Applications of the Flexilevel Test to assessment in Higher Education

Andrew Richard Pyper

A dissertation submitted to the University of Hertfordshire in partial fulfilment of the requirements of the degree of Doctor of Philosophy.

September 2015

ABSTRACT

The work reported in this dissertation investigates the potential for embedding Computerised Adaptive Testing (CAT) in students' and tutors' educational experiences. It seems that the tailored assessments that CAT can provide would be of real educational value in a range of contemporary Higher Education settings, however the resource requirements of some forms of CAT are prohibitive for making CAT assessments available to students across their studies.

A form of CAT that is less resource intensive than other forms, the Flexilevel test, was selected for this programme of research to investigate its effectiveness in real educational contexts and explore possible applications for the approach.

Ten empirical studies and a real data simulation study were conducted to test the effectiveness of the approach. It was found to show statistically significant correlations with other forms of assessment – in particular conventional Computer Based Testing (CBT) assessments, which is commonly used in contemporary educational settings.

Another strand of work concerned the attitudes of stakeholders to the approach. Part of this work was carried out through the empirical studies, and further studies including interviews were also undertaken to explore the views of academic staff and students to the use of the Flexilevel test. Both groups were positive about the use of the Flexilevel test and this was taken to support the idea that academic staff and students would accept the use of the Flexilevel test in their educational experiences.

In terms of both effectiveness and the acceptability of the approach to academic staff and students, the Flexilevel test was found to be a good candidate for embedding CAT in real educational contexts in Higher Education.

ACKNOWLEDGEMENTS

I am truly grateful to Mariana Lilley, my principal supervisor, for her guidance, expertise and insight.

Thanks to Paul Wernick, my second supervisor, for his encouragement, valuable advice and cogent critiques.

Thanks to Amanda Jefferies, my most recently appointed supervisor, for her support and advice about the work.

Thanks to Jonathan Meere for being so supportive of the work over the years.

Thanks to Sue Attwood for the critical reading, advice, encouragement and support.

Thanks to Lucy Bamwo, Anthony Herbland and Karen McCourt for the expert review work.

Finally, I'd like to thank my parents, Richard and Wendy and my sisters, Sarah, Liz and Helen for their unending support and encouragement over the years.

TABLE OF CONTENTS

1 INTRODUCTION

This dissertation reports on research conducted in the use of the Flexilevel test algorithm for objective testing in real educational contexts. The Flexilevel test was identified as a viable way of embedding assessment opportunities using a Computerised Adaptive Test (CAT) approach in students' learning activities. Indeed recently, its potential value for flexible learning and assessment has been identified in a Higher Education Academy report on Flexible pedagogies (Gordon, 2014).

CAT has been argued to provide more engaging assessments than conventional Computer Based Tests (CBT) (Wainer, 2000). Unlike conventional CBT tests, test questions – or items – that are considered too easy or too difficult for students are not presented to them in a CAT. In doing so, examinees do not have to answer lots of items that are too easy or too difficult for them (Carlson, 1994); examinees are challenged but not discouraged (Wainer, 2000). In terms of efficiency, CAT tests can be effective with fewer items than conventional CBT tests (Carlson, 1994).

This tailored approach to testing also offers a potentially useful way to support students in increasingly diverse learning and teaching contexts. However the resource requirements for some CAT approaches, for example Item Response Theory CAT techniques (Lord, 1980) and the Pyramidal approach, (for example, Larkin & Weiss, 1975) seem prohibitive for such contexts.

The Flexilevel test offers the potential to provide the benefits of a CAT without some of the more onerous resource requirements of other CAT techniques. This is of interest to this programme of research as while it seems in the past the challenge has been in finding sufficient technology to support CAT, the reverse seems to be true in contemporary Higher Education. Sophisticated computing resources are readily available, yet the time staff have to design and run the tests is in short supply (Wernick, 2010 personal communication).

To investigate the Flexilevel test in contemporary e-assessment contexts, e-assessment applications were developed – initially a desktop application and then a web based application. These were used to mediate assessments in diagnostic, formative and summative contexts, in classroom-based and online settings, with shorter and longer test lengths, and at different levels of study, including Levels 4, 6 and 7 Computer Science programmes of study.

The acceptance of the technique by primary stakeholders (Dix et al., 2004) was also of key interest to this programme of research and students and academic members of staff were involved from the beginning in evaluating the approach.

1.1 Motivation

The motivation for this programme of research was to gain an understanding of how adaptive test algorithms could be embedded in students' educational experiences, for example in providing more opportunities for adaptive assessment. CAT tailors assessments to individual examinees, and appears to have the potential to enhance the educational experience of students (Lord (1980), Wainer (2000), Carlson (1994), Lilley & Barker (2003)). However there seem to be resource implications with most CAT approaches that make the embedding of CAT algorithms into educational experiences a challenge. The programme of research reported in this dissertation is intended to investigate the effectiveness of a lighter approach to CAT, the Flexilevel test, and to evaluate it with academic staff and students.

1.2 Research Questions

This programme of research aims to answer the following research questions:

1. Does the Flexilevel test afford assessment opportunities in real Higher Education contexts?
2. What, if any, are the potential applications of the Flexilevel to Higher Education contexts?

1.3 Objectives

The following objectives have been derived from the research questions above:

1. To compare student performance in a conventional CBT and Flexilevel test.

2. To gain an understanding of the attitudes of staff to tests using the Flexilevel algorithm for item selection.
3. To gain an understanding of the attitudes of students to tests using the Flexilevel algorithm for item selection.
4. To identify contexts of use in which the Flexilevel tests may be applicable.
5. To evaluate the approach adopted in real educational contexts.
6. To apply Human Computer Interaction techniques to the development and design of e-assessment applications.

1.4 Overview of the chapters

This section provides an overview of the content of each subsequent chapter; there are seven chapters in total. It should be noted that there isn't a unitary literature review chapter, the relevant literature is discussed in the context of each section throughout the report.

1.4.1 Chapter 2: Educational context

Chapter 2 reports on work conducted to frame the real educational contexts that the Flexilevel test may be embedded within. The concept of learning ecosystems is used to describe the environment and interactions that may take place particularly between academic staff and students, and between students and their environment. This chapter also sets out, how the Flexilevel test may be ideally suited to providing assessments within such learning ecosystems.

1.4.2 Chapter 3: Applications of computers to assessment

In this chapter, assessment in general and objective testing in particular are considered. Forms of conventional CBT and adaptive testing are described including IRT-based CAT and fixed branch algorithm CATs. This chapter also sets out the rationale for the selection of the Flexilevel test for this programme of research.

1.4.3 Chapter 4: Studies Investigating the effectiveness of the Flexilevel test

The studies conducted to establish the effectiveness of the Flexilevel test in contemporary Higher Education settings are detailed in Chapter 4. The work conducted incorporated ten empirical studies that compared the performance of students in Flexilevel tests and conventional CBT tests. Comparisons were also carried out with students' performance on other related forms of assessment.

To investigate the effectiveness of shorter Flexilevel tests (with at least half the number of items of a comparison CBT test) a real data simulation study was used prior to studies with shorter Flexilevel tests in real educational contexts. Studies into shorter Flexilevel tests in real educational contexts followed. These are also reported in chapter 4.

To reflect the different methodologies adopted, the chapter is split into exploratory studies, two phases of live testing and a real data simulation study. Please see below for an overview of the studies.

