combination of number of questions and time limit to be the most suitable stopping condition. As for the use of standard error as a stopping condition in a summative assessment scenario, this should only be applied for those test-takers who performed well; lower scoring test-takers should be allowed to continue the test until a time or question limit is reached. In a formative assessment scenario, however, the use of standard error as a stop condition was more likely to be accepted, as this had the potential to lead to faster feedback.

An important issue uncovered during the focus session was participants' views about the inability to return to previously answered questions in a CAT. This is discussed in the next section of the thesis. Findings from the focus group session regarding test-takers' attitude towards the CAT approach in general will be reported later in section 4.2. Findings from the focus group session were also reported in Lilley et al. (2004a).

3.6 Reviewing previously entered responses

A practical concern in the design and implementation of a CAT application is whether or not test-takers should be permitted to review and modify previously

entered responses. Wainer (2000b) suggests that in most forms of assessment, test-takers are advised to use their time constructively. This often means that, once test-takers have completed the exam or test, they are frequently advised to use the time remaining to return to previous questions and check over their answers.

In a CAT, however, it is often assumed that test-takers should not be allowed to return to previous questions (Thissen & Mislevy, 2000; Wainer, 2000b; Vicino & Moreno, 2001; Guzmán & Conejo, 2004). This assumption arises from the potential to obtain artificially inflated scores, reduced testing efficiency and added complexity to the item selection algorithm. Such assumptions are summarised below.

Allowing item review could lead to artificially inflated scores (Vispoel, 1998; Olea et al., 2000). Vispoel et al. (2000) and Olea et al. (2000) cite what is known as the Wainer strategy. In such a strategy, test-takers would intentionally answer all questions incorrectly first. This would lead to less difficult questions being administered. Upon review, test-takers would answer all questions correctly and this would lead test-takers to answer a higher number of questions correctly than they would naturally. This could lead to artificially inflated scores, as ability estimates are based not only on the level of difficulty of the questions but also on the total number of questions answered correctly. Olea et al. (2000) discuss what is known as the Kingsbury strategy, in which test-takers evaluate whether or not the following question is harder than the previous one, and based on this evaluation they deduce whether or not the previous response was incorrect. This would, in turn, allow the test-taker to keep modifying their responses until the following question was a more difficult one (Olea et al. 2000). As one would expect, the Wainer and Kingsbury strategies will have different effects on test performance depending on issues such as stopping conditions and the algorithm employed for calculating ability estimates (Vispoel, 1998; Vispoel et al. 1999; Olea et al. 2000 and Vispoel et al., 2000).

Both the Wainer and Kingsbury strategies are somewhat risky for the test-taker, as very specific sequences of events are required for an increased score. In addition, both strategies assume that the test-taker has a profound knowledge within the subject domain as well as deep understanding of how the adaptive algorithm works. It is perhaps not surprising that there is no compelling evidence of the use of either strategy in a real testing setting.

Allowing item review could lead to increased testing times and consequent reduction in assessment efficiency (Vispoel, 1998; Olea et al., 2000). This issue is of greater relevance in a variable-length CAT setting, where the stopping condition is based on the standard error for the proficiency level estimate. This is because it might take longer to achieve the target measurement for the standard error for the proficiency level estimate if test-takers are allowed to review and change their responses at any given time. The impact of such an assumption in a fixed-length CAT setting can be controlled by, for example, limiting the amount of time available for item review.

Allowing item review could add complexity to item administration algorithms (Vispoel, 1998). This is true not only for item administration algorithms, but also for the algorithm responsible for estimating the test-takers' proficiency level. The complexity arises from the necessity take into account both sets of responses, namely before and after review.

Despite the common assumption that CATs should not support the review and change of previously entered responses, participants in a study conducted by Vicino & Moreno (2001) reported that the inability to go back to previous questions was perceived as a disadvantage of the CAT approach by test-takers. Lunz et al. (1992), Vispoel et al. (2000), Revuelta et al. (2000) and others also have argued that the inability to review and modify previously entered responses could lead to increase in test-taker anxiety levels and perceived loss of control over the application. Test-takers who used initial versions of the CAT prototype developed for the research where question review was disallowed, also reported their preference towards a CAT test

where question review was permitted. Moreover, it is argued that allowing participants to return to previous questions would offer greater resemblance to real educational settings, as in oral exams and paper-and-pencil tests test-takers are usually permitted to rectify previous answers.

In order to investigate the effect of item review on proficiency level estimates, an empirical study was carried out. The study was performed in a real educational setting, and is described below.

Method. As part of their summative assessment for a second year programming module, a group of 205 Level 2 Computer Science undergraduates took a test using the CAT software prototype developed for this research. The test took place in computer laboratories, under supervised conditions. The test consisted of 30 questions, within a 40-minute time limit.

In this study, test-takers were allowed to return to previous responses immediately after all questions had been answered. The CAT prototype was modified in such a way that once the test was finished and the reviewing process completed, the test-taker's proficiency level was recalculated using the final set of individual responses. A further modification to the CAT prototype was the addition of functionality to record in a database whether or not the review function had been used. In the event of a test-taker changing a response, all changes were also stored into the database. Note that these were additional database entries rather than overwriting previous entries for the same question.

Summary of test-taker performance. Mean and standard deviations before and after review are shown in Table 3-4. As it was discussed in section 2.2.4, in a CAT the focus is not only on the number of questions answered correctly by each individual test-taker, but on the level of difficulty of such questions.

CAT Performance indicator	Mean Before review	Mean After review
Proficiency level	-1.10	1.06
% Correct responses	54.75	56.06

Table 3-4: Summary of test-taker performance (N=205)

Findings. A One-Way Analysis of Variance (ANOVA) was performed on the data summarised in Table 3-4 to examine any significance of differences in the mean scores obtained by the test-takers. One-way ANOVA is a parametric technique that is appropriate in comparing means between one or more groups when the sample size is relatively large. Table 3-5 shows that there were no significant differences in the mean scores before and after review.

CAT Performance indicator	F	Sig.
Proficiency level	0.100	0.376
% Correct responses	1.405	0.118

Table 3-5: ANOVA results (N=205)

Table 3-6 illustrates the mean percentage of changed responses for this group of test-takers.

% Changed responses	Mean
% Changed responses (overall)	7.75
% Changed responses from right to wrong	2.01
% Changed responses from wrong to right	3.31
% Changed responses from wrong to wrong	2.42

Table 3-6: Test-takers' usage of review (N=205)

Although all test-takers in this study used the option to view previously entered responses, it can be seen from Table 3-6 that test-takers did not extensively use the option to change previously entered responses. In fact, whilst 79% of

the test-takers changed at least one response, the mean of changed responses was 7.75%. This is in line with the work by Vispoel et. al. (2000), in which it was found that a considerable proportion of test-takers changed their responses to at least one item, although the overall percentage of questions changed was small.

The data presented in Table 3-4, Table 3-5 and Table 3-6 relate to the whole group of test-takers (N=205). Table 3-7 illustrates the effect of test-takers' usage of review on proficiency level estimate, only for those test-takers who changed at least one response. It can be seen from Table 3-7 that test-takers who increased or maintained their proficiency level estimates after question review outnumbered those who had their results lowered (see Table 3-7). These findings are in line with those reported by Lunz et al. (1992), Vispoel et. al. (2000), and Revuelta et al. (2000).

Performance Indicator	Lower	Same	Higher
Proficiency Level	40	38	78
% Correct responses	54	11	91

Table 3-7: Test-takers' usage of review (N=156)

It was important to investigate if there were statistically significant differences between test-takers at different levels of ability. To this end, test-takers were divided into three groups, namely "low performing" (N=53), "average performing" (N=54) and "high performing" (N=49). The results shown in Table 3-8 illustrate an interesting finding, as these suggest that test-takers who performed less well in the test were less likely to benefit from the review option. In addition, the results shown in Table 3-8 weaken the argument that test-takers might employ the Wainer strategy (Vispoel et al., 2000; and Olea et al., 2000) to artificially inflate their scores. The Kingsbury strategy (Olea et al. 2000) is not of relevance here, as test-takers were only permitted to change their responses once all questions were administered.

% Changed responses	Low performing test-takers (N=53)	Average performing test-takers (N=54)	High performing test-takers (N=49)
	Mean	Mean	Mean
% Changed responses (overall)	11.70	10.62	8.10
% Changed responses from **right** to **wrong**	3.58	1.85	2.52
% Changed responses from **wrong** to **right**	3.90	5.31	3.81
% Changed responses from **wrong** to **wrong**	4.21	3.46	1.77

Table 3-8: Summary of review usage according to performance on the test (N=156)

Table 3-9 and Table 3-10 summarise test-taker performance before and after review only for those participants who changed at least one response. An Analysis of Variance (ANOVA) was performed on the data summarised in Table 3-9 and Table 3-10 to examine any significance of differences in the mean scores obtained for the three groups. The results of this statistical analysis are shown in Table 3-11 and Table 3-12.

Group	Proficiency Level Mean	Proficiency Level Mean	N
	Before review	After review	
Low performing test-takers	-2.29	-2.34	53
Average performing test-takers	-1.26	-1.12	54
High performing test-takers	0.32	0.36	49

Table 3-9: Proficiency level means according to performance on the test (N=156)

Group	% Correct responses Mean Before review	% Correct responses Mean After review	N
Low performing test-takers	46.42	46.49	53
Average performing test-takers	53.93	57.11	54
High performing test-takers	64.49	66.37	49

Table 3-10: Percentage of correct responses according to performance (N=156)

Table 3-11 and Table 3-12 show that the ability to modify previously entered responses did not lead to significant differences in the percentage of correct responses and/or proficiency level estimates for the low and high performing groups (*p>0.05*). Neither did it lead to statistically significant differences in the proficiency level estimate for the test-takers in the average group. The only significant difference between before and after review means found in this study concerns the percentage of correct responses for the adaptive section of the test for the average group.

Group	F	Sig.	N
Low performing test-takers	0.303	0.291	53
Average performing test-takers	5.925	0.085	54
High performing test-takers	0.070	0.396	49

Table 3-11: ANOVA results relating to the data summarised in Table 3-9 (N=156)

Group	F	Sig.	N
Low performing test-takers	0.001	0.487	53
Average performing test-takers	2.825	0.048	54
High performing test-takers	1.091	0.149	49

Table 3-12: ANOVA results relating to the data summarised in Table 3-10 (N=156)

The results presented here suggest floor and ceiling effects, where the amount of variability is reduced for those test-takers at the lower and higher levels of

performance. Hence, the review of previously entered responses did not have an effect on the final scores for test-takers in the low and high performing groups. Only those test-takers at the average range improved the percentage of correct responses significantly by reviewing their answers ($p<0.05$).

As part of the study into the effects of permitting test-takers to change previously entered responses, the standard error of 30 test-takers was examined. The random sample comprised 10 low performing, 10 average performing and 10 high performing test-takers who changed their responses. It can be seen from Figure 3-4 that, irrespective of performance, the standard error on their proficiency level estimates tended to decrease as the number of questions increased. This was taken to indicate that:

- the use of the review function had no negative impact on the accuracy of the proficiency level estimate;
- the level of difficulty of the tasks (i.e. questions) was appropriate for test-takers' proficiency levels.

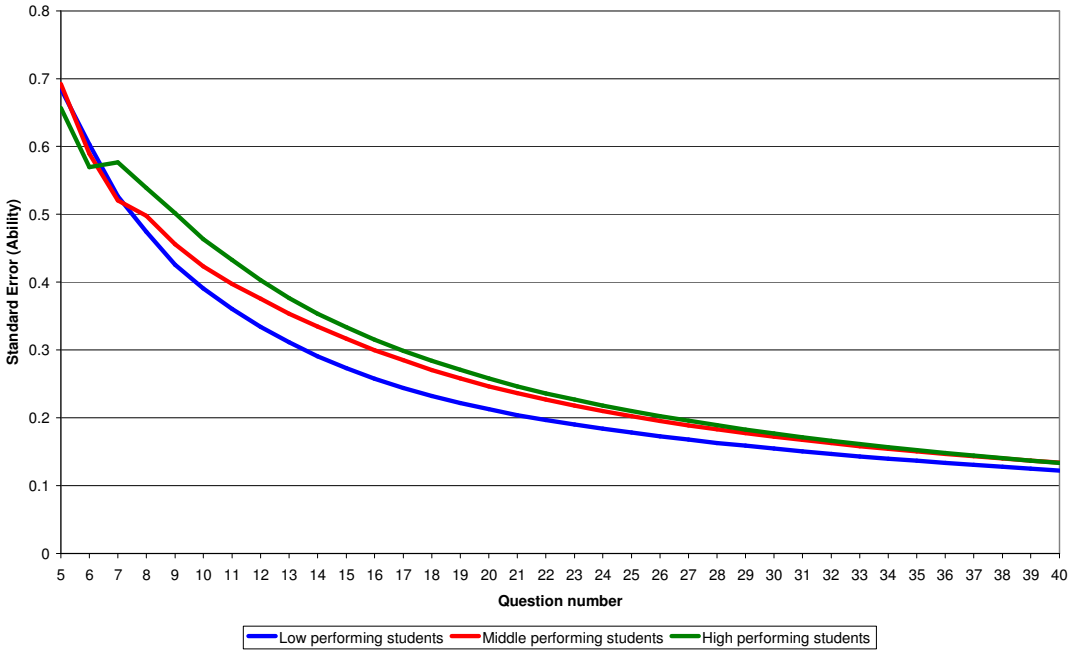

Figure 3-4: Standard Error for a random sample of test-takers who changed at least one response

The results reported here corroborate findings from previous studies (Vispoel et al., 2000; Olea et al., 2000) in that:

- test-takers only changed a small percentage of responses (7.75%);
- the mean performance values after review were higher than those before review;
- low performing test-takers were less likely to improve their scores after review;
- high performing test-takers were less likely to change responses (8.10%) than lower (11.70%) or average (10.62%) performing test-takers;
- despite most test-takers were able to increase their percentage of correct responses, such changes had little impact on their final proficiency level estimate.

In this study, test-takers seemed to amend responses that were incorrectly answered due to distraction or perhaps even assessment related anxiety. However, test-takers did not change answers from right to wrong to those questions that were above their proficiency levels. This is an important finding, as it supports the view that the CAT approach is effective at computing an estimate of a test-takers' proficiency level within a subject domain. It is also possible that test-takers managed to change answers from wrong to right by inferring "clues" from other questions or by being able to spend more time on those questions that they were less certain about. Both scenarios, however, could also occur in other forms of assessment such as written examinations.

The results reported here support the view that test-takers should be permitted to return to previously entered responses in a CAT. The option to return to previous questions seemed to have no adverse effect on proficiency level estimates and contribute towards a reduction in test-takers' anxiety. Test-takers expect to be provided with an opportunity to return to and change previously entered responses as this would hold greater resemblance to other assessments in which they have previously participated.

In the context of the research reported here, the effect of question review was also the focus of Lilley et al. (2003a), Lilley & Barker (2004) and Lilley & Barker (2005b).

3.7 Summary

A computer-adaptive test (CAT) application cannot exist without a calibrated item (i.e. question) database. The calibration of items in its conventional form, however, often proves to be an onerous – and on occasions too onerous – process for smaller applications of the CAT approach due to the need of pre-test studies involving large numbers of items and test-takers (see for example Huang, 1996). To overcome the problems related to the initial need of large groups of test-takers, a combination of expert calibration and MML parameter estimation methods were employed as part of this research. This combination approach proved to be useful to this research.

In addition to a calibrated item database, a CAT testing algorithm is required. The testing algorithm of a CAT can be broadly described as "start", "select the item to be administered next" and "stop". As part of the testing algorithm, examiners are also expected to determine whether or not test-takers should be permitted to change previously entered responses.

A CAT test can start with a question of difficulty based on prior information about the test-taker, a random question or a random question of middle difficulty. As prior information about test-takers is often unavailable and a completely random question could lead to a start question at either end of the difficulty scale, the option to start the test with a question of middle difficulty was employed in the work reported here. This choice is supported by previous research, such as the work reported by Wolfe et al. (2001).

In order to select the item to be administered next, there is more than one method that can be employed to select the most informative item for a test-taker's most recent proficiency level estimate. Due to its simplicity, the "difficulty-based" method (Guzmán et al., 2005) was chosen for this research.

It is important to note that item selection methods do not generally take into account parameters other than the ones that are part of the 3-PL model (Lord, 1980), namely difficulty, discrimination and pseudo-chance (see section 2.2.4). As part of this research, other factors were considered to be of relevance when selecting the question to be administered next, namely content balancing and item exposure.

The stop condition of a CAT can be based on a single stopping rule, or a combination of stopping rules. Examples of stopping rules include: a certain number of questions has been administered, a fixed time has elapsed or a predefined standard error for the proficiency level estimate has been attained. Work conducted as part of this research has shown that standard error for the proficiency level estimate is a valid CAT stop condition. However, a focus group study conducted as part of this research (Lilley et al., 2004a) suggests that, in a summative assessment setting, test-takers have a preference for CATs that do not employ a predefined standard error for the proficiency level estimate as a stop condition. This is because such a stop condition is likely to lead to variable-length tests and this could, in turn, cause some test-takers to question the fairness of the CAT approach. Findings from the focus group also suggest that CATs of variable lengths are more likely to be well-received in a formative assessment setting, as this could lead to faster feedback. There is also the issue of syllabus coverage, as a shorter test might mean that not all expected learning outcomes were covered.

The impact of allowing test-takers to change previously entered responses on proficiency level estimates was also investigated as part of this work (Lilley & Barker, 2004; Lilley & Barker, 2005b). Findings from this work corroborate published research by Vispoel (1998), Vispoel et al. (1999), Olea et al. (2000) and Vispoel et al. (2000), in which it was reported that only a small percentage of answers are changed, that most test-takers who changed their responses increased their proficiency level estimates after review and that review was more advantageous to test-takers at higher proficiency levels.

Stopping rules and permitting test-takers to change previously entered responses present an interesting dilemma to testers in an educational context. On the one hand, CATs of variable-length where changing previously entered responses is disallowed are likely to be more efficient, without jeopardising the accuracy of proficiency level estimates. On the other hand, test-takers' attitude towards CATs of variable-length where changing previously entered responses is disallowed is less favourable than towards CATs of fixed-length where changing previously entered responses is permitted.

In this work, it was assumed that test-takers' satisfaction and engagement was more important than test efficiency and, for this reason, a CAT of fixed-length where changing answers is permitted was found to be the most suitable combination in a real educational context.

Up to this point, the research focused on the design and implementation of the CAT software prototype. Two major groups of users for the application were identified:

- students, in their capacity as test-takers;
- academic staff, in their capacity as assessors.

The next stage of the research was concerned with the evaluation of the CAT software prototype by test-takers.

4. Test-taker evaluation of the CAT approach

The previous chapter was concerned with the establishment of testing conditions within the computer-adaptive test (CAT) approach. The next stage of the research was concerned with the evaluation of the CAT approach by test-takers, and this is the focus of this chapter. In order to perform the test-taker evaluation of the CAT approach, three user studies and one focus group session were carried out, and these are described in this chapter.

The chapter is organised into four main sections. The first describes the first user study, which was concerned with test-takers' perceived level of difficulty of a CAT within the domain of English as a second language. In addition, this first study included an observation study, where test-takers were observed whilst taking a test using the CAT software prototype developed for this research. The aim of the observation study was to uncover any usability issues that might affect test-taker performance. The second section is concerned with the focus group session. The aim of the focus group was also twofold: to examine usability issues related to the user interface of the CAT software prototype developed for this research, and to investigate test-takers' attitude towards the use of CATs as an assessment tool. The first user evaluation study and focus group session generated three papers: "The Development and Evaluation of a Computer-Adaptive Testing Application for English Language" (Lilley & Barker, 2002), "How computers can adapt to

knowledge: A comparison of computer-based and computer-adaptive testing" (Lilley et al., 2002c) and "The development and evaluation of a software prototype for computer adaptive testing" (Lilley et al., 2004a).

The third section introduces two empirical studies where test-takers' perceived level of difficulty of a CAT within the Computer Science subject domain was examined. These studies were also reported in the following two papers: "Learners' perceived level of difficulty of a computer-adaptive test: A case study" (Lilley et al., 2005c) and "Student attitude to adaptive testing" (Lilley & Barker, 2006b).

The fourth section is concerned with the changes made to the CAT application, in the light of the information gathered from the studies reported in this chapter.

In the next section, the first user study is described.

4.1 First user study

The first user evaluation involved 27 international students, who were studying English as a second language. The method, summary of test-taker performance and findings are presented next.

Method. The 27 participants took a test on the use of English language and grammar, using the CAT software prototype developed for this research. The test was carried out under supervised conditions, within a 40-minute time limit. As it can be seen from Figure 4-1, the test was organised into 2 parts: one of 10 dynamically selected CAT questions and the other of 10 static computer-based test (CBT) questions. The order in which the questions were presented was randomly selected, and participants were unaware of the presentation order.

Furthermore, participants were observed by two members of the research team whilst interacting with the CAT software prototype as part of an observation study. The observers took notes during the session for later analysis.

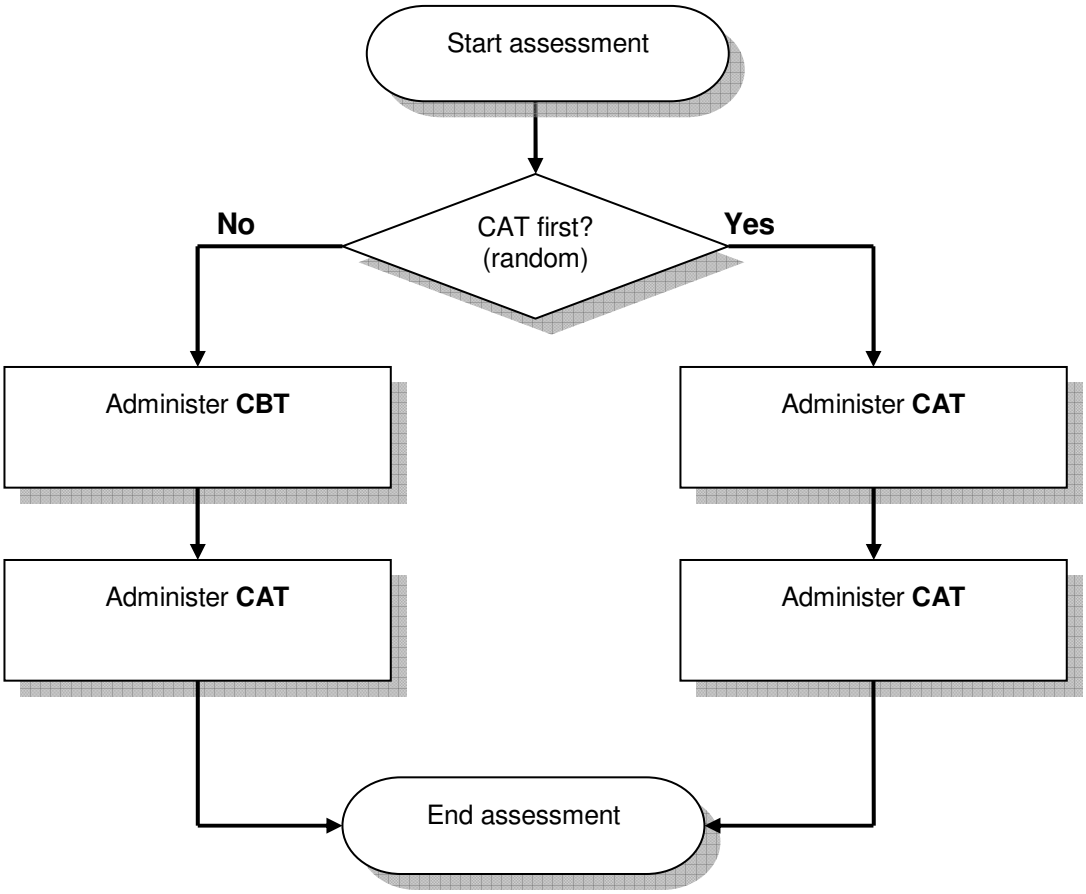

Figure 4-1: Configuration of the application used in the first user evaluation

Summary of test-taker performance. Mean and standard deviations for the CBT and CAT components of the test are presented in Table 4-1. In Table 4-1, the minimum attainable CAT score was -2 and the maximum was +2. As for the CBT section, the lowest score possible was 0 and the highest was 100.

Test type	Mean	Std. Dev.
Computer-Adaptive Test (CAT) proficiency level	0.02	1.02
Computer-Based Test (CBT)	63%	12.97

Table 4-1: Summary of test-taker performance (N=27)

A paired samples t-test performed on the data in Table 4-1 showed that there was no significant difference in performance by test-takers in CAT and CBT sections of the test, such as t=-0.71, df=26 and P(two tailed) =0.48.

Observation study findings. The two observers were knowledgeable about how the CAT application worked, and were able to provide the test-takers with help as to how the application worked on request. The observation guidelines used by the observers can be found in Appendix C. No test-taker requested help. At the end of the CAT session, the two observers analysed and compared their notes. The observers concurred in that no usability issues were uncovered during the observation study. The absence of test-takers' request for help in addition to the observers' notes were taken to indicate that the application was easy to use, and unlikely to affect test-takers' performance in an adverse way.

Electronic questionnaire findings. The first user evaluation involved the collection of data using an electronic questionnaire, and the format of the questions included in the questionnaire are shown in Figure 4-1.

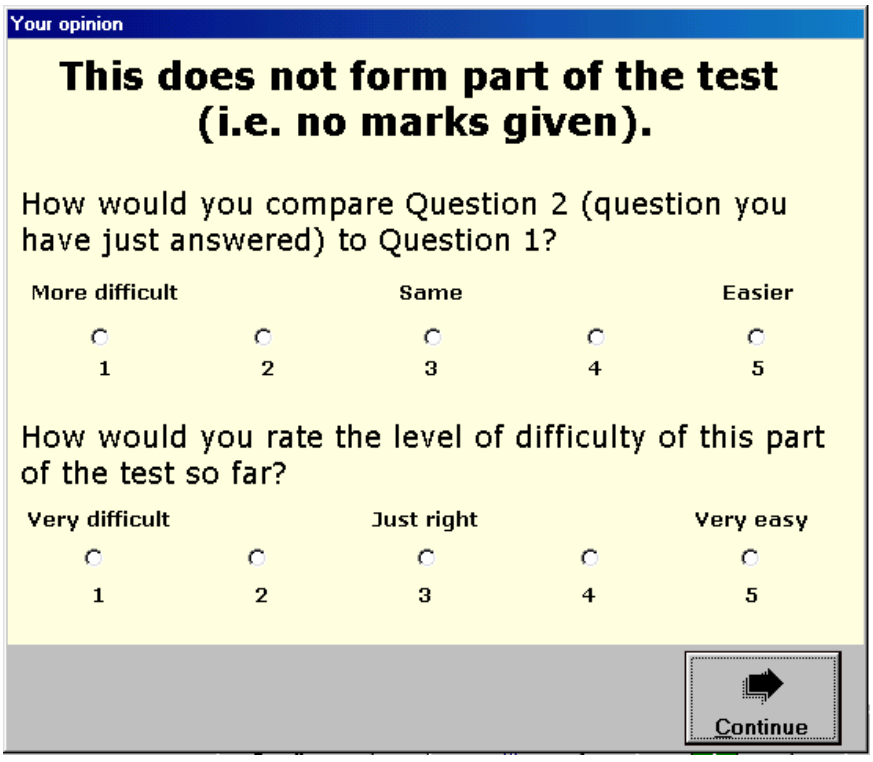

Figure 4-2: Electronic questionnaire screenshot

A summary of the data collected via the electronic questionnaire is presented in Table 4-2. For each question answered, all participants were asked to grade the level of difficulty of both the question they had just answered and the test up to that point, from 1 (more difficult) to 5 (easier). This electronic questionnaire was aimed at gathering information on how the participants perceived the level of difficulty of the questions. For the CAT section of the test, the mean values for level of difficulty of question and test were respectively 2.9 and 3.1, while for the CBT one these values were 2.5 and 2.7 within the 1 (more difficult) to 5 (easier) Likert scale. The findings from this study suggest that the test-takers perceived that within the CAT section the questions administered matched their level of ability.

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

Table 4-2: Electronic questionnaire results (N=27)

A Pearson's Product Moment correlation was also performed on the CBT score and CAT proficiency level obtained by the participants, such as $r=0.398$, $p<0.001$, $N=27$. The results obtained from this statistical analysis showed that there was a good correlation between the CAT proficiency level and the CBT score ($p<0.001$). These results showed that those test-takers who performed well on the CBT section also performed well on the CAT. As part of the work reported here, this was interpreted as showing that test-takers participants were not disadvantaged by the use of a CAT as compared to the CBT assessment.

In order to obtain qualitative data regarding the attitude of participants to the CAT approach, 12 randomly selected test-takers from the original group participated in a focus group session immediately after taking the CAT. The findings from the focus group are described next.

4.2 Test-taker attitude

In order to gather qualitative information about complex or sensitive issues regarding test-takers' attitude towards the CAT approach that were possibly overlooked when employing quantitative methods such as the electronic questionnaire described in the previous section, a focus group session was carried out. Moderating a focus group is a challenging task, and it was important to ensure that the moderator was experienced in dealing with group dynamics as well as individual differences among participants.

Findings from the focus group study regarding test-takers' attitude towards the CAT approach are presented next. The method employed in the focus group study has already been described in section 3.5.3, and for ease of reading this information is repeated below. The focus group guidelines can be found in Appendix B.

Method. The focus group was moderated by an expert in Human-Computer Interaction (HCI). The duration of the focus group session was 40 minutes. In order to facilitate later analysis, the session was recorded on video with the permission of the 12 participants. Participants were provided with a standard introduction to the session, where guidelines and issues of confidentiality were explained. As part of the introduction to the focus group session, participants were also given a copy of the 10 questions that were administered as part of the CBT part of the test. In addition, participants were briefed on how both the adaptive and non-adaptive elements of the test worked.

After the introduction, each participant gave a short summary of their attitude to the test they had just undertaken. This was followed by a general discussion led by the moderator.

Focus group findings. The findings from the focus group session are divided into 2 parts. The first part is concerned with usability issues related to the user interface. The second part focuses on test-takers' attitude towards the use of CATs as an assessment tool.

Usability issues. During the session, participants reported that they found the CAT high-fidelity prototype developed for this research easy to use, even without prior training. Participants perceived the user interface as being usable and easy to understand. Moreover, they reported that the user interface was unlikely to have adversely affected their performance during the test.

Test-taker attitude towards different assessment methods, including the CAT approach. Participants said that they considered the concept of a computer-adaptive test interesting. Indeed, many of them were already familiar with the concept, as they have previously heard of or encountered adaptive tests such as Test of English as a Foreign Language (TOEFL) (Glas et al., 2003) and Graduate Management Admission Test (GMAT) (Guo et al., 2006). Participants did not demonstrate any major concerns about the fact that a typical CAT would provide test-takers with a proficiency level rather than a raw score. They were then asked to expand on their views about the scoring method used within CATs.

Participants did not seem concerned about the scoring method itself as used within a CAT, nor did they seem overly concerned about the fact that test-takers are presented with different sets of questions. Interestingly, they indicated that it is reasonable to expect that answering a more difficult question correctly should score more marks than answering an easier question correctly or, in other words, that questions are weighted according to difficulty. Although participants considered the scoring method provided by a CAT to be fair, some members of the group expressed their concern about the fact that within a CAT test-takers are not allowed to go back and review their answers once they had been submitted. Review and modification of previous responses was not

permitted in this study. Issues regarding the ability to review and modify previously entered responses were discussed in section 3.6.

When prompted about what they thought of the level of difficulty of both tests they took on that day (i.e. CBT and CAT) and tests in general, participants said that the level of difficulty of the CAT questions was more likely to be "just right" or appropriate than for those questions in the CBT part of the test. All participants indicated that the CAT component of the test provided a more consistent assessment than the CBT component, which started too easy for many, and ended too difficult for all but one participant. This evidence corroborates the data collected through the electronically questionnaire, as summarised in Table 4-2. Moreover, the results shown in Table 4-2 were interpreted as an indication of the effectiveness of the adaptive algorithm implemented within the software application.

According to participants, the CBT part of the test was at some points very easy and at others, very difficult. This characteristic of CBTs was perceived as a weakness within this assessment method by the participants, who expected a well-designed test to consistently offer an appropriate level of challenge. Tests that were too easy were described as being "meaningless"; likewise, tests that were too difficult were reported as "frustrating". One participant suggested this was likely to lead to guessing, as he would not be able to base his responses on knowledge or reasoning ability.

Furthermore, it was suggested by several participants that a test tailored for the ability of an individual test-taker was valid and more likely to improve their enthusiasm and motivation during the assessment session than those that are static. This view would support Wainer's (2000a) perspective that a test in which the level of difficulty of the questions provides an appropriate level of challenge for each individual test-taker should lead to increased test-taker motivation. In a similar vein, Boyle (1997) states that an educational software application that comprises adaptive elements, such as tailoring the selection and presentation of the interface's contents for each individual user, is more likely to improve user motivation.

A further aspect related to administering different sets of questions for each individual test-taker is the potential to reduce unauthorised collaboration amongst test-takers during the assessment session. Although one enthusiastic participant suggested that there was "always a way around any safeguard", participants said that it would become more difficult to "cheat" during a session of assessment if the set of questions is not the same for all participants. In summary, the dynamic selection of questions was perceived by the participants as being capable of increasing test security, as the opportunities for test-takers to copy one another's responses are reduced.

The CAT high-fidelity prototype developed for this research was based on the use of objective questions, such as multiple-choice and multiple-response questions. Participants reported that CATs based on the use of objective questions are a fair assessment method. Despite the perceived fairness of objective questions, some participants indicated that they favoured coursework over examinations and tests, as in the former they have more time to prepare and review the work to be marked. Several participants reported that international students are likely to benefit more from coursework rather than timed assessments. The reason for this is additional time for preparation, and more time to consider their use of the English language. In addition, some participants reported that coursework has the potential to offer students an opportunity to demonstrate "everything" that they know, rather than simply whether or not they know a single answer. Participants also pointed out that although students have the opportunity to present their knowledge in a higher level of detail in a coursework, they indicated that this type of assessment was "slower" than a test.

When prompted to explore these issues, only a few participants agreed that examinations were a better method of assessment than tests. When further prompted, whether or not they would favour tests based on the use of objective questions over written exams, some participants suggested they preferred objective questions to examination type questions.

In summary, participants indicated that each assessment method has positive and negative aspects and a balance amongst written exams, tests and coursework is the most appropriate approach for summative assessments. As for formative assessments, they suggested that a combination of tests and coursework would be the most suitable option. Participants also reported that the CAT approach is likely to be favourably received by test-takers, when combined with other assessment methods.

Overall, the focus group participants exhibited a positive attitude towards the CAT approach. The first user study and the focus group session, however, were conducted with a group of students who had volunteered to take part. This posed interesting questions to the research team, as to whether similar results will be found in a scenario where test-takers were actually being assessed. The focus of the next section is the level of difficulty of a test based on the CAT approach, as perceived by the test-takers.

4.3 Perceived level of difficulty

Lord (1980) suggests that one of the advantages of the CAT approach is the matching of the difficulty of the items (i.e. questions) administered during a test to the proficiency level of individual test-takers. In order to examine whether the questions selected by the CAT software prototype are an appropriate match for a test-taker's ability, two empirical studies to investigate test-takers' perceived level of difficulty of a CAT were conducted, and are reported in this section. The first study is concerned with test-takers' perceived level of difficulty in a summative assessment. The second study also examines test-takers' perceived level of difficulty, but in a formative assessment setting.

4.3.1 Summative assessment

The study described in this section is concerned with the perceived level of difficulty of the CAT in a summative assessment setting, in a real educational

context. The method, summary of test-taker performance and findings from this study are reported next.

Method. A group of 113 Level 2 Computer Science undergraduates participated in a summative assessment session using the CAT application developed for this research. 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. At the end of the assessment session, all test-takers were asked to rate the difficulty of the test that they had just taken. A copy of the questionnaire used in this study can be found in Appendix E. Also after the test a group of five randomly selected test-takers participated in a short interview regarding the test. The guidelines for the interview can be found in Appendix D.

Summary of test-taker performance. Participants' performance on this assessment is summarised in Table 4-3. In Table 4-3, the value for the proficiency level ranged from -3 (lowest) to +3 (highest). As it was pointed out in section 2.2.4, in a CAT examiners are not concerned only with the number of correct responses. Indeed most test-takers one can argue that test-takers 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 not only on the number of questions answered correctly by each individual test-taker, but on the level of difficulty of such questions.

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

Table 4-3: Summary of test-taker performance (N=113)

As can be seen from Table 4-3, the mean for the CAT proficiency level was near zero (θ=0.08), and the proficiency level estimates were widely spread out (SD=1.07).

Findings. At the end of the assessment session, all test-takers 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). The test-takers' ratings, as summarised in Table 4-4, show that most test-takers found the level of difficulty of the test to be "just right" (N=72).

1	**2**	**3**	**4**	**5**
Very easy	Easy	Just right	Difficult	Very difficult
0	2	72	34	5

Table 4-4: Level of difficulty of the test as perceived by the participants (N=113)

After the test, five test-takers were randomly selected to participate in a short interview. The interviewer was a member of the research team. Due to the brevity of the interviews, capturing them on video or tape was not considered necessary. Instead, the interviewer took notes. During the interview, test-takers were asked whether the test was successful at assessing how much they have learned within the subject domain being tested. All interviewees agreed that the test was fair and they considered it an appropriate instrument for assessing their proficiency level within the subject domain being tested.

Furthermore, test-takers were asked to expand on the reasons why they rated the test in the way that they did. All interviewees rated the CAT as being "just right" using the Likert scale provided. In general, they reported the reasons for the choice were related to the fact that the CAT test was challenging and not "boring" as other tests that they have taken in the past. Interestingly, they reported that they liked to be assessed using the CAT application because they felt challenged rather than expected to answer "silly" test questions.