1.4.4 Chapter 5: Evaluation with students

Chapter 6 reports on studies carried out as part of the investigation into students' attitudes towards the use of the Flexilevel test in their educational experiences. Students' attitudes to the approach are key to understanding the extent to which the Flexilevel test would be acceptable to them. Studies reported in this chapter covered their views of the Flexilevel test and their attitudes to its use in their educational experiences, particularly as a formative assessment. Please see below for an overview of the studies conducted with students.

1.4.5 Chapter 6: Evaluation with academic staff

A critical factor in the application of the Flexilevel test in real educational contexts is stakeholders' attitudes towards the approach. For the research into part of the research, two main groups of primary stakeholders were identified: academic staff and students. Primary stakeholders are those groups of users that will make direct use of any system (Dix et al., 2004), including in this case an approach to assessment.

This chapter reports on studies conducted with academic members of staff both within the University of Hertfordshire and externally in the wider UK Higher Education sector. Studies were conducted as part of empirical studies and separately to investigate issues that may affect their views of the Flexilevel test approach. Please see below for an overview of the studies with academic staff.

1.4.6 Chapter 7: Summary, conclusions and future work

A summary of the work is presented and conclusions are drawn in terms of the contribution of the work. There are potentially interesting future directions that the work could take and these are also outlined in this chapter.

1.5 Overview of the studies

As noted above, both the effectiveness of the Flexilevel test approach and its acceptability to primary stakeholders was of interest in this programme of research. This generated two main strands of work: empirical studies comparing the Flexilevel test with conventional CBT testing and other forms of assessment; and studies that evaluated the approach with primary stakeholders.

Ethical approval was obtained for all studies that included components that were not directly related to students' normal educational experiences. This included the exploratory studies as well as subsequent studies that made use of surveys, interviews and focus groups.

Where the data were collected from a standard aspect of the students' educational experiences, they were covered by Appendix 2 of UPR RE01 concerning reflective practitioner work.

Moreover, the assessments used in this programme of research were worth no marks or represented a small proportion of the overall score for their containing modules. In cases where the tests were summative, they always included a standard test component. Had there been any issues with the adaptive test then the score of the conventional component of the test would be used. This was not necessary in any of the assessments.

1.5.1 Empirical studies

Different forms of study were employed during the research including exploratory studies, a real data simulation and two phases of live testing.

The exploratory studies and live testing studies conducted compared examinees' performance in the Flexilevel test and conventional CBT test. The core methodology is detailed below:

1. Present to the participant a test stage in which items are selected using the Flexilevel algorithm, and a stage in which items are selected using a conventional CBT algorithm. The order of presentation of the two test stages is randomised between participants.
2. Both stages present one item at a time. Each item must be answered in order to progress to the next item and examinees may not go back to previous items.
3. The scores achieved for each stage are statistically analysed to detect any correlation between them.

The exploratory studies. As part of this programme of research, studies were carried out that tested the application of the Flexilevel test in different assessment contexts.

Exploratory studies were carried out to establish the effectiveness of the Flexilevel test in a given context outside of students' assessment regimes. If there were any issues with the approach, they would not impact on students' formal studies.

The two exploratory studies included the pilot study and postgraduate study.

The first phase of live testing. Studies were carried out in real educational contexts in the first phase of live testing. The studies involved computer lab–based summative assessments. These studies investigated the Flexilevel test in a summative assessment context and compared performance of students on a Flexilevel test stage with their performance on a conventional CBT stage of the test.

Real data simulation study. This study intended to investigate the possibility of providing shorter Flexilevel tests that were as effective as longer conventional CBT tests. The methodology was to make use of the data from an existing conventional CBT test, selecting items using the Flexilevel algorithm to the answers provided by examinees, and marking them according to whether or not the examinee answered the item correctly or incorrectly in the actual test.

A statistical analysis of the results showed that the scores resulting from the selection of items using the Flexilevel algorithm were significantly correlated with the scores achieved by students in the live test.

The importance of carrying out the studies in real educational contexts led to the second phase of live testing.

Second phase of live testing. In the second phase, a web application was used that was based on the desktop application used previously. This allowed for studies to take place in a wider range of contexts: online, on mobile devices and in computer laboratories. Shorter Flexilevel tests were compared with longer conventional CBT tests as well as with other related assessment components.

1.5.2 Evaluation studies with primary stakeholders

This strand of the research work took place alongside the studies into the effectiveness of the Flexilevel test approach. The studies within the evaluation strand of work were concerned with issues of attitude and usability for the primary stakeholders.

The studies were separated into two main groups: studies with academic staff and studies with students.

Studies with academic staff. These studies covered the usability of the applications used for the Flexilevel tests and the views of academic staff about the approach.

They involved a survey published to the wider UK Higher Education sector to understand attitudes and usages of CAT in general. Additionally more specific studies about the use of the Flexilevel tests were conducted with experienced colleagues from with the University of Hertfordshire. These included a presentation about the Flexilevel test followed by an opportunity to try it out and then the presentation of a questionnaire for participants to feedback their views. Additionally, interviews were carried out with academic staff about the potential for use of the Flexilevel web application in formative mobile assessment contexts.

Studies with students. The students who participated in these studies were mainly Computer Science students at the University of Hertfordshire. They were from different programmes of study (both online and campus-based) and at different levels of study.

The studies with students involved investigations into both their views of the overall Flexilevel approach and the usability of the applications used to run the Flexilevel tests.

This involved the presentation of questionnaires about the Flexilevel approach as well as the usability of the application. The System Usability Scale (SUS) and the Usefulness, Satisfaction, and Ease of use (USE) questionnaires were also employed to evaluate the web application on mobile devices.

Finally, interviews were conducted to understand students' attitudes to the use of the Flexilevel test and its use as a web application on mobile devices.

The studies conducted as part of this programme of research are summarised in Table 1-1 and Table 1-2 below. Some of the studies applied to different aspects of the work and are so included in the dissertation more than once; for this reason they also appear in both Table 1-1 and Table 1-2 below. Table 1-1 contains a summary of the studies conducted to investigate the effectiveness of the Flexilevel test and Table 1-2 summarises the studies conducted to investigate the views of students and academic staff about the Flexilevel test and also to identify any usability issues with the software applications used to run the tests. Figure 1-1 gives an overall view of the approach of the programme of research. It sets out how different methodologies were employed in order to answer the research questions and how the results of the different forms of studies informed subsequent areas of research.