The interviewees were also asked to summarise their experiences using the CAT application. Overall, the interviewees were satisfied with the application. One interviewee mentioned that the mouse device on the computer that he had been assigned was not fully functional (i.e. intermittent failures) and that he had, at times, to use the keyboard instead. This was considered “really annoying” by the interviewee. Another interviewee suggested that the application should provide test-takers with information about the total number of questions to be answered and the number of questions answered so far, in addition to the time remaining. According to this interviewee, it would be helpful for test-takers to have this information available on the screen, as this would allow test-takers to pace themselves. This issue is further discussed in section 4.4.

Statistical Analysis. The correlation between participants' performance and their perceptions on the level of difficulty of the overall test were examined in order to identify whether or not this was statistically significant. Participants' results and their perception of test difficulty were subjected to a Spearman's rank order correlation. No statistically significant correlation was found between test-takers’ proficiency levels and the test's difficulty rating, such as rs=-0.1, 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. Mean ranks are shown in Table 4-5.

Group	N	Mean Rank
Low performing participants	38	58.96
Average performing participants	36	58.24
High performing participants	39	53.95

Table 4-5: Kruskal-Wallis mean rank results (N=113)

The Kruskal-Wallis test showed that there was no significant difference in the perceived level of difficulty that could be ascribed to the effect of test-takers' performance on the test. This is of particular importance, since one of the goals of the CAT prototype developed for this research was that test-takers

would be presented with tasks that are challenging and motivating, rather than tasks that are either too difficult and therefore bewildering, or too easy and thus uninteresting.

The results reported in this section are concerned with the application of the CAT approach in a real summative assessment setting. It is the experience of the research team that the motivation, strategy and preparation of test-takers in a summative assessment setting differ from that employed in a formative assessment one. In order to compare test-takers' attitude to the CAT approach in a formative assessment setting with a summative assessment one, a study was carried out and this is described next.

4.3.2 Formative assessment

The study described in this section is concerned with the perceived level of difficulty of the CAT in formative assessment setting. It was also considered to be of interest to report on the perceived level of difficulty of the CAT approach in a real summative assessment setting by the same group of test-takers. A copy of the questionnaire used in this study can be found in Appendix E. This study's method, summary of test-taker performance and findings are reported next.

Method. As part of their regular assessment for a programming module, a group of 76 Level 2 Computer Science undergraduates participated in two assessment sessions using the CAT software prototype developed for this research. The first assessment session was formative and therefore the scores obtained by the participants did not count towards their final grade. The second assessment session was summative. In both cases, participants had 40 minutes to answer 40 objective questions within the Visual Basic.NET subject domain.

Findings. Table 4-6 shows a summary of their assessment performance. In Table 4-6, the potential values for the proficiency level ranged from -3 (lowest) to +3 (highest).

Assessment	CAT proficiency level Mean	Std. Dev.
Formative	-0.03	1.02
Summative	0.21	1.42

Table 4-6: Summary of test-taker performance (N=76)

Test-takers' performance on the formative assessment (θ=-0.03) was slightly lower than that observed for the summative assessment (θ=0.21). In both tests, i.e. formative (SD=1.02) and summative (SD=1.42), the CAT proficiency levels were widely spread out.

At the end of each test test-takers were asked to rate the difficulty of the test that they have just taken from 1 (very easy) to 5 (very difficult). The test difficulty mean, as perceived by the test-takers, was 3.53 (SD=0.64, N=76) for the formative test and 3.46 (SD=0.59, N=76) for the summative one. Their ratings are illustrated in Table 4-7.

Assessment	1 Very easy	2 Easy	3 Just right	4 Difficult	5 Very difficult
Formative	0	2	36	34	4
Summative	0	2	39	33	2

Table 4-7: Perceived level of difficulty (N=76)

It was important to investigate whether or not the correlation between test-takers' performance and their perceptions on the level of difficulty of the overall test was statistically significant. To this end, test-takers' results and their perception of the test difficulty were subjected to Spearman's rank order correlations and Kruskal-Wallis tests. In addition, a paired-samples t-test was used to examine any significance of differences in their means between formative and summative assessment sessions.

Statistical Analysis: Formative assessment session. No statistically significant correlation was found between the test-takers' proficiency levels

and the test's difficulty rating (rs = -0.165, Sig. 2-tailed = 0.155, N = 76). The data gathered in this study was also subjected to a Kruskal-Wallis test (Chi-Square = 3.591, df = 2, Asymp. Sig. = 0.166). Mean ranks are presented in Table 4-8.

Group	N	Mean Rank
Low performing participants	25	44.54
Average performing participants	26	34.58
High performing participants	25	36.54

Table 4-8: Kruskal-Wallis test mean rank results: formative test (N=76)

The Kruskal-Wallis test showed that there was no significant difference in the perceived level of difficulty that could be attributed to the effect of test-takers' performance on the formative test.

Statistical Analysis: Summative assessment session. The findings for the summative assessment session were in line with those in the formative one. No statistically significant correlation was found between the test-takers' proficiency levels and the test's difficulty rating (rs = -0.025, Sig. 2-tailed = 0.829, N = 76). The Kruskal-Wallis test mean ranks are shown in Table 4-9 (Chi-Square = 4.336, df = 2, Asymp. Sig. = 0.114). The Kruskal-Wallis test showed that there was no significant difference in the perceived level of difficulty between the three groups that could be ascribed to the effect of test-takers' performance on the test.

Group	N	Mean Rank
Low performing participants	26	41.27
Average performing participants	24	31.65
High performing participants	26	42.06

Table 4-9: Kruskal-Wallis test mean rank results: summative test (N=76)

Comparisons between formative and summative assessment sessions. The absence of a statistically significant relationship between performance on the test and perceived test difficulty in both assessment settings (i.e. formative and summative) was an interesting finding. The perception of the difficulty of a test might be expected to relate in some way to performance. Although it is difficult to be certain of a reason for this finding, it is consistent with the view that the test generated using the CAT software prototype developed for this research was effective in establishing the appropriate level of difficulty for individual test-takers. This is of particular importance, since one of the goals of the CAT prototype was to provide individual test-takers with tasks that were engaging, rather than tasks that are uninteresting or frustrating. One can argue that establishing an appropriate level is necessary, though of course not sufficient, to achieve this objective.

A paired-samples t-test was used to examine any significant differences in the means for the perceived level of difficulty obtained for the two assessment sessions (i.e. formative and summative). No statistically significant difference was found (t = 0.799, df = 75, Sig. 2-tailed = 0.427). A paired-samples t-test was also used to examine any significant differences in the means for the proficiency level obtained for both assessment sessions. This test showed statistically significant differences between proficiency level means (t = 0-2.112, df= 75, Sig. 2-tailed = 0.038).

Whilst there were no statistically significant differences in the perceived level of difficulty means, there were statistically significant differences in the proficiency level means. The proficiency level mean for the summative test (θ=0.21, SD = 1.42, N = 76) was higher than the formative one (θ =-0.03, SD=1.02, N = 76). The fact that test-takers are more likely to revise for a summative test than for a formative one could explain the difference in performance. It is also possible that test-takers adopt different strategies and they are more meticulous when taking summative tests. Another possibility is that the formative test had a positive effect on test-takers' preparation for the summative test.

4.4 Changes to the CAT prototype

In the light of test-takers' reactions reported in sections 3.6 and 4.3, changes were made to the CAT software prototype developed for this research. Voluntary feedback from test-takers was also taken into account; a test-taker reported that the "Confirm Answer" button was "redundant", another cited the "Confirm Answer" button as an element on the user interface that had "no clear purpose".

Whilst Figure 4-3 illustrates the first iteration of the CAT software prototype, Figure 4-4 illustrates the most recent iteration.

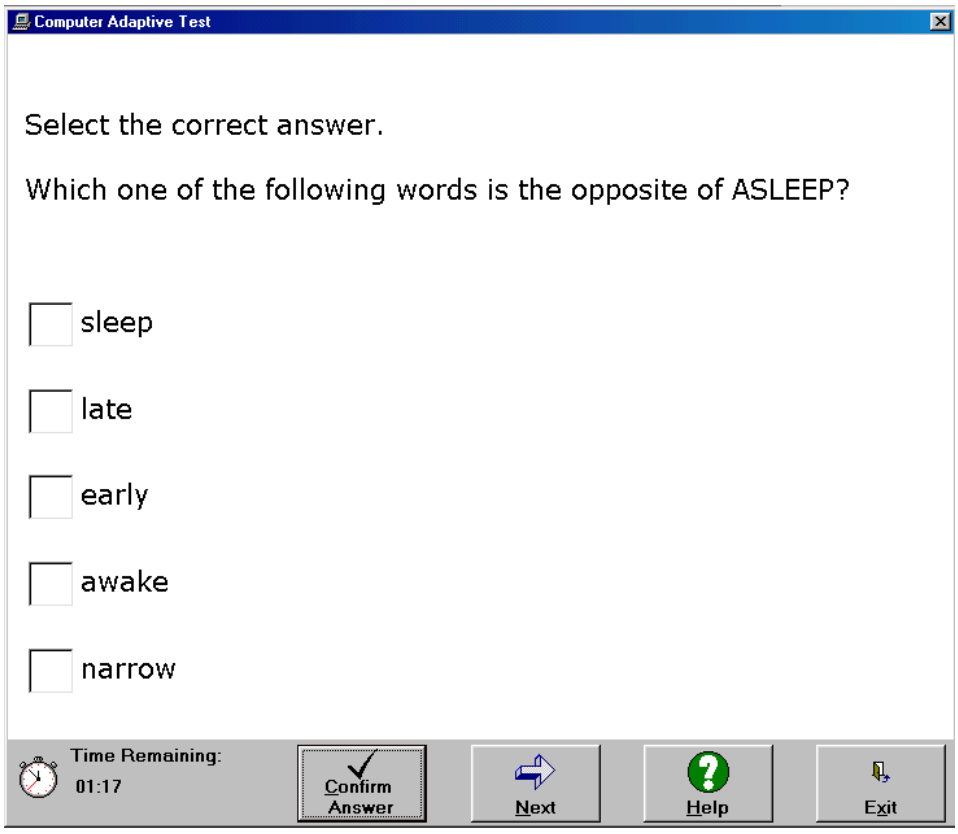

Figure 4-3: First iteration of the CAT software prototype

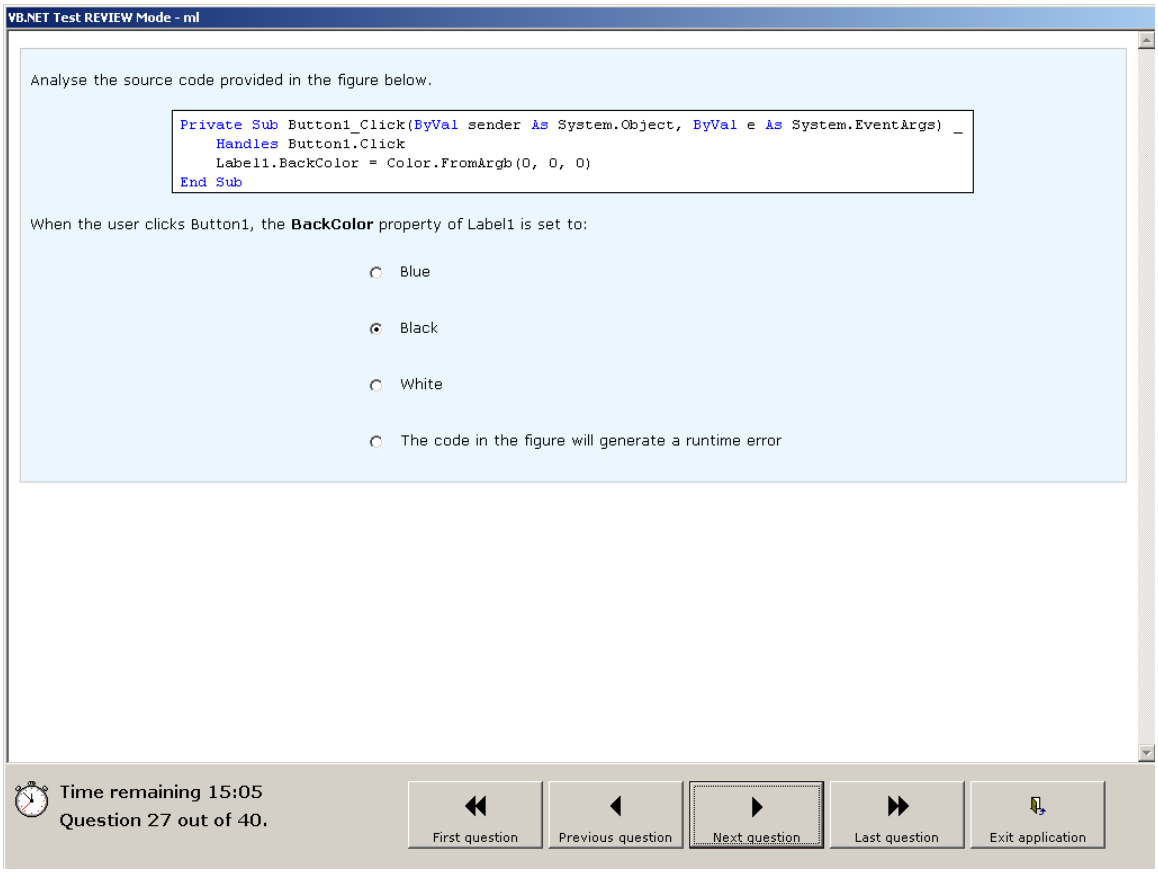

Figure 4-4: Most recent iteration of the CAT software prototype

The changes were made based on feedback from test-takers, and can be summarised as follows:

- the addition of “First question”, “Previous question”, “Next question” and “Last question” buttons to allow test-takers to review and modify previously entered responses. These 4 buttons are enabled once test-takers have answered all questions in the CAT. To change an answer, test-takers simply have to select their new answer. The need to review and modify previously entered responses was discussed in sections 3.6 and 4.2;
- the addition of a question counter on the bottom left-hand side of the screen, so test-takers can easily evaluate how many questions they have answered so far (see section 4.3);
- the removal of the “Confirm Answer” button. The original idea was that test-takers should click “Confirm Answer” and then “Next” in

order to proceed to the following question. Test-takers reported that they found the process tedious, as the application checked whether or not the correct number of options had been selected before proceeding to the next question. For example, in the case of a multiple response question where test-takers were expected to select 2 out of 4 options, test-takers were not permitted to proceed to the next question until 2 options had been selected. For this reason, it was the view of some test-takers who provided voluntary feedback that the "Confirm Answer" button was redundant. The research team agreed with this and the button was removed.

As can be seen from the list above, test-taker reactions to the CAT prototype resulted in a set of enhancements to the user interface. Although a series of mechanisms were in place in order to gather data about test-taker attitude towards the approach – i.e. focus groups, interviews, questionnaires – test-takers also provided voluntary (spontaneous) feedback by email. One test-taker sent the following email to the research team:

- "Would it be possible to issue us with the test questions and the correct answers?? I feel this would help us greatly for future revision as well as create a better understanding in any grey areas. Not only this, but it will also show us where we are going wrong and what areas need greater attention??? Thanks in advance,"

This type of feedback from test-takers led to specific suggestions as to how the CAT software prototype could be improved in order to more closely reflect the needs of test-takers. The issue of providing CAT test-takers with feedback that provides advice on individual development is discussed later in Chapter 7.

4.5 Summary

This chapter was concerned with test-takers' attitude towards the CAT approach. The aims of the work reported in this chapter were threefold:

- to examine usability issues related to the user interface of the CAT application developed for this research;
- to investigate test-takers' acceptance of the CAT approach as an assessment tool;
- to investigate test-takers' perceived level of test difficulty of a CAT.

In the case of usability issues, findings from the focus group session supported the findings from the observation evaluation study, in which it was reported that participants found the application easy to use, even without prior training. Similar findings were reported in subsequent empirical studies (Lilley et al., 2005c; Lilley & Barker, 2006b). Such findings were taken to indicate that the CAT application was easy to use and learn, and unlikely to affect test-takers' performance in an adverse way. To increase the application's usability, minor changes to the user interface were introduced based on the feedback provided by the application's main users (i.e. test-takers).

As part of the focus group session, test-takers' acceptance of the CAT approach was examined. Both the item selection and scoring methods were explored during the focus group session and, in general, test-takers presented no objection to the CAT format. A crucial product of the CAT approach is the generation, on the fly, of a test that is tailored to individual test-takers. In practical terms, this means that different test-takers will be presented with different sets of questions. Focus group participants did not appear to be concerned about this characteristic of the CAT approach. One can speculate that this is due to various reasons, such as:

- the increasing use of CATs in examinations such as TOEFL (Glas et al., 2003; Wainer & Eignor, 2000).
- the CAT scoring method takes into account the number of questions answered correctly and the level of difficulty of the questions, so test-takers presented with more difficult questions are rewarded for answering these correctly;

- it is also the case that in some traditional computer-based tests (CBTs), where random question selection is employed, test-takers are presented with different sets of questions (see for example Thelwall, 2000).

It should be noted that the participants in the focus group study relating to test-taker attitude reported in section 4.2 were students of the International Foundation Programme (IFP) at the University. Although the research focused on the use of adaptive testing for the assessment of Computer Science undergraduates, it was an assumption that participants of the focus groups were a fair representation of the student body as a whole. IFP students and Computer Science undergraduates alike identified some areas of concern relating to the use of the CAT software prototype for student assessment. The first was the inability to review and modify previously entered responses. This issue was investigated as part of this research, and the main findings from this investigation are reported in section 3.6. The second area of concern was related to different stopping conditions; this aspect of the research is discussed in section 3.5.

The CAT software prototype developed for this research was based on the use of objective questions, and it was important to examine test-takers' views about this question format. Focus group participants were familiar with the use of objective questions in Higher Education, and perceived such questions as being fair mostly due to the absence of bias when marking the question. With regards to the level of difficulty of the questions, participants concurred with the idea that objective questions at either end of the difficulty scale might introduce problems in the assessment process. Questions that are too difficult were thought to lead to guessing as test-takers would be unable to answer such questions based on their knowledge within the subject domain. On the other hand, questions that are too easy were described as "meaningless" or "silly". Participants agreed that, in principle, a test that is tailored to each individual test-taker could lead to increased test-taker motivation.

Interestingly, participants acknowledged that each assessment method has its own set of advantages and disadvantages. Objective tests, for example, are fair but they do not make it possible for students to "show" all they know. For this reason, the most appropriate approach for a summative assessment context would be a balance between written exams, tests and coursework. In a formative assessment context, a combination of tests (such as CATs and CBTs) and coursework was suggested by the participants as being the most appropriate option. It was the view of the focus group participants that the CAT approach as described in this research was likely to be favourably received by test-takers, in both summative and formative settings, when combined with other assessment methods. The main reasons for this are the potential for:

- tailored testing;
- timely feedback or, at least, faster than in other assessment methods such as written coursework;
- efficient testing (only in a formative assessment setting).

The level of test difficulty of a CAT was also the explored as part of the first usability study, and focus group session. In general, participants reported that the level of difficulty of the CAT questions was more likely to be "just right" or appropriate than for those questions in the CBT part of the test.

To investigate this issue further, the level of difficulty of a CAT was the focus of two further user evaluation studies.

The first of these two user evaluation studies was in the context of a real summative assessment within the Human-Computer Interaction domain. In this study, 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 test-takers, was 3.37 (SD=0.60, N=113). Statistical analysis of test-takers' results and their perceptions of the level of difficulty of the test showed that test-takers' performance on the test had no effect on the perceived difficulty of the test.

It was also important to examine test-takers' perceptions of the level of difficulty in a formative context. To this end, the second of these user evaluation studies examined the perceived level of difficulty of formative and summative CAT assessments within the Computer Science domain.

At the end of formative and summative tests, test-takers were asked to rate the difficulty of the test that they have just taken from 1 (very easy) to 5 (very difficult). The test difficulty mean, as rated by the test-takers, was 3.53 (SD=0.64, N=76) for the formative test and 3.46 (SD=0.59, N=76) for the summative one. Statistical analysis of the test-takers' results and their perceptions of the level of difficulty of the test showed:

- no statistically significant correlation between the test-takers' proficiency levels and the test's difficulty rating;
- no statistically significant differences in the perceived level of difficulty means for the formative and summative tests.

The results reported suggest that the CAT software prototype developed for this research was effective in tailoring the level of difficulty of the test to the proficiency level of individual test-takers. More importantly, these results were observed in three different subject domains, namely English as a second language, Human-Computer Interaction, and Visual Basic programming. This was taken to indicate that the approach can be transferred and generalised to different subject domains.

Overall, test-takers exhibited a positive attitude towards the CAT approach as proposed in this research. This is an important finding, as Jettmar & Nass (2002) and Georgiadou et al. (2006) suggest that test-takers' attitude towards the CAT approach is under-represented in the literature.

Test-takers, however, are not the only participants in the testing process and thus the attitude of academic staff towards the CAT approach is discussed in the next chapter.

5. Academic staff evaluation of the CAT approach

Evaluation studies reported in Chapter 4 provided evidence to support the claim that test-takers exhibited a positive attitude towards the CAT approach. This chapter is concerned with the evaluation of the CAT approach by academic staff.

A group of eleven members of academic staff attended a structured presentation about the underlying concepts of the CAT approach, as employed in the research. The duration of the presentation was 45 minutes, and at the end of the presentation all participants were able to ask questions. It was crucial to the research to ensure that the participants were able to understand the ideas underpinning the CAT approach, as well as recognise the differences between the CAT approach and the computer-based test (CBT) one. The participants concurred in that the CAT approach appeared to be valid, and agreed to take part in two different studies in order to examine the CAT prototype developed for this research. These two studies are described in the first and second sections of this chapter.

The first section focuses on usability issues, from the perspective of academic staff. It examines whether or not academic staff considered that the CAT software prototype was likely to hinder test-takers' performance. The second section explores the extent to which academic staff participants perceived the inclusion of the CAT approach in a Higher Education context as useful. The

work introduced here was also reported in the following paper: “The Development and Evaluation of a Computer-Adaptive Testing Application for English Language” (Lilley & Barker, 2002).

In the case of usability issues, a heuristic evaluation (Molich & Nielsen, 1990) based on structured expert reviewing was performed, and is described in the next section.

5.1 Heuristic evaluation

In Chapter 4, the CAT software prototype developed for this research was shown not to affect test-takers’ performance in an adverse way. In order to gather additional data concerned with the CAT prototype’s usability, a heuristic evaluation (Molich & Nielsen, 1990) based on structured expert reviewing was carried out as part of this research. This evaluation involved a group of eleven experts, formed by ten lecturers in Computer Science and one lecturer in English for Academic Purposes. The inclusion of a lecturer in English for Academic Purposes was important to the study, due to his extensive experience in the use and application of computerised tests for the assessment of English as a second language.

After watching the presentation about the CAT approach, the eleven experts were asked to undertake both a heuristic evaluation and an evaluation of the CAT prototype’s usefulness as a pedagogical tool. The usefulness of the CAT prototype as a pedagogical tool is described later in section 5.2.

In the heuristic evaluation described here, different elements of the interface were analysed by the experts and compared to usability principles (the heuristics). Each one of the eleven experts was provided with a copy of the CAT software prototype on disk, and they independently rated ten usability standards from 1 (Poor) to 5 (Excellent). A copy of the heuristic evaluation guidelines can be found in Appendix I.

Table 5-1 summarises the results of the heuristic evaluation, where all the usability principles evaluated obtained a mean score equal or greater than 3.9 on the 1 to 5 Likert scale.

	Poor				Excellent	
Heuristic	**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

Table 5-1: Heuristic evaluation results

Given that in a heuristic evaluation five evaluators could detect 75% of the usability problems within a system (Molich & Nielsen, 1990), the scores obtained from the eleven evaluators involved in the evaluation process were taken to indicate that the CAT prototype developed for this research presented no major usability problems.

The results shown in Table 5-1 were taken to indicate that:

- the CAT prototype's current state and available actions are made explicit to users through a simple dialogue;
- users need not be familiar with system-oriented jargon or remember long sequences of commands in order to satisfactorily operate the CAT prototype;
- the CAT prototype supports user control and freedom, and it is straightforward to move from an unwanted state (such as an option chosen by mistake) to the desired state;

- the location and meaning of buttons and associated actions are consistent throughout the prototype;
- the CAT prototype presents good error prevention, and users are presented with a confirmation option before the system performs any irreversible action (for example, to exit the test before it is completed);
- as all the available options are visible, there is no memory overload on the part of the users;
- the interface design is minimalist, and only contains elements that are relevant to the current state of the CAT prototype application;
- although the interface design attempts to prevent users from making errors, when errors occur the interface is error tolerant, and error messages are written in plain English;
- the interaction with the system is straightforward and clear.

The usability principles “flexibility and efficiency of use” and “help and documentation” obtained the lowest mean score, and merited further examination.

In the case of “flexibility and efficiency of use”, the lower score could be explained by the lack of functionality that allows the user to configure the way in which the questions are displayed. It was not possible, for example, to change the font size. One of the evaluators reported that it is usually more difficult to read on a computer monitor than on paper, and this factor becomes more evident when the items (i.e. questions) presented become more difficult. The “flexibility and efficiency of use” heuristic also refers to the use of accelerators that permit expert users to tailor frequent actions. Due to the simplicity of the interface, the use of accelerators was not considered relevant to the CAT prototype developed for this research.

As for the usability principle “help and documentation”, the evaluators recognised that the prototype offers a satisfactory context-sensitive help. However, they highlighted that it is not possible to obtain information on how the test is executed before it is started.

All in all, the prototype was evaluated as being easy to use and easy to learn, and unlikely to hinder test-takers' performance. This is in line with the findings from the usability studies, from the test-takers' perspective, described in the previous chapter.

The favourable findings from the usability evaluation fostered further research and the CAT prototype was subjected to a pedagogical evaluation. The pedagogical evaluation is described in the next section.

5.2 Pedagogical evaluation

After carrying out the heuristic evaluation, the eleven experts were asked to rate ten statements from 1 (Unlikely) to 5 (Likely) to gather data on the CAT prototype's pedagogical usefulness in a Higher Education setting. There was also a text box for free text entry, so participants had an area to give reasons for their ratings, should they choose to do so. A copy of the pedagogical evaluation guidelines can be found in Appendix J.

The statements shown in Table 5-2 were constructed by the research team, drawing from their collective experience of teaching in a Higher Education setting. Table 5-2 summarises the participants' responses.

As can be seen from Table 5-2, statements concerned with the ease of use of the system (mean=4.9) and students' interaction with the application (mean=4.5) scored the highest. Statements concerned with the usefulness of the CAT approach as a tool to enable students to detect their own educational needs in formative (mean=2.7) and summative (mean=2.6) settings scored the lowest, and possible reasons for the low scores are discussed later in this section. Academic views with regards to the speed and accuracy of marking on a CAT are discussed next.

	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

Table 5-2: Pedagogical evaluation results

Speed and accuracy of marking. The results shown in Table 5-2 indicate that the academic staff participants considered that the CAT prototype would be valuable in terms of both speed and accuracy of marking. It is important to note that such benefits are generic to computer-assisted assessment rather than exclusive to the CAT approach.

The CAT approach as a tool that would enable academic staff to detect students' educational needs. Brown et al. (1997) state that one of the purposes of assessment is to identify a student's strengths and weaknesses, in order to understand their educational needs. When asked if the CAT approach would enable academic staff to gauge information about students'

educational needs, the formative setting scored higher than the summative one (3.7 and 3.3, respectively). The participants suggested that formative assessments provide academic staff with more information regarding the students' strengths and weaknesses, since they can be taken on a regular basis. In addition, academic staff participants noted that although useful, CATs are more difficult to construct than traditional CBTs due to the need of a large and calibrated bank of items the use of the CAT approach is limited.

The CAT approach as a tool that would enable students to detect their educational needs. Regarding the prototype's ability to help students to detect their own potential educational needs, both summative and formative assessment settings received a mean score lower than 3. One can speculate that there are three main reasons for this. The first reason is associated with the use of objective questions. As pointed out in section 2.1, objective questions are not suitable for assessing higher cognitive skills such as synthesis and evaluation. This characteristic of objective questions would restrict the skills that can be assessed using the CAT prototype and consequently limit the detection of educational needs on the part of the test-takers. The second reason is concerned with the use of the adaptive algorithm. Test-takers might be unaware of the adaptive process and therefore possibly unable to understand that the questions presented are tailored to their current level of ability, but not necessarily indicative of the highest level of difficulty within the subject domain. The third reason – and perhaps the most important one – is that the only feedback provided by the CAT approach is an overall score. The feedback does not provide test-takers with any additional feedback on how they can improve within the subject domain being tested.

Academic staff perspective on how the CAT approach would be received by students. As part of this work, the CAT prototype was also tested by the University's Head of English Language Teaching Department. It was his view that the CAT prototype would have potential to be used as a tool to:

- support the process of testing English proficiency of overseas students in a summative assessment setting;
- be used in a formative assessment setting to allow overseas students to assess their progress (or lack of).

Despite the potential use of the CAT approach in formative and summative settings as identified above, it was important to investigate whether or not academic staff participants felt that students would be receptive to the approach. The academic staff participants considered that students would more receptive to use a CAT in a formative rather than in a summative assessment environment.

One can conjecture that there are two reasons for this, namely speed and the scoring method employed by the CAT approach. In the case of speed, this is due to the potential to provide test-takers with timely feedback on performance. This was also identified as a benefit of the CAT approach by participants, and these results were reported in the previous chapter. As for the scoring methods used by the CAT approach, the results reported in Table 5-2 suggest that academic staff participants foresee problems regarding the scoring method used within CAT. In a CAT, the final score given to a test-taker is calculated based on the number of questions answered correctly and incorrectly, as well as on the level of difficulty of these questions. As a result, test-takers who answered the same number of questions correctly would almost certainly have different final scores, and this could bring uncertainties about the "fairness" of the assessment. Interestingly, participants in the focus group study described in the previous chapter did not seem to share such concerns. These participants reported that it was reasonable to expect that a test-taker who answered more difficult questions would score higher than a test-taker who answered easier ones.

Students' interaction with the CAT prototype. The results shown in Table 5-2 suggest that academic staff participants felt that students' interaction with the CAT prototype would be straightforward and clear. Moreover, participants

considered the CAT prototype developed for this research easy to use, and unlikely to affect test-takers' performance in an adverse way.

5.3 Summary

Both academic staff and test-takers are crucial stakeholders in the assessment process and, for this reason, their attitude to the CAT approach in a Higher Education setting was investigated as part of this research. Chapter 4 is concerned with test-taker evaluation, whilst the current chapter focuses on the evaluation of the approach by academic staff.

The CAT prototype developed for this research was subjected to usability and pedagogical evaluations, and these evaluations involved a group of eleven academic staff. The findings reported in this chapter suggest that:

- the CAT prototype is easy to use, and unlikely to affect test-takers' performance in an adverse way;
- the CAT prototype does not hinder assessment by introducing extraneous variables, such as cognitive overload, due to the computer interface;
- academic staff participants considered that the CAT prototype would be valuable in terms of speed and accuracy of marking;
- the CAT approach in summative and formative assessment settings would help lecturers in detecting students' educational needs;
- the appropriateness of the CAT approach as a tool to help students to detect their own potential educational needs further research;
- CATs are more difficult to construct than traditional CBTs;
- academic staff participants considered the CAT approach to be of greater pedagogical value in a formative than in a summative setting.

In this chapter evidence was provided that academic staff in general exhibited a favourable attitude towards the CAT approach. This is an important finding, as academic staff attitude towards the CAT approach is an issue that has not been adequately explored in the literature. Much of the work in the CAT field (see for example Hambleton & Rogers, 1991; Swaminathan, 1991; Guzmán & Conejo, 2005; Guo et al., 2006) concentrates on practical application issues such as test construct, whereas the research reported in this thesis also considers academic staff attitude towards the CAT approach.

An issue identified in this chapter that merits further investigation was the appropriateness of the CAT approach as an educational tool aimed at promoting and supporting student learning. As indicated by the members of staff who participated in the study reported here, the provision of a score alone is not sufficient to help students detect their own potential educational needs. However, before this line of investigation was pursued, it was essential to the research to examine whether or not the CAT approach was both valid and reliable. The validity and reliability of the CAT approach is discussed next.

6. Validity and Reliability of the CAT approach

This chapter addresses issues of validity and reliability which are of crucial importance to all stakeholders in the student assessment process, including students, academic staff, educational institutions and prospective employers. In Chapters 4 and 5, test-taker and academic staff attitude towards, and acceptance of, the computer-adaptive test (CAT) approach were examined. The findings reported as part of this work indicate that the CAT approach is likely to be positively received by these two groups of stakeholders. Such favourable findings alone, however, do not provide sufficient evidence to support the use of the CAT approach as part of student assessment in Higher Education since "it is important for all stakeholders in the assessment process that the measurement of performance is valid and reliable" (Dunn et al., 2003: p. 17).

The purpose of this chapter is to evaluate the extent to which the CAT approach, as described in this research, was valid and reliable. Findings related to the validity and reliability of the CAT approach were published as part of this research in Lilley et al. (2002c), Barker & Lilley (2003), Lilley & Barker (2003b), Lilley & Barker (2006a), and Lilley et al. (2007).

The chapter is divided into two main sections. The first section focuses on the validity of the approach, and the second section is concerned with reliability issues. Specific points of interest are discussed in the relative sections.

6.1 Validity of the approach

The American Psychological Association (1999: p. 9) states that “validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests”. This definition applies to a wide range of tests, such as tests constructed to measure depression as well as tests devised to measure academic achievement. Definitions of validity within the context of student assessment in Higher Education are largely available in the related literature. Miller et al. (1998: p. 233), for instance, state that “a test is said to be valid when it measures the extent to which the objectives of the teaching programme have been achieved”. In a similar vein, Dunn et al. (2003: p. 17) describe a valid assessment as one that is meaningful, useful, and measures “the performance of the intended learning outcomes specified”.

There are different types of validity (Miller et al., 1998), and the types that were considered to be of interest to this research are face validity, content validity and construct validity. These are discussed next.

6.1.1 Face validity

Miller et al. (1998: p. 234) state that “an assessment task is said to have face validity if a number of judges – ranging from experts in the field to students – agree that the test item is valid”. Face validity is concerned with the extent to which, academic staff and students alike, agree that a test is a valid method to measure what it is intended to measure.

Reports from test-takers, as described in sections 4.2 and 4.3, support the view that a test based on the CAT approach "looked valid" to them. Furthermore, Lunz et al. (1992) suggest that CATs where the review of previously entered responses is allowed, such as the CAT software prototype developed for this research, are likely to have greater face validity than those CATs where review is not permitted.

Findings from the academic staff evaluation reported in Chapter 5, were taken to indicate that the CAT approach was valid in both formative and summative assessment settings, with a greater face degree of validity in the former.

Although Miller et al. (1998) amongst others recognise the importance of face validity, doubts have been expressed about its rigour. Anastasi (1988, p. 144), for instance, argues that face validity is "not validity in the technical sense" and proposes that other forms of validity testing, such as content validity, are required. Content validity, as applied in the research, is discussed next.

6.1.2 Content validity

Content validity is concerned with the extent to which the content of a test satisfactorily represents the subject domain (or syllabus) being assessed (American Psychological Association, 1999). One way to evaluate whether a test has sufficient content validity for its purpose would be the analysis, by subject domain experts, of the relationship between the test content and the intended learning outcomes. Hambleton & Rogers (1991, p. 18) state that "expert judgement is the main mode of investigation of a test's content validity". Content validity is of particular importance in order to avoid the inclusion of irrelevant elements, the under-representation of core components, and the overemphasis of certain elements within the subject domain being tested.

Validity based on test content is often a laborious task in the context of CATs, as the recommended number of questions required in the question bank is, at least, 4 times the number of questions to be administered in a test sitting. It should be noted that questions should be evenly distributed across the different ability levels. Validity based on test content is a well established technique, and it is often part of the regular internal and external moderation processes in Higher Education institutions (Miller et al., 1998; Rhodes & Tallantyre, 2003).

The CAT approach, as implemented as part of this research, was based on the use of objective questions such as multiple-choice and multiple-response. Ward (1980) identified contributing factors that relate to the validity of objective tests in general, such as: “good syllabus coverage” (p. 9), “consistent syllabus coverage from year to year” (p. 11), “compulsory questions” (p. 13), “results less influenced by irrelevant abilities” (p. 12) and “precise questions” (p. 13). Such factors can also be applied to support the view that the CAT approach, as implemented in this work, has content validity.

Good syllabus coverage. Ward (1980) argues that objective tests can make it possible to assess a greater range of the syllabus by presenting test-takers with more questions in a given period of time than it would be possible with non-objective questions such as essay type questions. In addition, the CAT software prototype developed for this research allows the examiner to specify, within a subject domain, the number of topics being assessed as well as the number of questions per topic. This would, in turn, make it possible to ensure content balancing. It should be noted that content balancing is not a factor that is taken into account by the Three-Parameter Logistic (3-PL) Model (Lord, 1980), as described in section 3.4. A further characteristic of this work intended to increase its content validity is the stopping condition. As described in section 3.5, the CAT prototype developed for this research is of fixed, rather than of variable length. This means that, assuming that the maximum time for the test had not elapsed, the test-taker would be presented with a predefined number of questions in order to cover all intended topics within the subject domain.

Consistent syllabus coverage from year to year. The CAT approach, as implemented as part of this research, supports the construction of tests that present consistent levels of syllabus coverage from year to year. It should be noted that the total number of questions as well as the number of questions per topic is determined by the examiner.