Study	Participants	Assessment Context	Methodology	Results
Pilot Study Section 4.1.1 Reported in Lilley & Pyper (2009)	24 Level 6 campus-based Computer Science students	Exploratory study PC laboratory on campus. Exam conditions.	Comparison of examinee performance in conventional CBT stage and Flexilevel test stage (10 items in each stage)	Pearson's correlation coefficient (r =0.764, N=24, significant at p≤0.01) Paired t-test (t(23) = -0.954, p=0.350 (no significant difference))
Postgraduate Study Section 4.1.2 Reported in Pyper et al. (2014b)	29 Level 7 Computer Science students	Exploratory study PC laboratory on campus. Exam conditions	Comparison of examinee performance in conventional CBT stage and Flexilevel test stage (10 items in each stage)	Pearson's correlation coefficient (r=0.614, N=29, significant at p≤0.01) Paired t-test (t=-1.279, df=28 p≤0.212 (no significant difference))
Study A Section 4.2.1	131 Level 4 campus-based Computer Science students	Summative PC laboratory on campus. Exam conditions.	Comparison of examinee performance in conventional CBT stage and Flexilevel test stage (20 items in each test stage)	Pearson's correlation coefficient (r=0.461, N=131, significant at p≤0.01). Paired t-test (t=-1.836, df=130, p=.069 no significant difference)

Study B Section 4.2.2 Reported in Pyper & Lilley (2010)	180 Level 4 campus-based Computer Science students	Summative PC laboratory on campus. Exam conditions.	Comparison of examinee performance in conventional CBT stage and Flexilevel test stage (20 items in each test stage)	Pearson's correlation coefficient (r=0.646, N=180, p≤0.01) Paired t-test (t= 1.488, df=179, p=0.139 – no significant difference)
Real data simulation study Section 4.3 Reported in Pyper et al. (2014a)	11 Level 4 Computer Science students 65 Level 6 Computer Science students	Simulation based on summative assessment	Level 4: Comparison of examinee performance in live summative conventional CBT test (22 items) and simulated Flexilevel test (11 items) Level 6: Comparison of examinee performance in live summative conventional CBT test (17 items) and simulated Flexilevel test (9 items)	Level 4: Pearson's correlation coefficient (r=0.946, N=11, p≤0.01) Level 6: Pearson's correlation coefficient (r=0.941, N=65, p≤0.01).
Study C Section 4.4.1	21 Level 6 Online Computer Science students	Diagnostic Online test	Comparison of examinee performance in conventional CBT stage (20 items) and Flexilevel test stage (10 items)	Pearson's correlation coefficient (r= 0.645, N=21, significant at p≤0.01)

Study D Section 4.4.2	18 Level 6 Online Computer Science students	Low stakes summative Online test	Comparison of examinee performance in conventional CBT stage (20 items) and Flexilevel test stage (10 items)	Pearson’s correlation coefficient (r= 0.839, N=18, significant at p≤0.01)
			Comparison of examinee performance between tests:	
			Flexilevel (Study C-Study D)	Pearson’s correlation coefficient (r=0.696 p≤0.01)
			CBT (Study C- Study D)	Pearson’s correlation coefficient (r=0.683 p≤0.01)
Study E Section 4.4.3 Elements of this work were reported in Pyper et al. (2015a)	15 Level 4 Online Computer Science students	Summative Online test. Exam conditions	Conventional CBT stage (20 items) and Flexilevel test stage (10 items)	Pearson’s correlation coefficient (r=0.574, N=15, p≤=0.05)

Study F Section 4.4.4 Elements of this work were reported in Pyper et al. (2015b)	58 Level 4 campus-based Computer Science students	Summative PC laboratory. Exam conditions	Conventional CBT stage (25 items) and Flexilevel test stage (15 items)	Comparison of examinees' performance in both tests: Pearson's correlation coefficient (R=0.62, N=58, significant at p≤0.01). Comparison of examinees' performance in Flexilevel test stage of in-class test and practical programming assignment: Pearson's correlation coefficient (R=0.394, N=48, significant at p≤0.01) Comparison of examinees' performance in conventional CBT test and practical programming assignment: Pearson's correlation coefficient (R=0.292, N=48, significant at p≤0.05)
Study G Section 5.3	22 Level 4 campus-based Computer Science students	Formative Online test	Paper Based Test (PBT) compared with Flexilevel test	Pearson's correlation coefficient (r=0.916, N=21, p≤0.01).

Study H Section 4.4.6	22 Level 6 online Computer Science students	Test 1 Diagnostic Flexilevel test (20 items) Test 2 low-stakes summative Flexilevel test (20 items)	Flexilevel test from Test 1 compared with over time (14 days)	Paired t-test (t(21) = -0.905, p=0.375 (not significant)

TABLE 1-1: STUDIES CONDUCTED TO INVESTIGATE THE EFFECTIVENESS OF THE FLEXILEVEL TEST

Study	Participants	Description	Methodology
Pilot Study & Postgraduate study Section 5.2	24 Level 6 campus-based Computer Science students 28 Level 7 Computer Science students	Collection of views of the approach (pilot study) and usability of the desktop application (both pilot study and postgraduate study)	Questionnaire presented at end of test Direct observation of use of application in test
Study G Section 5.3	29 Level 4 campus-based Computer Science students	A Flexilevel test was made available online using an adapted version of the web application. Students were invited it to take the test on a device of their choice	Data about the user agent (browser and device) used for test was collected in addition to the test data in order to see the proportion of participants who used a mobile device for the test.
Study I Section 5.4	8 Level 7 and 2 Level 4 Computer Science students	Discussion about participants’ use of mobile device, their studies and to discuss views of participants’ usage of mobile devices, Flexilevel test generally and possibilities for using Flexilevel test in mobile formative assessment contexts	Online interview

Study J Section 5.5	10 level 7 online Computer Science students	Students tried a Flexilevel test on a mobile device as many times as requested and then completed two questionnaires about the usability of the web based Flexilevel test application as used on mobile devices	Online SUS and USE questionnaires.
Study K Section 6.1 Reported in (Lilley, Pyper & Wernick, 2011)	Academic staff	An online survey was publicized to academic staff about their attitudes and usage, if any, of CAT.	Online survey
Study L Section 6.2	Academic staff	Following a presentation and a use of the Flexilevel test, academic staff were asked to fill in a questionnaire about their views of the Flexilevel approach and the usability of the desktop application they had used.	Presentation, use of the test followed by a questionnaire.
Study M Section 6.3	Academic staff	Discussion about the Flexilevel approach, perceptions about its usefulness in mobile contexts and the potential for providing tailored feedback.	Face to face interviews