Compulsory questions. Ward’s (1980: p. 12) argument on increased test validity by the use of compulsory questions is mostly based on the idea that, in

such a scenario, “all students answer questions on the same syllabus”. In addition, Ward (1980) cites the example of a typical paper where students have to answer 6 out of 10 questions and any 2 students may answer questions on totally different topics. Although the questions on a CAT are dynamically selected, the approach to content balancing used in the prototype developed for this research ensures that all test-takers will be presented with a fixed number of questions per topic, within the subject domain being assessed.

Results less influenced by irrelevant abilities. There are circumstances when skills such as writing, are not relevant to the learning outcomes being assessed and therefore should not affect a test-taker’s score. An example of such a learning outcome could be knowledge and understanding of Visual Basic.NET programming terms. Ward (1980) argues that in such a scenario, the choice of objective questions over non-objective ones (for example, essay type questions) can increase a test’s validity. Ward (1980: p. 13), however, warns that “the objective test’s independence of such skills as drawing, writing English and selection of relevant information is, of course, only an advantage if they are indeed irrelevant to the abilities being assessed”.

Precise questions. Like the four factors listed above, this factor is generic to objective tests rather than exclusive to the CAT approach. Well-devised objective questions can be very precise, and therefore leave little – if any – room for misinterpretation on the part of the test-taker about what is being asked. Such precision can also be seen as a means to increase a test’s content validity.

6.1.3 Construct validity

Construct validity is “the measure of the underlying theory or construct of a particular test or examination” (Brown, 1997: p. 241). Construct validity is concerned with the degree to which a test assesses the underlying theoretical construct it is intended to measure. In this research, construct validity is concerned with the extent to which CAT proficiency level estimates are interrelated to scores obtained by other traditional assessment methods

intended to measure similar learning outcomes. To investigate the construct validity of the CAT approach, an empirical study was conducted in which a group of test-takers participated in three different assessment methods, namely computer-adaptive test, computer-based test and practical programming test. The questions employed in this study were analysed by two subject experts with the purpose of ensuring content validity. This study was also published in Lilley & Barker (2006a), and is presented next.

Method. As part of their regular assessment for a programming module, a group of 125 Level 2 Computer Science undergraduates participated in three assessments. The assessments are summarised in Table 6-1. All assessments took place in computer laboratories, under supervised conditions.

Assessment	Brief description
Computer-based test (CBT)	Test-takers were asked to answer 10 predefined questions
Computer-adaptive test (CAT)	Test-takers were asked to answer 30 dynamically selected questions
Programming Test	Test-takers were asked to write a computer program using Visual Basic, based on an unseen program specification.

Table 6-1: Summary of assessments undertaken by participants

Summary of test-taker performance. Test-takers' performance in three assessments is summarised in Table 6-2. In Table 6-2, the possible scores for the CBT and practical programming test ranged from 0 (lowest) to 100 (highest). The possible scores for the computer-adaptive test ranged from -3 (lowest) to +3 (highest).

Assessment	Mean	Std. Dev.
Computer-based test	36.96	18.41
Computer-adaptive test	0.16	1.23
Practical programming test	44.52	25.38

Table 6-2: Summary of test-taker performance (N=125)

Findings. In order to investigate the correlations between CAT proficiency level estimates and other assessment methods intended to measure similar learning outcomes (i.e. CBT and programming test), the results shown in Table 6-2 were subjected to a Pearson's Product Moment correlation. This is shown in Table 6-3.

Assessment		Practical programming test	CBT
CAT	Pearson Correlation	0.428	0.548
	Sig. (2-tailed)	0.000	0.000
CBT	Pearson Correlation	0.221	*
Test	Sig. (2-tailed)	0.013	

Table 6-3: Pearson's Product Moment correlation results (N=125)

The significant correlation observed between the CAT and the practical programming test ($r=0.43$, $p<0.001$) and between the CAT and the CBT ($r=0.55$, $p<0.001$) are an important finding, and were taken to support the claim that the CAT approach has construct validity. The results shown in Table 6-3 show that those performing well on the CAT test also performed well on the other two test formats. The correlation between the CBT and the practical programming test, although significant was smaller than either correlation with the CAT ($r=0.22$, $p<0.01$). This supports the view that the test-takers were not disadvantaged by the CAT approach.

Up to this point, this chapter has focused on validity issues. However, a test that is valid is not necessarily reliable and vice-versa. Reliability issues were

also of importance to this research, and the next section of this chapter is concerned with these issues.

6.2 Reliability of the approach

Reliability is "the degree to which test scores for a group of test-takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test-taker" (American Psychological Association, 1999: p. 180). Ward (1980: p. 9) adds that an assessment is reliable when "it applies a consistent standard of measurement to all students and in all years". In general terms, one can argue that reliability refers to the extent to which assessments are consistent. On the topic of test reliability, Miller et al. (1998: p. 237) warn that "it is unrealistic to expect to achieve 100 percent reliability" and that the aim should be to construct tests that are "as reliable as possible".

In a similar vein to test validity, there are factors that contribute towards test reliability that are generic to objective tests rather than exclusive to the CAT approach. These factors are explored next.

6.2.1 Contributing factors

Ward (1980) identified three factors that contribute to the reliability of objective tests. Two of these: "reliable marking" (p. 14) and "assessment of student's own work" (p. 14) are of relevance to this research. These factors are discussed next.

Reliable marking. In the implementation of the CAT approach employed for this research, all questions are marked consistently and objectively by the software application.

Assessment of student's own work. Ward (1980) argues that objective tests are often conducted under supervised conditions, and this can increase assessment reliability. The reason for this is that such a scenario would

involve some form of authentication, and therefore it would be relatively straightforward to ensure that results obtained by test-takers were based solely on their own work.

The two factors above both contribute to reliability rather than measuring it. The next section of this chapter, discusses how one approach to measuring reliability, namely test-retest reliability, was applied to this work.

6.2.2 Test-retest reliability study

In a test-retest reliability study, the same group of participants are subjected to two different forms of the same test. The reliability is considered to be the correlation between the scores of both tests. In order to investigate the reliability of the CAT approach, an empirical study was performed as part of this work. This empirical study was published in Lilley & Barker (2003b), and its method and main findings are presented next.

Method. A group of 133 Level 2 Computer Science undergraduates enrolled on a programming module took part in two sessions of summative assessment using the CAT software prototype developed for this research. The characteristics of these two sessions are summarised in Table 6-4.

The CAT software prototype developed for this research was modified to include a traditional computer-based test (CBT) component, in order to administer a predefined set of questions to all participants. Prior to the first session of assessment using the modified CAT software prototype, test-takers were given a brief introduction to the use of the software, but were not informed of the existence of two sections within the test (i.e. CBT followed by CAT). In both sessions of assessment, the order in which the CBT questions were presented was randomly selected, as an attempt to minimise unauthorised collaboration amongst test-takers.

In addition to the two computer-delivered assessments, participants were required to undertake two additional assessments as part of their programming module. These two assessments are also summarised in Table 6-4.

Assessment	Brief description
1. In-Class Test 1	10 predefined questions (i.e. CBT mode) followed by 10 questions dynamically selected (i.e. CAT mode).
2. In-Class Test 2	10 predefined questions (i.e. CBT mode) followed by 20 questions dynamically selected (i.e. CAT mode).
3. In-Class Programming Test	Test-takers were asked to write a computer program using Visual Basic, based on an unseen program specification.
4. Practical project	Participants were asked to produce a straightforward high fidelity software prototype, according to a brief, over a period of 4 weeks.

Table 6-4: Summary of assessment employed for the group of participants

With exception of the practical project (i.e. Assessment 4), all assessment sessions listed in Table 6-4 were conducted under supervised conditions in computer laboratories.

Summary of test-taker performance. A summary of the test-takers' performance in each of the assessments is presented in Table 6-5.

Assessment		Mean score
Assessment 1	CBT 1	51.5%
	CAT 1 (proficiency level)	-0.832
Assessment 2	CBT 2	42.3%
	CAT 2 (proficiency level)	-0.909
Assessment 3		
In-Class Programming Test		49.7%
Assessment 4		
Practical project		71.7%

Table 6-5: Summary of test-taker performance (N=133)

In Table 6-5, the potential CAT scores ranged from -2 (lowest) to +2 (highest). The remaining scores ranged from 0% (lowest) to 100% (highest).

Findings. An Analysis of Variance (ANOVA) was performed on the data summarised in Table 6-5, in order to test the significance of any differences in the means. The results of this ANOVA are shown in Table 6-6.

Between groups	Probability (p)
CBT Assessment 1 and Assessment 2	0.001
CAT Assessment 1 and Assessment 2	0.607
Assessment 3 (Programming Test) and Assessment 4 (Coursework)	0.001

Table 6-6: ANOVA results relating to the data summarised in Table 6-5 (N=133)

The results in Table 6-6 show that there was a significant difference between the number of questions answered correctly in the CBT element of assessments 1 and 2 ($p=0.001$). However, there was no significant difference between the CAT levels obtained by test-takers in assessments 1 and 2 ($p>0.60$). This is an interesting result, especially in consideration of the finding that the mean CBT performances in assessment 1 and 2 were significantly different ($p<0.001$). These results were taken to indicate that the CAT level is a reliable measure of test-taker ability, and possibly a better and more consistent measure than a simple test score.

There was also a significant difference observed in the performance of students on the two off-computer assessments (assessments 3 and 4, $p=0.001$). In order to further understand the implications of these findings, a Pearson's Product Moment correlation was also performed on the data collected from the four assessments, and the results of this analysis are shown in Table 6-7.

		CAT 1	CBT 1	CAT 2	CBT 2	Programming test	Practical project
CAT 1	Pearson Correlation	*	.849(**)	.617(**)	.548(**)	.552(**)	.377(**)
	Sig. (2-tailed)		.000	.000	.000	.000	.000
CBT 1	Pearson Correlation	*	*	.552(**)	.467(**)	.445(**)	.300(**)
	Sig. (2-tailed)			.000	.000	.000	.000
CAT 2	Pearson Correlation	*	*	*	.816(**)	.571(**)	.407(**)
	Sig. (2-tailed)				.000	.000	.000
CBT 2	Pearson Correlation	*	*	*	*	.527(**)	.398(**)
	Sig. (2-tailed)					.000	.000
Programming test	Pearson Correlation	*	*	*	*	*	.528(**)
	Sig. (2-tailed)						.000

Table 6-7: Pearson's Moment Correlation results (N=133)
(**) Correlation is significant at the 0.01 level (2-tailed).

The results of the Pearson's test shown in Table 6-7 indicate that the scores obtained by participants in assessments 1, 2, 3 and 4 (see Table 6-5) are well correlated with each other ($p<0.001$). This was interpreted as indicating that a score obtained by a participant in one assessment is a reasonable and fair predictor of performance in any other. It can also be seen that there is a high correlation between scores in the CBT and the CAT sections of assessments 1 and 2. On average, participants who performed well in the CBT sections also performed well in the CAT sections and vice versa ($p<0.001$).

It was also found that the CAT proficiency levels achieved by the participants in assessment 1 were highly correlated with the CAT levels in assessment 2. This was taken to indicate that:

- the CAT test was a fair reflection of participants' ability in the assessment;
- the CAT assessment was at least as good an indicator of the ability of a test-taker as the CBT component of the prototype;
- no participant was disadvantaged by the CAT approach.

6.3 Summary

This chapter discussed issues related to the validity and reliability of the CAT approach: face validity, content validity, construct validity and test-retest reliability. It was of relevance to this work to show that the CAT approach complies with these well-established standards since it is crucial to all stakeholders in the student assessment process that assessment methods are both valid and reliable. As part of this work, two empirical studies were carried out are and reported in this chapter. Both studies were performed in a real educational context, as recommended by Laurillard (1993) and Barker & Barker (2002). The findings from these two empirical studies provided evidence to support the claims that:

- the CAT approach is, at least, as fair and accurate as other traditional computer-assisted assessment methods in measuring a test-taker's proficiency level within a subject domain,
- test-takers are not disadvantaged by the CAT approach,
- the CAT approach is both valid and reliable.

Furthermore, it was shown that several factors that contribute to the validity and reliability of objective tests can also be applied to the CAT approach.

There is an increasing body of research supporting the validity and reliability of the CAT approach; for instance, Segall (2001), Wolfe et al. (2001b) and Segall et al. (2001) report on the validity of the CAT approach. Other research, such as the work by Schoonman (1989) and Moreno & Segall (2001), report on the reliability of the approach. Such research, however, focuses mostly on the validity and reliability of the CAT approach when compared with traditional objective tests using a paper-and-pencil format. The studies published as part of this research – Lilley et al. (2002c), Barker & Lilley (2003), Lilley & Barker (2003b), Lilley & Barker (2006a) – are a useful addition to this body of research since they examined test interrelations between CAT proficiency level estimates and scores obtained using other forms of computer-assisted assisted assessments, rather than paper-and-pencil tests.

Up to this point in the research, the only feedback provided to test-takers was their overall CAT proficiency level. In spite of the accuracy and potential usefulness of CAT proficiency level estimates, academic staff who participated in the pedagogical evaluation (see section 5.2) reported that such a performance indicator alone would not be sufficient to help students obtain valuable information about how to improve.

The aim of the next stage of the research was to investigate how the information about a test-taker's proficiency level gathered during a CAT test could be employed to provide feedback that is timely, individual and meaningful. This is described in the next chapter.

7. The automated feedback prototype

In the previous chapters of this thesis, the research focused on the design, implementation and evaluation of the computer-adaptive test (CAT) software application developed as part of this work. Evidence was provided to support the claims that:

- the CAT approach offers a measurement of test-taker performance which is as fair and accurate as that provided by the computer-based test (CBT) approach;
- the CAT approach supports a more interactive and challenging assessment experience, given that the questions are dynamically selected to match each individual test-taker's proficiency level within the subject domain;
- both test-takers and academic staff exhibited positive attitude towards the CAT approach.

Findings from the pedagogical evaluation reported in section 5.2, impacted on the direction of the research and led to the investigation of ways as to how the CAT approach can be applied to provide personalised feedback to individual CAT test-takers.

This chapter is divided into three main sections. The first section provides a brief description of the technologies employed in the implementation of the

automated feedback prototype. In the second section, common approaches to the provision of student feedback are outlined. The third section focuses on the provision of feedback on performance for the CAT software prototype. It covers a pilot study conducted by the research team and a description of the web-based automated feedback prototype developed as part of this work.

The following section focuses on practical implementation issues relating to the automated feedback prototype.

7.1 Implementation overview

The automated feedback prototype was implemented as a web application, so test-takers would be able to access feedback on test performance from any location, and in their own time, pace and frequency.

The implementation of the automated feedback prototype was divided into two main stages. The first stage was concerned with the implementation of the database (i.e. back-end), and the second stage with the development of the graphical user interface (i.e. front-end).

In the first stage, a batch VB program was written in order to extract data relating to test-taker performance from the CAT database as summarised in Table 7-1.

Data extracted	Usage
Overall proficiency level estimate	Data employed to provide test-takers with information about their overall performance.
Test-taker responses to items grouped by topic area	Data employed to calculate a proficiency level estimate per topic area, which is then translated into one of Bloom's cognitive skills. This process is described in section 7.3.
Test-taker incorrect responses to items	Data employed to select revision tasks that are appropriate for each test-taker's proficiency level. This process is also described in section

Table 7-1: Summary of data extracted from the CAT database

The sets of data listed in Table 7-1 are then imported into a Microsoft Access back-end database. The database was hosted on one of the University's web

servers. In the case of the automated feedback database, no split mirror database copies were created. There are three main reasons for this. First, no study carried out as part of the research involved more than 255 participants at one time. Second, the number of database read and write operations performed in the automated feedback application is very low. Third, the access to the automated feedback application was scattered over time and this minimised the risk of technical problems related to a high number of concurrent users.

The front-end of the automated software prototype was implemented in Active Server Pages (ASP) version 3.0. ASP is Microsoft's server-side script engine for dynamically-generated web pages (Microsoft Corporation, 2007c). The ASP pages were written in VBScript. Screenshots of the application can be found in section 7.2.

The following section introduces different approaches to the provision of student feedback.

7.2 Approaches to the provision of student feedback

Much has been written on the crucial role of feedback in student learning. Gibbs (2003: p. 46), for instance, state that “learners require feedback in order to learn”; Sambell et al. (1999) suggest that lack of feedback can lead to student de-motivation. Although feedback can occur without assessment, much of the literature focuses on the importance of feedback on assessment performance (see for example Brown et al. 1998; Miller et al., 1998; Bull & McKenna, 2004) and this is also the focus of this section. It should be noted that the section aims to provide a brief introduction to student feedback issues that were considered as part of the work reported here, rather than provide an extensive literature review.

Timeliness and usefulness are two factors that have been shown to contribute towards the effectiveness of feedback, and such factors are outlined below.

Timeliness. Brown et al. (1998), Miller et al. (1998), Dunn et al. (2003) and Gibbs (2003) argue that feedback must be timely to be useful. Gibbs (2003) suggests that increased student numbers often lead to slow feedback, with students receiving feedback on performance when the course has moved on or they are working on other assessment activities. Interestingly, Gibbs (2003) also suggests that there are circumstances when the quality of the feedback is not as important as its frequency and speed. In fact, Dunn et al. (2003) indicate that even detailed and valuable feedback is of little use if not returned within reasonable time, so students can act upon it.

Usefulness. Brown et al. (1998) argue that feedback has been shown to be more effective when it is useful; feedback is useful when it (1) is designed to help students learn more effectively and (2) shows the ways in which their performance can be improved. Dunn et al. (2003) warn that increased student numbers often mean that feedback on performance is restricted to an overall score, followed by a short comment such as "good work", or sometimes even no comment at all. As Brown et al. (1998) and Miller et al. (1998) point out, such an approach does not encourage students to engage in learning activities, nor does it offer suggestions for improvement that are within a student's grasp.

A number of authors including Brown et al. (1998), Miller et al. (1998), Dunn et al. (2003) and Bull & McKenna (2004) have reported on the importance to student learning of the provision of timely and useful feedback. However, the provision of such feedback is not always within reach of even the most conscientious academic staff, and those teaching large groups of students in particular. Various feedback techniques have been suggested in order to improve timeliness and usefulness, such as peer assessment (see for example Dunn et al. 2003), face-to-face feedback to whole classes (see for example Race et al., 2004) and electronic feedback (see for example Race et al., 2004; Bull & McKenna, 2004).

In the case of electronic feedback, English & Siviter (2000), Denton (2003), Dunn et al. (2003), Race et al. (2004) amongst others have reported on the

increased use of computer applications for providing student feedback. Such applications range from the creation of a statement pool to facilitate the storage and re-use of common statements that can be easily tailored to individual students (see for example Dunn et al., 2003), to the use of commercial computer-assisted assessment (CAA) software applications such as Question Mark Perception for the provision of formative feedback (see for example Steven & Hesketh, 1999).

The work introduced in this thesis is based on the use of objective questions, and Bull & McKenna (2004) identify five types of feedback that can be useful and particularly suitable for this type of assessment:

- to provide information about whether the test-taker's response to a question was correct or incorrect;
- to provide the correct answer (for example, "You have chosen 'red', but the correct answer is 'green' ");
- to explain why a response is correct (for example, "This is correct. 'Myself' is an example of a reflexive pronoun");
- to provide non-directive feedback to encourage the test-taker to find the correct answer (for example, "Remember that verbs in Spanish not only contain information about tense (i.e. when the action took place), but also about the subject (i.e. who performed the action)";
- to provide directive feedback, in order to assist the test-taker to find the correct answer (for example, "See Chapter 1 from 'Taxonomy of Educational Objectives' for an introduction to Bloom's taxonomy of cognitive skills").

Steven & Hesketh (1999) and Bull & McKenna (2004) indicate that various commercial CAA applications not only automate the marking of objective tests, but also provide functionality that allows academic staff to provide automated feedback using one or more of the types of feedback listed above.

In the case of the research reported here, the CAT software prototype does not provide functionality to support the provision of automated feedback and

an automated feedback prototype was designed and implemented to this end. The underlying idea was to create an application that would be able to generate feedback that was both timely and useful, based on the information about test-takers gathered via the CAT software prototype. In summary, although the CAT and the automated feedback applications are different, it is intended that they will complement each other. The following section introduces the main aspects of the automated feedback prototype.

7.3 Approach to automated feedback used in the research

When designing the automated feedback that is the focus of this section, an important assumption of this work was that a face-to-face feedback session led by a member of academic staff would typically comprise the provision of:

- an overall score;
- general comments about proficiency level per topic;
- recommendations on which concepts within the subject domain should be revised in the form of directive feedback (Bull & McKenna, 2004).

A pilot study was conducted in order to explore how the elements above could be implemented as part of the feedback for a CAT, and this study is described next.

7.3.1 Pilot study

Feedback on CAT test-takers' performance had so far been limited to an overall score. Test-takers' sores were sent directly to their individual email accounts, using a mail merge program. The mail merge program generated a simple report on test performance in Microsoft Word format (file extension .doc), using the template shown in Figure 7-1. The report on test performance was sent to test-takers by email, as an attachment.

To: <<Student_Name>>

Your score for the <<Assessment_Title>> was <<Student_Score>>%.

This is an automated email from
The <<Module_Title>> team

Figure 7-1: Overall score template

The values used for the Student_Name, Assessment_Title, Student_Score and Module_Title fields were retrieved from the actual CAT database.

Test-takers appeared to value the convenience of receiving their reports on test performance via email. Nonetheless, voluntary feedback sent by email to the research team by some test-takers suggested that the score on its own, although useful, was unlikely to help students improve their future work. Test-takers' reactions to the feedback (in the form of an overall score) were in line with the views of the experts who participated in the pedagogical evaluation discussed in section 5.2. In this pedagogical evaluation, the experts reported that the score provided by the CAT prototype alone was unlikely to help test-takers improve their future performance.

As an attempt to exploit the potential of the CAT approach to provide feedback to test-takers, a pilot study was conducted. In this pilot study, the report on test performance was extended to include two additional sections: feedback according to topic area, and a list of topics for revision.

Feedback according to topic area. The aim of this section was to provide test-takers with a summary (up to 100 words) of their performance in each topic area. In the research reported here, it has been assumed that proficiency level estimates can be used as an indicator of test-taker's performance according to Bloom's taxonomy of cognitive skills (Bloom, 1956; Anderson & Krathwohl, 2001). Thus, feedback sentences were constructed in the light of Bloom's taxonomy of cognitive skills, and Table 7-2 illustrates how these sentences were structured.

Proficiency level	Skill	Brief description
$-2 \leq \theta < -0.6$	Knowledge	In this section of the assessment, you demonstrated awareness of relevant terminology relating to usability goals. We recommend that you now concentrate on identifying which usability goals are most likely to be relevant for your Semester B project.
$-0.6 \leq \theta < 0.8$	Comprehension	Your performance in this section of the assessment suggests an understanding of the role of usability goals in the software development process. With the importance of usability goals and user experience goals in mind, start planning how you are going to apply these concepts to your Semester B multimedia project.
$0.8 \leq \theta \leq 2$	Application	You showed knowledge and understanding of fundamental principles relating to usability goals. Your performance in this section of the assessment suggests an ability to apply these principles to your multimedia project.

Table 7-2: Example of feedback statements used in the pilot study

In the event of a test-taker answering all questions incorrectly, a sentence such as the one below would be used instead:

- None of your responses provided in this section of the test were correct. This is clearly an area where you need to work hard. If you need any help, please ask.

In the pilot study, all responses for each individual test-taker were selected from the CAT database. Test-taker responses were then grouped by topic and a proficiency level was calculated using the response likelihood function shown in Equation 2-3 (p. 46). A feedback statement matching the test-taker's performance (i.e. knowledge, comprehension, or application) was then selected based on the proficiency level estimate per topic area, and added to the report on test performance.

List of topics for revision. This section of the report on test performance consisted of a list of points for revision, based on the questions answered incorrectly by each individual test-taker. Each question in the CAT database

had a feedback sentence associated with it. This feedback sentence did not reproduce the question itself, but listed specific sections within the recommended reading that should be reviewed. The same feedback sentence could be used for more than one question in the database. For instance, test-takers who answered incorrectly a question about usability goals will be directed to a specific section within the textbook, whether the question was about memorability or safety. The format of the extended report on test performance is illustrated in Figure 7-2.

To: <<Student_Name>>

1. Overall score

Your score for the <<Assessment_Title>> was <<Student_Score>>%.

2. Performance per topic area

Usability goals	You showed knowledge and understanding of fundamental principles relating to usability goals. Your performance in this section of the assessment suggests an ability to apply these principles to your multimedia project.
…	…
Topic n	

3. List of topics for revision

Usability goals are discussed in section 1.3 of …

…

After reading section 2.3 of … answer the following question: what is the main difference between 'instructing' and 'conversing' when interacting with a system?

This is an automated email from

The <<Module_Title>> team

Figure 7-2: Extended report on test performance

The extended report on test performance was sent to test-takers by email, as an attachment. In order to examine the usefulness of the extended report, the pilot study included an evaluation of the approach by test-takers. This evaluation is described next.

Method. A group of 122 Level 2 BSc Computer Science undergraduates enrolled in a programming module participated in a session of assessment session using the CAT software prototype. Test-takers had 30 minutes to answer 20 questions. The test was within the subject domain of Human-Computer Interaction (HCI) and covered six different topic areas, which are listed in Table 7-3.

Summary of test-taker performance. Table 7-3 shows the overall CAT proficiency level as well as the CAT proficiency level per topic area. The potential proficiency level values ranged from -2 (lowest) to +2 (highest).

Topic	Mean	Std. Dev.
Overall proficiency level	-1.13	1.55
1. Issues related to the use of sound at interfaces	-0.70	1.61
2. Graphical representation at interfaces, focusing on the use of colour and images	-1.19	1.33
3. User-centred approaches to requirements gathering	-1.25	1.35
4. Design, prototyping and construction	-0.49	1.78
5. Usability goals and User experience goals	-0.77	1.58
6. Evaluation paradigms and techniques	-0.97	1.65

Table 7-3: Summary of test-taker performance (N=122)

Test-taker attitude towards the extended report on test performance. In order to gather information about test-taker attitude towards the extended performance report, the research team sent an email to all test-takers asking them to classify the report that they had just received as being: ‘not useful’, ‘useful’ and ‘very useful’.

Findings. A total of 58 test-takers replied to the email sent by the research team. The results were split 50%/50% between "very useful" and "useful". No test-taker classified the report on test performance as being "not useful". The following two statements are examples of comments that test-takers sent by email:

- "Rather then giving just the mark the document gives very positive feedback";
- "The hints section at the end was very useful, nice to know what I need to work on".

The findings from the pilot study in addition to the degree of personalisation afforded by the CAT approach made it worthwhile to design and implement a web-based automated feedback software prototype. The prototype is described in the following section. It should be noted that the findings from the pilot study are supported by Lilley et al. (2004b) and Lilley et al. (2005d).

7.3.2 Prototype overview

As part of this work, an automated feedback prototype based on the CAT approach was designed, implemented and evaluated. This section of the thesis focuses on the design and implementation; the evaluation is described in Chapters 8 and 9.

The user interface for the automated feedback prototype was built based on the general principles for user interface design developed by Nielsen (2005), as these were found useful in the design of the CAT software prototype. Whilst the CAT software prototype was a Windows-based application, the automated feedback was web-based. The underlying idea was to design and implement an automated feedback prototype that could provide students with opportunities to learn in their own time, pace, location and frequency.

The automated feedback generated by the application consists of three sections in order to reflect the features identified above: overall score,

summary of performance per topic area and personalised revision plan. These three sections are described below.

The screenshots produced in Figure 7-3, Figure 7-4 and Figure 7-5 are taken from the web-based feedback application, showing results obtained by one test-taker who took a test using the CAT software prototype. In this test, test-takers had 40 minutes to answer 40 objective questions within the Visual Basic.NET subject domain. The questions were organised into five topic areas, namely 'Representing data', 'Classes and Controls', 'Functions and Procedures', 'Controlling program flow' and 'ADO.NET'.

Overall score. The overall score was obtained by employing the Three-Parameter Logistic Model from IRT (Lord, 1980), as described in section 2.2.4. Figure 7-3 illustrates how this information was displayed to test-takers.

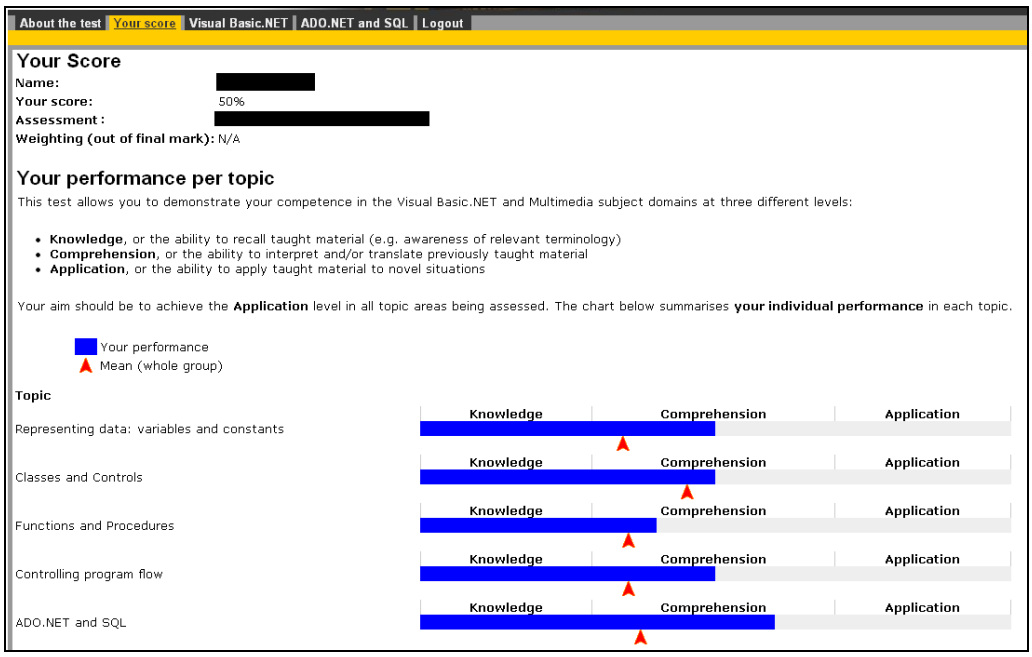

Figure 7-3: Automated feedback prototype
Screenshot illustrating how overall score and performance per topic were displayed within the feedback application. The test-taker's name and module have been omitted.

Summary of performance per topic area. The Three-Parameter Logistic Model from IRT (Lord, 1980) was also employed to estimate a proficiency level

per topic, in the same way as in the pilot study. An important assumption of this work is that test-takers' proficiency levels per topic could be mapped into Bloom's taxonomy of cognitive skills. For instance, a proficiency level between -3 and -1 would indicate that the cognitive skill knowledge has been demonstrated. A proficiency level between -1 and 1 would be taken as evidence that the cognitive skills knowledge and comprehension have been achieved. Finally, a proficiency level between 1 and 3 would denote that the test-taker has demonstrated the cognitive skills knowledge, comprehension and application. Higher level cognitive skills are considered to include all lower level skills. So, a question classified as application is assumed to embrace both comprehension and knowledge. This is illustrated in Figure 7-3 below.

The red arrow in Figure 7-3 shows how the test-taker's performance per topic compares with their peers or, in other words, whether their score is above or below the group's average. This feature was not included in the original version of the automated feedback prototype designed by the research team, but it was included after receiving voluntary feedback from test-takers. Test-takers appeared to measure their performance not solely based on their individual scores, but also based on the performance of their peers. This finding is further discussed in Chapter 8.

Personalised revision plan. Both Figure 7-4 and Figure 7-5 show examples of personalised revision plans. For each question answered incorrectly by a test-taker, the relevant revision task is retrieved from the database and listed as part of the personalised revision plan. Although based on the question's stem, revision tasks do not duplicate test questions.

Providing test-takers with a copy of all questions they answered incorrectly was a simple practical solution from a software development's perspective. However, such solution presented important pedagogical limitations, namely that providing a copy of the questions and respective key answers would not foster reflection and research (Ellis & Ratcliffe, 2004). Moreover, it is often argued that increased exposure of questions would jeopardise their use in future assessment sessions. The possibility of reusing questions is one of the

expected benefits of the creation and maintenance of a database of questions (Freeman & Lewis, 1998).

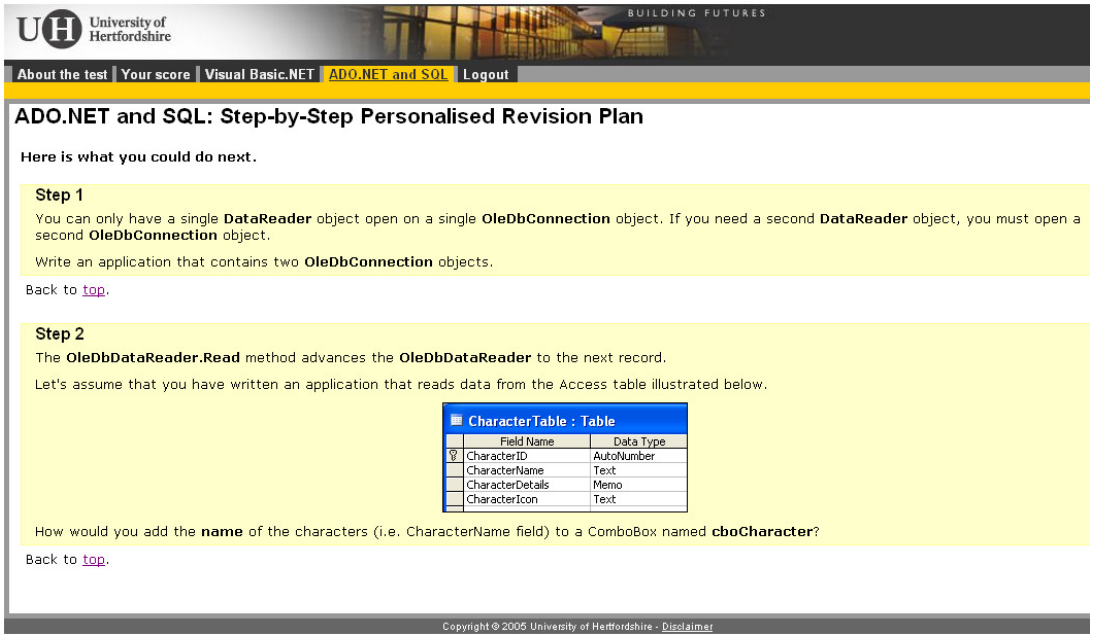

Figure 7-4: Screenshot illustrating a personalised revision plan.

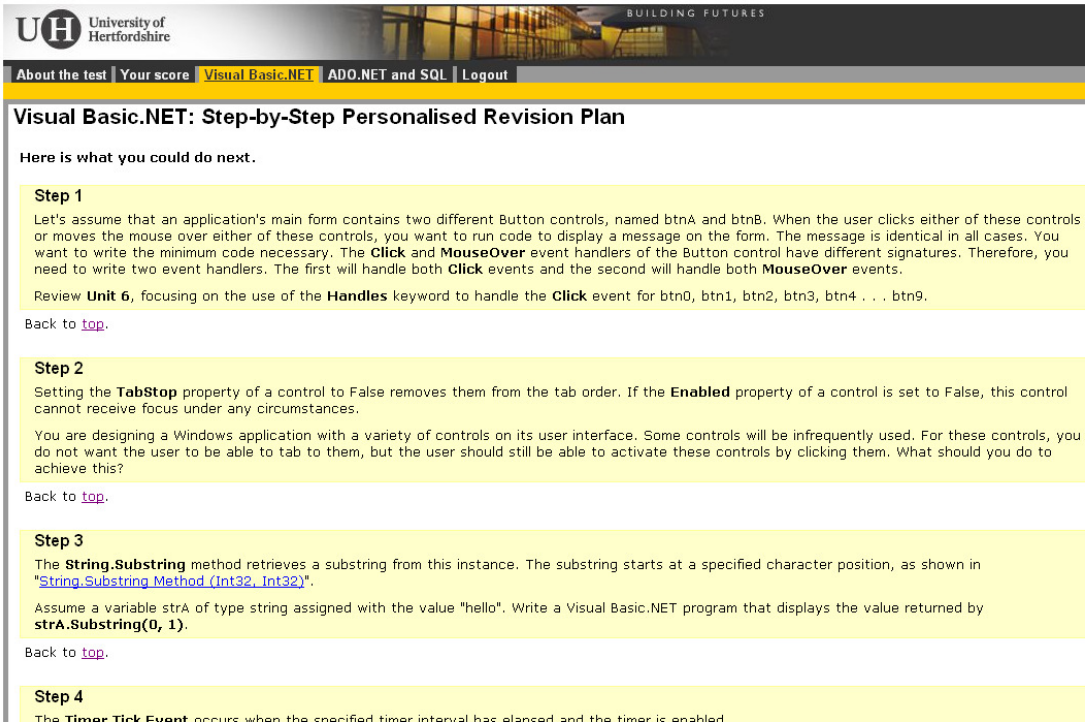

Figure 7-5: Screenshot illustrating a personalised revision plan.

It can be seen from Figure 7-4 and Figure 7-5 that the revision tasks involve a range of activities including: writing programs from scratch, reviewing specific lecture or tutorial learning materials and using external resources such as the software vendor online library. In so doing, it is expected that test-takers will be encouraged to learn in different ways.

As discussed in 2.2.4, one of the aims of a CAT is to match the level of difficulty of the questions to the proficiency level of individual test-takers. Because test-takers differ in proficiency levels, they are presented with a personalised set of questions. By having one revision task per question, the automated feedback prototype introduced here is capable of offering individual test-takers with a set of revision tasks that match their current level of ability within the subject domain. This ensures that less able test-takers are not provided with revision tasks that are too hard and therefore bewildering or frustrating. Similarly, more able test-takers are not presented with revision tasks that are unchallenging and therefore de-motivating. The underlying idea is to provide test-takers with realistic challenges, given that one of the aims of assessment is to direct test-takers to go beyond their current boundaries of knowledge (Yorke, 2003).