TABLE 1-2: STUDIES CONDUCTED ABOUT THE VIEWS EXPRESSED BY STAKEHOLDERS ABOUT THE FLEXILEVEL TEST

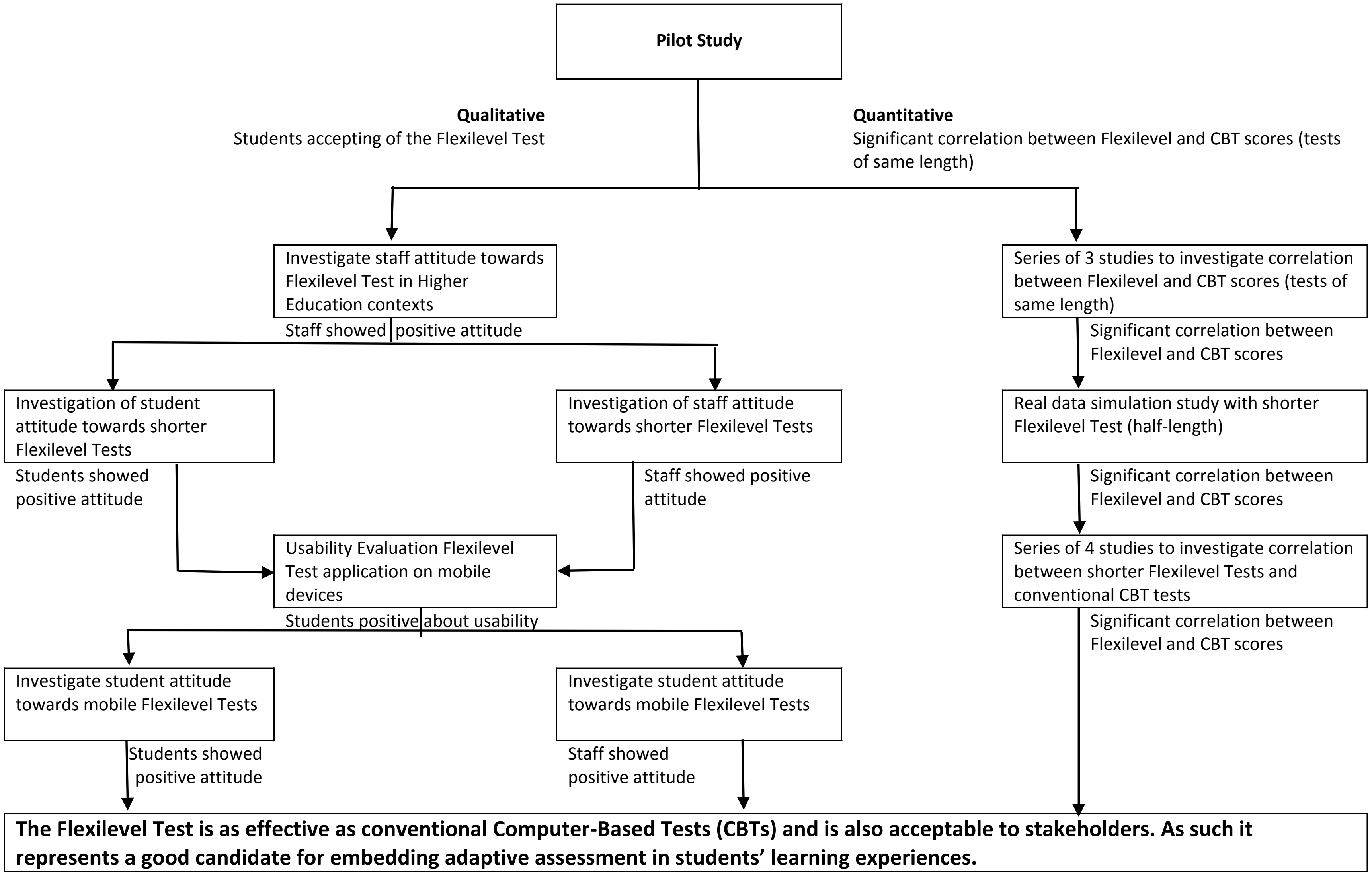

Pilot Study
Qualitative
Students accepting of the Flexilevel Test
Quantitative
Significant correlation between Flexilevel and CBT scores (tests of same length)
Investigate staff attitude towards Flexilevel Test in Higher Education contexts
Staff showed positive attitude
Series of 3 studies to investigate correlation between Flexilevel and CBT scores (tests of same length)
Significant correlation between Flexilevel and CBT scores
Investigation of student attitude towards shorter Flexilevel Tests
Students showed positive attitude
Investigation of staff attitude towards shorter Flexilevel Tests
Staff showed positive attitude
Real data simulation study with shorter Flexilevel Test (half-length)
Significant correlation between Flexilevel and CBT scores
Usability Evaluation Flexilevel Test application on mobile devices
Students positive about usability
Series of 4 studies to investigate correlation between shorter Flexilevel Tests and conventional CBT tests
Significant correlation between Flexilevel and CBT scores
Investigate student attitude towards mobile Flexilevel Tests
Students showed positive attitude
Investigate student attitude towards mobile Flexilevel Tests
Staff showed positive attitude
The Flexilevel Test is as effective as conventional Computer-Based Tests (CBTs) and is also acceptable to stakeholders. As such it represents a good candidate for embedding adaptive assessment in students' learning experiences.

FIGURE 1-1: OVERVIEW OF THE METHODOLOGIES EMPLOYED

2 Educational Context

Work reported in this chapter was intended to inform the educational basis for this programme of research. This work has also been discussed as part of Pyper & Lilley (2007), Pyper (2011), Pyper & Lilley (2008a), Pyper & Lilley (2008b), Pyper et al. (2007), Pyper et al. (2009) and Pyper et al. (2011).

With the increased accessibility to the Internet an overarching information ecosystem has emerged, causing a revolution in the ability of communities to form and share knowledge. If properly exploited, this revolution has the potential to transform students' educational experiences at all stages.

There is increasing diversity in the ways in which students can engage in this information ecosystem in terms of the control they may exercise over what they study, when they study, how long they study for as well as the media they are able to use in order to engage in the information ecosystem for their studies.

It has been argued that there is a disjunction between the ways students are learning outside of Higher Education Institutions (HEIs) and the way in which they are learning within HEIs (Siemens, 2007). It may be argued that this distinction is blurring with the increasing range of online learning opportunities of varying degrees of formality and structure. Nonetheless the relative formality of education in HEIs increasingly does not resemble the environments in which students are learning outside of HEIs (Williams et al., 2011). Indeed it has been argued that there should be a move from traditional classroom-based education to learning ecologies, as well as a change in the structures of learning from hierarchical to networked (Siemens, 2007).

Learning ecologies or ecosystems are made up of the environment of technical tools, utilities and infrastructure to support learning and the transactions of the participants within them (Change & Guetl, 2007). Educational transactions (please see section 2.1.3 for a more detailed discussion) involve the interaction between participants and other participants, as well as that between participants with their environment. They are made up of cognitive, communicative and emotional elements.

Ideally a learning ecosystem should allow access to a broad range of resources and tools for the students, enable knowledge to be shared, reinterpreted and repurposed, and allow for experimentation with its attendant risk of failure (Siemens, 2007). There are also risks for the development of students studying within such ecosystems. Sources of information may be unreliable, even to the extent that they are pedagogically noxious, for example promulgating misconceptions about a given topic. Such environments may be confusing to students (Anderson & Dron, 2010), something that may be couched in terms of extraneous cognitive load, whereby the complexity of the environment acts as an extraneous load and detracts from the student's capacity to process the information that is intrinsic to the learning (i.e. intrinsic load (Sweller, 2010)). This issue will be considered within this dissertation in relation to mobile contexts for learning and assessment. Moreover, students' interactions with other members of the information ecosystem may not support them in attaining the higher order learning skills in the way that an educational dialogue between tutor and student could (Laurillard, 2001).

The potential of such environments seems compelling, although their design is non-trivial. To foster the kinds of transactions that are likely to be of educational value, a framework (Pyper et al. (2007), Pyper et al. (2009)) was developed to set out the narrative structure and interactions that are held to support students. This was subsequently refined to incorporate the idea of learning ecosystems more explicitly (Pyper et al., 2011).

The framework (Pyper et al., 2011) provides an approach to learning and teaching activity that whilst grounded on and connected to the information ecosystem, is intrinsically designed to support the student. A further development here is to reframe the interactions of the framework as transactions and to apply the theory of Transactional Distance to them. The design of the learning ecosystem, in terms of the environment and the transactions within it, is exposed in the narrative account of the tutor and will be considered next.

2.1.1 THE NARRATIVE STRUCTURE

The narrative account of the learning experience is intended to put the educational experiences engaged in by the students into a broader context, and at a more specific level to express the tutor's own experience of the individual learning transactions being undertaken by students (Pyper et al., 2009). There is a risk however that this could have a pernicious effect on the development of meta-cognitive skills of the students. A narrative in the form of the narration that Freire (1970) warns against would risk turning students into passive consumers of information. As such, it is important that students themselves produce their own narrative for their studies. Tutors' narratives provide a model for how students can approach their learning experiences and thus produce their own narrative accounts of their own studies that may be shared with the community, but these are not a substitute for students' own narratives.

The use of narrative in the framework (Pyper et al., 2011) is intended to model the kind of critical and reflective thinking about a given topic that students are expected to engage in. It is also intended to represent the presence of the tutor within a given educational experience. Narrative accounts have an affective role as well as a cognitive one. Lowenthal & Dunlap (2010) and Plowman et al. (1999) point out the value of storytelling in establishing social presence in a learning community. It is intended within the framework (Pyper et al., 2011) that the provision of a narrative account would foster students' active engagement with the tutor and the wider learning community.

The narrative account provided by the tutor is intended to provide students with insights into varying accounts of the same issues as well as the tutor's own view of the issues and how they came to form them. The narrative is embedded within the transactions designed at a low level as well as separately from them at a higher level. This provides context for the transactions set down for students as well as a detailed sense of why they should engage with them.

2.1.2 The environment of a learning ecosystem

It has previously been noted that a learning ecosystem is comprised of a range of content and software applications that support different forms of transaction within the ecosystem. Together with the software applications, this learning content forms the environment of the learning ecosystem. In terms of contexts, the environment includes a range of learning and teaching contexts, from the more formal and structured transactions within computer lab tests through to more informal and potentially self-directed mobile, ad hoc contexts.

In designing the content and software applications for the environment of the learning ecosystem, designers are stating what they think is important about their topic domain both in terms of what is known and how it is known. To place a topic domain in context, the idea here is that it is presented as a whole, interconnected as necessary with other topic domains. Ideally tutors would design for topics rather than for curricula, supporting the preservation of the context within which the topic domains exist, including interconnections with other topic domains. In short, it would balance the value of closely designed and aligned curricula (for example Biggs & Tang, 2007) with the messy, interconnected nature of knowledge construction (Siemens, 2007).

Thus the formal structure involving scheduled synchronous sessions and criterion-referenced summative work would make up a key part of the learning ecosystem; indeed it would support students in engaging with the wider learning ecosystem. An example would involve the sharing of tools and resources used by students in a module within a Virtual Learning Environment (VLE) or some other shared environment for that course. Here, elements of courses that have been subject to formal quality control procedures sit alongside tools and resources whose quality control is informal. However, it should be noted that much of the learning ecosystem is designed so that it is pedagogically useful.

Those aspects of the learning ecosystem that are not so closely designed are still largely contextualized within the structure of the learning ecosystem – for example through the use of narrative, either of a student or a tutor. This aspect of the learning ecosystem would either model or include unaltered elements of the wider information ecosystem. Students who are more autonomous in their interactions in the learning ecosystem would be increasingly guided to use this wider ecosystem. Overall, the design of a given educational experience would be intended to foster the development of students' skills of self-regulation that would support them studying autonomously.

Content also represents an important part of the learning ecosystem. Learning content can be produced by tutors, students and through their interaction with each other. In the process of writing an essay, a student might produce large quantities of notes, engage in conversations with peers and tutors about the work, and write drafts and notes to themselves before finally submitting the essay.

The way in which content may be produced was previously considered by Pyper & Lilley (2007) and Pyper & Lilley (2008b) as it was hypothesised that the way in which learning content was produced had unintended effects on the educational properties of the learning. In short, there are two main forms of learning content: durable learning content and ephemeral learning content. This represents a change from the term "disposable" learning content used in Pyper & Lilley (2007) and Pyper & Lilley (2008b). The term "ephemeral" is preferred to avoid any potential negative interpretations. In this example, the essay itself is the durable learning content and everything else is ephemeral learning content.

Like most other methods, this method of producing learning content (i.e. essay writing) generates both types of learning content and, as with most other methods, it serves significantly different purposes and has different properties. Nonetheless it is suggested that in contemporary educational contexts, such methods are often used in the same way as learning resources (Pyper & Lilley (2007) and Pyper & Lilley (2008a)). It seems that a key reason for this is the use of technology in education. Without technology, capturing ephemeral learning content is not usually feasible, so ephemeral learning content would rarely be used as a learning resource. However, the use of technology has made it increasingly easy to capture and distribute ephemeral learning content. Indeed the storage of learning content is often necessary for the technology to function, so the natural distinction between the two learning content types is eroded to a significant extent when technology is involved.

Without technology, much of the ephemeral learning content is lost because it is verbal. Indeed, it was for this reason that in the early work conducted (Pyper & Lilley, 2008a), a dialogic learning environment was identified as being among the most likely to show the distinction between disposable and durable learning content. The research conducted shows support for this idea and has also led to the production of a model of resource production (Pyper & Lilley, 2008b) as shown in Figure 2-1.

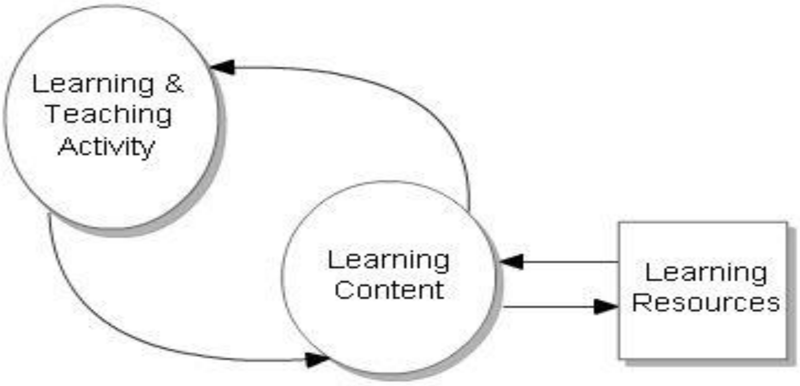


FIGURE 2-1: THE INTERACTION OF THE PROCESS (LEARNING AND TEACHING ACTIVITY) AND THE LEARNING CONTENT (OUTPUT) THAT MAY RESULT IN THE GENERATION OF A PRODUCT (LEARNING RESOURCES)

Figure 2-1 shows the interaction of learning and teaching activity, the creation of learning content inherent in this activity and the idea that learning resources are created from the learning content. When technology is involved in the process, the ease with which learning content can be turned into learning resources makes it more likely that the activity would lead to the creation of learning resources. Without the use of technology to capture ephemeral learning it might feed back into the teaching and learning activity, but not be made into a learning resource. If the teaching and learning activity was aimed at producing learning resources, then durable learning content would be created and go towards creating the resource.

To understand the properties that allow the differences between ephemeral and durable learning content to be identified, four dimensions were identified (Pyper & Lilley, 2008b) as shown in Figure 2-2 below. The properties of a given item of learning content can be mapped along these dimensions to show the extent to which it is disposable or durable learning content. This is intended to support creators of learning content in identifying the type of learning content they have and enable them to make conscious decisions about how to use it, particularly its suitability for use as a learning resource. After the table, each dimension is discussed in detail.

Ephemeral	**← Dimension →**	**Durable**
By-product of learning activity	Intent	Product of learning activity
Context sensitive	Context	Context independent
Incomplete	Completeness	Complete
Time sensitive	Longevity	Time insensitive

FIGURE 2-2: THE PROPOSED DIMENSIONS THAT MAY BE USED TO DISTINGUISH BETWEEN DISPOSABLE AND DURABLE LEARNING CONTENT

Content generated as a by-product of learning activity ←→ Content generated as a final product of learning activity.