The automated feedback prototype was also described in the following papers: Lilley et al. (2005a), Lilley et al. (2005b), Barker & Lilley (2006) and Lilley & Barker (2006c).

7.4 Summary

Whilst assessment is often referred to as an important driving force in student learning, given its substantial impact on when, what and how students learn (Freeman & Lewis, 1998; Brown et al., 1998; Miller et al., 1998; Biggs, 2002; Race et al., 2004), feedback on assessment helps students improve (Brown et al., 1997; Miller et al., 1998; Bull & McKenna, 2004).

Increased student to staff ratios, however, often mean that academic staff are unable to provide feedback on student performance that is timely and useful.

Miller et al. (1998: p. 113), for instance, point out that “lecturers are often criticized for failing to produce sufficient feedback on the quality of a student’s work or the level of attainment reached by the student”.

Feedback must be timely to be useful. It is the experience of the research team that when large-scale computerised objective testing is used, feedback in the form of scores is usually returned in a timely fashion, as a result of automated methods of marking. Feedback on how students can improve, however, can be slow and not delivered until the course has moved on when it is of less use. In some cases, such feedback is absent; this is because it is time consuming to produce individual feedback for a large group of test-takers. Issues related to the timeliness of feedback, from the perspective of academic staff, are also discussed in Chapter 9.

As has been discussed in Chapter 2, substantial investments in computer technology by Higher Education institutions and high student to staff ratios have led to an increased pressure on staff and students to incorporate electronic methods of learning and teaching. This includes a growing interest in the use of computer-aided assessment and automated feedback, not only to make the technological investment worthwhile but also to explore the opportunities presented by the computer technology available.

Current computer technology allows the provision of automated feedback to test-takers who participate in a traditional CBT by, for example, making the questions answered correctly and incorrectly available electronically. The level of personalisation in such a scenario is, however, low. This is because all test-takers have been presented with the same fixed set of questions, regardless of their proficiency level within the subject domain. Such automated feedback approach would present similar problems to those encountered in other forms of feedback such as face-to-face feedback to whole classes, where the level of personalisation is also low.

Brusilovsky (2004) cites the CAT approach 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. In

spite of the predicted benefits of the CAT approach, there is very little evidence in the literature of the provision of feedback other than a CAT overall proficiency score (see for example Julian, 1993; Fitzgerald, 1999).

In this chapter, it was shown that CAT proficiency levels as well as the questions dynamically selected during a CAT test, can be employed to support the generation of automated feedback that is timely and tailored to each individual test-taker. The combination of adaptive testing and automated feedback provides an opportunity to individualise feedback to a far greater extent that than supported by traditional CBTs, where all test-takers are presented with the same set of predefined questions. In addition to greater individualisation, the approach to automated feedback described here should enable academic staff to obtain valuable information about test-taker's progress. An important assumption of the CAT approach is that questions that are too difficult or too easy provide little valuable information regarding a test-taker's knowledge within the subject domain. Only those questions exactly at the boundary of the test-taker's knowledge provide academic staff with valuable information about the level of a test-taker's proficiency level.

The automated feedback prototype built as part of this research consisted of three elements: (1) overall proficiency level, (2) overall proficiency level per topic area and (3) revision tasks. The design and implementation issues resulting from the inclusion of these three elements can be divided into three broad areas: (1) database creation and maintenance, (2) software algorithm, and (3) user interface design.

The underlying design of the feedback database was relatively straightforward. Overall proficiency level, proficiency level estimates and questions answered correctly, for each test-taker, were imported from the CAT database. An additional table was created to store revision tasks; each question contained a revision task associated with it. In this work it was found that the most onerous, albeit vital, undertaking was the production of revision tasks by academic staff.

The design and implementation of the software algorithm was also comparatively straightforward; complex proficiency level estimates and question selection procedures were carried out by the CAT algorithm.

The automated feedback algorithm was used to map proficiency level estimates per topic area into Bloom's taxonomy of cognitive skills, as described in section 7.3. In addition, the automated feedback algorithm was used to select and display the information stored for each individual test-taker in the feedback database, upon the submission of a valid username and password. Each test-taker was issued a unique username and password to access the automated feedback prototype. It was important to ensure test-takers' privacy; this is in spite of Gibbs' (2003: p. 46-47) view that as "students care about others think about them" removing some aspects of confidentiality when providing feedback could lead to better performance. It is possible that the "social dimension" identified by Gibbs (2003: p. 46) was the factor that led students to request the addition of the group's mean performance to the feedback; it appears that students found it useful to know how their performance related to the rest of the group.

In section 7.2, it was reported that there are various types of feedback that can be provided for objective questions: (1) informing students about the correctness of the response, (2) providing students with the correct answer, (3) providing students with the correct answer followed by explanation, (4) prompting students with relevant hints so they can find or construct the correct answer, and (5) providing students with directive feedback so they can find or construct the correct answer. All the types listed above can be applied to the provision of feedback for a CAT. In this work, type (1) was not employed, as it simply provides a score per question (i.e. correct/incorrect). Types (2) and (3) were not employed either, as there is evidence that they might not foster important graduate skills such as reflection and research (Ellis & Ratcliffe, 2004). Bull & McKenna (2004: p. 62), for instance, report on a case study where students employed "a 'smash and grab' technique, 'punching any key' to 'strip off' the feedback and correct answers". Type (4) and (5) were both regarded as suitable candidates. Type (5) was chosen as the model for the

revision tasks as it was considered the most useful, and the most likely to promote reflection and research.

Based on the results from the pilot study and the combined teaching experience of the research team, the automated feedback as described in this chapter was considered timely and useful. It was therefore of relevance to this research to investigate test-takers' attitude towards the automated feedback, and this is the focus of the next chapter.

8. Test-taker evaluation of the automated feedback prototype

In the previous chapter, the method adopted by this research to the generation of automated feedback based on the computer-adaptive test (CAT) approach was described. It was of relevance to this work to investigate the attitude towards and acceptance of the automated feedback approach by test-takers. To this end, three empirical studies involving test-takers in a real educational setting were conducted and these are the focus of this chapter.

The empirical studies reported in sections 8.1.1 and 8.1.2 are concerned with the application of the automated feedback prototype in the context of summative assessment. The empirical study included in section 8.1.3 is concerned with test-takers' attitude towards the feedback approach in the context of formative assessment.

8.1 Test-taker attitude

This section reports on three empirical studies concerned with the usefulness of the automated tool that was developed as part of this research.

8.1.1 Summative assessment

The study described in this section is concerned with the application of the automated feedback prototype in order to provide feedback on performance to a group of test-takers who participated in a session of summative assessment within the Human Computer-Interaction (HCI) domain. The test-takers were assessed using the CAT software prototype developed for this research. The automated feedback was generated using the software application described in section 7.3.2.

The findings reported in this section were also published in Lilley et al. (2005a). The method employed in the study is presented next.

Method. A group of 113 Level 2 Computer Science undergraduates participated in a session of summative assessment using the CAT prototype developed for this research. The test-takers had 40 minutes to answer 24 questions within the subject domain. The questions were organised into 4 different HCI topics, namely 'Identifying needs and establishing requirements', 'Design, prototyping and construction', 'Implementation issues' and 'Evaluation paradigms and techniques'. The CAT assessment session took place in computer laboratories, under supervised conditions. A summary of test-taker performance in the CAT assessment session is provided below.

Feedback on CAT performance was provided using the automated feedback prototype. In order to elicit test-takers' views of the automated feedback provided by the prototype, they were required to rate a series of statements regarding using a five point Likert Scale. A copy of the questionnaire can be found in Appendix F.

Summary of test-taker performance. Table 8-1 shows the overall CAT proficiency level as well as the CAT proficiency level per topic area. The potential proficiency level values ranged from -3 (lowest) to +3 (highest). It can be seen from Table 8-1 that proficiency level means were all near zero.

Performance Indicator	Mean	Std. Dev.
Overall proficiency level	0.08	1.08
1. Identifying needs and establishing requirements	-0.04	1.83
2. Design, prototyping and construction	0.13	1.59
3. Implementation issues	-0.07	1.77
4. Evaluation paradigms and techniques	-0.26	1.94

Table 8-1: Summary of test-taker performance (N=113)

All test-takers received feedback on test performance via the automated feedback prototype described in section 7.3.2. Test-takers were then required to complete a questionnaire in which they rated a series of statements regarding the usefulness of the feedback using a Likert Scale from 1 (Strongly disagree) to 5 (Strongly agree). Their views are summarised next.

Findings. In order to investigate test-takers' attitude towards the automated feedback approach, all test-takers were asked to complete a questionnaire, which was completed by 97 out of the 113 test-takers. Their responses are summarised in Table 8-2.

It can be seen from Table 8-2 that the automated feedback generated by the software application was positively received by the test-takers who participated in this study. The effectiveness of the application in providing feedback on performance scored the highest (mean=3.99), followed by the effectiveness of the application in providing helpful advice for individual development (mean=3.93). Interestingly, the usefulness of the "Overall Score" section at providing information on how successfully test-takers had learned scored the lowest (mean=3.68) and indeed lower than the two other feedback sections. This is an important result, as it adds to the findings from the academic staff evaluation reported in Chapter 5, in that a CAT proficiency level estimate alone would not be sufficient to provide students with informative feedback on assessment performance.

Statement	Strongly disagree 1	2	3	4	Strongly agree 5	Mean	Std. Dev.
Overall, the feedback application was effective at providing helpful advice for individual development.	4	5	15	43	30	3.93	1.02
Overall, the feedback application was effective at providing feedback on performance.	4	4	13	44	32	3.99	1.01
The “Overall Score” section was useful at providing information on how successfully I have learned.	6	9	23	31	28	3.68	1.17
The “Performance Summary per Topic” was useful at providing information on how successfully I have learned in each topic area.	6	6	19	34	32	3.82	1.15
The "Step-by-Step Personalised Revision Plan" was useful at providing information on how successfully I have learned.	8	9	14	35	31	3.74	1.24
The content of the feedback was appropriate for my individual performance.	6	6	20	39	26	3.75	1.11

Table 8-2: Test-taker attitude towards the automated feedback provided (N=97)

Building on the findings from this study, further investigation concerning test-takers' attitude towards the automated feedback approach was carried out. In particular, the issue of whether performance on the test did affect the perceived usefulness of the automated feedback merited consideration and this is the focus of the next section.

8.1.2 According to performance

The empirical study reported here investigates the perceived usefulness of the feedback provided by the automated feedback prototype according to performance. It should be noted that the findings from this study were also published in Lilley & Barker (2005a).

Method. A group of 188 Level 2 Computer Science undergraduate students participated in a summative assessment session using the CAT application developed for this research. The assessment session took place in computer laboratories, under supervised conditions. The participants had 40 minutes to answer 30 questions within the Visual Basic.NET domain. The questions were organised into the following 5 topics: 'ADO.NET and SQL', 'Classes and Controls', 'Representing data: Variables and Constants', 'Functions and Expressions' and 'Program Flow'. A summary of test-taker performance in the CAT assessment session is provided next.

Feedback on performance was generated using the web-based automated feedback software prototype developed for this research, as described in section 7.3.2. Immediately after the feedback on performance was made available, test-takers were required to rate a series of statements regarding the usefulness of the feedback provided using a 1-5 Likert Scale.

Summary of test-taker performance. Table 8-3 shows the overall CAT proficiency level as well as the CAT proficiency level per topic area. The potential proficiency level values ranged from -3 (lowest) to +3 (highest).

Topic	Mean	Std. Dev.
Overall proficiency level	0.37	1.07
1. ADO.NET and SQL	-0.12	1.75
2. Classes and Controls	0.64	1.40
3. Representing data: Variables and Constants	-.067	1.85
4. Functions and Expressions	0.31	1.71
5. Program Flow	0.33	1.70

Table 8-3: Summary of test-taker performance (N=188)

The objective of the automated feedback software prototype was to provide test-takers with timely feedback that was useful for individual development. In order to investigate test-takers' attitude towards the automated feedback

approach employed, 80 volunteers from the original group were asked to rate the usefulness of the feedback application from 1 (Not Useful) to 5 (Very Useful). Their responses are summarised in the next section. A copy of the questionnaire can be found in Appendix G.

Findings. Table 8-4 summarises test-takers' ratings. It can be seen from Table 8-4 that the automated feedback generated by the software application was perceived as being useful by the test-takers who participated in this study. One participant reported on the increased need for automated feedback that is as useful as the one used in the study, given the reduction in the number of face-to-face sessions.

Not Useful 1	2	Useful 3	4	Very Useful 5	Mean	Std Dev
0	1	14	31	34	4.22	0.78

Table 8-4: Usefulness of the feedback application as perceived by the participants (N=80)

One of the aims of this study was to investigate if the correlation between test-takers' performance and their perceptions of the usefulness of the automated feedback generated by the software application was statistically significant. To this end, test-takers' results and their perception of the usefulness of the web-based feedback prototype were subjected to a Spearman's rank order correlation. No statistically significant correlation was found between test-takers' proficiency levels and the test's difficulty rating, such as rs = 0.034, Sig. (2-tailed) = 0.762, N=80.

In addition, test-takers were ranked and assigned to one of three groups – namely 'low', 'average' and 'high' performing – on the basis of their performance in the test. Kruskal-Wallis test procedures were carried out in order to determine whether there were significant differences between the usefulness ratings for each of the three groups that could be ascribed to performance on the test.

Table 8-5 shows the mean ranks of the Kruskal-Wallis test. The Kruskal-Wallis test showed that there was no significant difference in the usefulness of the feedback that could be ascribed to the effect of test-takers' performance on the test (Chi-Square = 0.353, df = 2, Asymp. Sig. = 0.838).

Group	N	Mean Rank
Low performing	24	40.79
Average performing	28	38.68
High performing	28	42.07

Table 8-5: Kruskal-Wallis test mean rank results: summative assessment (N=80)

Reactions from test-takers. In addition to rating the usefulness of the feedback, test-takers were able to add free text comments. The examples below provide a sample of test-takers' views:

- "Very good - able to get your results when you want them, 24hrs / 7 days a week."
- "Good feedback with links to relevant topics."
- "Helps identify what you need to revise the least."
- "I am now aware the I struggle on functions and expressions, so I can delegate my time to improving my knowledge on this topic in particular."
- "I can find out where I am under achieving."
- "Indicates reading carefully the question, because some question I already knew but I answered wrong"
- "The individual performance table enables you to know which aspect of the VB you got correct and what you need to improve on."
- "You can receive results immediately and maintain your anonymity."

As can be seen from the list above, test-takers' reactions to the approach were positive. It should be noted that test-takers were not explicitly informed that the revision tasks were related to the questions they answered incorrectly in

the test; the webpage heading read: “Here is what you could do next”. The comments above suggest that most test-takers recognised the link between revision tasks and questions answered incorrectly, possibly because it was clear that the application was built to provide tailored feedback on test performance. Nonetheless, the comments from two test-takers suggest that not all were fully aware of the link between revision tasks and questions answered incorrectly:

- “Recommended points of revision... excellent source of information!!! (seems very relevant to my performance in the test). ”
- “Recommended points for revision. I am not sure if these are related to the questions you got wrong in the exam, if it is then this sort of feedback is awesome. “

Interestingly, a test-taker saw the revision plan as a useful tool to save time:

- “Further points of revision... extremely useful source of revision guide, save the student from having to find themselves!”

Although it is fair to say that some revision tasks contained links to resources, as opposed to requiring test-takers to find these resources by themselves, it is argued that this approach was more constructive and useful than simply providing test-takers with the correct answers. In addition, depending on the performance of the test-taker, many of the revision tasks were complex and required elements of independent research in order to be completed.

The study described in this section supports the study reported in section 8.1.1, as it is concerned with the correlation between test-takers’ CAT proficiency levels and their perceived usefulness of the automated feedback generated by the application. Furthermore, both studies are concerned with test-takers’ perceived usefulness in a summative assessment setting. In order to provide a complete picture of the usefulness of the automated feedback prototype, it was important to conduct a study in a formative assessment setting. This is the focus of the next section.

8.1.3 Formative assessment

It was of relevance to this work to examine test-takers' attitude towards the automated feedback based on the CAT approach in a formative assessment setting. To this end, an empirical study was carried out and is described in this section. The findings from this study were also published in "Students' perceived usefulness of formative feedback for a computer-adaptive test" (Lilley & Barker, 2006c). The method, summary of test-taker performance and findings are reported next.

Method. A group of 76 Level 2 Computer Science undergraduates participated in a formative assessment session using the CAT software prototype developed for this research as part of their regular assessment for a programming module. The participants had 40 minutes to answer 40 objective questions within the Visual Basic.NET subject domain. The questions were organised into five topic areas, namely 'Representing data', 'Classes and Controls', 'Functions and Procedures', 'Controlling program flow' and 'ADO.NET'.

Summary of test-taker performance. Table 8-6 shows the overall CAT proficiency level as well as the CAT proficiency level per topic area. The potential proficiency level values ranged from -3 (lowest) to +3 (highest).

Topic	Mean	Std. Dev.
Overall proficiency level	-0.03	1.02
1. Representing data	-0.121	1.54
2. Classes and Controls	-0.007	1.38
3. Functions and Procedures	-0.087	1.64
4. Controlling program flow	-0.31	1.61
5. ADO.NET	-0.02	1.47

Table 8-6: Summary of test-taker performance (N=76)

The 76 test-takers received feedback on performance using the automated feedback prototype described in section 7.3.2. The next stage of the study, which is described in the following section, was to obtain information about test-takers' reactions to the automated feedback prototype.

Findings. In order to investigate the perceived usefulness and ease of use of the automated feedback prototype, test-takers were required to complete a questionnaire in which they were asked to rate a series of statements using a Likert scale from 1 (Strongly disagree) to 5 (Strongly agree). A copy of the questionnaire employed in this study can be found in Appendix H.

Forty-nine out of 76 test-takers participated in the evaluation and their responses are summarised in Table 8-7. The results presented in Table 8-7 show that on average test-takers rated the automated feedback prototype as being able to provide feedback that was useful, capable of identifying a student's strengths and weaknesses as well as fast. In addition, the application was perceived as easy to use (mean=4.29).

Statement	Strongly disagree 1	2	Agree 3	4	Strongly disagree 5	Mean	Std. Dev.
The "Your Score" section would be useful at providing information on how successfully I have learned.	0	3	9	25	12	3.94	0.82
The "Your performance per topic area" diagram would be useful at providing information on how successfully I have learned.	0	3	8	25	13	3.98	0.82
The "Step-by-Step Personalised Revision Plan" section would be useful at providing feedback for individual development.	0	2	10	18	19	4.10	0.87
Using the application would enable me to receive feedback on performance more quickly.	0	5	10	12	22	4.04	1.04
Using the application would be effective in identifying my strengths and weaknesses.	0	1	12	15	21	4.14	0.86
I would find the application easy to use.	0	1	9	14	25	4.29	0.84

Table 8-7: Test-taker attitude towards the automated feedback provided (N=49)

Similarly to the study reported in section 8.1, the performance per topic area (mean=3.98) was considered by test-takers as a better indicator of how successfully they have learned than the score alone (mean=3.94). One can speculate that this is due to the fact that the former is divided into different topic areas, providing a clearer indication of what has been achieved. However, anecdotal evidence from test-takers suggests that – although they were unaware of the meaning of Bloom's taxonomy of cognitive skills (Bloom, 1956; Anderson & Krathwohl, 2001) – a contributing factor to the higher score is the possibility to gauge how well they have performed in comparison with other test-takers, see Figure 8-1 below.

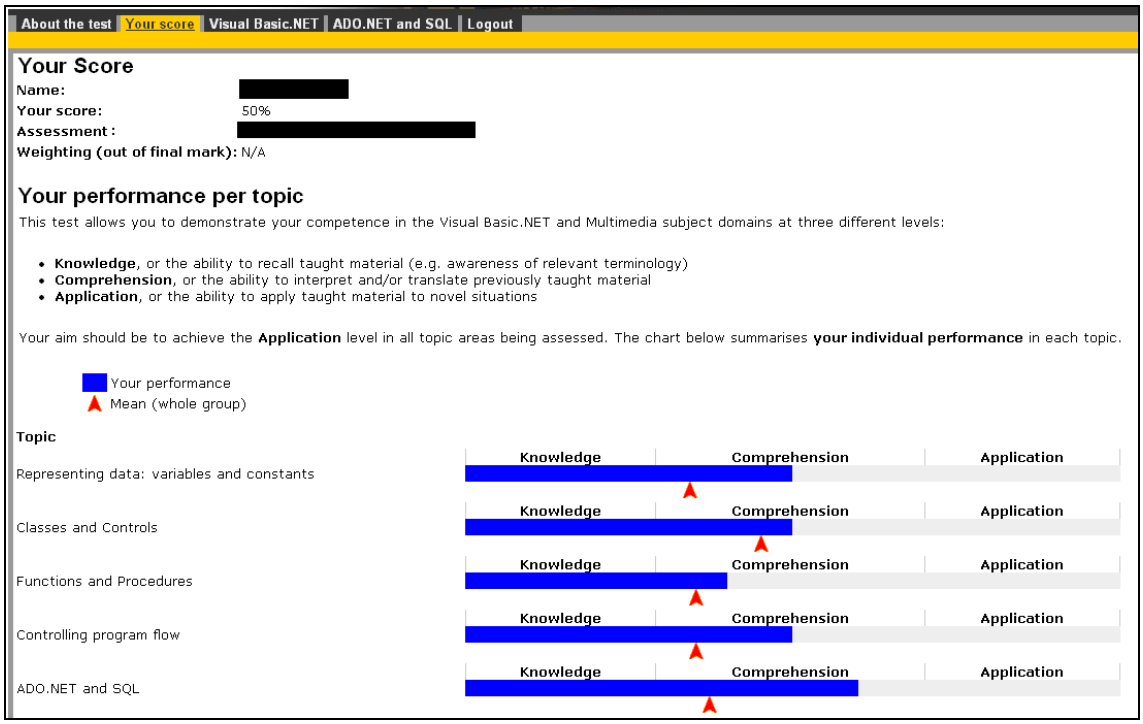

Figure 8-1: Automated feedback prototype
Screenshot illustrating the performance per topic section. The red arrow shows the group's mean, and therefore makes it possible to assess how well a particular test-taker has performed in comparison with their group.

An assumption of the study reported here was that test-taker attitude towards the feedback was independent of test performance. To test this assumption, test-takers' proficiency level estimates and the automated feedback prototype's usefulness ratings were subjected to a Spearman's rank order correlation. The results in Table 8-8 show that there is no statistically significant correlation between test-taker performance and perceived usefulness of the application. This was taken to indicate that test-taker attitude towards the feedback did not depend on test performance.

Statement	Proficiency Level	
The "Your Score" section would be useful at providing information on how successfully I have learned.	Correlation Coefficient	0.000
	Sig. (2-tailed)	0.998
The "Your performance per topic area" diagram would be useful at providing information on how successfully I have learned.	Correlation Coefficient	-0.065
	Sig. (2-tailed)	0.658
The "Step-by-Step Personalised Revision Plan" section would be useful at providing feedback for individual development.	Correlation Coefficient	0.110
	Sig. (2-tailed)	0.453
Using the application would enable me to receive feedback on performance more quickly.	Correlation Coefficient	0.129
	Sig. (2-tailed)	0.378
Using the application would be effective in identifying my strengths and weaknesses.	Correlation Coefficient	0.031
	Sig. (2-tailed)	0.834

Table 8-8: Spearman's rho correlation between perceived usefulness of the feedback provided and assessment performance (N=49)

In addition, test-takers were ranked and assigned to one of three groups – namely 'low', 'average' and 'high' performing – on the basis of their performance in the test. A Kruskal-Wallis test was carried out to assess the significance of any differences in attitude between these groups. The results of this statistical analysis are shown in Table 8-9 and Table 8-10 below.

Statement	Chi-Square	Asymp. Sig.
The "Your Score" section would be useful at providing information on how successfully I have learned.	0.235	0.889
The "Your performance per topic area" diagram would be useful at providing information on how successfully I have learned.	1.309	0.520
The "Step-by-Step Personalised Revision Plan" section would be useful at providing feedback for individual development.	0.924	0.630
Using the application would enable me to receive feedback on performance more quickly.	0.440	0.803
Using the application would be effective in identifying my strengths and weaknesses.	0.369	0.832

Table 8-9: Kruskal-Wallis test results: formative assessment (N=49, df=2)

Statement	Test-taker Performance	N	Mean Rank
The "Your Score" section would be useful at providing information on how successfully I have learned.	Low	17	25.44
	Average	18	25.69
	High	14	23.57
The "Your performance per topic area" diagram would be useful at providing information on how successfully I have learned	Low	17	26.35
	Average	18	26.36
	High	14	21.61
The "Step-by-Step Personalised Revision Plan" section would be useful at providing feedback for individual development	Low	17	22.47
	Average	18	26.28
	High	14	26.43
Using the application would enable me to receive feedback on performance more Quickly	Low	17	24.38
	Average	18	24.03
	High	14	27.00
Using the application would be effective in identifying my strengths and weaknesses	Low	17	23.65
	Average	18	26.39
	High	14	24.86

Table 8-10: Kruskal-Wallis test mean rank results: formative assessment (N=49)

It can be seen from Table 8-9 and Table 8-10 that no significant differences were found between the attitudes of test-takers with poor, average and high performances. These results support the view that the automated feedback

prototype was perceived as being useful, regardless of test-taker performance. This means that low performing test-takers rated the usefulness of the automated feedback prototype no differently than average and high performing test-takers. This was taken to indicate that the automated feedback proposed in this research was effective at providing feedback that was tailored to each individual test-taker.

It should be noted that in a summative assessment undertaken by the same group of test-takers on the same topic two weeks later, using the same CAT software and feedback application, the proficiency level mean for the summative assessment was 0.21 (SD=1.42, N=76). The mean performance was therefore higher in the summative assessment than that in the formative assessment shown above, (θ=-0.03, SD=1.02, N=76). A paired-samples t-test was used to examine any significant differences in the means for the proficiency level obtained for both assessment sessions. The results of this analysis showed that the observed differences between the proficiency level means were significant and that the differences could not be ascribed to chance alone (t = -2.112, df= 75, Sig. 2-tailed = 0.038). One can speculate possible reasons for this difference in performance between the formative and summative assessment sessions. It is likely that test-takers considered the formative assessment as a way of identifying strengths and weaknesses and providing them with information on which topics they should prepare for the summative assessment. In this case, it may be argued that the formative assessment had achieved its objectives, as performance was shown to be improved in the later summative test. It is also possible, of course, that test-takers were more likely to revise for a summative test than for a formative one. Another possibility is that test-takers adopt different strategies during the test and that they are more meticulous in their approach when taking summative tests.

The following section contains a summary of the findings reported in this chapter.

8.2 Summary

This chapter is concerned with test-takers' attitude towards and acceptance of the automated feedback software prototype developed for this research. This software tool aims to provide feedback on performance to test-takers of CATs. The automated feedback was made available via a web-based application, as described in section 7.3.2. The feedback consisted of three sections:

- overall score;
- performance topic per area;
- personalised revision plan.

In order to assess the attitude towards and acceptance of the automated feedback approach by test-takers, a series of empirical studies involving test-takers in a real educational setting were conducted. These studies were conducted in summative and formative assessment settings, as well as in two different subject domains.

The findings from these studies were taken to indicate that the automated feedback prototype was perceived as being easy to use. On average, test-takers found the automated feedback software prototype capable of providing feedback that was timely, useful for individual development and effective at identifying strengths and weaknesses.

The performance per topic area was perceived as more useful than the overall score at indicating how effectively test-takers have learned. This was an interesting result, especially when one takes into account that it corroborated the views of academic staff in that a score alone would not be sufficient to promote test-takers' individual development (see section 5.2).

The personalised revision plan was also valued by test-takers. The following examples illustrate test-takers' views on this section of the automated feedback (Lilley et al., 2004b):

- "I found it useful, gave me an idea of what to revise and work harder on."

- "I now know where I'm going wrong and know how to find out more about the areas of which I scored low marks."
- "I rated this as very useful this is because this does not only allows you check your results but this contain enough updated information on required main topics with useful information, where it can be very useful for revisions. "
- "This feedback page is good because it gives you an insight as to what questions you failed on. It also gives you links to pages that can help you with the questions you did not answer correctly. "
- "This is very useful. It is good to know the exact areas in which I need to work harder. "

The personalised revision plan consisted of a series of tasks that should be completed by the test-takers in order to improve their proficiency levels within the subject domain, as described in section 7.3.2. One can argue that the personalised revision plan was considered to be useful by the test-takers as a result of the application of the CAT approach to the provision of automated feedback. The application of this approach meant that the revision plan was tailored to each individual test-taker, and contained tasks that matched their proficiency levels within the subject domain.

Such a scenario, where individual proficiency levels are built into the design and delivery of feedback on assessment performance, would be difficult to obtain with traditional computer-based testing (CBT). This is because in a CBT the fixed set of questions to be administered during a session of assessment are typically selected in such a way that all ability levels, ranging from low to advanced, are included (Pritchett, 1999). Presenting test-takers with a set of tasks related to the questions answered incorrectly in the CBT could pose problems to individual test-takers. For example, low-performing test-takers might be presented with one or more tasks that are above their level of ability. Such tasks might lead to de-motivation, rather than enable test-takers to improve their proficiency levels within the subject domain. It should also be noted that, from the perspective of academic staff, when a high-

performing test-taker successfully completes a task that is far below his or her current ability (and therefore unchallenging), this provides little valuable information about this individual. Similarly, little valuable information is provided when a low-performing test-taker is unable to complete a revision task that is far above his or her current proficiency level within a given subject domain.

Statistical analysis of the correlations between test-takers' proficiency level estimates and their attitude towards the automated feedback showed no statistically significant correlations. This was a finding of great importance to the research, since it had been a concern that attitude to feedback was affected by performance on the assessment. Performing well or badly on an assessment might influence attitude to the feedback provided which could, in turn, introduce bias in the score. For instance, someone who performed poorly might be less impressed than someone who performed well. The lack of any relationship between performance and attitude supported the view that the feedback was acceptable to all test-takers irrespective of their proficiency level within the subject domain. As part of this work, test-takers were also ranked and assigned to one of three groups – namely 'low', 'average' and 'high' performing – on the basis of their performance in the test. Kruskal-Wallis tests were carried out to examine the significance of any differences in attitude between these groups. No significant differences were found between the attitudes of test-takers with poor, average and high performances. There was no significant effect of test-taker performance on perceived usefulness of the automated feedback approach.

The fact that the feedback was delivered via a web-based application was also valued by test-takers. Feedback on test performance could be accessed at any time from any location on or off campus; in addition, test-takers were able to use the application at their own pace.

All in all, test-takers exhibited a positive attitude towards and acceptance of the automated feedback, regardless of their proficiency level within the subject domain. Related work to support this view was published as part of this

research in Lilley et al. (2004b), Lilley & Barker (2005a), Lilley et al. (2005a), Lilley et al. (2005b), Lilley et al. (2005d), Lilley & Barker (2006c) and Lilley & Barker (2007). In the next chapter, reactions to the automated feedback prototype from the other main group of stakeholders, i.e. academic staff, are examined.

9. Academic staff evaluation of the automated feedback prototype

The evaluation of the automated feedback prototype was divided into two major stages. The first stage was concerned with the evaluation of the feedback by test-takers, and this is the focus of the previous chapter. In Chapter 8, it is shown that test-takers exhibited positive attitude towards the automated feedback prototype based on the computer-adaptive test (CAT) approach.

The second stage was concerned with the evaluation of the automated feedback by academic staff. As part of the evaluation of software for educational purposes, it is essential that views of academic staff are taken into account. To this end, three studies involving members of academic staff were conducted and are reported in this chapter. These studies aimed to gather qualitative data regarding academic staff attitude to the automated feedback approach used in the application developed for this research.

Each of the studies involved a presentation of the automated feedback prototype, including an overview of the feedback approach, examples of feedback output screens, research data related to test-takers' performance and attitude to the feedback provided. After each presentation, a semi-structured question and answer session was conducted, where the research

team and academic staff could exchange ideas. Sessions were led by an experienced facilitator and discussion topics were focused, based upon a previously prepared script. The sessions, however, were semi-structured, since open discussion was encouraged on any issue related to the relevant topics. Sessions were recorded on video and later transcribed in full by the researcher. The transcripts were analysed using QSR N6 software (QSR International, 2007), in order collate and link together themes and ideas.

This chapter is organised into three main sections: outline of the studies involving academic staff regarding the automated feedback prototype, a report on academic staff views of the automated feedback prototype and a summary of academic staff responses to the questionnaire.

9.1 Overview of the three studies conducted

This stage of the research comprised three studies involving members of academic staff in order to examine their attitude towards the automated feedback prototype.

The first study involved a group of 10 Computer Science lecturers, experts in systems and software design and implementation, with an interest in the provision of online educational systems. The study consisted of a 30 minute presentation followed by a 30 minute semi-structured discussion. A copy of the guidelines to the semi-structured section can be found in Appendix K.

The second study involved a group of 50 University lectures at an academic conference concerned with the use of a Managed Learning Environment (MLE). The study included a 25 minute presentation followed by a 5 minute question session and a short questionnaire. A copy of the questionnaire used in this session can be found in Appendix L.

A group of 20 experienced University lecturers interested in online and blended teaching learning participated in the third study, which comprised a 30 minute presentation followed by a 30 minute semi-structured discussion.

Appendix K contains a copy of the guidelines used during the semi-structured section.

The next section focuses on academic staff views of the automated feedback prototype, as gathered during the discussion sessions.

9.2 Findings from the discussion sessions

In all, three discussion sessions were employed as part of this study, based on methods described by Barker & Barker (2002). The focus of the second session was primarily to administer the questionnaire, which is presented later in section 9.3. As one would expect, there was little opportunity for discussion in the second session and therefore it contributed little to the collection of qualitative data.

The bulk of the qualitative data reported in this section was therefore collected during the first and third studies. As discussed above, 10 Computer Science experts participated in the first study and 20 University lecturers in the second. Data related to this study was also published in Barker & Lilley (2006).

In all sessions, after the presentation of the ideas underlying the automated feedback prototype, printed copies of screenshots of the actual feedback – where test-takers' names were omitted for anonymity reasons – were distributed for inspection. The discussion topics for the sessions are presented in Table 9-1.

Discussion topics
1. What are the most common feedback methods used at present?
2. How do you assess the quality of feedback provided at present?
3. What are the benefits and limitations of the feedback provided at present?
4. What is your view of the CAT approach for formative and summative assessment?
5. What is your opinion of the CAT approach to automated feedback?
6. What are the benefits and limitations of automated feedback based on the CAT approach?
7. How could the automated approach be improved?
8. What should be the role of the lecturer in the automated feedback system?
9. What is the need for monitoring and how might this be achieved? What, if any, are the ethical issues in the method?

Table 9-1: Discussion topics

In general terms, during these sessions one member of the research team played the role of presenter and another member of the team played the role of session facilitator. As soon as the presentation of the automated feedback prototype was concluded, the session facilitator introduced the semi-structured discussion session. This included a short scripted introduction, where the objectives of the discussion and ethical issues, such as confidentiality and the video recording, were described to the academic staff present.

In the first instance, the facilitator started the discussion session by asking the first question listed in Table 9-1, which is related to the type of feedback currently provided by academic staff. The level of discussion generated in the first and third studies was good. The session facilitator encouraged all present to engage in the discussion when possible in addition to checking that all the intended topics had been covered adequately. When discussion moved far from the focus, or sufficient time had been spent on a thread, new topics were introduced by the facilitator as unobtrusively as possible. Each of the items listed in Table 9-1 is discussed in greater detail next.

Feedback methods used at present. At present, feedback methods employed are mostly classroom and lecture theatre based sessions lasting approximately one hour. Such sessions are not tailored to individual students;

generally each question is worked through by the lecturer. In some cases, general problems identified by lecturers are covered in greater depth. If a question is well answered by most students, then less time is spent on this question. Problem questions are dealt with more fully by most lecturers. Other methods include providing only the questions and worked answers online (either through a web-based system, or by electronic mail). One lecturer was using a spreadsheet to attempt to individualise feedback, which amounted to personally typing in comments to the answer sheet for each student. For essay type questions, feedback was usually given as comments on the essay script, either written in pen or added electronically. At times feedback was provided in small group sessions where topics were discussed, rather than questions analysed in detail. One lecturer reported that she used one-to-one sessions to provide feedback on rare occasions. The feedback method appeared to be related to the type of test. For objective tests, most of the methods were employed, with the obvious exception of writing directly on scripts. The main purpose of feedback was to provide advice on individual development. Few reported providing feedback on summative assessments other than a final score.

Quality of feedback provided at present. Lecturers emphasised the necessity to be able to interact directly with students and, based upon experience, provide directed and tailored feedback. It was possible to "gauge" how a test had gone, and to provide the necessary feedback in an appropriate format. When pressed as to how this was possible, given large class sizes and the small amount of time devoted to feedback, some lecturers agreed that it was not always possible. The quality of feedback provided did indeed vary according to some lecturers and less experienced colleagues might on occasions provide feedback that was variable. When asked to think about the problems of high performing and very low performing students, most agreed that feedback was usually focused at "the average" student, with an account taken of general problems that appeared in the test itself. Several lecturers expressed the opinion that that the quality of the individualised automated feedback as proposed in the research was likely to be high, citing the

relationship between the performance per topic area (see section 7.3.2), Bloom's taxonomy of cognitive skills (Bloom, 1956; Anderson & Krathwohl, 2001) and, most importantly, the provision of remedial work tasks that are tailored to each individual test-taker. The fact that the automated feedback prototype was web-based was also valued by the lecturers in the discussion session, due to its increased availability.

Benefits and limitations of the feedback provided at present. The benefits of current methods for providing feedback (for example, in-class feedback sessions) can be summarised as the possibility of direct control and monitoring of test performance and feedback. Lecturers liked the ability to be able to "keep a finger on the pulse" when providing feedback. Some concern was expressed that an automated approach would lead to potential problems going un-noticed. This could not happen when lecturers themselves gave feedback. Feedback normally occurs some time after the test, ranging from six weeks to several months. Some lecturers, however, realised that un-timely feedback was far less useful than feedback provided quickly. One lecturer asked the reason for the delay between the CAT test and the release of the automated feedback as, in theory, it was possible to release the feedback immediately after the test. It was then explained that the delay between the CAT test and the release of the automated feedback was introduced in order to allow the research team to verify that the automated feedback was generated correctly. This was necessary due to technical and ethical reasons. In future, it would be possible to release the feedback without verification. Most lecturers concurred in that the speed of the automated feedback was a major benefit.

The CAT approach for formative and summative assessment. The CAT approach was not the main focus of the discussion, as academic staff attitude to it had been the subject of earlier studies (see Chapter 5). It was important, however, to discuss the CAT in context of the feedback. Most academic staff were familiar with the CAT approach, as it has been in use in the University for several years. Benefits of a CAT in terms of efficiency, motivation and potential prevention of collusion were already well known. Some lecturers who participated in the discussion suggested that the automated feedback

prototype could also be used for non-adaptive tests. Although this is somewhat true, many agreed that this would not necessarily be the most effective way of employing the automated feedback prototype as the ability to tailor the revision tasks to the proficiency level of individual test-takers would be severely hindered. The use of the CAT approach in a summative assessment setting was generally less well received than for a formative setting, which was in accordance with earlier findings reported in section 5.2 and the questionnaire data summarised later in section 9.3. It was noted by one lecturer, however, that the use of a CAT for summative assessment would facilitate the provision of timely feedback to all test-takers.

The CAT approach to automated feedback. As part of the discussions, it was realised that the use of automated feedback was an important benefit of the CAT approach. Although some lecturers wanted to discuss the CAT approach in greater detail, this was resisted by the session facilitator. Some lecturers expressed the fact that they realised that individual student profiles obtained from a CAT, containing information on performance in topic areas, as well as cognitive levels could be used in a variety of different ways. It was noted that the use of a CAT in automated feedback involves two issues that were closely linked in the study, a CAT and automated feedback. It was the belief of the research team, expressed in the presentation, that a CAT was essential to provide individualised and rich automated feedback. It is fair to say that some lecturers were not entirely convinced of this link.

Benefits and limitations of automated feedback based on the CAT approach. The most important benefit cited was the speed of feedback possible with the automated approach introduced here. The most important limitation identified during the sessions was related to the loss of control by lecturers. Providing automated feedback was liable to remove an important "human aspect" of the lecturer's role. Other limitations expressed related to the use of objective testing as the only method with the approach and to issues related more to the CAT approach than the feedback itself. Other potential benefits cited included the motivational aspects of CAT and how this might be used in order to help students do extra work, either remedially, or as extra

challenges. This was seen as an important aspect by some lecturers. It was emphasised in the presentation prior to discussion that the CAT level obtained represented an important boundary for an individual between what they knew and what they did not know. Providing feedback at this boundary was important and this view was expressed by some lecturers present at both sessions. Efficiency of the method was also cited as a benefit. Providing feedback in traditional ways, such as during tutorial sessions, was difficult and often slow. An automated system, once in operation could process test results efficiently with the minimum of human intervention. Once the bulk of the implementation had been completed, the generation of feedback using the application should not make excessive demands on lecturer time. Admittedly some lecturers saw the reduction in human intervention as a disadvantage, though this view was in the minority at both sessions.

Suggested improvements of the automated feedback. There were a few suggested improvements to the system. One lecturer expressed the opinion that the CAT feedback might be used as the focus point for either group seminars or small remedial classes. It would be possible to obtain useful summaries of strong and weak points in the tests in each topic area from the CAT. Such summaries might be useful to lecturers in their teaching and for providing remedial materials or planning lectures. The speed of the CAT would be likely to provide such information quickly and certainly in time for action. Patterns of feedback might be identified in this way and the item database could be analysed to identify problem areas (and areas of strength) in all topics.

The role of the lecturer in the automated feedback. It is fair to say that a concern of some lecturers was that automated feedback was another step on the road to an uncertain impersonal future. This was rarely expressed in an open way, though it was apparent from some questions that it was a concern. Others expressed the view that there was an opportunity in the approach to develop useful systems that would provide lecturers with more time to develop interesting online and off-computer activities related to the outcome of tests, for example activities related to performance on tests. One lecturer suggested

that tests could be developed where feedback could be directly incorporated into the CAT and that this might provide a learning opportunity within a CAT. Although beyond the scope of the research, this was an interesting idea for future applications of the CAT approach.

Monitoring of the CAT automated feedback. The approach to making sure students were not disadvantaged either by the CAT approach as proposed in this research or by the way automated feedback was generated by the automated feedback prototype included statistical analysis of test-takers' performance (see for example section 4.3) and manual verification by the main researcher of the automated feedback generated. No lecturer expressed the feeling that students would be disadvantaged either by the CAT or by the method of providing feedback as proposed in the research. Most stated the view that it would be important to monitor the CAT and feedback systems to ensure that they were performing properly and fairly. One lecturer suggested a method of sampling, both for CAT results and feedback to ensure fairness.

For ease of reading, Table 9-2 (p. 187) and Table 9-3 (p. 188) provide a summary of the findings listed above.

All in all, academic staff present at these sessions raised interesting points about the automated feedback prototype that proved to be common concerns. For instance, there were concerns over the retention of control by academic staff of the feedback provided to students, should feedback to students be generated by a software application. Interestingly, academic staff appeared to pay little attention to the fact that the remedial tasks are still devised by members of the teaching team. It is the selection of tasks to individual test-takers that is performed by the automated feedback prototype.

In spite of their aims, many attendees reported that feedback on assessment performance is often limited to the provision of a score and opportunities to provide advice on how students can improve are rare or not tailored to individuals. The majority of the attendees recognised the value of the provision of automated feedback and, in particular, of the approach suggested by the research. Issues such as the perceived usefulness of the approach by

academic staff were also investigated through a questionnaire, and this is the focus of the next section.

Feedback methods used	• face-to-face feedback to the whole group (lecture); • face-to-face feedback to small groups of students; • face-to-face feedback to individual students; • provision of model answers; • written comments on essays; • electronic feedback, such as computer-delivered feedback (e.g. web-based applications, email) and spreadsheets containing feedback sentences; • the main aim of feedback is to provide students with useful advice for individual development; • in the case of summative assessment, it is not uncommon for the feedback to be limited to an overall score.
Quality of feedback provided	• quality of the feedback is variable; • it might take from a few weeks to several months for feedback to be available to students.
Benefits of current approaches to feedback	• one of the major benefits of current methods is that academic staff have direct control over the feedback provided; • it is feared that automated feedback methods can lead to potential problems going un-noticed.
Limitations of current approaches to feedback	• as a result of increased student to staff ratios, it is not always possible to provide timely feedback; • although the aim is to provide feedback that is tailored to individual students, feedback tends to focus on the "average student".
CAT as an assessment tool	• in the study, participants were familiar with the CAT assessment format and its expected benefits (e.g. improved efficiency, increased student motivation and potential prevention of collusion); • the approach is more likely to be favourably received in a formative assessment setting than in a summative one.

Table 9-2: Summary of discussion topics 1-4 (see Table 9-1 for a list of discussion topics)

The CAT approach to automated feedback	• participants' views were mixed; • some agreed with the view of the research team in that the CAT approach supports the provision of automated feedback that is tailored to each individual student; • others were uncertain of the added benefits of the CAT approach as suggested by the research team.
Benefits of automated feedback based on the CAT approach	• provision of timely feedback; • potential increased student motivation, as feedback tasks are within students' grasp; • increased efficiency, due to the use of automated methods for the provision of feedback; • increased availability, as feedback is made available via a web-based application.
Limitations of automated feedback based on the CAT approach	• potential removal of "human aspect" in student feedback; • lack of academic staff control over the feedback provided to students.
Suggested improvements of the automated feedback	• to use the CAT feedback as the focus point for face-to-face feedback to small groups of students; • to generate reports to help academic staff to evaluate student learning and their own teaching; • to incorporate feedback directly into the CAT.
The role of the lecturer in automated feedback	• some lecturers may feel that they have lost control over aspects of the feedback process; • others may find the use of automated feedback beneficial (and somewhat liberating).
Monitoring of the automated feedback	• in the study, participants agreed that students were unlikely to be disadvantaged by the approach to automated feedback proposed in this research; • current monitoring techniques include statistical analysis of test-taker performance and manual verification of the automated feedback generated by the main researcher; • in the study, participants suggested that the automated feedback generated by the prototype should be monitored, with 'sampling' being cited as a possible solution.

Table 9-3: Summary of discussion topics 5-9 (see Table 9-1 for a list of discussion topics)

9.3 Questionnaire responses

As part of the second study all 50 members of academic staff who attended the conference presentation were asked to complete a short questionnaire. The questionnaire was completed by 19 members of staff. Data related to this study was also published as part of Lilley et al. (2005a) and Barker & Lilley (2006).

The questionnaire was organised into two sections. In the first section, the respondents were asked to rate statements regarding the usefulness of the automated feedback approach using a 1-5 Likert scale. Their responses are summarised in Table 9-4.

Statement	Not useful 1	2	Useful 3	4	Very useful 5	Mean	Std. Dev.
In the context of summative assessment, the automated feedback approach that I have just seen is:	1	1	10	1	6	3.53	1.17
In the context of formative assessment, the automated feedback approach that I have just seen is:	0	0	8	3	8	4.00	0.94
In the context of objective testing (i.e. multiple-choice questions), the automated feedback approach that I have just seen is:	0	1	7	2	9	4.00	1.05
In the context of written assignments, the automated feedback approach that I have just seen is:	6	5	5	0	3	2.42	1.39

Table 9-4: Academic staff perceived usefulness of the automated feedback prototype (N=19)

In the second section of the questionnaire, the respondents were asked to rate statements regarding the speed, accuracy and appropriateness of the automated feedback approach using a 1-5 Likert scale. Their responses are summarised in Table 9-5.

Question	Poor		Good		Very good	Mean	Std. Dev.
	1	2	3	4	5		
With regards to its *speed*, the automated feedback approach that I have just seen is:	0	0	4	3	12	4.42	0.84
With regards to its *quality*, the automated feedback approach that I have just seen is:	1	1	8	4	5	3.58	1.12
With regards to its *appropriateness* to enhance students' learning experience, the automated feedback approach that I have just seen is:	1	0	6	4	8	3.95	1.13

Table 9-5: Academic staff perceived speed, quality and appropriateness of the automated feedback provided by the prototype (N=19)

As can be seen from Table 9-4 and Table 9-5, in general academic staff considered the automated approach to be a useful method for the provision of feedback. This was an important finding, since it was of crucial importance to the research that academic staff as well as test-takers valued the automated feedback approach. Table 9-4 shows that the automated feedback was valued more highly in the context of formative, rather than summative, assessment. The use of the automated feedback for written assignments was considered the least useful. It was not clear whether this was because of the difficulty of providing automated feedback for written work, or because academic staff felt that providing feedback themselves was a better approach. Table 9-5 shows that, on average, academic staff thought the automated approach to be fast, appropriate and of good quality, though the quality dimension achieved the lowest mean score. All in all, academic staff exhibited a favourable attitude towards the automated feedback approach developed for the research.

The following section presents a summary of the chapter.

9.4 Summary

A major problem for most – if not all – educational institutions is the provision of feedback that is timely and useful to individual students. This has become increasingly difficult due to growing student to staff ratios. Feedback on performance is often limited to the provision of an overall score. Advice on how to improve is provided in various formats, including group feedback sessions and the provision of electronic copies of worked examples. Such feedback methods, although useful, are mostly designed to address the needs of the "average student". Opportunities for tailored feedback, such as traditional face-to-face feedback sessions with a lecturer, are exceedingly rare especially in those modules that attract large numbers of students.

In this work it was found that when feedback on how students can improve is provided, it is often limited to generic worked examples and a list of questions answered correctly and incorrectly. Other approaches to the provision of feedback to groups of students, such as in-class sessions where all questions from an objective test are presented by a member of academic staff, are likely to remain as important feedback methods. Such in-class approaches offer high quality information about the test and each of the questions, often providing students with an opportunity to work through the questions. They do not, however, address the individual needs of many of the students. Explaining a question that is set at a difficulty level that is too low for most students will not be of interest for the majority of the group. Similarly, it can be argued that discussing questions that only one or two students are capable of answering will not be the most efficient way of employing academic staff and student time. It can also be argued that, in order to make feedback more useful, it has to be tailored for each individual student. There is also the issue that such in-class sessions are of no benefit to those students who, for whatever reason, miss these sessions.

Significant efforts have been made as part of this research to develop and implement an alternative feedback method based on the CAT approach. These efforts were produced in the light of an increased demand for the

development of software applications that would enable the provision of timely and tailored feedback, especially to those students who are assessed via computer-aided assessment applications.

The automated feedback prototype developed for this research proved to be of value, allowing individual test-takers to receive useful advice for individual development (see Chapter 8). Barker & Barker (2002) noted the importance of all major stakeholders in design, implementation and evaluation of projects related to the use of technology in teaching and learning. For this reason, it was important to also consider the views and attitudes of academic staff to the provision of automated feedback based on the CAT approach. The three studies described in this chapter were carried out in order to obtain detailed views and suggestions related to the automated feedback prototype. These studies comprised a questionnaire and focused discussion sessions.

Data gathered via the questionnaire suggests that academic staff perceived the automated feedback prototype as being capable of providing timely and useful feedback. During the sessions, a complex range of issues related to the provision of automated feedback were discussed. Lecturers were able to explore a range of topics related to how feedback was currently provided by themselves and colleagues and compare with the way in which feedback was provided in this work. In general, the automated feedback approach proposed as part of this research was well received and lecturers were receptive to the ideas that underpinned the work. In addition, lecturers were able to appreciate the potential benefits in terms of speed and efficiency, as well as the ability to personalise feedback at a time when online learning is becoming increasingly important in Higher Education, and staff time for providing individual feedback is decreasing.

Concerns related to the provision of automated feedback were general in nature, rather than specifically directed at the system developed as part of this work. These concerns tended to be focused on the loss of human input into the feedback process. There was no evidence from these sessions that feedback currently provided by lecturers was of a universally high standard or

that it was individualised. In fact, there was exceedingly little evidence that any form of individualised feedback is taking place as a matter of course. Subsequent analysis of the sessions using qualitative data analysis methods showed that lecturers in general were receptive to the idea of generating automated feedback based on the CAT approach.

Academic staff recognised that in order to enhance engagement and motivation, students require feedback that is individual, timely and meaningful. In addition, any remedial tasks should be well chosen, challenging and relevant. The views expressed by some lecturers suggest that they would like to be in control when choosing such tasks, but they acknowledged that this is not always possible due to increasing student numbers.

Some academic staff who participated in the discussion session introduced in section 9.2 suggested that the automated feedback prototype could also be employed in non-adaptive test assessment settings. Although it is fair to say that the basic engine of the prototype could be used to provide students with feedback on traditional non-adaptive tests, such use of the automated feedback prototype would present two important limitations. First, it would not be possible to provide test-takers with feedback on performance per topic area according to Bloom's taxonomy of cognitive skills. Second, the revision tasks would not be tailored to the proficiency level of individual test-takers. In Chapter 8, it was shown that the automated feedback prototype in an adaptive assessment setting was effective at providing individual test-takers with tailored feedback, and also that this approach was valued by test-takers in general.

All in all, academic staff involved in the studies introduced here recognised that the automated feedback approach as proposed in the research is useful, and that the combination of the CAT approach with the automated feedback prototype is capable of generating individual feedback that promotes learning.

The following chapter presents the conclusions drawn from the findings reported in the current and previous chapters.

10. Conclusion

This chapter is organised into three main sections. The first section provides a summary of the research. The second section highlights the outcomes of the research, including answers to the two research questions introduced in section 1.1. In the final section, possible directions for future work are presented.

10.1 Summary of the research

Literature in the field of student assessment in Higher Education points towards an increased use of computer-assisted applications (CAA) (Conole & Bull, 2002; Joy et al., 2002; Warburton & Conole, 2003; Bull & McKenna, 2004; Warburton & Conole, 2004). With the expansion of the use of CAA applications, the need to consider a broader range of assessment methods has also increased. Indeed Joy et al. (2002), Brusilovsky (2004), Challis (2005) and others have argued that there has been a demand for interactive CAA applications that dynamically adapt to their users, such as computer-adaptive tests (CATs).

CATs are a type of CAA in which a computer algorithm dynamically selects the items (i.e. questions) to be administered to individual test-takers according to their performance during the test. The CAT approach originates from the

assumption that very little is learned about a test-taker's ability if the questions presented during an assessment session are either too difficult or too easy for that individual. Hence, CAA applications developed using the CAT approach aim to present test-takers with questions that match their abilities within the subject domain.

CATs are typically based on Item Response Theory (IRT) (Lord, 1980). IRT is a general statistical theory that relates characteristics of items (item parameters) and characteristics of individuals (latent traits) to the probability of a correct response. The CAT literature is dominated by work relating to the psychometric properties of various IRT models; reports on practical applications of the CAT approach in Higher Education environments are rare. Moreover, little research has been carried out that investigates academic staff and test-taker attitudes towards the CAT approach in such environments.

The work presented in this thesis was undertaken to add to the knowledge base of CAT, by increasing the understanding of the fundamental issues and concerns relating to the appropriate use of the CAT approach in Higher Education environments, in particular for the assessment of and provision of feedback to Computer Science undergraduates.

The research can be divided into two main phases. The first phase is concerned with the design, implementation and evaluation of the CAT software prototype. The second phase relates to the design, implementation and evaluation of an automated feedback software prototype based on the CAT approach.

10.1.1 CAT prototype

In this work, the CAT high-fidelity prototype was initially built based on ideas drawn from the literature, in particular Lord (1980), Wainer (2000a), Wainer (2000b), Wainer & Mislevy (2000), Wolfe et al. (2001) and Guzmán et al. (2005). The prototype was evaluated and refined based on feedback from the

two main groups of stakeholders, namely test-takers (students) and academic staff.

The CAT software prototype developed as part of this work was subjected to a series of empirical studies, concerned with: (1) database calibration, (2) stopping conditions, (3) the effect of item review, (4) usability, (5) test-taker attitude, (6) the level of difficulty of a CAT as perceived by test-takers, (7) academic staff attitude, and (8) validity and reliability.

10.1.1.1 Database calibration

The pedagogical experience of the research team guided the construction of the item database. Accurate estimation of item parameters (i.e. difficulty b, discrimination a, and pseudo-chance c) is vital in the implementation of the CAT approach; however, this can often be an expensive and cumbersome process. In this work, a combination of expert calibration and the marginal maximum likelihood (MML) item parameter estimation (Gierl & Ackerman, 1996) method was employed to calibrate the item (i.e. question) database as follows:

- expert calibration was used for newly-written items (i.e. items with no historical data);
- test-taker responses and the MML parameter estimation method, as estimated by the XCalibre software application (Assessment Systems Corporation, 2007), were employed to recalibrate existing items.

The expert item calibration, as implemented in this work, was based on Bloom's taxonomy of cognitive skills (Bloom, 1956; Anderson & Krathwohl, 2001) and difficulty within the subject domain. Bloom's taxonomy of cognitive skills is well explored in the literature as a tool to classify objective questions; nevertheless, this particular use of the taxonomy had not been reported prior to this research. Section 3.2.2 presents the findings from an empirical study where the initial expert calibration and subsequent MML calibrations were subjected to statistical analysis. This approach to item calibration was found to

be useful for smaller applications of the CAT approach, such as the one reported in this work.

10.1.1.2 Stopping conditions

The CAT software prototype built as part of the research was of fixed-length, and its stopping condition was a combination of the number of questions answered and the elapsed time. In section 3.5.2, the standard error for the proficiency level estimate was investigated as a possible stopping condition for a variable-length version of the CAT software prototype. Findings from this study show that the standard error for the proficiency level estimate decreases as the number of questions increases. The standard error for the proficiency level estimate was found to be a valid stopping condition, with 15-16 questions required to achieve an acceptable accuracy in proficiency level estimates.

An important outcome of this research was an understanding of test-taker attitudes towards different stopping conditions. A variable-length CAT could lead to tests of different lengths; this is because the target standard error for the proficiency level estimate could be achieved after 14 questions for one test-taker and 19 questions for another. The main benefit of such a variable-length approach would be higher test efficiency, and lower exposure of the questions in the database.

Although the participants of the focus group study reported in section 3.5.3 appeared to understand the ideas underpinning different stopping conditions, their attitude towards different stopping conditions varied considerably. On the one hand, findings from the focus group study suggest that test-takers would exhibit a positive attitude towards variable-length CATs in a formative assessment setting.

On the other hand, focus group participants suggested a combination of number of questions and time limit as the most suitable stopping condition in a summative assessment setting. Perhaps not surprisingly, test-takers would only favour CATs of variable-length in summative assessment for high-performing test-takers. It was a common belief among the focus group

participants that test-takers who started the test displaying a poor performance should not be subjected to a variable-length CAT, so they would be provided with an opportunity to improve and achieve a higher score. This common belief, however, was shown to be mistaken and evidence to this effect provided in section 3.5.2.

Much of the CAT literature focuses on the merits of variable-length CATs, and how these can lead to more efficient testing by making it possible to achieve accurate proficiency level estimates with shorter tests (see for example Jacobson, 1993; Carlson, 1994; Wainer 200a; Wainer 200b). Prior to this research, there was a lack of compelling evidence on which stopping condition is most suited in a Higher Education environment. An important outcome of this research was to show that variable-length CATs are less suitable than fixed-length ones for the assessment of Computer Science undergraduates. There are two main reasons for this. First, variable-length CATs might lead to questions of equality and fairness and affect the face validity of the assessment in an adverse way. Second, shorter tests might mean that not all intended learning outcomes are assessed, which could have a detrimental effect on both content and face validities. In this work, it was found that the most suitable stopping condition was a combination of the number of questions answered and the elapsed time (see section 3.5).

An important assumption of this work was that the CAT software prototype should behave in the same way in formative and summative assessment settings. This is because it was assumed that formative assessment sessions were not only important for pedagogical reasons (shorter tests, as predicted in a variable-length CAT, could jeopardise syllabus coverage), but also to provide test-takers with additional opportunities to get familiar with the software (a form of “rehearsal”).

In short, even though section 3.5.2 presented evidence that the accuracy of proficiency level estimates are unaffected by different stopping conditions, findings from the study reported in section 3.5.3 indicated that different stopping conditions may influence test-takers’ reactions to the CAT approach

as well as affect syllabus coverage in an adverse way. Hence, the stopping condition employed in this work was a combination of the number of questions answered and the elapsed time.

10.1.1.3 Effect of item review

In a CAT, test-takers are not typically permitted to go back to and modify previously entered responses. The reasons for this range from the potential to obtain artificially inflated scores, reduced testing efficiency, to added complexity in the item selection algorithm. However, participants in the focus group session reported in section 3.5.3, indicated that they would value the opportunity to go back to and alter previously entered responses. Similar concerns were expressed by participants in studies conducted by Vispoel et al. (2000), Olea et al. (2000) and Vicino & Moreno (2001). Linz et al. (1992: p. 34) warn that CAT test-takers "feel at a disadvantage when they cannot review and alter their responses".

The effect of reviewing items and altering responses on proficiency level estimates was explored as part of this work. Test-takers were allowed to return to previous responses immediately after all questions had been answered, and their responses pre- and post-review were subjected to statistical analysis. Evidence was provided in section 3.6 to support the view that the option to return to previous items and alter responses as implemented in this work, had no adverse effect on proficiency level estimates, and contributed towards a reduction in test-takers' anxiety.

10.1.1.4 Usability

The importance of good interface design has been stressed by Preece et al. (1994), Boyle (1997), Preece et al. (2002) and others. In this work, it was assumed that a poor interface design could hinder student performance on the test. The ten general principles for user interface design developed by Nielsen

(2005) were found useful in guiding the design of the interface of the CAT software prototype.

Findings from the observation, focus group and interview studies reported in Chapter 4 showed that the user interface of the CAT software prototype developed for this research was unlikely to affect test-takers' performance in an adverse way. Findings from the heuristic evaluation presented in section 5.1, and the pedagogical evaluation reported in section 5.2 also support this view.

10.1.1.5 Test-taker attitude

In this work, test-taker attitude towards the CAT approach was examined.

Sections 4.2 and 4.3.1 present evidence that test-takers in general exhibited a positive attitude towards the CAT approach. This finding was very important in the context of the overall research project, since it supported the use of the CAT approach in the assessment of Computer Science undergraduates.

In a CAT, each person takes a test that is tailored to his or her proficiency level within the subject domain. Tailored testing was valued by the participants in the studies reported in sections 4.2 and 4.3.1, as they felt that they were challenged by test items at an appropriate level, rather than discouraged by items that are far above or below their proficiency levels. In addition, section 4.2 presents evidence to support the idea that tailored testing is likely to lead to increased levels of test-taker motivation.

As a result of the tailored testing afforded by the CAT approach, test-takers will be presented with different sets of questions. Interestingly, no evidence was found to suggest that test-takers reacted negatively to different sets of questions. In this work, it is argued that the lack of a negative reaction can be ascribed to three factors. First, the scoring method used in the CAT approach takes into account the number of correct responses and the level of difficulty of the questions. As a result, a test-taker's proficiency level estimate will be higher if he or she correctly answers more difficult questions. Second, some

participants were already familiar with the concept of a CAT, based on previous exposure to this assessment format (for example, in the form of TOEFL). Third, it is not uncommon for academic staff to set up objective tests where test-takers are required to answer different question sets.

Other findings from the focus group study reported in section 4.2 include the participants' views that objective questions are a fair assessment method, and that each assessment method has positive and negative aspects. Thus, in a summative assessment setting, they would favour a balance amongst different assessment methods, including objective tests (for example, CAT), written exams and coursework.

Evidence was found that differences in the number of questions administered (and hence stopping conditions) could affect test-taker attitude towards the CAT approach, and these are discussed in section 3.5. In this work, it was also found that preventing test-takers from navigating freely within a test to review and alter responses was likely to produce negative reactions from test-takers, and this is discussed in sections 3.6 and 4.2.

10.1.1.6 Perceived level of difficulty

Tailored testing or, in other words, the ability to dynamically match the level of the difficulty of the question to the proficiency level of a test-taker is a well known benefit of the CAT approach.

As part of this research, the level of difficulties of a CAT and a linear computer-based test (CBT) were compared. Sections 4.1 and 4.2 present evidence that test-takers in general found the level of difficulty of the CAT component more likely to be "just right" than the CBT component of the test.

Two further empirical studies were conducted to examine the perceived level of difficulty of the CAT approach, and are described in section 4.3. The first study was carried out in a summative assessment setting, and the second in a formative one. In both cases, test-takers were asked to rate the level of difficulty of the test they had just taken using a five point Likert scale. For each

test, test-taker performance and level of difficulty ratings were subjected to statistical analysis. In both studies, no statistically significant difference in the perceived level of difficulty that could be ascribed to the effect of test-takers' performance on the test was found. Section 4.3 provides evidence to support the claim that the CAT approach is effective at tailoring the level of difficulty of the test to the proficiency level of individual test-takers in both summative and formative assessment settings.

10.1.1.7 Academic staff attitude

In Chapter 5, academic staff attitude towards the CAT approach was investigated. Academic staff acknowledged that the CAT approach would be valuable in terms of speed and accuracy of marking; this characteristic, however, is generic to CAA rather than exclusive to the CAT approach.

The CAT approach was valued by academic staff in both summative and formative assessment settings, with a greater preference for its application in formative assessment. In this work, it is argued that the reason for this is threefold. First, formative assessment plays a key role in student learning and the use of CATs would support the provision of timely feedback. Second, the CAT approach as proposed in this work is based on the use of objective questions, which restricts the type of tasks that can be undertaken by students. Third, academic staff felt that students would more receptive to the CAT approach in a formative rather than in a summative assessment environment. One can speculate that one of the reasons for this is the complexity of the CAT scoring method, especially when compared to more traditional methods such as CBTs. Interestingly, the scoring method was not perceived as a drawback by test-takers as predicted by academic staff. This is reported in sections 4.2 and 5.2.

An important barrier to the implementation of the CAT approach in Higher Education environments was identified by academic staff: the fact that CATs are more difficult to construct than linear CBTs due to the need of an adaptive

algorithm and a calibrated question database. Issues related to the calibration of the database are discussed in section 3.2.

An important issue that emerged from the pedagogical evaluation reported in section 5.2 was that the provision of feedback to test-takers in the form of a score alone is not sufficient to help students detect their own potential educational needs. The provision of feedback on performance to CAT test-takers is explored in Chapters 7, 8 and 9.

All in all, academic staff exhibited a positive attitude towards the CAT approach and its potential applications in the assessment of Computer Science undergraduates.

10.1.1.8 Validity and reliability

As with the introduction of any assessment method, it was crucial to the research to examine the validity and reliability of the CAT approach.

Sections 4.2, 5.2 and 6.1.1 present evidence that the CAT approach, as proposed in this research, has face validity. In section 6.1.2, it was shown that the CAT approach has content validity. In a CAT, a database containing at least 4 times the number of questions to be administered is required; this means that in practical applications of the CAT approach, ensuring content validity can be a very laborious task. Evidence was presented in section 6.1.3 that the CAT approach has construct validity. Importantly, findings reported in sections 4.1 and 6.1.3 support the view that the CAT approach is fair and that test-takers were not disadvantaged by the approach.

A test-retest reliability study was conducted to examine the reliability of the CAT approach, and is reported in section 6.2.2. Results from this study showed that the CAT approach is reliable, and also that CAT proficiency level estimates were, at least, as good an indicator of the ability of a test-taker as other traditional forms of assessment such as CBT scores.

In summary, in Chapter 6 it was shown that the CAT approach is both valid and reliable. These findings were of great importance in the context of this

research, as they provided evidence to support the view that the CAT approach is a viable alternative for the assessment of Computer Science undergraduates.

10.1.2 Automated feedback prototype

Findings from the pedagogical evaluation reported in section 5.2 and voluntary feedback from test-takers reported in section 4.4 provided an important new direction for the research: the provision of automated feedback. The CAT literature has failed to provide compelling evidence of how the CAT approach can be employed to provide feedback to test-takers other than an overall proficiency score. This is in spite of the predicted benefits of the CAT approach, including that of Brusilovsky (2004), who cites the CAT approach as an example of a paradigm shift in educational technology, from "one size fits all" to one capable of offering higher levels of personalisation.

As part of the research, an automated feedback prototype was designed, implemented and evaluated. Early in the research it became apparent to the research team that the provision of automated feedback was important not only to build on the information about strengths and weaknesses of test-takers obtained through the CAT approach, but also to release some of the pressure associated with high student to staff ratios. Barker (1999), Bull & McKenna (2004) and others have also suggested the use of computer technology as a potential solution to address issues relating to the reduction in contact hours.

The automated feedback devised as part of this research consisted of 3 sections: (1) an overall score, (2) feedback on proficiency level per topic, based on Bloom's taxonomy of cognitive skills (Bloom, 1956; Anderson & Krathwohl, 2001), and (3) recommendations on which concepts within the subject domain should be revised in the form of directive feedback. The overall score and proficiency level per topic were estimated using the response likelihood function (see Equation 3-2, p. 69), as described in section 7.3.2. Performance topic per area was then mapped into one of the three following cognitive skills: knowledge, comprehension and application (see section 7.3.2).

The directive feedback consisted of revision tasks, which could optionally be preceded by cues (for example, the definition of a term). The revision tasks were created by the research team using questions' stems as a starting point. For each test-taker, only tasks relating to questions answered incorrectly were selected. In the same vein as the CAT test, the underlying idea was to provide test-takers with revision tasks that offer a suitable degree of challenge, rather than tasks that are far above (or below) their ability levels.

The goal of matching learning tasks to the proficiency levels of students, in order to provide students with learning opportunities at an appropriate level, is not new. Barker (1999) and Barker et al. (2002), for instance, describe an attempt to configure a multimedia educational system based on a co-operative psychological student model. In this work, information and the level of tasks and questions presented to users were adapted co-operatively based upon their proficiency levels. Although useful, the work of Barker (1999) and Barker et al. (2002) was unable to support a fully automated feedback as required in this research. There was also the issue of timeliness. Co-operative student models can, at times, be slow and timing of feedback on test performance is continually stressed as being crucial both by test-takers and by academic staff. Mitrovic & Martin (2004) and Martin & Mitrovic (2005), for example, propose an Intelligent Tutoring System (ITS) where the matching between students and tasks is fully automated, and based on a set of constraints for both domain and student models. Although this was a valuable approach, it did not explore the richness of information about test-takers provided by the CAT approach and resulting test-taker profile.

A series of empirical studies were carried out to examine test-taker and academic staff attitude towards the automated feedback approach proposed by this research. These are summarised next.

10.1.2.1 Test-taker attitude

Three empirical studies were conducted in order to investigate test-takers' attitude towards the automated feedback approach. Findings reported in

section 8.1.2 provide evidence that the automated feedback prototype was easy to use.

The empirical studies reported in Chapter 8, show that in general test-takers found the automated feedback software prototype capable of providing feedback that was timely, useful for individual development and effective at identifying strengths and weaknesses.

No statistically significant correlations were found between test-takers' proficiency level estimates and their attitude towards the automated feedback, and this result was taken to indicate that attitude to feedback was not affected by performance on the test. This was an interesting finding, as it was possible that test-takers' actual performance on the test may influence their attitude towards subsequent feedback. For example, poor performance on the test could lead to negative attitude towards the automated feedback approach.

The results reported in Chapter 8 present evidence that the automated feedback tool was effective at providing test-takers with timely and useful feedback in both, summative and formative assessment settings, regardless of their proficiency level within the subject domain.

10.1.2.2 Academic staff attitude

Academic staff attitude towards the automated feedback approach was examined in Chapter 9.

Four important factors regarding the current provision of feedback on assessment performance to students were reported in section 9.1. First, increased student to staff ratios often mean that feedback on assessment performance is limited to an overall score. Second, feedback mechanisms currently in place (for example, electronic copies of worked examples) are typically designed with the "average student" in mind. Third, there is no compelling evidence to suggest that the current provision of feedback is of a consistently high standard, or tailored to individual students. Fourth, feedback has to be timely and tailored to individual students to be useful. Yet

this is not always possible, especially for those members of academic staff teaching large groups.

Reports from academic staff presented in section 9.1 suggest that whilst some members of staff will find the process of delivering feedback via automated means beneficial, others may fear that this process threatens the human interaction tutors value in face-to-face teaching. It should be noted, however, that academic staff play a crucial role in the success of the provision of automated feedback: without carefully designed and pedagogically sound revision tasks, its usefulness is drastically reduced. Regardless of their views on this issue, academic staff recognised the value of the approach as proposed in this research for the provision of automated feedback that is timely, useful and tailored to individual test-takers.

In section 9.2, it was found that the automated feedback prototype was evaluated by academic staff as effective in providing CAT test-takers with timely and useful feedback.

10.2 Outcomes of the research

The principal contribution to existing knowledge made by this work was to demonstrate the ways in which:

- the CAT approach can be applied to the assessment of Computer Science undergraduates;
- the tailored test generated by the CAT approach can be used to identify the strengths and weaknesses of individual students, and to support the adaptive selection of learning resources that meet their educational needs.

The work presented in this thesis addressed two major themes: (1) the application of the CAT approach in a real educational setting for the assessment of Computer Science undergraduates and (2) the use of

information about test-takers obtained through the CAT approach to provide students with feedback on test performance.

10.2.1 Assessment

The first research question formulated in section 1.1 was:

- What are the potential applications of the CAT approach in the assessment of Computer Science undergraduates?

The answer proposed in this thesis is that the CAT approach is both valid and reliable in the assessment of Computer Science undergraduates in summative and formative settings. In this work, it was established that the CAT approach is a viable and useful alternative to extend the range of methods currently employed for the assessment of Computer Science undergraduates.

The research reported in this thesis provided evidence that the two main groups of stakeholders in the assessment process, students (test-takers) and academic staff, exhibited a positive attitude towards the CAT approach, with academic staff displaying a preference for its use in a formative assessment context.

In terms of learning outcomes that can be assessed using the CAT approach, it was found that the CAT approach was effective at assessing the three lowest levels of Bloom's taxonomy of cognitive skills (Bloom, 1956; Anderson & Krathwohl, 2001), namely knowledge, comprehension and application. Hence, the CAT approach must be combined with other forms of assessment in order to assess higher cognitive skills, and learning outcomes that are not suitably assessed with objective questions. It is argued that the unsuitability of the CAT approach as proposed in this research to assess higher cognitive skills was one of the factors that led academic staff to favour this assessment method in formative rather than summative assessment context.

The work reported in this thesis demonstrated that the CAT approach was effective at tailoring the level of difficulty of the test to individual students. It is argued that the CAT approach can be used to identify an important boundary

between what the student knows and does not know in a subject area or, in other words, the unique boundary between what is challenging and motivational, and what is too difficult or too easy. The information about students obtained through the application of the CAT approach can be used in a variety of ways, one of them being the provision of tailored feedback. This is the focus of the following section.

10.2.2 Feedback

The second research question formulated in section 1.1 was:

- In which ways can the CAT approach be used to provide automated feedback to students that is timely and useful?

As with any CAA application based on the use of objective questions, a CAT can be scored immediately after completion, providing test-takers with instantaneous feedback on performance in the form of an overall score. In the work reported in this thesis it was shown that, although test-takers (students) value the possibility of receiving test scores immediately, this is not sufficient to enhance their learning and future performance.