A key influence on the type of learning content being created is the intent behind its creation. If learning content is being produced with the intention of using it to be made into or contribute towards a learning resource – for example being subjected to quality assurance – then clearly it is much more likely to have the properties that make it durable learning content. Conversely if learning content is being created that is intended only to support a current learning activity (e.g. a diagram sketched on a whiteboard to exemplify some concept), it is likely that this will have the properties of ephemeral learning content.

Context Dependent ←→ Context Independent

The extent to which the learning content can be used meaningfully outside of the context in which it was created is another dimension that can discriminate between disposable and durable learning content. Discussion forum posts that are written informally and contain colloquialisms may not make much sense when removed from the context in which they were created. By contrast, a student’s essay should still make sense when removed from the context in which it was created.

To provide resources that are durable and reusable, context independence has been considered important, for example in terms of learning objects (Irlbeck & Mowat, 2007), but it has also been identified as impairing the educational value of a resource (Polsani, 2006).

Incomplete ←→ Complete

The distinction between incomplete and complete rests on the extent to which the learning content contains sufficient information on its own to be comprehensible to someone else. Ephemeral learning content is much more likely to contain partial information, perhaps an unfinished sentence as the conversation changes direction or someone is interrupted. Lecture slides stored on a server somewhere for later retrieval may lack the additional learning content of the lecture for which they were used – for example what the tutor said. Durable learning content is much more likely to be complete, so there should be no incomplete sentences in a student’s essay or unintended ambiguity in a text book.

Time sensitive ←→ Time insensitive

Time sensitivity refers to the extent to which the learning content becomes irrelevant, meaningless or lost entirely over time. Ephemeral learning content is much more susceptible to this than is durable learning content. Traditionally, much of the learning content created by learning and teaching activity would be verbal in the form of educational dialogues. Inherently the words used exist only briefly and only endure in some processed form in the memory of participants.

This is the dimension upon which technology has the greatest inherent impact, since technology can take learning content that is time-sensitive and make it time-insensitive simply by storing it and making it available. No other action is required by the author(s) of the content; this is simply a function of the technology.

Implications of this categorisation

The current distinction between disposable and durable learning content rests on the dimensions identified above. Durable learning content will tend to the right of the scale, would usually result from a conscious effort to produce learning content, and would be relatively insensitive to the context in which it was created. The information it contains is coherent and complete and is packaged in a way that permits it to endure in its meaning and usefulness over time. Ephemeral learning content by contrast is usually created as a by-product of learning activity. It is commonly time sensitive and rooted in its context. Educational dialogues are the principal transactions that generate ephemeral learning content, but this is also the case for any learning transaction in which the end goal is loosely defined. It is important to note that these are quite often features of tertiary level learning activities (for example design problems (Jonassen, 1997)).

It may be argued that the two types of content are educationally useful in different ways. However, the use of technology in education makes it more likely that they are used in similar ways, as a learning resource; specifically when technology is used, as part of its function it stores learning content. This has the effect of skewing the time-sensitive/time-insensitive dimension such that all learning content, whether disposable or durable, is made to endure. However, it does not inherently alter the other dimensions thereby giving ephemeral learning content the appearance of durable learning content, but in only one of its properties.

Part of the reason this is pertinent to the design of the environment in a learning ecosystem is that it is in flux: elements are constantly being added, removed or amended and the greater the level of interaction between the different elements (or nodes, to use the connectivist term (Siemens, 2007)), the greater the level of flux in a given section of the ecosystem. This may result in durable changes to the ecosystem, even though these may be by-products of a transaction between students. The value of these changes to the environment may be limited to the interaction that produced them, but this interaction is an important part of the learning ecosystem. Information that is partial, that may be inaccurate or is contested, is a feature of any information ecosystem and it is important that students are able to critically evaluate the value of such information. To revisit an earlier point, significant parts of a learning ecosystem following the framework should be designed or structured in such a way that the framework has more of the features of durable learning content (Pyper & Lilley (2007), Pyper & Lilley (2008b).

The technologies or media may also have a significant effect on the way in which transactions may take place. These tools may be generic software applications, such as productivity applications (word processors for example), or they may be specific tools; see for example Pyper (2011). It is increasingly a feature of the learning ecosystem approach that it contains a diversity of tools. Indeed mappings between pedagogical approaches and media have been produced (for example Laurillard, 2001)) that can usefully be applied to the design of transactions. Some tools may be very specific to a single function and it is here that tools such as the ones that mediate Flexilevel tests seem to offer most promise.

Summary. In this section, the factors that may play a role in the design of the learning ecosystem environment have been considered. Potential implications of the method of production of resources has been detailed and the media and software applications that are available to participants within the learning ecosystem have been identified as important elements of the learning ecosystem overall.

In addition to their effect on the environment, of further relevance to this work is the extent to which media can facilitate transactions that involve dialogues between participants. Transactions are central to the framework and the way in which they may be mediated by technology has substantial implications for the nature of the transaction. This aspect of the framework will be discussed next.

2.1.3 Transactions

Transactions represent the interactions between participants and other participants, as well as interactions with their environment. They are based on Dewey's notion of a transaction as interpreted by Moore (1993) and Garrison (2011). Transactions incorporate both the individual and social elements of an educational experience – the individual's understanding within the context of the group's understanding of a given topic. The term also encompasses emotional as well as cognitive and communicative components (Elkjaer, 2009).

Transactions exist within a limited period of time and are often treated synonymously with interactions. However in this analysis, interactions lack the interdependence of the participants and/or their environment in a transaction. The idea of student–teacher and teacher–student participants within an educational dialogue, for example as influentially argued by Freire (1970), provides an example of how the role of participants depends on the interplay between those participants.

There is also an element of change in the relationship between tutor and student whereby roles change depending on the nature of the transaction. By extension, the relationship between a student and their environment may also generate change, not only in the student and what the student contributes to the environment, but also, in the case of adaptive systems, in the way the environment may adapt to the student's contributions. In the case of an adaptive test, the performance of the student changes the composition of the test that they are taking and the nature of the test may change the student's affective response to the test – for example a test the student finds difficult may have a demoralising effect (Weiss (1974), Carlson (1994), Wainer, (2000)) These issues will be discussed in more detail in terms of Computerised Adaptive Testing (CAT) later in this dissertation.

Whilst transactions are distinct from interactions, the types of transactions that may take place follow a similar typology to those identified for interactions. Garrison & Anderson (2003) identify six types of interaction: student–teacher, student–content, teacher–content, student–student, teacher–teacher and content–content. The framework (Pyper et al., 2011) focuses on the student-centric interactions, identifying how students may engage within the learning ecosystem; independently, cooperatively, collaboratively or competitively. Moreover, transactions may involve many to one, one to many, and one to one relationships.

As well as in the design of the environment of the learning ecosystem, it is in the design of the transactions that students can be supported to varying degrees in their work. Transactions can be very closely sequenced and designed, be set out probabilistically or they can be left open for students themselves to decide on how they will interact.

The identification of different modes of transaction is based on work conducted in Computer Supported Collaborative Learning (CSCL) (Dillenbourg (1999), Paulus (2005)) and includes individual, cooperative and collaborative transactions. Additionally, a competitive transaction is included. The modes of transaction are now discussed in more detail.

Cooperative transactions. In cooperative transactions, students contribute to a common effort, but do so with relative independence from other members of a learning community as compared to collaborative transactions. There is similarity between the ways in which an individual and cooperative transaction may take place. With cooperative learning however there is likely to be more active student–student transaction. Common forms of cooperative transactions include students interacting with learning content to address a common goal, or to respond to one another’s queries

Both cooperative and collaborative modes of transaction may happen at the level of a one-to-one transaction, although it has been suggested that an optimal number for such forms of transaction in group learning is around five members as a group (Paulus, 2005). Nonetheless, the scope of the framework goes beyond group learning and is intended to help describe and design for how individuals may interact with the learning ecosystem; this would include group work but is not limited to it.

Collaborative Transactions. The distinction between cooperative and collaborative learning is considered unclear (Jones et al., 2007). However, it has also been suggested that there is more of an affective aspect to collaborative learning (Stacey, 2005) and that the group works together more or less synchronously on a given task. The extent to which students actually work collaboratively is unclear and it may be argued that the main form of collaborative transaction is tutor–student. Laurillard pointed out that the assumption of value in student–student transactions may be problematic (Laurillard, 2001). Whilst collaborative transactions may be centrally important in education, designing for them is non-trivial, and where design is sufficiently difficult, learning activities are in any case likely to need significant mediation by tutors; the environment or tools are unlikely to provide the necessary level of guidance.

Competitive transactions. Whilst there is an underlying competitiveness to education, for example in norm-referencing, competitive transactions seem less well represented explicitly in learning design. There is potential for negative educational experiences (for example the impact on a student of seeing that their score for a test has put them in the bottom 5% of their class), but carefully designed competitive learning experiences (for example combining cooperative transactions with competitive transactions (Attle & Baker, 2007)) may guard against such negative experiences.

Additionally, competitive transactions may be individual with students competing with themselves ipsatively as well as comparing themselves to others normatively. Further competition can be made optional or even separate from students' studies and take, for example, the form of games (Gredler, 2004). Whilst competitive transactions need careful consideration, it may be argued that if students' education is to be grounded on authentic contexts, then a competitive element within their learning experiences would be beneficial as well as motivational (Attle & Baker, 2007).

Individual transaction. Individual transactions include the students' transaction with the content and structure of the environment they find themselves in. The extent to which this is designed by tutors feeds into the support they receive from tutors. This is implicit in the design of the learning experiences and also explicit in the provision of the tutors' narrative.

In this sense, whilst a transaction is ostensibly individual, it is important to note that the environment is designed, and as such this individual transaction is supported by that environment. As such, the transaction would involve a clear and evident sense of the tutor’s presence and is intended to engage students actively and reflectively, so it is intended to go beyond transmission or the passive consumption of information. Indeed, one of the most important individual transactions is the transaction the student has with themselves, in the form of reflection. The majority of a student’s individual transactions will involve student–content transactions, reflective and reflexive transactions, such as those seen in reflective and reflexive practice (Prpic, 2005).

Within the framework (Pyper et al., 2011) , all educational experiences should be consequential in some way; transactions that make up an educational experience contribute to the consequence of that educational experience. For individual transactions, assessment and feedback are key methods of achieving this. Feedback is commonly provided through the use of criterion-based feedback in the form of a score; it can also be extended in a range of ways to encourage students to engage in critical reflection about their studies. Although assessment will be discussed in more detail later, the studies into the Flexilevel test approach for this programme of research have mainly taken the form of summative and formative objective tests taken by individual examinees. As such they have been designed as individual transactions.

So, individual transactions are consequential, something that can be effectively mediated through assessment and feedback. In terms of educational experience design they represent an important way of providing the balance between support and flexibility that is intrinsically important within the framework. Individual interactions provide students with flexibility in how they learn; they are free to engage with the learning experiences in a way that is flexible in both time and place. In terms of support, it is a contention of this programme of research that adapting tests to their performance would provide students with such support.

One factor that is held to influence all transactions is the cognitive and communicative distance between the participants within them. This will be considered next.

Exploring transactional distance in education. The way in which transactions in educational experiences are mediated has an impact on those transactions. This can usefully be framed in terms of Transactional Distance theory (Moore, 1993). Transactional Distance is influential in distance education, although it applies more broadly to educational transactions generally. Importantly, the distance is pedagogical (Moore, 1993) and based on the immediacy of communication (Wheeler, 2007) rather than geographical, something that is a principal contribution of the theory (Gorsky & Caspi, 2005).

Distance relates to the space in understanding between the student and others in their environment, particularly tutors as well as the structure of the environment itself. It is within this space that misunderstandings and confusion can occur; the variables that have been held to influence transactional distance include the dialogue between participants (usually student and tutor), the extent to which the learning experience is structured and the extent to which the student has the self-regulatory skills to be able to study.

There exists a distance in the extent to which tutors understand students' conceptions of a given topic and learners understand the explanation of it. So when opportunities for dialogue between tutors and learners are relatively low – for example when dialogue is conducted infrequently via a text-based medium – then there is considered to be a relatively high transactional distance (Moore, 1993). The distance seems to be substantially explicable in terms of the opportunities for dialogue (in practical terms tutor–student dialogue). Indeed, dialogue may be the main factor (Gorsky & Caspi, 2005) in influencing the transactional distance of a given educational experience. Nonetheless, other elements of Transactional Distance are also of interest to this programme of research, in particular the transaction between a learner and their environment.

A contributory factor to the transactional distance of a given educational experience is the extent to which that educational experience structured (Benson & Samarawickrema, 2009). Where learners' educational experiences are closely structured and unchanging in the transaction with students, this is considered to contribute to a relatively high transactional distance.

It has been noted that Transactional Distance applies regardless of the context for the learning. This is important for the work reported here as the concern is for applying effective formative assessment experiences within different educational contexts.

A key motivation of devising the learning framework (Pyper et al., 2011) was to incorporate the use of the learning ecosystem idea whilst still explicitly providing support for students in educational experience design. In providing this support, the students' understanding of how they are progressing, as well as the tutors', is fundamental and thus assessment clearly is important in educational experiences for this purpose.

It seems that the Flexilevel test lends itself well to the kind of embedded and numerous assessment opportunities that would support sound transactions within this framework. One of the ways it could do this is in reducing the structure of an educational experience; where a conventional CBT would not change in a transaction with the learner, an adaptive test, such as the Flexilevel Test, that could tailor the test to the learner would be able to adapt.

This chapter has discussed the pedagogical background to this programme of research and located the Flexilevel test approach within it. In the next Chapter, the focus will focus on assessment and approaches to adaptive assessment.

3 APPLICATIONS OF COMPUTERS TO ASSESSMENT

One of the aims of this programme of research was to investigate the extent to which the Flexilevel test can provide adaptive tests that may be applied within contemporary educational contexts. Having set out the broad educational context in the previous chapter, this chapter will consider different levels of involvement of technology in assessment, contexts for assessment and objective testing. Furthermore, different approaches to CAT will be detailed and the rationale for the choice of the Flexilevel test will be set out. It is intended that this chapter will inform the discussion of the findings of this programme of research in the subsequent chapters.