In this research, a method for the provision of automated feedback based on the CAT approach was designed, implemented and evaluated. The automated feedback consisted of three elements: (1) overall score, (2) performance per topic area and (3) tailored revision plan. Within the automated feedback proposed in this work, CAT proficiency estimates per topic covered in the test, in addition to overall CAT proficiency estimates are computed for each individual test-taker. This approach was shown to be useful at providing students not only with an overall score, but also at identifying areas of strengths and weaknesses according to topic area within the subject domain. In a CAT the questions administered during a test are tailored to each individual test-taker, allowing a tailored revision plan based on questions answered incorrectly to be provided. This approach was proven to

be useful for identifying which revision tasks are the most suitable for each student, giving them an individual revision plan, tailored to their needs.

The strengths of the automated feedback as proposed in this research are many. First, it was shown that students (test-takers) found the tailored feedback to be useful in improving future performance, and exhibited positive attitude towards the automated approach. Second, it was shown that the automated approach not only allows the provision of feedback that is tailored and within each individual student's grasp, but also timely. Gibbs & Habeshaw (1993: p. 95), for instance, stress that when feedback is not timely students "have neither the time nor the interest to take feedback to heart". Third, it was shown that academic staff perceived the automated feedback prototype as being capable of providing timely and useful feedback, and promoting learning. Fourth, the feedback proposed in this research is based on the provision of directive feedback in the form of revision tasks, rather than simply providing a copy of questions and correct answers. It is argued that, although the initial creation of revision tasks demands considerable amounts of time and effort from academic staff, the process is worthwhile, as tasks can be re-used. Fifth, the type and quality of feedback is consistent to all students. Miller et al. (1998), for example, indicates that the type of feedback provided to students (for example, in essay submissions) varies significantly, even in those cases where there is only one marker. Sixth, as the automated feedback is provided through a web-based application, the feedback is available from any location and it can be used at any time and frequency. Additionally, the process of going through the revision tasks is non-threatening and can be self-paced. Seventh, the automated feedback approach can be combined with traditional feedback methods. For example, a face-to-face feedback session can be arranged so students can discuss with a tutor their solutions to the revision tasks provided.

In summary, it was found that the automated feedback prototype developed for this research support the provision of feedback that is timely and effective at matching revision and learning tasks to the proficiency levels of individual students.

10.2.3 Research objectives

The list of objectives formulated to explore the two research questions was introduced in section 1.1, and is highlighted below:

(a) to identify the main issues in designing and implementing a CAT software application to be used in the assessment of Computer Science undergraduates;

(b) to design and implement a CAT software application;

(c) to identify the key issues in evaluating a computer-assisted assessment (CAA) application;

(d) to evaluate the CAT software application;

(e) to identify the key components of the CAT approach that are useful in the provision of feedback to students;

(f) to design and implement an automated feedback software application based on the CAT approach;

(g) to evaluate the automated feedback software application.

In order to achieve these objectives, this research sought to understand the fundamental issues and concerns in the appropriate use of the CAT approach for the assessment of, and provision of feedback to Computer Science undergraduates. All of the objectives listed above were achieved; CAT fundamentals were applied to the design and implementation of the CAT and automated feedback software prototypes, as summarised in section 10.1. The evaluation of both software prototypes is discussed below.

A summary of the issues that need to be considered in the evaluation of CAA applications, and CAT in particular, is presented in this section. These issues can be divided into four broad areas: (1) identification of the purpose of the evaluation, (2) identification of the main groups of stakeholders, (3) selection of evaluation methods, and (4) authenticity.

In this work, it was found that the evaluation of CAA applications, especially those that introduce new concepts such as CATs, require a great amount of

time and effort on the part of the academic staff. It is possible that this is one of the key reasons why CATs have made relatively little impact on student assessment in Higher Education, in spite of their potential.

Evaluation of the CAT software prototype. The purpose of the evaluation was twofold. First, to assess the usability of the CAT prototype for the target audience. Second, to assess whether the CAT software prototype met its educational objectives.

In this work, it was considered that the participation of the main groups of stakeholders was crucial to the success of the evaluation. Preece et al. (2002) suggest that any group of people who might be affected by the success or failure of a system – in this case the CAT software prototype – should be classified as stakeholders.

Two groups of people were identified as the main stakeholders. The first group of stakeholders is formed by test-takers (students). Students are the largest group of stakeholders, and the intended users of both software prototypes. The second group of stakeholders consists of academic staff. Although academic staff are not end-users of the software prototypes, they have a great influence on whether or not these will be used in practice. In addition, it was important to gather academic staff reflections on the pedagogical value of the CAT approach as proposed by this research in a Higher Education environment.

Several authors including Laurillard (1993), McAteer & Shaw (1994), Boyle (1997), Barker & Barker (2002), and Bull & McKenna (2004) have advocated using qualitative and quantitative methods in the evaluation of educational software, due to its complexity. Such a hybrid approach was found to be effective in this research, as different methodologies are useful depending on the specific objectives of the evaluation.

The first aim of the evaluation was to assess the usability of the CAT prototype. In this work, the user interface was found to be usable, and unlikely

to hinder students' performance on the test. The different stages in the usability evaluation are summarised in Table 10-1.

Stage	Participants	Method	Section	Outcome
(1)	Test-takers	Observation study	4.1	No usability problems were found.
(2)	Test-takers	Focus group	4.2	No usability problems were found.
(3)	Academic staff	Heuristic evaluation	5.1	No usability problems were found.
(4)	Academic staff	Pedagogical evaluation	5.2	No usability problems were found.
(5)	Test-takers	Interview	4.3.1	No usability problems were found. Changes to user interface - addition of question counter.
(6)	Test-takers	Voluntary feedback (email sent to research team)	4.4	Changes to user interface - removal of the "Confirm answer" button.

Table 10-1: Usability evaluation findings

As can be seen from Table 10-1, quantitative methods were employed to gather information about the usability of the prototype from the perspective of test-takers. The underlying idea was to prompt end-users (i.e. test-takers) to discuss ideas regarding the user interface that may otherwise have been overlooked. Stages (5) and (6) were the only two stages that originated suggestions on how the user interface could be modified to enhance user satisfaction. One can speculate that the reason for this is that these two stages involved test-takers in a real assessment setting, and therefore they have a greater interest in ensuring that the user interface meets their needs.

In the case of the stages involving academic staff, a range of design and pedagogical issues relating to the usability of the CAT software prototype were covered, ranging from "Are error messages helpful?" (see Appendix I) to "Students' interaction with the system would be simple and clear." (see Appendix J). In this stage of the evaluation, academic staff were asked to rate

a series of statements related to the usability of the CAT software prototype using a five point Likert scale. This method was found useful in obtaining meaningful quantitative data, and thus was employed in all questionnaires employed in the research reported in this thesis.

In this work, it was important to evaluate the use of the CAT prototype in an educational context, rather than evaluating the software per se. Hence, the second aim of the evaluation was to assess whether the CAT software prototype met its educational objectives. The definition of educational objectives, however, is a complex task and often includes a combination of factors. In this work, it is argued that the educational objective of the CAT approach is to provide an assessment method that is valid, reliable and tailored for the assessment of Computer Science undergraduates.

The validity of the CAT approach was examined and this process consisted of 7 stages, which are illustrated in Table 10-2. As can be seen from Table 10-2, a combination of quantitative and qualitative methods was employed. Similarly to the usability evaluation, the validity studies involved both groups of key stakeholders. In this work, it was found that the CAT approach is valid and the main findings relating to the validity of the approach are reported in Chapter 6 of this thesis. The empirical study relating to the calibration of the item (i.e. question) database was included in Table 10-2, as the quality of the item database affects the validity of CAT scores.

Stage	Participants	Method	Section(s)	Evidence provided
(1)	Test-takers and experts (data only)	Statistical analysis	3.2.2	The database calibration employed in the research was found to be appropriate and useful. Note: Study conducted to support (5) and (7).
(2)	Test-takers	Focus group	3.5.3 4.2	In summative assessment settings, test-takers favour fixed-length CATs. Variable-length CATs are acceptable in formative assessment settings. Test-takers favour CATs in which question review is permitted. Test-takers exhibited a positive attitude towards the CAT approach.
(3)	Test-takers (data only)	Statistical analysis	3.6	Study conducted as a result of (3). Question review, as implemented in this research, has no adverse effect on the accuracy of proficiency level estimates.
(4)	Academic staff	Questionnaire	5.2	Academic staff participated in a pedagogical evaluation, and exhibited a positive attitude towards the CAT approach. Provision of feedback in the form of a score is useful. Feedback should, however, be enhanced in order to help students improve.
(5)	Test-takers and Academic staff	Various	6.1.1	Findings from stages (2)-(4) support the view that the CAT approach had face validity.
(6)	Academic staff	Expert review	6.1.2	The CAT approach had content validity. An important contributing factor to content validity is the use of content balancing for CAT item selection.
(7)	Test-takers (data only)	Statistical analysis	6.1.3	The CAT approach had construct validity.

Table 10-2: Validity of the CAT approach

Whilst the validity studies summarised in Table 10-2 entailed a combination of qualitative and quantitative methods, the investigation of the reliability of the CAT approach was based on a test-retest reliability study. This quantitative study employed real assessment data from actual test-takers, and it showed that the CAT approach was found to be reliable (see Table 10-3).

Stage	Participants	Method	Section	Evidence provided
(1)	Test-takers (data only)	Statistical analysis	6.2.2	The CAT approach was found to be reliable.

Table 10-3: Reliability of the CAT approach

An important assumption in this work was that the CAT approach was capable of matching the level of difficulty of the test to the ability level of individual test-takers. In order to verify this assumption, a series of empirical studies were conducted and are summarised in Table 10-4. As can be seen from Table 10-4, this evaluation consisted of 4 stages. Stages (1), (2) and (4) were concerned with the analysis of quantitative data, obtained through the use of questionnaires where test-takers were required to rate the level of difficulty of a CAT using a five point Likert scale. In order to obtain a more in-depth insight of test-takers' perceived level of difficulty, a random sample of test-takers were invited to participate in an interview.

Stage	Participants	Method	Section	Evidence provided
(1)	Test-takers	Electronic questionnaire	4.1	The CAT approach was effective at tailoring the level of difficulty of the test to the ability of individual test-takers.
(2)	Test-takers	Statistical analysis	4.3.1	The CAT approach was effective at tailoring the level of difficulty of the test to the ability of individual test-takers, in a summative assessment setting.
(3)	Test-takers	Interview	4.3.1	The CAT approach was effective at tailoring the level of difficulty of the test to the ability of individual test-takers, in a summative assessment setting.
(4)	Test-takers	Statistical analysis	4.3.2	The CAT approach was effective at tailoring the level of difficulty of the test to the ability of individual test-takers, in a formative assessment setting.
(5)	Test-takers	Voluntary feedback (email sent to research team)	4.4	Provision of feedback in the form of a score was useful. Feedback should, however, be enhanced in order to help students improve.

Table 10-4: Test-taker attitude towards the CAT approach

The findings from the stages (1)-(4) reported in Table 10-4 support the view that the CAT approach was effective at adapting the level of difficulty of the

test to individual test-takers. The impact of stage (5) on the evaluation carried out as part of this research is discussed in section 10.1.2.

An important aspect of the evaluation work is that of authenticity. Draper (1997) and Barker (1999) argue that tightly controlled experiments have little relevance to real educational settings. The evaluation work conducted as part of this research took place in a Higher Education environment and, as can be seen from Table 10-1, Table 10-2, Table 10-3 and Table 10-4, the evaluation studies resulted in useful and usable findings that would possibly have been otherwise overlooked or even remained unknown to the research. Findings from the evaluation were analysed and, as part of an iterative process, used to refine the CAT software prototype. Input from academic staff and students was necessary in order to obtain an in-depth view of both, the user interface and the CAT approach.

The evaluation of the CAT software prototype reported in this thesis involved over 700 Computer Science undergraduates and 11 members of academic staff during a period of 5 academic years. Findings from this evaluation were published in numerous conference proceedings, and a list of publications can be found in Appendix M.

Evaluation of the automated feedback software prototype. Based on the evidence reported in Chapters 4, 5 and 6 it was appropriate to design, implement and evaluate an automated feedback prototype based on the CAT approach. In Table 10-2 and Table 10-4, it was shown that the provision of feedback in the form of a score alone was not sufficient to help students improve.

The purpose of the evaluation reported in this section was therefore twofold. First, to assess the usability of the automated feedback prototype. Second, to assess whether the automated feedback prototype met its educational objectives.

The evaluation of the automated feedback prototype also involved two main groups of stakeholders. The first group consisted of students, who were also

the target end-users of the prototype. The second group of stakeholders was academic staff.

The prototype was first evaluated from the student's perspective. The evaluation of the usability of the automated feedback prototype was less detailed than that conducted for the CAT software prototype. This is because the main purpose of the automated feedback prototype was less sensitive than the CAT one. A poor CAT software interface could affect test-takers' performance in an adverse way and, consequently, have a negative impact on their grades for the module. In the case of the automated feedback, the underlying idea was to present students with a list of revision tasks that should be completed using a different computer application (for example, to write a Visual Basic.NET program using the integrated development environment provided by Visual Studio). Thus, the user interface, although important, was less critical.

As part of the evaluation of the automated feedback prototype, test-takers were required to rate the statement "I would find the application easy to use" using a five point Likert scale. The use of this quantitative method to gather information about perceived ease of use was found useful in this research. It was found that the application was usable, and Table 10-5 summarises the usability the study.

Stage	Participants	Method	Section	Outcome
(1)	Test-takers	Questionnaire	8.1.3	The application was found to be easy-to-use.

Table 10-5: Usability evaluation finding

The second aim of the evaluation was to assess whether the automated software prototype met its educational objective of providing students with tailored feedback that is timely and useful.

Test-takers participated in real CAT formative and summative assessments; feedback on CAT performance was provided using the automated feedback prototype. Questionnaires were used to gather information about test-takers'

attitude towards the approach; test-takers' scores and their ratings were subjected to statistical analysis. Table 10-6 illustrates that test-takers exhibited a positive attitude towards the automated feedback approach in a variety of real assessment settings.

Stage	Participants	Method	Section	Evidence provided
(1)	Test-takers	Statistical analysis	8.1	Test-takers exhibited a positive attitude towards the automated feedback approach in a summative assessment setting.
(2)	Test-takers	Statistical analysis	8.1.2	Test-takers exhibited a positive attitude towards the automated feedback approach in a summative assessment setting. No significant difference in the perceived usefulness of the feedback that could be ascribed to the effect of test-takers' performance on the test was found.
(3)	Test-takers	Statistical analysis	8.1.3	Test-takers exhibited a positive attitude towards the automated feedback approach in a formative assessment setting.

Table 10-6: Test-taker attitude towards the automated feedback

Evaluating the usefulness of the automated feedback was a complex and challenging task. It became apparent in the initial stages of the evaluation that greater input from academic staff was required than that necessitated in the evaluation of the CAT software prototype. Whilst the format of objective questions (as employed in the CAT software prototype) and the CAT adaptive algorithm were initially defined based on findings from the literature (and later refined based on findings from the evaluation), in the case of the automated feedback the entire system was designed from scratch. Most importantly, no evidence of the provision of feedback to students on CAT performance other than the provision of an overall score was found in the literature.

In order to elicit reactions from academic staff to the automated feedback application, a series of empirical studies were conducted and are summarised in Table 10-7.

Stage	Participants	Method	Section	Evidence provided
(1)	Academic staff	3 semi-structured discussions	9.2	Overall, academic staff exhibited a positive attitude towards the automated feedback approach. Some academic staff fear that the automated feedback may threaten the human interaction tutors value in face-to-face teaching.
(2)	Academic staff	Questionnaire	9.3	Overall, academic staff found the feedback effective at providing feedback that is timely and useful.

Table 10-7: Academic staff attitude towards the automated feedback

The data that emerged from the semi-structured discussions was rich and informative, and provided a more comprehensive view of the perceived benefits and limitations of the automated feedback approach than that afforded by the questionnaire. Based on this evidence, it is argued that academic staff in general exhibited a positive attitude towards the CAT approach.

Authenticity was also a crucial issue in the evaluation of the automated feedback prototype. The evaluation of the automated software prototype reported in this thesis involved over 400 Computer Science undergraduates and over 40 members of academic staff during a period of 3 academic years. Findings from the evaluation of the automated feedback were the focus of several conference papers, and these are listed in Appendix M.

10.3 Future directions for the research

This section presents some future directions for the research, including the ways in which the CAT and automated software software prototypes developed as part of this work can be improved.

The CAT and automated feedback software prototypes developed for this research include full functionality from the perspective of the test-taker. However, from the perspective of academic staff, the current prototypes do not support the creation of reports based on a set of criteria, nor do they support the execution of create, read, update and delete (CRUD) operations via a

graphical user interface (for example, to add a new question to the database). At present, report creation and CRUD operations are performed by manipulating the database directly, or by writing ad-hoc small software programs. However, in order to become more widely usable, it would be critical to develop software applications that provide a graphical user interface for the tasks that academic staff are likely to perform when using the CAT and automated feedback prototypes.

An important future direction of this work is to examine the issue of student motivation on a CAT. A common assumption in the CAT literature, which was accepted in this work, is that tailoring the level of difficulty of the tasks to individual proficiency levels will lead to increased levels of student motivation. In section 4.3.1, the research reported here provides some evidence to this effect. Chan et al. (1997: p. 301) suggest that test performance is a "joint function of ability and motivation" and therefore "ability and motivation should play a nontrivial role in determining test performance". It will therefore be important to examine the possible relationship between performance on a CAT and motivation.

Another important future direction is to investigate how the Three-Parameter Logistic (3-PL) model from Item Response Theory (IRT) (Lord, 1980) can be enhanced. Wainer et al. (2000) rightly indicate that one of the limitations of IRT and Classical Test Theory is the assumption that the complexity of a person's proficiency level in a domain can be represented by such one-dimensional models. In this work, it was shown that the 3-PL model, although one-dimensional, was effective at supporting the assessment of Computer Science undergraduates in a real educational setting. However, it will be interesting to examine the ways in which different approaches to learning, assessment strategies, motivation and cognitive skills (as defined by Bloom, 1956) can be employed in the development of a multi-dimensional model to be used in adaptive testing. The issue of IRT multi-dimensionality is not new but, as Wainer et al. (2000) point out, existing research has not yet delivered a definite answer to this problem.

Another avenue to be pursued is the automated generation of test questions. In this work, the generation of question was found to be an onerous process, and it will be interesting to investigate how an existing calibrated question can be used as the basis for the automated generation of a new question; item parameters from the original question can then be used to assign values for the difficulty b, discrimination a and pseudo-chance c parameters of the new one. As an example, assume that there is question about while loops, consisting of a source code snippet and options (i.e. key and distractors) regarding the value assigned to a variable after "n" loop iterations. A computer application in the form of an automated question generator could be employed to change sections of the source code (for example, to change the condition that causes the while loop to end), as well as compute the key answer and distractors of the new question.

Finally, it will be important to improve the quality of the automated feedback provided by providing students with feedback on questions answered correctly; it is possible that, on occasions, students answer questions correctly by chance or even for the wrong reasons. In addition, the automated feedback as proposed in this study is updated only when a student takes a test; it will be important to investigate other mechanisms that could be used to update the student profile more regularly, possibly by combining the CAT approach with a co-operative student model such as the one described by Barker (1999), in order to increase its effectiveness.

The work reported in this thesis illustrates the ways in which academics can use CATs effectively in the assessment of the Computer Science undergraduates. In addition, it provides key information to academics and practitioners on the main issues relating to the design, implementation and evaluation of CATs in a Higher Education setting. Finally, and most important, this research demonstrates how the information about students' proficiency levels obtained through the CAT approach can be used in the creation of systems that support the dynamic selection of learning materials that are pitched at the right level for individual students. It is hoped that the work

reported in this PhD thesis will foster future research, based on the ideas highlighted above.

References

Alfonseca, E.; Carro, R. M.; Freire, M.; Ortigosa, A.; Pérez, D., & Rodríguez, P. (2005) 'Authoring of Adaptive Computer Assisted Assessment of Free-text Answers', *Educational Technology & Society*, 8 (3), 53-65.

American Psychological Association (1999) *Standards for Educational and Psychological Testing*. American Educational Research Association.

Anastasi, A. (1988) *Psychological testing*. New York: Macmillan.

Anderson, L. W. & Krathwohl, D. R. (Eds.) (2001) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Addison Wesley.

Assessment Systems Corporation (2007) *XCALIBRE - Marginal Maximum-Likelihood Estimation*, Available: http://assess.com/xcart/product.php?productid=270&cat=0&page=1 [11 Jun 2007]

Baker, F. B. & Kim, S. (2004) *Item Response Theory: Parameter Estimation Techniques*. Marcel Dekker Ltd.

Barker, T. (1999) *The use of a student model in a multimedia application to configure learning,* Thesis (PhD). University of Hertfordshire.

Barker, T. & 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.

Barker, T. & Lilley, M. (2003) 'Are Individual Learners Disadvantaged By The Use Of Computer-Adaptive Testing In Higher Education?', *Proceedings of the 8th Learning Styles Conference*, European Learning Styles Information Network (ELSIN), University of Hull, United Kingdom.

Barker, T. & Lilley, M. (2004) 'The development and evaluation of computer-adaptive testing software in a UK university', *Proceedings of the 2004*

Learning and Teaching Conference, University of Hertfordshire, United Kingdom.

Barker, T. & Lilley, M. (2006) 'Measuring staff attitude to an automated feedback system based on a Computer Adaptive Test', *Proceedings of Computer-Assisted Assessment 2006 Conference,* Loughborough University, July 2006.

Barker, T., Jones, S., Britton, C. & Messer, D. (2002) 'The use of a co-operative student model of learner characteristics to configure a multimedia application', *User Modelling and User Adapted Interaction*, 12 (2/3), pp. 207-241.

Barker, T.; Lilley, M & Britton, C. (2006a) 'Computer Adaptive Assessment and its use in the development of a student model for blended learning', *Annual Blended Learning Conference*, University of Hertfordshire, July 2006.

Barker, T.; Lilley, M. & Britton, C. (2006b) 'A student model based on computer adaptive testing to provide automated feedback: The calibration of questions', *Paper presented at the Association for Learning Technology (ALT) 2006*, Herriot-Watt University, September 4-7, 2006.

Baydoun, R. & Neuman, G. (1998) 'Computerization of Paper-and-Pencil Tests: When are They Equivalent?', *Applied Psychological Measurement*, 22(1) pp. 71-83.

Biggs, J. B. (2002) *Teaching for Quality Learning at University*. Buckingham: Society for Research into Higher Education & Open University Press.

Bloom, B. S. (1956) *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*. New York: David McKay Co Inc.

Boyle, T. (1997) *Design for Multimedia Learning*. Prentice-Hall.

Brosnan, M. (1999) 'Computer anxiety in students: should computer-based assessment be used at all?', in S. Brown, J. Bull & P. Race (Eds.) (1999), *Computer-Assisted Assessment in Higher Education*, London: Kogan Page Ltd.

Brown, G. (1997) *Assessing Student Learning in Higher Education*. Lodnon: Routledge Falmer.

Brown, G., Bull, J. & Pendlebury, M. (1998) *Assessing Student Learning in Higher Education*. Routledge.

Brown, J. D. (1988) *Understanding Research in Second Language Learning: A Teacher's Guide to Statistics and Research Design*. Cambridge University Press.

Brown, S. (2003) Institutional Strategies for Assessment, in S. Brown & A. Glasner (Eds.) (2003), *Assessment Matters in Higher Education: Choosing and Using Diverse Approaches*. Society for Research into Higher Education, Open University Press.

Brusilovsky P (2004) 'Knowledge Tree: A Distributed Architecture for Adaptive E-Learning', *Proceedings of WWW 2004*, May 17-22, New York, New York, USA, pp. 104-113.

Bull, J. & McKenna, C. (Eds.) (2004) *Blueprint for Computer-assisted Assessment*. London: Routledge Falmer.

Bull, J. (1999) 'Update on the National TLTP3 Project: The implementation and evaluation of computer-assisted assessment', *Proceedings for 3rd Computer-Assisted Assessment Conference 1999*, Available: http://www.caaconference.com/pastConferences/1999/proceedings/keynote.pdf [26 Nov 2006].

Callear, D.; Jerrams-Smith, J. & Soh, D. (2001) 'CAA of short non-MCQ answers', *Proceedings of the 5th International Computer Assisted Assessment Conference*. Available: https://dspace.lboro.ac.uk/dspace/handle/2134/1791 [Accessed 21 Aug 2007].

Cann, A. J. & Pawley, E. L. (1999) Automated online tutorials: new formats for assessment on the WWW, in S. Brown, J. Bull & P. Race (Eds.) (1999), *Computer-Assisted Assessment in Higher Education*, London: Kogan Page Ltd.

Carlson, R. D. (1994) 'Computer-Adaptive Testing: a Shift in the Evaluation Paradigm', *Journal of Educational Technology Systems*, 22(3), pp 213-224.

Chalhoub-Deville, M.; Milanovic, M. & Weir, C. J. (Eds.) (2000) *Issues in Computer-Adaptive Testing of Reading Proficiency: Studies in Language Testing 10 (Studies in Language Testing)*. Cambridge University Press.

Challis, D. (2005) 'Committing to quality learning through adaptive online assessment', *Assessment & Evaluation in Higher Education*, 30(5), pp. 519-527.

Chan, D.; Schmitt, N.; DeShon, R. P.; Clause, C. S. & Delbridge, K. (1997) 'Reactions to Cognitive Ability Tests: The Relationships Between Race, Test Performance, Face Validity Perceptions, and Test-Taking Motivation', *Journal of Applied Psychology*, 82 (2), pp. 300-310.

Chin, C. H. L.; Donn, J. S. & Conry, R. F. (1991) 'Effects of Computer-Based Tests on the Achievement, Anxiety, and Attitudes of Grade 10 Science Students', *Journal of Educational and Psychological Measurement*, 51 (3), pp. 735-745.

Conejo, R.; Millán, E.; Pérez-de-la-Cruz, J. L.; Trella, M. (2000) 'An Empirical Approach to On-Line Learning in SIETTE', *Lecture Notes in Computer Science*, 1839, pp. 604-614, 2000.

Conole, G. & Bull, J. (2002) 'Pebbles in the Pond: Evaluation of the CAA Centre', *Proceedings for 6th Computer-Assisted Assessment Conference 2002*, Available: http://www.caaconference.com/pastConferences/2002/proceedings/conole_g1.pdf [26 Nov 2006].

Cristea, P. & Tuduce, R. (2005) 'Automatic Generation of Exercises for Self-testing in Adaptive E-Learning Systems: Exercises on AC Circuits', *Third*

International Workshop on Authoring of Adaptive and Adaptable Educational Hypermedia Amsterdam, The Netherlands, 19 July 2005.

Davies, P. (2001) 'Computer Aided Assessment must be more than multiple-choice tests for it to be academically credible?', *Proceedings of the 5th Computer-Assisted Assessment Conference 2001*, Available: http://www.caaconference.com/pastConferences/2001/proceedings/e2.pdf [10 Jan 2007].

Denton, P. (2003) 'Evaluation of the 'Electronic Feedback' Marking Assistant and Analysis of a Novel Collusion Detection Facility', *Proceedings of the 7th Computer-Assisted Assessment Conference 2003*.

Divgi, D. R. (1986) 'Does the Rasch model really work for multiple choice items? Not if you look closely', *Journal of Educational Measurement*, 23, pp. 283-298.

Douce, C.; Livingstone, D.; Orwell, J.; Grindle, S. & Cobb, J. (2005) 'A Technical Perspective on ASAP – Automated System for Assessment of Programming', *Proceedings of the 9th Computer-Assisted Assessment Conference 2005*, Available: http://www.caaconference.com/pastConferences/2005/proceedings/DouceC_LivingstoneD_OrwellJ_GrindleS_CobbJ.pdf [10 Jan 2007].

Draper, S. (1997). 'Prospects for summative evaluation of CAL in Higher Education', *Association for Learning Technology Journal*, 5(1), pp. 33-39.

Dunn, L.; Morgan, C.; O'Reilly, M. & Parry, S. (2003) *The Student Assessment Handbook: New Directions in Traditional and Online Assessment*. Lodnon: Routledge Falmer.

Eggen, T. J. H. M. (2004) *Contributions to the theory and practice of computerized adaptive testing*. Citogroep Arnhem, Netherlands.

Ellis, W. & Ratcliffe, M. (2004) 'Improving Results with Positive Directed Feedback in Summative Assessments', *Proceedings of the 8th Computer-Assisted Assessment Conference 2004*.

English, J. & Siviter, P. (2000) 'Experience with an automatically assessed course', *Proceedings of the 5th annual SIGCSE/SIGCUE ITiCSE conference on Innovation and technology in computer science education*, Helsinki, Finland, pp. 168–171.

Fernandez, G. (2003) 'Cognitive Scaffolding for a Web-Based Adaptive Learning Environment', *Lecture Notes in Computer Science*, 2783, Advances in Web-Based Learning - ICWL 2003, pp. 12-20, 2003.

Fitzgerald, C. (1999) 'Adaptive Testing Works for You', *Microsoft Certified Professional Magazine Online*, June 1999. Available: http://mcpmag.com/columns/article.asp?EditorialsID=264 [20 May 2007].

Flaugher, R. (2000) Item Pools, in H. Wainer (2000), *Computerized Adaptive Testing: A Primer*, Lawrence Erlbaum Associates Inc.

Foxley, E.; Higgins, C.; Hegazy, T.; Symeonidis, P. & Tsintsifas, A. (2001) 'The CourseMaster CBA System: Improvements over CEILIDH', *Proceedings of*

the 5th Computer-Assisted Assessment Conference 2001, Available: http://www.caaconference.com/pastConferences/2001/proceedings/f1.pdf [10 Jan 2007].

Freeman, R. & Lewis, R. (1998) *Planning and implementing assessment.* London: Kogan Page.

Georgiadou, E.; Triantafillou, E. & Economides, A. A. (2006) 'Evaluation parameters for adaptive testing', *British Journal of Educational Technology*, 37(2), 2006, pp. 261-278.

Gibbs, G. & Habeshaw, T. (1993) *Preparing to Teach: An Introduction to Effective Teaching in Higher Education*. Technical and Educational Services.

Gibbs, G. (2003) Using Assessment Strategically to Change the Way Students Learn in S. Brown & A. Glasner (Eds.) (2003), *Assessment Matters in Higher Education: Choosing and Using Diverse Approaches*. Society for Research into Higher Education, Open University Press.

Gierl, M. J. & Ackerman, T. (1996) 'Software Review: XCALIBRE — Marginal Maximum-Likelihood Estimation Program, Windows Version 1.10', *Applied Psychological Measurement*, 20(3), September 1996, pp. 303-307.

Glas, C. A. W.; Wainer, H. & Bradlow, E. T. (2003) MML and EAP estimation in testlet-based adaptive testing, in W. J. Van der Linden & C. A. W. Glas (Eds.) (2003), *Computerized Adaptive Testing: Theory and Practice*, Kluwer Academic Publishers.

Gonçalves, J. P.; Aluisio, S. M.; de Oliveira, L. H. M. & Oliveira Jr, O. N. (2004) 'A Learning Environment for English for Academic Purposes Based on Adaptive Tests and Task-Based Systems', *Lecture Notes in Computer Science*, 3220, pp. 1-11, 2004.

Guo, F.; Rudner, L. M. & Talento-Miller, E. (2006) 'Differential Impact as an Item Bias Indicator in CAT and Other IRT-based Tests', Research Reports, RR-06-09, July 17, 2006, Available: http://www.gmac.com/NR/rdonlyres/434119CC-6D15-49B1-A146-5D8817547EC8/0/RR0609_DifferentialItemImpact.pdf [19 Nov 2006].

Guzmán, E. & Conejo, R. (2004) 'A Brief Introduction to the New Architecture of SIETTE', *Lecture Notes in Computer Science*, 3137, Adaptive Hypermedia and Adaptive Web-Based Systems, pp. 405-408, 2004.

Guzmán, E. & Conejo, R. (2005) 'Towards Efficient Item Calibration in Adaptive Testing', *Lecture Notes in Computer Science*, User Modeling 2005, 3538, pp. 402-406, 2005.

Guzmán, E., Conejo, R., & García-Hervás, E. (2005) 'An Authoring Environment for Adaptive Testing', *Educational Technology & Society*, 8 (3), 66-76.

Hambleton, R. K. & Cook, L. L. (1983) Robustness of Item Response Models and Effects of Test Length and Sample Size on the Precision of Ability

Estimates, in D. J. Weiss (Ed.) (1983), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, Academic Press Inc.

Hambleton, R. K. & Rogers, H. J. (1991) Advances in criterion-referenced measurement, in R. K. Hambleton & J. C. Zaal (Eds.) (1991), *Advances in Educational and Psychological Testing: Theory and Applications* (Evaluation in Education & Human Services), Kluwer Academic Publishers.

Hambleton, R. K. & Swaminathan, H. (1990) *Item Response Theory: Principles and Applications*. Kluwer-Nijhoff Publishing.

Hambleton, R. K., & Murray, L. (1983) Some goodness of fit investigations for item response models, in R. K. Hambleton (Ed.) (1983), *Applications of Item Response Theory*, Educational Research Institute of British Columbia.

Harvey, J. & Mogey, N. (1999) Pragmatic issues when integrating technology into nether assessment of students, in S. Brown, J. Bull & P. Race (Eds.) (1999), *Computer-Assisted Assessment in Higher Education*, London: Kogan Page Ltd.

He, Q. & Tymms, P. (2004) 'The Development of A Computer Assisted Design, Analysis and Testing System for Analysing Students' Performance', *Proceedings for 8th Computer-Assisted Assessment Conference 2004*, Available: http://www.caaconference.com/pastConferences/2004/proceedings/He_Quingping.pdf [19 Nov 2006].

He, Q. & Tymms, P. (2005) 'A computer-assisted test design and diagnosis system for use by classroom teachers', *Journal of Computer Assisted Learning* 21, pp. 419–429.

Heaning, G. (1989) 'Does the Rasch Model Really Work for Multiple-Choice Items? Take Another Look: A Response to Divgi', *Journal of Educational Measurement*, Spring 1989, Vol. 26, No. 1, pp. 91-97.

Hetter, R. D., & Sympson, J. B. (2001) Item exposure control in CAT-ASVAB, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing: From inquiry to operation*, Washington DC: American Psychological Association.

Ho R.G. & Yen, Y. C. (2005) 'Design and Evaluation of an XML-Based Platform-Independent Computerized Adaptive Testing System', *IEEE Transactions On Education*, Volume 48(2), May 2005.

Hornke, L. F. (2000) 'Item Response Times in Computerized Adaptive Testing', *Psicológica* (2000) 21, 175-189.

Huang, S. X. (1996) 'A Content-Balanced Adaptive Testing Algorithm for Computer-Based Training Systems', *Lecture Notes in Computer Science*, 1086, pp. 306-314, 1996.

Jacobson, R. L. (1993) 'New Computer Technique Seen Producing a Revolution in Educational Testing', *Chronicle of Higher Education*, 40(4), pp. A22-23,26 September 15 1993.

Jettmar, E. & Nass, C. (2002) 'Adaptive Testing: Effects on User Performance', *Proceedings of the 2002 Conference on Human Factors in*

Computer Systems, Minneapolis, Minnesota USA, 20-25 April 2002, pp. 129-134.

Joy, M.; Muzykantskii, B. & Evans, M. (2002) 'An Infrastructure for Web-Based Computer-Assisted Learning', *ACM Journal of Educational Resources* 2(4), December 2002, pp. 1–19.

Julian, E. (1993) CAT: 'What feedback?', *Rasch Measurement Transactions* 6 (4), 1993, p. 246.

Kaburlasos, V. G.; Marinagi, C. C. & Tsoukalas, V. S. (2004) 'PARES: A Software Tool for Computer-Based Testing and Evaluation Used in the Greek Higher Education System', *Proceedings of the IEEE International Conference on Advanced Learning Technologies* (ICALT'04).

Kontio, J.; Lehtola, L. & Bragge, J. (2004) 'Using the Focus Group Method in Software Engineering: Obtaining Practitioner and User Experiences', *Proceedings of the 2004 International Symposium on Empirical Software Engineering* (ISESE'04)

Krimpen-Stoop, E. M. L. A. van & Meijer, R. R. (2003) Detecting person misfit in adaptive testing using statistical process control techniques, in W. J. Van der Linden & C. A. W. Glas (Eds.) (2003), *Computerized Adaptive Testing: Theory and Practice*, Kluwer Academic Publishers.

Laurillard, D. M. (1993) *Rethinking University Teaching: A Framework for the Effective Use of Educational Technology*. Routledge, London.

Lee, J. A.; Moreno, K. & Sympson, J. B. (1986) 'The Effects of Mode of Test Administration on Test Performance', *Journal of Educational and Psychological Measurement*, 46(2), pp. 467-474.

Lilley, M. & 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 University, United Kingdom, pp. 169-184.

Lilley, M. & Barker, T. (2003a) 'An Evaluation of a Computer-Adaptive Test in a UK University Context', *Proceedings of the 7th Computer-Assisted Assessment Conference*, Loughborough University, United Kingdom.

Lilley, M. & Barker, T. (2003b) 'Comparison between Computer-Adaptive Testing and other assessment methods: An empirical study', *Proceedings of the 10th International Conference of the Association for Learning Technology* (ALT-C), University of Sheffield, United Kingdom.

Lilley, M. & Barker, T. (2004) 'A Computer-Adaptive Test that facilitates the modification of previously entered responses: An empirical study', *Lecture Notes in Computer Science*, 3220, 7th International Conference ITS 2004, Volume 3220/2004, pp. 22-33, 2004.

Lilley, M. & Barker, T. (2005a) 'The Use of Item Response Theory in the Development and Application of a User Model for Automatic Feedback: A Case Study', *Proceedings of the 19th British HCI Group Annual Conference*, Napier University, Edinburgh, United Kingdom.

Lilley, M. & Barker, T. (2005b) 'An empirical study into the effect of question review in a computer-adaptive test', *Proceedings of the 6th Annual Higher Education Academy Subject Network for Information Computer Science Conference*, University of York, United Kingdom.

Lilley, M. & Barker, T. (2005c) 'Computer-adaptive testing: A case study', *Proceedings of the 6th Annual Higher Education Academy Subject Network for Information Computer Science Conference*, University of York, United Kingdom.

Lilley, M. & Barker, T. (2006a) 'Computerised adaptive testing: extending the range of assessment formats in a Computer Science course', *Proceedings of ICL2006 Conference*, September 27 -29, 2006 Villach, Austria.

Lilley, M. & Barker, T. (2006b) 'Student attitude to adaptive testing', *Proceedings of HCI 2006 Conference*, Queen Mary, University of London, 11-15, September 2006.

Lilley, M. & Barker, T. (2006c) 'Students' perceived usefulness of formative feedback for a computer-adaptive test', *Proceedings of ECEL 2006: The European Conference on e-Learning*, University of Winchester, 11-12 September 2006.

Lilley, M. & Barker, T. (2007) 'Students' perceived usefulness of formative feedback for a computer-adaptive test', Electronic *Journal of e-Learning* (EJEL), 5(1), February 2007, Special Issue (ECEL 2006) Available: http://www.ejel.org/Volume-5/v5-i1/v5-i1-art-5.htm [17 May 2007].

Lilley, M.; Barker, T. & Britton, C. (2003a) 'Review and Modification of Responses in a Computer-Adaptive Test: Preliminary Considerations', *Proceedings of the 2nd International Conference on Information and Communication Technologies in Education*, Junta de Extremadura Consejería de Educación, Ciencia y Tecnología, Badajoz, Spain.

Lilley, M.; Barker, T. & Britton, C. (2004a) 'The development and evaluation of a software prototype for computer adaptive testing', *Computers & Education Journal* 43(1-2), pp. 109-123.

Lilley, M.; Barker, T. & Britton, C. (2004b) 'The generation of automated student feedback for a computer-adaptive test', *Proceedings of the 8th Computer-Assisted Assessment Conference*, Loughborough University, United Kingdom.

Lilley, M.; Barker, T. & Britton, C. (2005a) 'Learners' perspectives on the usefulness of an automated tool for feedback on test performance', *Proceedings of the 4th European Conference on E-Learning*, Royal Netherlands Academy of Arts & Sciences, Amsterdam, Netherlands.

Lilley, M.; Barker, T. & Britton, C. (2005b) 'Automated feedback for a computer-adaptive test: A case study', *Proceedings of the 9th Computer-Assisted Assessment Conference*, Loughborough University, United Kingdom.

Lilley, M.; Barker, T. & Britton, C. (2005c) 'Learners' perceived level of difficulty of a computer-adaptive test: A case study', Proceedings of the 10th International Conference on Human-Computer Interaction, *Lecture Notes in Computer Science*, 3585, pp. 1026-1029.

Lilley, M.; Barker, T. & Britton, C. (2005d) 'The generation of automated learner feedback based on individual proficiency levels', Proceedings of the 18th International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, *Lecture Notes in Artificial Intelligence,* 3533, pp. 842-844.

Lilley, M.; Barker, T. & Britton, C. (2007) 'Computer Adaptive Testing in Higher Education: A case study', *Proceedings of 2007 Annual Solstice Conference*, Edge Hill University, United Kingdom.

Lilley, M.; Barker, T. & Maia, M. (2002a) 'The use of Objective Items in Higher Education: Potential and Limitations', *Proceedings of the 37th Asamblea del Consejo Latinoamericano de Escuelas de Administración (CLADEA)*, Porto Alegre, Brazil.

Lilley, M.; Barker, T. & Maia, M. (2002b) 'Web-based adaptive testing in distance learning: an overview', *Proceedings of the 5th Simpósio de Administração da Produção, Logística e Operações Internacionais (SIMPOI)*, Fundação Getulio Vargas Escola de Administração de Empresas de São Paulo, Brazil.

Lilley, M.; Barker, T. & Maia, M. (2003b) 'Computer-Adaptive Testing in Higher Education: the way forward?', *Proceedings of the 38th Asamblea del Consejo Latinoamericano de Escuelas de Administración (CLADEA)*, Lima, Peru.

Lilley, M.; Barker, T. & Maia, M. (2003c) 'Do Cognitive Styles of Learning Affect Student Performance in Computer-Adaptive Testing?', *Proceedings of the 6th Simpósio de Administração da Produção, Logística e Operações Internacionais (SIMPO I)*, Fundação Getulio Vargas Escola de Administração de Empresas de São Paulo, Brazil.

Lilley, M.; Barker, T. & Maia, M. (2003d) 'The Evaluation of a Computer-Adaptive Test', *Proceedings of the 27th Encontro da Associação Nacional dos Programas de Pós-Graduação em Administração (ANPAD),* Atibaia, Brazil.

Lilley, M.; Barker, T., Bennett, S. & Britton, C. (2002c) ,How computers can adapt to knowledge: A comparison of computer-based and computer-adaptive testing', *Proceedings of the 1st International Conference on Information and Communication Technologies in Education*, Junta de Extremadura Consejería de Educación, Ciencia y Tecnología, Badajoz, Spain.

Linacre, J. M. (2000) A measurement approach to computer-adaptive testing of reading comprehension, in M. Chalhoub-Deville, M. Milanovic & C. J. Weir (Eds.) (2000), *Issues in Computer-Adaptive Testing of Reading*

Proficiency: Studies in Language Testing 10 (Studies in Language Testing) Cambridge University Press.

Litosseliti, L. (2003) *Using Focus Groups in Research*. Continuum International Publishing Group Ltd.

Lord, F. M. (1971a) 'The Self-Scoring Flexilevel', *Journal of Educational Measurement*, 8(3), September 1971, pp. 147-151.

Lord, F. M. (1971b) 'A theoretical study of two-stage testing', *Journal Psychometrika*, 36(3), September 1971, pp. 227-242.

Lord, F. M. (1980) *Applications of Item Response Theory to Practical Testing*. Lawrence Erlbaum Associates Inc.

Lord, F. M. (1983) Small *N* Justifies Rasch Model, in D. J. Weiss (Ed.) (1983), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, Academic Press Inc.

Lunz, M. E.; Bergstrom, B. E. & Wright, B. D. (1992) 'The Effect of Review on Student Ability and Test Efficiency for Computerized Adaptive Tests', *Applied Psychological Measurement*,16(1), March 1992, pp. 33-40.

Lütticke, R. (2004) 'Problem Solving with Adaptive Feedback', Adaptive Hypermedia and Adaptive Web-Based Systems 2004, *Lecture Notes in Computer Science*, 3137, pp. 417-420, 2004.

Martin, B. & Mitrovic, A. (2005) 'Using Learning Curves to Mine Student Models', *Lecture Notes in Computer Science*, 3538, pp. 79 – 88, 2005.

McAteer, E. & Shaw, R. (1994) *Courseware in Higher Education Evaluation 1: Planning, Developing and Testing*. EMASHE Project, University of Glasgow.

McBeath, R. J. (Ed.) (1992) Instructing and Evaluating Higher Education: a guidebook for planning learning outcomes, Englewood Cliffs, Educational Technology Publications, in J. Bull & C. McKenna (Eds.) (2004) *Blueprint for Computer-assisted Assessment*. London: Routledge Falmer.

McBride, J. R. & Martin, J. T. (1983) Reliability and Validity of Adaptive Ability Tests in a Military Setting, in D. J. Weiss (1983), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, Academic Press Inc.

McBride, J. R. (2001a) Research Antecedents of Applied Adaptive Testing, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing: From inquiry to operation*, Washington DC: American Psychological Association.

McBride, J. R. (2001b) The Marine Corps Exploratory Development Project: 1977-1982, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing: From inquiry to operation*, Washington DC: American Psychological Association.

McBride, J. R. (2001c) Technical Perspective, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing: From inquiry to operation*, Washington DC: American Psychological Association.

Microsoft Corporation (2006) *Exam and Testing Procedures*, Posted: April 15, 2006. Available: http://www.microsoft.com/learning/mcpexams/faq/procedures.asp [19 Nov 2006].

Microsoft Corporation (2007a) *ADO*, Available: http://msdn2.microsoft.com/en-us/library/ms805098.aspx [11 Nov 2007]

Microsoft Corporation (2007b) *Preparing Your Access 2003 Database for Deployment*, Available: http://msdn2.microsoft.com/en-us/library/aa662933(office.11).aspx [11 Nov 2007]

Microsoft Corporation (2007c) *Microsoft Active Server Pages: Frequently Asked Questions*, Available: http://msdn2.microsoft.com/en-us/library/ms972347.aspx [11 Nov 2007]

Miller, A.; Imrie, B.W. & Cox, K. (1998) *Student Assessment in Higher Education: A Handbook for Assessing Performance*. London: Routledge Falmer.

Mitrovic, A. & Martin, B. (2004) 'Evaluating Adaptive Problem Selection', *Lecture Notes in Computer Science*, 3137, pp. 185–194, 2004.

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

Moreno, K. E. & Segall, D. O. (2001) Validation of the Experimental CAT-ASVAB System, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing: From inquiry to operation*, Washington DC: American Psychological Association.

Nielsen, J. (2005). *Ten Usability Heuristics* (Online), Available http://www.useit.com/papers/heuristic/heuristic_list.html [Accessed 21 Aug 2007].

Olea, J.; Revuelta, J.; Ximénez, M.C. & Abad, F. J. (2000) 'Psychometric and psychological effects of review on computerized fixed and adaptive tests', *Psicológica* (2000), 21, pp. 157-173.

Pérez, D. & Alfonseca, E. (2004) Adapting the Automatic Assessment of Free-Text Answers to the Students, *Proceedings for 8th Computer-Assisted Assessment Conference 2004*, Available: http://www.caaconference.com/pastConferences/2005/proceedings/PerezD_AlfonsecaE.pdf [19 Nov 2006].

Preece, J.; Rogers, Y. & Sharp, H. (2002) *Interaction Design: Beyond Human-Computer*. John Wiley and Sons Ltd.

Preece, J.; Rogers, Y.; Sharp, H.; Benyon, D.; Holland, S. & Carey, T. (1994) *Human-Computer Interaction*. Addison Wesley.

Pritchett, N. (1999) Effective Question Design, in S. Brown, J. Bull & P. Race (Eds.) (1999), *Computer-Assisted Assessment in Higher Education*, London: Kogan Page Ltd.

QSR International (2007) *Software for Qualitative Research – From Academic and Social Research to Market Research*, Avialable from http://www.qsrinternational.com/ [Accessed 11 Jun 2007]

Race, P.; Brown, S. & Smith, B. (2004) *500 Tips on Assessment*. Routledge Falmer.

Redmond-Pyle, D. & Moore, A. (1995) *Graphical User Interface Design and Evaluation: A Practical Process*. Prentice-Hall BCS Practitioner.

Revuelta, J., & Ponsada, V. (1998) 'A comparison of item exposure control methods in computerized adaptive testing', *Journal of Educational Measurement*, 35, pp. 311-327.

Revuelta, J., Ximénez, M. C. & Olea, J. (2000) 'Psychometric and psychological effects of review on computerized testing', *Educational and Psychological Journal*, 63 (5), pp. 791-808.

Rhodes, G. & Tallantyre, F. (2003) Assessment of Key Skills in S. Brown & A. Glasner (Eds.) (2003), *Assessment Matters in Higher Education: Choosing and Using Diverse Approaches*. Society for Research into Higher Education, Open University Press.

Robinson, J. M. (1999) Computer-assisted peer review, in S. Brown, J. Bull & P. Race (Eds.) (1999), *Computer-Assisted Assessment in Higher Education*, London: Kogan Page Ltd.

Romero, C.; Ventura, S.; Hervás, C. & De Bra, P (2006) 'An Authoring Tool for Building Both Mobile Adaptable Tests and Web-Based Adaptive or Classic Tests', *Lecture Notes in Computer Science*, 4018, pp. 203–212, 2006.

Rudner, L. M. (2001) *Measurement decision theory* (SuDoc ED 1.310/2:457164) U.S. Dept. of Education, Office of Educational Research and Improvement, Educational Resources Information Center.

Sambell, K.; Sambell, A. & Sexton, G. (1999) Student perceptions of the learning benefits of computer-assisted assessment: a case study in electronic engineering, in S. Brown, J. Bull & P. Race (Eds.) (1999), *Computer-Assisted Assessment in Higher Education*, London: Kogan Page Ltd.

Samejima, F. (1969) 'Estimation of latent ability using a response pattern of graded scores', *Psychometrika Monograph Supplement*, 34(4), pp. 100-114.

Sands, W. A. & Waters, B. K. (2001) Introduction to ASVAB and CAT, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing: From inquiry to operation*, Washington DC: American Psychological Association.

Schoonman, W. (1989) *Applied Study on Computerized Adaptive Testing*. Swets & Zeitlinger.

Segall, D. O. & Moreno, K. E. (2001) Current and Future Challenges, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing: From inquiry to operation*, Washington DC: American Psychological Association.

Yorke M (2003) 'Formative assessment in Higher Education: Moves towards theory and the enhancement of pedagogic practice', *Higher Education*, 45, pp. 477–501.

Segall, D. O. (2001) The Psychometric Comparability of Computer Hardware, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing: From inquiry to operation*, Washington DC: American Psychological Association.

Segall, D. O., Moreno, K. E., Kieckhaefer, W. F., Vicino, F. L. & McBride, J. R. (2001) Validation of the Experimental CAT-ASVAB System, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing: From inquiry to operation*, Washington DC: American Psychological Association.

Steven, C. & Hesketh, I. (1999) Increasing learner responsibility and support with the aid of adaptive formative assessment using QM designer software, in S. Brown, J. Bull & P. Race (Eds.) (1999), *Computer-Assisted Assessment in Higher Education*, London: Kogan Page Ltd.

Stocking, M.L. & Lewis, C. (2003). Methods of controlling the exposure of items in CAT. in W. J. Van der Linden & C. A. W. Glas (Eds.) (2003), *Computerized Adaptive Testing: Theory and Practice*, Kluwer Academic Publishers.

Swaminathan, H. (1991) Analysis of Covariance Structures, in R. K. Hambleton & J. C. Zaal (Eds.) (1991), *Advances in Educational and Psychological Testing: Theory and Applications* (Evaluation in Education & Human Services), Kluwer Academic Publishers.

Thelwall, M. (2000) 'Computer-based assessment: a versatile educational tool', *Computers and Education*, 34(1), January 2000, pp. 37-49(13) Available: http://www.ingentaconnect.com/content/els/03601315/2000/00000034/00000001/art00037 [1 Jul 2007].

Thissen, D. & Mislevy, R. J. (2000) Testing Algorithms, in H. Wainer (2000), *Computerized Adaptive Testing: A Primer*, Lawrence Erlbaum Associates Inc.

Thissen, D. (1983) Timed Testing: An Approach Using Item Response Theory, in D. J. Weiss (1983), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, Academic Press Inc.

Trentin, G. (1997) 'Computerized adaptive tests and formative assessment', *Journal of Educational Multimedia and Hypermedia*, 6(2), 1997, pp. 201-220.

Tzanavari, A.; S. & Pastellis, P. (2004) 'Giving More Adaptation Flexibility to Authors of Adaptive Assessments', Adaptive Hypermedia and Adaptive Web-Based Systems 2004, *Lecture Notes in Computer Science*, 3137/2004, pp. 340–343, 2004.

Van der Linden, W. J. & Hambleton, R. K. (Eds.) (2000) *Handbook of Modern Item Response Theory*. Springer-Verlag New York Inc.

Veerkamp, W. J. J. & Berger, M. P. F. (1999) 'Optimal Item Discrimination and Maximum Information for Logistic IRT Models', *Applied Psychological Measurement*, 23(1), March 1999, pp. 31-40.

Vicino, F. L. & Moreno, K. E. (2001) Human Factors in the CAT System: A Pilot Study, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing: From inquiry to operation*, Washington DC: American Psychological Association.

Vispoel, W. P. (1998) 'Reviewing and Changing Answers on Computer-adaptive and Self-adaptive Vocabulary Tests', *Journal of Educational Measurement*, Winter 1998, 35(4), pp. 328-347.

Vispoel, W. P.; Hendrickson, A. B. & Bleiler, T. (2000) 'Limiting Answer Review and Change on Computerized Adaptive Vocabulary Tests: Psychometric and Attitudinal Results', *Journal of Educational Measurement*,37(1), Spring 2000, pp. 21-38.

Vispoel, W. P.; Rocklin, T. R.; Wang, T. & Bleiler, T. (1999) 'Can Examinees Use a Review Option to Obtain Positively Biased Ability Estimates on a Computerized Adaptive Test?', *Journal of Educational Measurement*, 36 (2), 141–157.

Vogel, L A. (1994) 'Explaining Performance on P&P versus Computer Mode of Administration for the Verbal Section of the Graduate Record Exam', *Journal of Educational Computing Research*, 11(4), pp. 369-383.

Vos, H. J. (2000) 'A Bayesian Procedure in the Context of Sequential Mastery Testing', *Psicológica* (2000), 21, pp. 191-211.

Wainer, H. & Eignor, D. (2000) Caveats, Pitfalls, and Unexpected Consequences of Implementing Large-Scale Computerized Testing, in H. Wainer (2000), *Computerized Adaptive Testing: A Primer*, Lawrence Erlbaum Associates Inc.

Wainer, H. & Mislevy, R. J. (2000) Item Response Theory, Item Calibration, and Proficiency Estimation, in H. Wainer (2000), *Computerized Adaptive Testing: A Primer*, Lawrence Erlbaum Associates Inc.

Wainer, H. (2000a) Introduction and History, in H. Wainer (2000), *Computerized Adaptive Testing: A Primer*, Lawrence Erlbaum Associates Inc.

Wainer, H. (2000b) 'CATs: Whither and whence', *Psicológica* (2000), año/vol. 21, número 1, pp. 121-133.

Wainer, H.; Dorans, N. J.; Green, B. F.; Mislevy, R. J.; Steinberg, L. & Thissen, D. (2000) Future Challenges, in H. Wainer (2000), *Computerized Adaptive Testing: A Primer*, Lawrence Erlbaum Associates Inc.

Warburton, B. & Conole, G. (2003) 'Key Findings from recent literature on Computer-aided Assessment', *Proceedings of the Association for Learning Technology Conference* (ALT-C 2003)

Warburton, B. & Conole, G. (2004) 'Whither E-Assessment?', *Proceedings for 9th Computer-Assisted Assessment Conference 2005*, Available:

http://www.caaconference.com/pastConferences/2005/proceedings/WarburtonB_ConoleG.pdf [26 November 2006].

Ward, C. (1981) *Preparing and Using Objective Questions*. Handbooks for Further Education. Nelson Thornes Ltd.

Ward, W. C. (1988) 'The College Board Computerized Placement Tests: An Application of Computerized Adaptive Testing', *Journal of Machine-Mediated Learning*, v2, pp. 271-82.

Weiss, D. J. & Yoes, M. E. (1991) Item Response Theory, in R. K. Hambleton & J. C. Zaal (Eds.) (1991), *Advances in Educational and Psychological Testing: Theory and Applications* (Evaluation in Education & Human Services), Kluwer Academic Publishers.

Weiss, D. J. (Ed.) (1983) *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, Academic Press Inc.

Wheadon, C. & He, Q. (2006) 'An Investigation of the Response Time for Maths Items in a Computer Adaptive Test', *Proceedings of the 10th Computer-Assisted Assessment Conference 2006,* Available: http://www.caaconference.com/pastConferences/2006/proceedings/Wheadon_C_He_Q_j3.pdf [10 Jan 2007].

Williams, J. H. S.; Maher, J.; Spencer, D.; Barry, M. D. J. & Board, E. (1999) Automatic test generation from a database, in S. Brown, J. Bull & P. Race (Eds.) (1999), *Computer-Assisted Assessment in Higher Education*, London: Kogan Page Ltd.

Wolfe, J. H., Moreno, K. E. & Segall, D. O. (2001a) Evaluating the Predictive Validity of CAT-ASVAB, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing: From inquiry to operation*, Washington DC: American Psychological Association.

Wolfe, J. H.; McBride, J. R. & Sympson, J. B. (2001b) Development of the Experimental CAT-ASVAB System In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized adaptive testing: From inquiry to operation. Washington DC: American Psychological Association.

Yao, T. (1991) 'CAT with a Poorly Calibrated Item Bank', *Rasch Measurement Transactions*, 5 (2), 1991, p. 141.

Yong, C. F. & Higgins, C. (2004) 'Self-assessing with Adaptive Exercises', *Proceedings for 8th Computer-Assisted Assessment Conference 2004*, Available: http://www.caaconference.com/pastConferences/2004/proceedings/Yong.pdf [19 Nov 2006].

Appendix A Glossary

Glossary

Ability A theoretical value indicating the level of a test-taker on the ability or trait measured by the test. In this work, the terms 'ability' and 'proficiency level' are used interchangeably.

Bloom's taxonomy of cognitive skills A six level classification system for categorising the level of abstraction of questions that occur in educational settings: knowledge, comprehension, application, analysis, synthesis and evaluation. The taxonomy was proposed in 1956 by Benjamin Bloom, an educational psychologist at the University of Chicago.

Computer-adaptive test A computer-assisted assessment application where the questions administered during a test are dynamically selected based on test-taker performance.

Computer-assisted assessment Also known as Computer-Aided Assessment and e-Assessment. A general term used to describe the use of computers to support student assessment.

Computer-based test A computer-assisted assessment application where test-takers are presented with a set of fixed questions.

Construct validity The degree to which a test instrument measures what is intended to measure.

Content validity The extent to which the test samples the content and the objectives set out in the specification as determined by experts.

Diagnostic assessment A special form of summative assessment that measures a student's current knowledge and skills for the purpose of identifying a suitable program of learning.

Difficulty parameter A measure of a question's complexity. In the 3-PL model, the item difficulty parameter is denoted by b.

Discrimination parameter A term referring to an item's potential to differentiate between test-takers. In the 3-PL model, the item discrimination parameter is denoted by a.

Face validity A subjective measure that represents the extent to which the test 'appears valid'.

Formative assessment The primary purpose of formative assessment is to help students improve. Formative assessment should not be used for grading purposes.

Guessing parameter In the 3-PL model, the guessing (or pseudo-chance) parameter is denoted by c. It represents the probability of a test-taker answering an item correctly by chance alone.

Item Characteristic Curve "The curve that portrays the probability of a correct response to a test item as a function of trait levels that would give rise to those probabilities" (Weiss, 1983: p. 2). If a test-taker's ability is the same as the difficulty level of the item, that test-taker has a 50-50 chance of answering that item right. If the ability is less, that probability decreases. The relationship between the test-takers' item performance and the abilities underlying item performance is described in an item characteristic curve (ICC).

Item exposure The number of times a question is presented to test takers.

Item Response Theory A Latent Trait Model which is based on a relationship between the observable test performance of examinees and the unobservable traits or abilities which underlie that performance. It uses statistical techniques to estimate the probability of an examinee with an unknown ability θ answering an item correctly.

Item A common term referring to an individual question used within a test.

Latent trait The knowledge dimension on which test items rely, to some extent, for their correct response.

Objective item See objective question.

Objective question Type of question which has a single correct answer. Examples of objective questions include: true/false, multiple-choice and multiple-response.

Objective test A test instrument containing only objective questions.

One-Parameter Logistic Model This is the simplest IRT model for dichotomously scored items. It has only one parameter, i.e. item difficulty b.

Proficiency level In this work, the terms 'proficiency level' and 'ability' are used interchangeably. For a definition, see ability.

Rasch Model The Rasch model for dichotomous data is often regarded as a special case of the 2-PL model and thus the 3-PL model, where a=1 and c=0.

Reliability The extent to which a test's results are repeatable and fair from one examinee to the next, and from one occasion to the next. A reliable assessment is one which consistently achieves the same results with the same (or similar) cohort of examinees.

Self-assessment A special form of formative assessment which involves students assessing themselves.

Summative assessment The main purpose of this form of assessment is to make a judgement regarding each student's performance. Summative assessment results are typically used for grading purposes, or pass/fail decisions.

Test-taker A general term used in this work to refer to a student taking part in a test.

Three-Parameter Logistic Model IRT model employed in this work. As its name implies, this model has three parameters: item difficulty b, item discrimination a and pseudo-chance c.

Two-Parameter Logistic Model This IRT model has two parameters, i.e. difficulty b and item discrimination a.

Validity The degree to which the test instrument actually measures what it purports to measure. It relates to the appropriacy of the inferences made on the basis of the test scores. In this work, the following types of validity were considered: face, content, and construct.

XCalibre Software application produced by Assessment Systems Corporation (USA) for IRT item parameter estimation. XCalibre uses marginal maximum-likelihood estimation methods for obtaining IRT parameter estimates.

θ The Greek letter θ (Theta) is used to represent a test-taker's ability estimate.

Appendix B Focus group guidelines

This appendix contains the focus group guidelines used in the study concerned with test-taker attitude towards the CAT approach reported in sections 3.5.3 and 4.2.

Test-taker attitude towards the CAT approach: Focus group guidelines

1. Introduction

- Welcome all participants.
- Introduce focus group moderator.
- Clarify purpose of the focus group, i.e. to investigate (1) usability issues related to the user interface and (2) participants' attitude towards the use of computerised adaptive testing in summative and formative assessments.
- Discuss ground rules with participants, i.e. all participants should contribute equally; all contributions are equally important and therefore should be valued by all participants; what is discussed in the focus group, should remain within the focus group.

2. Confidentiality issues

- Assure participants that what they say during the focus group will be kept confidential to the research team and, if published, will not be identifiable as belonging to them.
- Inform participants that the session will be recorded on video.
- Inform participants about their right to withdraw from the focus group session at any time.

3. Exploratory questions

- Ask participants what they thought of the test that they had just taken.
- Provide students with a copy of the CBT (fixed) questions that they have answered.

- Ask participants to describe the difference between the CBT and CAT sections of the test.
- Explain the difference between the CBT and CAT approaches.
- Ask participants whether they heard about the CAT approach before.

4. Key questions

- Introduce the following components of the CAT approach: question selection method, scoring method and stopping condition.
- Explore the following issues:
  - Different stopping conditions (i.e. fixed and variable-length CAT) in formative and summative assessment settings;
  - Different sets of questions and scoring;
  - Different sets of questions and cheating;
  - Preferred types of assessments (e.g. coursework, exam, CAT, and CBT) in formative and summative assessment settings.

5. Ending questions

- Summarise the issues that emerged from the discussion.
- Ask participants whether the summary is adequate.
- Ask participants of all the issues discussed during the focus group session, which was the most important.

Appendix C Observation study guidelines

This appendix contains the guidelines used in the observation study discussed in section 4.1.

CAT software prototype: Observation study

1. Prior to the observation study

- Load the CAT software prototype application on enough machines (one computer per participant).
- Ensure that monitor, keyboard and mouse are all operational.
- Ensure that room lighting and temperature are good.

2. Introduction

- Welcome all participants.
- Introduce observer(s).
- Clarify purpose of the observation, i.e. to investigate usability issues relating to the user interface of the CAT software prototype.
- Inform participants that the observation should take no longer than 40 minutes.
- Inform participants that, should they need support in using the CAT software prototype, they can request help from the observers at any point.

3. Confidentiality issues

- Assure participants that their behaviour during the session will be kept confidential to the research team and, if published, will not be identifiable as belonging to them.
- Inform participants that the observers will be taking notes during the session.
- Inform participants about their right to withdraw from the observation study at any time.

4. Guidance notes to observers

- Assign participants to computers randomly.
- Try to be as unobtrusive as possible.
- Note down any relevant events; in particular, any questions from the participants.
- Try to be aware of any factors that might be affecting users' interaction with the system.
- Write down your first impressions immediately after the observation session is concluded.

Appendix D Interview guidelines

This appendix highlights the interview guidelines used in the study reported in section 4.3.

Test-taker attitude towards the CAT approach: Interview guidelines

1. Introduction

- Welcome interviewee.
- Briefly introduce the topic of the interview (i.e. to explore issues relating to the CAT test they had just taken), and the interviewer.

2. Confidentiality issues

- Assure interviewees that what they say during the interview will be kept confidential to the research team and, if published, will not be identifiable as belonging to them.
- Inform interviewees about their right to withdraw from the interview at any point.

3. Key questions

- What do you think of the test that you had just taken?
- Do you think that the test was good at assessing how much you have learned as part of the course (to date)? Do you think that any topics have been omitted or over emphasised?
- What do you think of the level of difficulty of the test that you have just taken? How does it compare to other test that you had taken in the past?
- I can see that you have rated the overall difficulty of the test as being … Can you say more about the reasons why you rated the test this way?
- What do you think of the application that you have just used? Was it easy to use?

Appendix E Perceived level of difficulty

This appendix shows a copy of the questionnaire used in the studies reported in sections 4.3.1 and 4.3.2.

PC_LAB_001

University of Hertfordshire

School of Computer Science

Full name: :		SRN:	
Test No.:		Module code:	

Please note that:

- Any information that you provide will be treated confidentially and, if published, will not be identifiable as belonging to you.
- You have the right to withdraw from the research at any time.

Please rate the difficulty of the test that you have just taken:

1 Very difficult	**2 Difficult**	**3 Just right**	**4 Easy**	**5 Very easy**

Please add any comments here (please continue overleaf if necessary):

Thank you for your participation!

Appendix F Automated feedback evaluation questionnaire (1)

This appendix shows a copy of the questionnaire used in the study reported in section 8.1.1.

University of Hertfordshire

School of Computer Science

Full name: :		SRN:	
Test No.:		Module code:	

Please note that:

- Any information that you provide will be treated confidentially and, if published, will not be identifiable as belonging to you.
- You have the right to withdraw from the research at any time.

Thank you for participating in this evaluation. Please follow the steps below:

1. Log into the application to be evaluated (http://chico/review/2com0062/);
2. Inspect your individual feedback;
3. Rate the statements below as you work through the application;
4. Add any additional comments to the text box provided.
5. Log out.

Please rate the statements below.

1. Overall, the feedback application was effective at providing helpful advice for individual development.

1	**2**	**3**	**4**	**5**
Strongly disagree	**Disagree**	**Neither agree, nor disagree**	**Agree**	**Strongly agree**

2. Overall, the feedback application was effective at providing feedback on performance.

1	**2**	**3**	**4**	**5**
Strongly disagree	**Disagree**	**Neither agree, nor disagree**	**Agree**	**Strongly agree**

3. The “Overall Score” section was useful at providing information on how successfully I have learned.

1	**2**	**3**	**4**	**5**
Strongly disagree	**Disagree**	**Neither agree, nor disagree**	**Agree**	**Strongly agree**

4. The “Performance Summary per Topic” was useful at providing information on how successfully I have learned in each topic area.

1	**2**	**3**	**4**	**5**
Strongly disagree	**Disagree**	**Neither agree, nor disagree**	**Agree**	**Strongly agree**

5. The "Step-by-Step Personalised Revision Plan" was useful at providing information on how successfully I have learned.

1	**2**	**3**	**4**	**5**
Strongly disagree	**Disagree**	**Neither agree, nor disagree**	**Agree**	**Strongly agree**

6. The content of the feedback was appropriate for my individual performance.

1	**2**	**3**	**4**	**5**
Strongly disagree	**Disagree**	**Neither agree, nor disagree**	**Agree**	**Strongly agree**

Please add any comments here (please continue overleaf if necessary):

Thank you for your participation!

Appendix G Perceived usefulness of the automated feedback

This appendix shows a copy of the questionnaire used in the study reported in section 8.1.2.

University of Hertfordshire

School of Computer Science

Full name: :		SRN:	
Test No.:		Module code:	

Please note that:

- Any information that you provide will be treated confidentially and, if published, will not be identifiable as belonging to you.
- You have the right to withdraw from the research at any time.

Thank you for participating in this evaluation. Please follow the steps below:

1. Log into the application to be evaluated (http://chico/review/2com0062/);
2. Inspect your individual feedback;
3. Rate the statements below as you work through the application;
4. Add any additional comments to the text box provided.
5. Log out.

How would you rate the usefulness of this feedback page?

1	**2**	**3**	**4**	**5**
Not useful		**Useful**		**Very useful**

Use the space provided below to tell us why you rated the feedback page this way or any other comments (please continue overleaf if necessary).

Thank you for your participation!

Appendix H Automated feedback evaluation questionnaire (2)

This appendix shows a copy of the questionnaire used in the study reported in section 8.1.3.

University of Hertfordshire

School of Computer Science

Full name: :		SRN:	
Test No.:		Module code:	

Please note that:

- Any information that you provide will be treated confidentially and, if published, will not be identifiable as belonging to you.
- You have the right to withdraw from the research at any time.

Thank you for participating in this evaluation. Please follow the steps below:

- Log into the application to be evaluated (http://chico/review/2com0062/);
- Inspect your individual feedback;
- Rate the statements below as you work through the application;
- Add any additional comments to the text box provided.
- Log out.

Please rate the statements below.

1. The "Your Score" section would be useful at providing information on how successfully I have learned.

1	**2**	**3**	**4**	**5**
Strongly disagree	**Disagree**	**Neither agree, nor disagree**	**Agree**	**Strongly agree**

2. The “Your performance per topic area” diagram would be useful at providing information on how successfully I have learned.

1	**2**	**3**	**4**	**5**
Strongly disagree	**Disagree**	**Neither agree, nor disagree**	**Agree**	**Strongly agree**

3. The “Step-by-Step Personalised Revision Plan” section would be useful at providing feedback for individual development.

1	**2**	**3**	**4**	**5**
Strongly disagree	**Disagree**	**Neither agree, nor disagree**	**Agree**	**Strongly agree**

4. Using the application would enable me to receive feedback on performance more quickly.

1	**2**	**3**	**4**	**5**
Strongly disagree	**Disagree**	**Neither agree, nor disagree**	**Agree**	**Strongly agree**

5. Using the application would be effective in identifying my strengths and weaknesses.

1	**2**	**3**	**4**	**5**
Strongly disagree	**Disagree**	**Neither agree, nor disagree**	**Agree**	**Strongly agree**

6. I would find the application easy to use.

1	**2**	**3**	**4**	**5**
Strongly disagree	**Disagree**	**Neither agree, nor disagree**	**Agree**	**Strongly agree**

Please add any comments here (please continue overleaf if necessary):

Thank you for your participation!

Appendix I Heuristic evaluation questionnaire

This appendix shows a copy of the heuristic evaluation guidelines pertaining to the study reported in section 5.1.

Heuristic evaluation

Instructions

Thank you for attending the presentation about our Computer-Adaptive Test (CAT) software prototype and accepting to participate in its evaluation. We are interested in drawing up on your expertise in order to conduct a heuristic evaluation of the prototype. To this end, you have been provided with:

- A copy of the CAT software prototype on disk;
- A copy of Nielsen's heuristics;
- A questionnaire[1].

Please contact Mariana Lilley (m.lilley@herts.ac.uk) or Dr Trevor Barker (t.1.barker@herts.ac.uk) should you:

- have any questions about this evaluation;
- wish to return the completed questionnaire.

Thank you very much for your contribution to this research.

[1] The questionnaire is based on Preece et al. (2002: pp. 408-409).

Heuristic evaluation

	Poor 1	2	3	4	Excellent 5
1. Visibility of system status Are users kept informed about what is going on? Is appropriate feedback provided within reasonable time about a user's action?					
2. Match between system and the real world Is the language used at the interface simple? Are the words and phrases used familiar to the user?					
3. User control and freedom Are there ways of allowing users to easily escape from places they unexpectedly find themselves in?					
4. Consistency and standards Are the ways of performing similar actions consistent?					
5. Error prevention Is it easy to make errors? If so, where and why?					
6. Recognition rather than recall Are actions and options always visible?					

Heuristic evaluation

	Poor 1	2	3	4	Excellent 5
7. Flexibility and efficiency of use Have accelerators been provided that allow more experienced users to carry out tasks more quickly?					
8. Aesthetic and minimalist design Is any unnecessary and irrelevant information provided?					
10. Help users recognize, diagnose, and recover from errors 11. Are error messages helpful? Do they use plain language to describe the nature of the problem and suggest a way of solving it?					
10. Help and documentation Is help information provided that can be easily searched and easily followed?					

Please use this box to expand on the information you have provided above (please continue overleaf if necessary). Thank you.

Ten Usability Heuristics by Jakob Nielsen [2]

These are ten general principles for user interface design. They are called "heuristics" because they are more in the nature of rules of thumb than specific usability guidelines.

Visibility of system status

The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

Match between system and the real world

The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

User control and freedom

Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

Consistency and standards

Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

Error prevention

Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.

Recognition rather than recall

Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

[2] The list of heuristics was extracted from http://www.useit.com/papers/heuristic/heuristic_list.html.

Flexibility and efficiency of use

Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

Aesthetic and minimalist design

Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

Help users recognize, diagnose, and recover from errors

Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

Help and documentation

Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Appendix J Pedagogical evaluation questionnaire

This appendix shows a copy of the questionnaire used in the pedagogical evaluation discussed in section 5.2.

Pedagogical evaluation

Instructions

Thank you for attending the presentation about our Computer-Adaptive Test (CAT) software prototype and accepting to participate in its evaluation. We are interested in drawing up on your expertise in order to evaluate the pedagogical value of the prototype. To this end, you have been provided with:

- A copy of the CAT software prototype on disk;
- A questionnaire.

Please contact Mariana Lilley (m.lilley@herts.ac.uk) or Dr Trevor Barker (t.1.barker@herts.ac.uk) should you:

- have any questions about this evaluation;
- wish to return the completed questionnaire.

Thank you very much for your contribution to this research.

Pedagogical evaluation

	Unlikely 1	2	3	4	Likely 5
1. Summative assessment					
CAT would enable lecturers to mark summative assessments more quickly.					
CAT would enable lecturers to mark summative assessments more accurately.					
CAT as summative assessment tool would enable lecturers to detect students' educational needs.					
Students would be receptive to using CAT in a summative assessment environment.					
CAT as summative assessment tool would enable students to detect their educational needs.					
2. Formative assessment					
CAT as formative assessment tool would enable lecturers to detect students' educational needs.					
Students would be receptive to using CAT in a formative assessment environment.					
CAT as formative assessment tool would enable students to detect their educational needs.					

Pedagogical evaluation

	Unlikely 1	2	3	4	Likely 5
3. Students' interaction with the system Students' interaction with the system would be simple and clear.					
Students would find the system easy to use.					

Please use this box to expand on the information you have provided above (please continue overleaf if necessary). Thank you.

Appendix K Semi-structured discussion guidelines

The semi-structured discussion sessions discussed in Chapter 9 adhered to the format shown in this Appendix.

Academic staff attitude towards the CAT approach: Semi-structured discussion session guidelines

1. Introduction

- Welcome participants.
- Introduce the research and research team.
- Introduce the topic of the session (i.e. CAT approach, or CAT automated feedback approach).
- Introduce the purpose of the session (i.e. to explore issues relating to…)

2. Confidentiality issues

- Inform participants that the semi-structured session is part of a programme of research, and that the data collected during the session will be used to inform future iterations of the CAT/automated feedback prototype.
- Inform participants that any information they provide will be treated confidentially and, if published, will not be identifiable as belonging to them.
- Request participants' permission to video the session.

3. Presentation

- Provide participants with an overview of the research to date.
- Describe the method employed by the research team to computerised adaptive testing/provision of automated feedback to a CAT.
- Provide participants with screenshots of actual questions/automated feedback.

4. Discussion

- Ask participants to share their views on the topic that has just been presented. Encourage participants to express their agreement/disagreement with the ideas presented by the research team.
- Include the following discussion topics:
  - What are the most common feedback methods used at present?
  - How do you assess the quality of feedback provided at present?
  - What are the benefits and limitations of the feedback provided at present?
  - What is your view of the CAT approach for formative and summative assessment?
  - What is your opinion of the CAT approach to automated feedback?
  - What are the benefits and limitations of automated feedback based on the CAT approach?
  - How could the automated approach be improved?
  - What should be the role of the lecturer in the automated feedback system?
  - What is the need for monitoring and how might this be achieved? What, if any, are the ethical issues in the method?

Appendix L Automated feedback evaluation questionnaire (3)

This appendix shows a copy of the questionnaire used in the study involving academic staff reported in section 9.3.

Automated feedback evaluation questionnaire

Instructions

Thank you for attending the presentation about our automated feedback software prototype and accepting to participate in its evaluation. We would be very grateful if you could complete the questionnaire below.

Please contact Mariana Lilley (m.lilley@herts.ac.uk) or Dr Trevor Barker (t.1.barker@herts.ac.uk) should you have any questions about this evaluation.

Thank you very much for your contribution to this research.

Please note that:

- Any information that you provide will be treated confidentially and, if published, will not be identifiable as belonging to you.
- You have the right to withdraw from the research at any time.
- This questionnaire contains 8 questions, and completing this questionnaire should not take you longer than 10 minutes.

Please rate the statements below.

1. In the context of summative assessment, the automated feedback approach that I have just seen is:

☐	☐	☐	☐	☐
1	**2**	**3**	**4**	**5**
Not useful		**Useful**		**Very useful**

2. In the context of formative assessment, the automated feedback approach that I have just seen is:

1	**2**	**3**	**4**	**5**
Not useful		**Useful**		**Very useful**

3. In the context of objective testing (i.e. multiple-choice questions), the automated feedback approach that I have just seen is:

1	**2**	**3**	**4**	**5**
Not useful		**Useful**		**Very useful**

4. In the context of written assignments, the automated feedback approach that I have just seen is:

1	**2**	**3**	**4**	**5**
Not useful		**Useful**		**Very useful**

5. In the context of written assignments, the automated feedback approach that I have just seen is:

1	**2**	**3**	**4**	**5**
Not useful		**Useful**		**Very useful**

6. With regards to its *speed*, the automated feedback approach that I have just seen is:

1	**2**	**3**	**4**	**5**
Poor		**Good**		**Very good**

7. With regards to its *quality*, the automated feedback approach that I have just seen is:

1	**2**	**3**	**4**	**5**
Poor		**Good**		**Very good**

8. With regards to its *appropriateness* to enhance students' learning experience, the automated feedback approach that I have just seen is:

1	**2**	**3**	**4**	**5**
Poor		**Good**		**Very good**

Please add any comments here (please continue overleaf if necessary):

Thank you for your participation!

Appendix M Research publications

The research reported in this thesis resulted in the publication of several papers. These are listed in this Appendix.

1 **Lilley, M. & Barker, T. (2002). The Development and Evaluation of a Computer-Adaptive Testing Application for English Language In Proceedings of the 6th Computer-Assisted Assessment Conference, Loughborough University, United Kingdom.**

Abstract This paper reports on research undertaken at the University of Hertfordshire into the development and initial expert evaluation of a computer-adaptive testing programme based on Item Response Theory (IRT). The paper explains how the Three-Parameter Logistic model was implemented in the prototype. The underlying theory and assumptions of the model used in its development are also explained, along with the limitations and benefits of the computer-adaptive test (CAT) approach compared to traditional computer-based test (CBT) methods. In this paper use of the prototype as an alternative to the current method used by the University is evaluated by experts, and summaries of their reports and recommendations are presented. This paper also describes plans for developing this work further, including its use in computer-based student modelling where an accurate estimation of performance within a subject domain can be used to inform and adapt the choice of presentation of learning materials. Considerations for extending the CAT model to encompass other types of questions rather than multiple-choice or multiple-response questions are also presented.

2 **Lilley, M., Barker, T. & Maia, M. (2002). The use of Objective Items in Higher Education: Potential and Limitations In Proceedings of the 37th Asamblea del Consejo Latinoamericano de Escuelas de Administración (CLADEA), Porto Alegre, Brazil.**

Abstract This paper reports on research undertaken at the University of Hertfordshire (United Kingdom) and at the Fundação Getúlio Vargas (Brazil) regarding the use of objective questions in Higher Education.

Objective questions are questions designed in such a way that the marking process does not depend on any subjective judgement on the part of the marker (Ward, 1980). The most popular types of objective questions are both multiple-choice and multiple-response questions.

The use of objective questions has been increasing in the last twenty years due to an increased popularisation of computer-assisted assessments (Pritchett, 1999). These computer-assisted assessments have been used for both formative and summative assessments. In both cases, computer-assisted assessments and OMRs, the use of objective questions has various practical benefits: (1) large numbers of students can be simultaneously assessed more accurately and quickly, (2) students can be provided with immediate feedback on their performance

immediately after a given session of assessment, (3) statistical reports on students' performance can be produced with less effort, and (4) assessments can be easily stored and reused.

Notwithstanding the advantages mentioned earlier, many lecturers are unwilling to accept the use of objective questions, since they feel that the use of objective questions is not indicated in the context of Higher Education. The main reason for this opposition is the fact that objective questions are often considered less effective than other types of formats, such as essays and exam papers, when assessing which learning outcomes have been achieved by the students.

It is our belief that the assessment process should enhance students' learning and, as such, it should be an integral part of the students' learning process and not an isolated activity at the end of the academic year. In order to achieve this goal, it is crucial that the assessment process is a regular activity through the academic year in which different delivery media (e.g. paper, computer-based tests) and different formats (e.g. essays, objective questions) are involved. In addition to its contribution to the diversification of assessment, a further benefit within the use of objective questions would be the fact that students tend to be positive about this assessment format. Students consider that objective questions are fair, since the score does not depend on any interpretative element on the part of the marker and thus is identical for all the students (Ward, 1980).

At the University of Hertfordshire, objective questions have been successfully used as part of both formative and summative assessments of more than 100 students enrolled per year in one of the core modules in the MSc in Computer Science course since 2000. This paper outlines simple techniques on how objectives questions can be designed more efficiently, taking into account Bloom's taxonomy of cognitive skills.

3 Lilley, M., Barker, T., Bennett, S. & Britton, C. (2002). How computers can adapt to knowledge: A comparison of computer-based and computer-adaptive testing In Proceedings of the 1st International Conference on Information and Communication Technologies in Education, Junta de Extremadura Consejería de Educación, Ciencia y Tecnología, Badajoz, Spain.

Abstract This paper describes research on computer-adaptive testing undertaken at the University of Hertfordshire, in which a prototype of a computer-adaptive test (CAT) based on the Three-Parameter Logistic Model from Item Response Theory (IRT) was designed and developed. After a positive evaluation by experts, the prototype was submitted for a two-part student evaluation. The first part of the evaluation comprised a

student evaluation followed by a focus group and the second part was a different student evaluation. This paper introduces and discusses the information gathered from the student evaluation, ranging from the subjects' perception of the level of difficulty of an adaptive test to their perception of its fairness. In addition, our plans for future research on computer-adaptive testing as well as a brief discussion on the advantages and disadvantages of an adaptive algorithm are presented.

4 Lilley, M., Barker, T. & Maia, M. (2002). Web-based adaptive testing in distance learning: an overview In Proceedings of the 5th Simpósio de Administração da Produção, Logística e Operações Internacionais (SIMPOI), Fundação Getulio Vargas Escola de Administração de Empresas de São Paulo, Brazil.

Abstract This paper reports on research undertaken at Fundação Getúlio Vargas in Brazil on Distance Learning and at the University of Hertfordshire in the UK on computerised adaptive testing (CAT). While 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 a CAT the questions are selected dynamically for each student, based on his or her individual performance during the assessment. In order to select questions dynamically, one of the techniques available is Item Response Theory (IRT). The central element of IRT is a family of mathematical functions that calculates the probability of a specific student answering a particular question correctly. The main characteristics of IRT are introduced in this paper. This paper also introduces some of the issues relating to the implementation of web-based CATs in Distance Learning education in Operations Management.

5 Lilley, M. & Barker, T. (2003). Comparison between Computer-Adaptive Testing and other assessment methods: An empirical study In Proceedings of the 10th International Conference of the Association for Learning Technology (ALT-C), University of Sheffield, United Kingdom.

Abstract This paper describes the development and evaluation of a computer-adaptive test (CAT). The application is based on Item Response Theory, and was used to assess 133 students enrolled in a Visual Basic programming module. The findings from a comparison between the CAT, conventional computer-based tests (CBTs) and off-computer coursework, suggest that students were not disadvantaged by the use of a CAT. These findings also suggest that the CAT approach has the potential to provide teachers with valuable information on learners' ability. Further issues, such as students' attitude, potential benefits of the

approach and future work are also explored.

6 Lilley, M., Barker, T. & Maia, M. (2003). The Evaluation of a Computer-Adaptive Test In Proceedings of the 27th Encontro da Associação Nacional dos Programas de Pós-Graduação em Administração (ANPAD), Atibaia, Brazil.

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.

7 Lilley, M. & Barker, T. (2003). An Evaluation of a Computer-Adaptive Test in a UK University Context In Proceedings of the 7th Computer-Assisted Assessment Conference, Loughborough University, United Kingdom.

Abstract This paper reports on work undertaken at the University of Hertfordshire into the development and evaluation of a computer-adaptive test (CAT) for English language based on Item Response Theory (IRT). It also reports on how this work was extended, including the development of software to perform two large-scale computer-adaptive tests for a second year Visual Basic programming module at the University of Hertfordshire.

The CAT application we developed used an adaptive algorithm based on the Three-Parameter Logistic Model. The application selects the most appropriate questions to be presented to each individual student based upon their ability, as measured by performance in the test. The main purpose of a CAT is to present students with questions that are fitted for their individual level of ability. The underlying principle is that questions

that reflect the student's skills provide more valuable information about the student and motivate more than those that are either too difficult or too easy. One of the consequences of the dynamic selection of questions according to each individual student performance is that is unlikely that one student will be answering the same set of questions as any other.

This characteristic may bring both advantages and disadvantages. Students may feel more motivated during the test, given that they are not presented with questions that are either too difficult, and thus frustrating, or too easy, and therefore uninteresting. Some students, however, may consider that the fairness of the test is jeopardised, since the set of test questions is not identical for all participants. One student may answer the same number of questions correctly as another student, yet achieve a lower level, and hence a lower grade.

The first stage of this work was intended to show that the application was of pedagogical interest to teachers and the interface did not impose any barriers to assessment. To this end, academic staff and students evaluated the prototype and a group of international students compared the software with a non-adaptive computer-based test (CBT), and took part in a focus group session. During this session, students discussed issues relating to computer-adaptive tests, ranging from their perception that very easy tests are "meaningless" to their insights into the fairness of such computer-assisted assessments. The findings of the focus group are reported in the first part of this paper.

In the second part of the paper, the results of a study of performance with 132 participants for a second year Visual Basic programming course at the University of Hertfordshire in two computer-adaptive assessments are reported. In this study, we made a comparison between CBT and CAT. We report the results of the assessments and also students' attitude to the testing at debriefing sessions following the tests. We were able to show, using statistical analysis of the data obtained in the tests, that participants were not disadvantaged by computer-adaptive testing.

Finally, the benefits and potential limitations of this method of assessment are also presented.

8 Lilley, M., Barker, T. & Maia, M. (2003). Computer-Adaptive Testing in Higher Education: the way forward? In Proceedings of the 38th Asamblea del Consejo Latinoamericano de Escuelas de Administración (CLADEA), Lima, Peru.

Abstract This paper marks a further progression on research previously done by Fundação Getúlio Vargas (Brazil) on distance learning and the University of Hertfordshire (United Kingdom) on the use of

computer-adaptive tests in Higher Education (HE). In this work, the growing interest in Business Administration distance learning within the Brazilian scenario, in addition to how this growth has led to an increased interest from both teaching staff and educational researchers as to the potential benefits and limitations of computer-assisted assessments were discussed.

Findings from our most recent research suggest that the distance learning pedagogical model has the potential to play a role of increasing importance in widening the access to Brazilian HE. Although some valuable progress has been made mainly in the area of design and development of computer-aided instruction, there are still areas for further development, such as computer-delivered assessments and educational software evaluation. The development of these areas is vital, as it could lead to the full exploration of the technological resources available. In the first part of this paper, we outline our experience at the University of Hertfordshire regarding the development and software evaluation of a computer-adaptive test for two large-scale summative assessments in the "Program Development" module. We then compare computer-adaptive tests with other assessment methods, namely traditional computer-based tests, practical projects and exams. Moreover, we present perceived advantages and disadvantages of each assessment method in HE in general and more specifically in Business Administration distance learning. In the final part of the paper, we describe how our work can be developed further.

9 Barker, T. & Lilley, M. (2003). Are Individual Learners Disadvantaged By The Use Of Computer-Adaptive Testing In Higher Education? In Proceedings of the 8th Learning Styles Conference, European Learning Styles Information Network (ELSIN), University of Hull, United Kingdom.

Abstract This paper presents ongoing research at the University of Hertfordshire on the use of computer-adaptive tests in Higher Education. Computer Adaptive tests are a form of computer-based testing where the difficulty of the test is tailored to the individual learner. In general terms, the test starts with a question of medium difficulty. If the student answers the question correctly, a more difficult question is next presented. Conversely, if the question is answered incorrectly, an easier question follows. The statistical process that supports the selection of the next question is based on Item-Response Theory (IRT).

The main purpose of CAT is to present the student with questions that are challenging for his or her level of ability. Questions that reflect a student's skills provide more information about the student and motivate more than those that are either too difficult or too easy. One of the consequences of

the dynamic selection of questions is that no two students will answer the same set of questions. This may bring both advantages and disadvantages. Although students may feel motivated, some students may consider that the fairness of the test is jeopardized, since the set of test questions is not the same for all participants. One student may answer the same number of questions correctly as another student, yet achieve a lower level, and hence a lower grade. It is therefore important to be sure that students are not disadvantaged by the CAT approach.

The research described in this paper therefore relates to the design, development and evaluation of computer-adaptive testing software for a Visual Basic programming course at the University of Hertfordshire in a real educational context. In previous research, academic staff and students evaluated the CAT software introduced here. The academic staff performed an expert evaluation of the software to ensure that it was usable and pedagogically sound. A group of international students compared the software with a traditional computer based test, and took part in a focus group session. During this session, students discussed issues related to computer-adaptive tests, ranging from their perception that very easy tests are “meaningless” to their insights into the fairness of such computer-assisted assessments. In a later study, 133 second year computer programming students at the University of Hertfordshire took the CAT test as part of their normal coursework assessment. This assessment consisted of two theory tests each having a traditional CBT component and CAT component and off-computer project work. Performance on the CAT and CBT parts of the course was compared, using an Analysis of Variance and Pearson’s correlation. The results of this suggested that the CAT test was a better measure of learner ability than the CBT component.

We also compared the CAT and CBT tests with the off-computer assessments. We were able to conclude from this that the CAT approach was a fair measure of learner ability. Students were also measured using Riding’s CSA test. We present the results of this test and discuss some interesting differences in learner performance related to individual cognitive style.

10 Lilley, M.; Barker, T. & Britton, C. (2003). Review and Modification of Responses in a Computer-Adaptive Test: Preliminary Considerations In Proceedings of the 2nd International Conference on Information and Communication Technologies in Education, Junta de Extremadura Consejería de Educación, Ciencia y Tecnología, Badajoz, Spain.

Abstract Findings from ongoing research at the University of Hertfordshire on the use of computerised adaptive testing suggest that the

approach represents a fair assessment method in addition to the potential to offer a more consistent and accurate measurement of student ability than that supported by traditional computer-based tests. Despite the fact that it is usually expected that within a computer-adaptive test (CAT) test-takers should not be allowed to review and modify previously entered responses, some participants from two different empirical studies expressed their concern about this assumption.

In the first part of this paper the two empirical studies and their main findings are summarised. We also present findings from our most recent empirical study, which involved a modified version of the application that allowed students to return to and modify previously entered responses. Findings from this latter study suggested that allowing students to review and modify previously entered responses was unlikely to have a significant impact on their final score. However, it seems to the authors that it could lead to a reduction in student anxiety as well as an increase in student confidence in this assessment method. Further issues, such as student attitude, potential benefits of the approach and future work are explored in final section of this paper.

11 Lilley, M., Barker, T. & Maia, M. (2003). Do Cognitive Styles of Learning Affect Student Performance in Computer-Adaptive Testing? In Proceedings of the 6th Simpósio de Administração da Produção, Logística e Operações Internacionais (SIMPO I), Fundação Getulio Vargas Escola de Administração de Empresas de São Paulo, Brazil.

Abstract This paper marks a further progression on research previously done by Fundação Getúlio Vargas on distance learning and University of Hertfordshire on the use of computer-adaptive tests in Higher Education. In this paper we provide a brief introduction to cognitive styles of learning and computerised adaptive testing. We then investigate whether or not cognitive styles of learning have the potential to be an important factor influencing student performance when participating in a computer-adaptive test. In the final section of this paper, we discuss how the findings from this study can be applied within the domain of Business Administration distance learning.

12 Lilley, M.; Barker, T. & Britton, C. (2004). The generation of automated student feedback for a computer-adaptive test In Proceedings of the 8th Computer-Assisted Assessment Conference, Loughborough University, United Kingdom.

Abstract This paper marks further progression on research previously

undertaken at the University of Hertfordshire on the use of computer-adaptive tests (CATs) in Higher Education. Findings from two previous empirical studies by the authors suggested that the CAT approach was a fair assessment method, capable of offering accurate and consistent measurement of student abilities. Participants in a pedagogical evaluation of the application indicated that one of the limitations of the approach was the type of the feedback provided to students. According to the evaluators, the sole provision of a score would not help students to detect their educational needs. Providing students with a copy of all questions they got wrong did not seem an attractive option either, as it could jeopardise the re-use of these questions in future assessment sessions. Furthermore, it seemed unlikely that providing students with the questions alone, without any comment or explanation, would foster research and/or reflection skills.

This paper reports on our most recent empirical study, in which the ability estimate for each student in each section of the CAT test was used to generate automated feedback based on Bloom's taxonomy of cognitive abilities. The feedback was then sent directly to individual students via personal email. In the first section of this paper, we present an overview of our CAT research followed by the main characteristics of the feedback tool we designed and implemented. In the final section of this paper, we present the results a summary of how learners performed on the CAT, along with student attitude towards the automated feedback. In addition, we present our views on how the work described here can be developed further.

13 Lilley, M. & Barker, T. (2004). A Computer-Adaptive Test that facilitates the modification of previously entered responses: An empirical study. Lecture Notes in Computer Science 3220, 7th International Conference ITS 2004, Volume 3220/2004, pp. 22-33.

Abstract In a computer-adaptive test (CAT), learners are not usually allowed to revise previously entered responses. In this paper, we present findings from our most recent empirical study, which involved two groups of learners and a modified version of a CAT application that provided the facility to revise previously entered responses. Findings from this study showed that the ability to modify previously entered responses did not lead to significant differences in performance for one group of learners ($p>0.05$), and only relatively small yet significant differences for the other ($p<0.01$). The implications and the reasons for the difference between the groups are explored in this paper. Despite the small effect of the modification, it is argued that this option is likely to lead to a reduction in student anxiety and an increase in student confidence in this assessment method.

14 Lilley, M.; Barker, T. & Britton, C. (2004). The development and evaluation of a software prototype for computer adaptive testing. Computers & Education Journal 43(1-2), pp. 109-123.

Abstract This paper presents ongoing research at the University of Hertfordshire on the use of computer-adaptive tests (CAT) in Higher Education. A software prototype based on Item Response Theory has been developed and is described here. This application was designed to estimate the level of proficiency in English for those students whose first language is not English. Academic staff and students evaluated the prototype introduced here and we summarise their attitude to the user interface and to pedagogical aspects of the prototype. We provide evidence that learners are not disadvantaged by the CAT approach, based on a comparison of performance between CAT and computer-based tests (CBTs). A group of international students also took part in a focus group session after using the software. During this session, students discussed issues related to computer-adaptive tests, ranging from their perception that very easy tests are "meaningless" to their insights into the fairness of such computer-assisted assessments.

In addition, this paper outlines how our current work will be developed further by implementing multimedia resources, developing more subjective tests and adding a stop condition associated with the calculation of standard error. Finally, the benefits and potential limitations of this method of assessment are also presented here.

15 Barker, T. & Lilley, M. (2004). The development and evaluation of computer-adaptive testing software in a UK university In Proceedings of the 2004 Learning and Teaching Conference, University of Hertfordshire, United Kingdom.

Abstract The use of computers within the educational sector as a tool to support the assessment of students has been growing. This growth has led to an increased interest from both academic staff and educational researchers as to the potential benefits of a computer-adaptive (CAT) approach as an alternative to traditional computer-based tests (CBTs).

In a CBT the questions presented during a given assessment session are not tailored to the proficiency level of individual learners and thus all learners are typically presented with the same set of questions. In contrast, in a CAT the questions are selected dynamically for each learner, based on his or her individual performance during the assessment.

In this study, we describe the design, implementation and evaluation of a CAT software prototype for the assessment of Computer Science

undergraduates.

16 **Lilley, M. & Barker, T. (2005). The Use of Item Response Theory in the Development and Application of a User Model for Automatic Feedback: A Case Study In Proceedings of the 19th British HCI Group Annual Conference, Napier University, Edinburgh, United Kingdom.**

Abstract At the University of Hertfordshire we have developed a computer-adaptive test (CAT) prototype. The prototype was designed to select the questions presented to individual learners based upon their ability. Earlier work by the authors had shown benefits of the CAT approach, such as increased learner motivation. It was therefore important to investigate the fairness of this assessment method. Statistical analysis of test scores from 310 participants show that in all cases scores were highly correlated between CATs and other assessment methods ($p<0.05$). This was taken to indicate that learners of all abilities were not disadvantaged by our CAT approach.

17 **Lilley, M.; Barker, T. & Britton, C. (2005). Automated feedback for a computer-adaptive test: A case study In Proceedings of the 9th Computer-Assisted Assessment Conference, Loughborough University, United Kingdom.**

Abstract This paper reports on an empirical study regarding the generation of automated feedback for a computer-adaptive test (CAT) application. In the study reported here, two groups of Computer Science undergraduate students participated in a session of assessment using our CAT application (N=106 and N=82).

Participants had 40 minutes to answer 30 questions organised into 5 topics within the Visual Basic.Net subject domain. Participants were provided with feedback on CAT performance via a web-based application specially designed and implemented for this purpose. The feedback provided was divided into three sections: overall proficiency level, performance in each topic and recommended topics for revision. Thirty-one participants from the first group and 25 participants from the second group rated the usefulness of the feedback provided from 1 (not useful) to 5 (very useful). The mean values obtained for the usefulness of the feedback provided were respectively, 4.10 and 3.52. These results were taken to indicate that learners' attitude towards the feedback approach employed was positive overall.

18 Lilley, M.; Barker, T. & Britton, C. (2005). Learners' perspectives on the usefulness of an automated tool for feedback on test performance In Proceedings of the 4th European Conference on E-Learning, Royal Netherlands Academy of Arts & Sciences, Amsterdam, Netherlands.

Abstract Computer-adaptive tests (CATs) are computer-aided assessment applications in which Item Response Theory is employed to adapt the level of difficulty of the test to each individual learner's proficiency level within a subject domain. This paper is concerned with the initial evaluation of an automated feedback tool for a CAT. In the empirical study introduced here, a group of 113 Computer Science undergraduate students participated in a session of summative assessment using our CAT application.

Participants were expected to answer 24 objective questions within a 40-minute time limit. The 24 questions were organised into 4 topics within the Human Computer Interaction subject domain. Participants were provided with feedback on CAT performance via a web-based application specially designed and constructed for this purpose. The feedback provided was divided into three sections: overall proficiency level, performance summary per topic and recommended topics for revision. A group of 97 students favourably evaluated the automated feedback tool introduced in this study. In addition to the learners' evaluation, a group of 19 Higher Education lecturers positively assessed the feedback tool. These results were taken to indicate that our approach to the provision of automated feedback was a valid one, capable of offering useful advice for individual development.

This paper is organised into five sections: (1) CAT prototype overview; (2) overview of the automated feedback tool employed in this study, (3) learners' attitude towards the feedback approach; (4) tutors' attitude towards the feedback approach and (5) our views on how the work presented here can be developed further.

19 Lilley, M. & Barker, T. (2005). An empirical study into the effect of question review in a computer-adaptive test In Proceedings of the 6th Annual Higher Education Academy Subject Network for Information Computer Science Conference, University of York, United Kingdom.

Abstract Interactive software applications that adapt to their users have been gaining rapidly in importance within the computer-aided education field. Computer Adaptive Tests (CATs) are an example of such adaptive systems. In a CAT, the level of difficulty of the questions administered is dynamically adapted to the proficiency level of individual

users. A common assumption within CATs is that users should not be permitted to review and modify previously entered responses.

Relevant literature, however, provides evidence that some users view the inability to return to previous questions as a disadvantage of the CAT approach. In the empirical study reported here, 205 Computer Science undergraduates took a test using our CAT prototype. After answering a predefined number of questions, users were allowed to review and modify previously entered responses. Findings from this study showed that the ability to modify previously entered responses did not lead to significant differences in performance for low and high performing groups ($p>0.05$), and only relatively small yet significant difference in the percentage of correct responses for the intermediate group ($p<0.05$). The results reported here support the view that learners should be permitted to return to previously entered responses in the context of summative assessments using the CAT approach.

20 Lilley, M. & Barker, T. (2005). Computer-adaptive testing: A case study In Proceedings of the 6th Annual Higher Education Academy Subject Network for Information Computer Science Conference, University of York, United Kingdom.

Abstract In a computer-adaptive test (CAT), the questions to be administered during an assessment session are dynamically selected according to individual student performance. Statistical analysis of test scores from 205 participants show that scores between CATs and other assessment methods were highly correlated ($p<0.01$). This was taken to indicate that learners were not disadvantaged by the CAT approach adopted in this study.

21 Lilley, M., Barker, T. & Britton, C. (2005). The generation of automated learner feedback based on individual proficiency levels In Proceedings of the 18th Internati

onal Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, Lecture Notes in Artificial Intelligence 3533, pp. 842-844.

Abstract Computer-adaptive tests (CATs) are software applications that adapt the level of difficulty of test questions to the learner's proficiency level. The CAT prototype introduced here includes a proficiency level estimation based on Item Response Theory and a questions' database. The questions in the database are classified according to topic area and difficulty level. The level of difficulty estimate comprises expert evaluation

based upon Bloom's taxonomy and users' performance over time. The output from our CAT prototype is a continuously updated user model that estimates proficiency in each of the domain areas cov-ered in the test. This user model was employed to provide automated feedback for learners in a summative assessment context. The evaluation of our feedback tool by a group of learners suggested that our approach was a valid one, capable of providing useful advice for individual development.

22 Lilley, M., Barker, T. & Britton, C. (2005). Learners' perceived level of difficulty of a computer-adaptive test: A case study In Proceedings of the 10th International Conference on Human-Computer Interaction, Lecture Notes in Computer Science 3585, pp. 1026-1029.

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.

23 Lilley, M. & Barker, T. (2006). Student attitude to adaptive testing In Proceedings of HCI 2006 Conference, Queen Mary, University of London, 11-15, September 2006.

Abstract A computer-adaptive test (CAT) is a computer-assisted assessment application in which the test dynamically adapts itself to the proficiency level of individual students. To enhance student engagement, CAT software applications aim to provide students with tasks that are sufficiently challenging, and yet not so difficult that could lead to boredom or frustration.

The CAT prototype introduced here comprised a graphical user interface, a database of questions and an adaptive algorithm based on the Three-Parameter Logistic Model from Item Response Theory. A group of 76 Computer Science undergraduate students participated in a summative and a formative assessment session using our CAT prototype. At the end

of each 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 the participants were subjected to Spearman's rank order correlations. Findings from this statistical analysis suggest that participants' perceptions of difficulty were not related either to performance or to the type of test undertaken.

24 Barker, T. & Lilley, M. (2006). Measuring staff attitude to an automated feedback system based on a Computer Adaptive Test In Proceedings of Computer-Assisted Assessment 2006 Conference, Loughborough University, July 2006.

Abstract In Higher Education today, increasing reliance is being placed upon the use of online learning and assessment systems. Often these are used to manage learning, present information and test learners in an entirely undifferentiated way, all users having exactly the same view of the system. With the development of increasingly large and complex computer applications and greater diversity in learner groups, consideration of individual differences and greater efficiency in learning and testing have become important issues in designing usable and useful applications.

Computer Adaptive Tests (CAT) are software applications that adapt the presentation of test questions to the learner's proficiency level. In our earlier research, we have shown that CATs provide an efficient individual motivational test for each leaner, based on his or her individual abilities. An important feature of our CAT was the development of a student model upon which the delivery of automated feedback could be based. The student model employed and developed in our CAT prototype included a proficiency level estimation based on Item Response Theory and a database of questions calibrated according to Bloom's taxonomy, initially by experts and then updated according to user performance. The output from our CAT prototype is therefore, a continuously updated student model that estimates proficiency in each of the domain areas covered in the test.

Our initial findings, reported at CAA 2005, suggested that students valued this approach to providing automated feedback and considered it to be a fast, effective and reliable method. In the study presented in this paper, the attitude of staff to our automated feedback tool is presented. Three presentation sessions involving more than 50 staff were undertaken and their views of the feedback tool were captured using video recordings. Subsequent analysis of the sessions using qualitative data analysis methods showed that teachers in general were receptive to the idea of automated feedback based on CAT. Several interesting ideas arose from the discussions, which are presented here. Computer based testing and

automated feedback are becoming increasingly important in Higher Education. It is important that the views of teachers are considered when developing and implementing such systems if they are to be accepted and hence effective.

25 Lilley, M. & Barker, T. (2006). Students' perceived usefulness of formative feedback for a computer-adaptive test In Proceedings of ECEL 2006: The European Conference on e-Learning, University of Winchester, 11-12 September 2006.

Abstract In this paper we report on research related to the provision of automated feedback based on a computer adaptive test (CAT), used in formative assessment.

A cohort of 76 second year university undergraduates took part in a formative assessment with a CAT and were provided with automated feedback on their performance. A sample of students responded a short questionnaire to assess their attitude to the quality of the feedback provided. In this paper, we describe the CAT and the system of automated feedback used in our research and also present the findings of the attitude survey. On average students reported that they had a good attitude to our automated feedback system. Statistical analysis was used to show that attitude to feedback was not related to performance on the assessment ($p>0.05$). We discuss this finding in the light of the requirement to provide fast, efficient and useful feedback at the appropriate level for students.

26 Barker, T.; Lilley, M & Britton, C. (2006). Computer Adaptive Assessment and its use in the development of a student model for blended learning. Annual Blended Learning Conference, University of Hertfordshire, July 2006.

Abstract This paper presents an overview of our work on the development and testing of an automated feedback tool based on Computer-Adaptive Testing. Computer-adaptive tests (CATs) are software applications that adapt the presentation of test questions to the learner's proficiency level, so that those performing well are given more difficult questions and vice versa. In this paper, we present and describe the development of the models used in a feedback tool based on this approach. The model includes a proficiency level estimation based on Item Response Theory and also a questions' database. The questions in the database are classified according to topic area and difficulty level. The difficulty level is initially set by expert evaluation based upon Bloom's

taxonomy and adapted according to students' performance over time.

The output from our adaptive test is a continuously updated student model that estimates proficiency in each of the domain areas covered in the test, relating not only to performance, but also to cognitive ability, based on Bloom's levels. Earlier work has shown that the approach we adopt is reliable and fair to students and provides useful and important measures of ability. Potentially these measures may be used, not only in formative and summative assessment, but also to help in the delivery of learning or remedial activities based on individual ability. We describe our student model based on adaptive testing and show how it was used to provide automated feedback for students in a summative assessment context. The evaluation of our feedback tool by groups of learners and teachers suggested that our approach was a valid one, capable of providing useful advice for individual development. The results of these evaluations are presented in this paper. In the concluding section of the paper we suggest ways that the student profiles created by our method are likely to be useful in a variety of learning contexts.

27 Barker, T.; Lilley, M. & Britton, C. (2006). A student model based on computer adaptive testing to provide automated feedback: The calibration of questions. Presented at Association for Learning Technology, ALT 2006, Herriot-Watt University, September 4-7, 2006.

Abstract In Higher Education today, increasing reliance is being placed upon the use of online learning and assessment systems. Often these are used to manage learning, present information and test learners in an entirely undifferentiated way, all users having exactly the same view of the system. With the development of increasingly large and complex computer applications and greater diversity in learner groups, consideration of individual differences and greater efficiency in learning and testing have become important issues in designing usable and useful applications.

We have produced a Computer Adaptive Testing system that not only provides an estimate of student performance, but also generates a student model based on Bloom's Taxonomy. This system is used to provide automated feedback to learners, not only on their performance in tests, but also on their cognitive levels. The research reported in this short paper relates to the development of our modelling approach and how we use it to provide feedback. An important feature in our model is the use of Computer Adaptive Testing to establish performance levels for learners. Question databases in our tests are calibrated for difficulty by experts in the first instance, and later adapted according to performance. The result of an analysis of the calibration of our adaptive model is presented. We were able to show that the method of calibration using

experts was accurate and effective. We also present the results of a Computer Adaptive Test involving 139 students and discuss the results of this test in the context of the calibration method employed. The potential of this approach in the establishment of managed Learning Environments is also discussed.

28 Lilley, M. & Barker, T. (2006). Computerised adaptive testing: extending the range of assessment formats in a Computer Science course In Proceedings of Conference ICL2006, September 27-29, 2006 Villach, Austria.

Abstract A computer-adaptive test (CAT) is a computer-assisted assessment application that makes use of Item Response Theory (IRT) to create a test that is tailored to individual students. 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 group of 125 Computer Science undergraduates participated in three assessment sessions: a traditional computer-based test, a computer-adaptive test and a practical programming test. Their scores in these assessments were subjected to a Pearson's Product Moment correlation. The results of this statistical analysis suggest that students were not disadvantaged by the CAT approach. The implications of this finding are discussed in the concluding section of the paper.

29 Lilley, M. & Barker, T. (2007). Students' perceived usefulness of formative feedback for a computer-adaptive test. Electronic Journal of e-Learning (EJEL) Volume 5 Issue 1 February 2007 Special Issue (ECEL 2006). Available: http://www.ejel.org/Volume-5/v5-i1/v5-i1-art-5.htm

Abstract In this paper we report on research related to the provision of automated feedback based on a computer adaptive test (CAT), used in formative assessment.

A cohort of 76 second year university undergraduates took part in a formative assessment with a CAT and were provided with automated feedback on their performance. A sample of students responded a short questionnaire to assess their attitude to the quality of the feedback provided. In this paper, we describe the CAT and the system of automated feedback used in our research and also present the findings of the attitude survey. On average students reported that they had a good attitude to our automated feedback system. Statistical analysis was used to show that attitude to feedback was not related to performance on the

assessment ($p>0.05$). We discuss this finding in the light of the requirement to provide fast, efficient and useful feedback at the appropriate level for students.

30 Lilley, M.; Barker, T. & Britton, C. (2007). Computer Adaptive Testing in Higher Education: A case study In Proceedings of 2007 Annual Solstice Conference, Edge Hill University, United Kingdom.

Abstract At the University of Hertfordshire we have developed a computer-adaptive test (CAT) prototype. The prototype was designed to select the questions presented to individual learners based upon their ability. Earlier work by the authors during the last five years has shown benefits of the CAT approach, such as increased learner motivation. It was therefore important to investigate the fairness of this assessment method. In the study reported here, statistical analysis of test scores from 320 participants show that in all cases scores were highly correlated between CATs and other assessment methods ($p<0.05$). This was taken to indicate that learners of all abilities were not disadvantaged by our CAT approach.