Bull & McKenna (2003) indicate that there are different levels of involvement of technology within assessment and this provides a basis upon which different forms of assessment can be identified. Furthermore, Bull & McKenna (2003) suggest a distinction between Computer Assisted Assessment (CAA) and Computer Based Assessments (CBA).

CAA can include anything that has a technical component, for example paper based tests that are scanned for marking, or essays that are written using word processors. Where computers or computational devices become more intrinsic to the assessment, for example in the context of a practical programming test, the assessments may be considered to be CBA. Here the assessment would not be feasible or even possible without the technical component.

Where the internet is involved, a test undertaken via a web browser or other network-enabled program is said to be an online assessment (Warbuton & Conole, 2005). This is often considered synonymous with e-assessment but as noted, for this work e-assessment is considered to be broader than online assessment. Indeed, it may be taken to subsume all forms of technology-enhanced assessment.

Additionally, as noted earlier in terms of Transactional Distance, distance may be conflated with method of delivery. For example online delivery may involve remote self-assessment but the distance comes from the format and context of the assessment rather than its means of delivery. As an example, empirical studies carried out as part of the work reported here were conducted both using online subjects and in a computer laboratory. In this regard, it is perhaps more useful to consider forms of assessment that are available rather than how they may be deployed using technology.

As the range of forms and approaches to the use of computers in assessment has become wider and more diverse, something that is evidenced in the examples provided below, the distinction between CAA and CBA has lost prominence and, more recently, the term e-assessment has been applied to technology-enhanced assessment approaches overall. (See, for example (Warbuton & Conole, 2005).) Indeed it may be anticipated that the term e-assessment itself will be subsumed into assessment more generally (Warburton, 2013). This seems increasingly likely as the application of technology to assessment becomes ubiquitous. An indication of this is perhaps seen in the diversity of e-assessment in different assessment contexts, and these are discussed next. In this chapter, students are also referred to as examinees to reflect the fact that they are undertaking a test.

3.1 Contexts for e-assessment

Clearly the purpose of an assessment is instrumental to how it is designed and implemented, and also to the attitudes of academics and students taking part. There may not always be a clear distinction between different contexts for assessment; however common contexts for assessment include diagnostic assessment, self-assessment, formative assessment and summative assessment (Ward (1981), Bull & McKenna (2003)).

3.1.1 Diagnostic assessment

Diagnostic assessment occurs prior to learning (for example at the beginning of a course) and is designed to identify the examinee’s current level of proficiency in a given domain. Results provide an overview of where the examinee may need to improve, where they may have misconceptions that need to be addressed and areas in which they are strong. It provides a basis to support further instruction as to what students need to focus on in their studies. As an example, Lazarinis et al. (2010) report on an adaptive system that uses diagnostic testing to support students in their studies.

3.1.2 Self-assessment

This form of assessment provides opportunities for students to test themselves and gain a greater understanding of their own proficiency in a given domain. It may be provided automatically by an assessment system – for example by means of criterion-referenced feedback, which is likely to include a score and feedback – or it may be more general and encourage reflection on progress. There is less pressure in self-assessment tests than might be present in other contexts of assessment since these tests need not be part of a formal assessment regime as the other forms of assessment identified here are likely to be. In this regard, Brusilovsky & Sosnovsky (2005) describe a web-based application that generates parameterised programming exercises for the C language, automatically evaluating the correctness of student responses.

3.1.3 Formative assessment

Formative assessment is usually voluntary and carries no marks. It is intended to provide feedback to students about their level of proficiency – for example in terms of criterion-referenced feedback in the form of a score, and often more detailed guidance for future work. It is often included under the approach of "assessment for learning" in which assessment is an integrated part of students' learning experiences – for example objective tests before and/or after scheduled sessions in order for students to gain a better sense of their own understanding and to identify where they may need to focus their efforts (JISC, 2014).

Students should typically encounter less pressure about making mistakes in formative assessment (Biggs, 1999), the idea being to represent accurately their proficiency to support them in their learning. For example, Higgins & Bligh (2006) report on a CBA system for the formative assessment of diagrams that eases the marking load on academics whilst allowing students to still obtain regular feedback on their work.

Another approach to supporting formative assessment is through the use of peer assessment (Sitthiworachart & Joy, 2003); again the resource required for feedback is eased for academics, and students are provided with an opportunity to both receive and provide feedback, something that is held to support their learning for example through the development of their critical review of their own work (Brown, Bull, & Race, 1999).

3.1.4 Summative assessment

Summative assessment is used to generate a grade or score for students that contributes in part or whole to a form of accreditation – for example a grade in a module or a pass/fail grade in a certification programme. As with other components of a formal course of study, the assessment is typically mapped to a given learning outcome or set of learning outcomes (Biggs, 1999). Also, as with other components of a course, it seems there is the opportunity for making existing approaches more efficient as well as introducing innovation. As an example, Daly & Waldron (2004) used an automated marking system, Roboprof in laboratory exams that allowed students to submit their programming solutions to the system during the exam. This allowed for rapid feedback while also easing the marking burden on tutors.

3.1.5 Implications of different assessment contexts

Different forms of assessment are motivated by different assessment needs, and this has significant implications for the requirements they place on assessments and the resulting ways in which they may be designed and implemented. For summative assessments these are likely to be more stringent than they might be for a formative or self-assessed evaluation (Knight, 2001).

Moreover, the attitudes of both examinees and assessors is influenced by the purpose for assessment; in general terms it has been suggested that summative assessment is treated more seriously than formative assessment and as such, there is more conservatism in the approaches adopted for summative assessment than for formative assessment (Bull & McKenna (2003), Knight (2002)). Indeed, this is evident in work conducted for this programme of research (Lilley & Pyper, 2009). Given the importance of summative assessment (Boud, 1995), this is perhaps not surprising;

nonetheless, it is an issue of interest, since attitudes of stakeholders are an important concern in this programme of research.

Sometimes the distinction between formative and summative assessment is somewhat eroded by the use of low-stakes assessment. Formative assessments that do not count towards the grade achieved for a given student, but are required to pass a given learning unit or module, may be considered to be low-stakes, whereas summative assessments that do contribute significantly to a grade are considered high-stakes assessments (Knight, 2001). A further distinction is the use of low-stakes summative assessments that typically contribute only a relatively small proportion of the marks available for a given learning unit or module.

It might be a requirement of an educational experience that students take the formative assessment opportunities available to them in order to pass the educational experience. The rationale for doing this may be to ensure engagement with more of the learning activities available but also practically to ensure that students are familiar with an assessment format prior to taking a summative assessment using that format.

For the purposes of this research, formative assessment is not a requirement for passing a module, nor does it contribute any marks to an overall grade. Assessments that are required but carry no marks and assessments contributing 10% or less of a score for a module are considered to be low-stakes summative assessment. Assessments that contribute more than 10% of the score for a module are considered high-stakes summative assessments, which is consistent with institutional practice.

As can be seen from the discussion above, there is a wide range of computing technologies being applied in different assessment contexts. Higgins and Bligh (2006) assert that computing technology can be used to improve the productivity and efficiency of existing assessment, and this in turn can provide more opportunities for embedding formative assessment in students' educational experiences. When less effort is expended on preparing for and executing assessments, it is suggested that there is more resource available and it is also less onerous to embed formative assessments as part of educational experience design work. Equally, the potential for innovation is another driver, to provide entirely new approaches to assessment using computing technology, something that is part of a wider discussion about the use of technology in education generally (for example, Noss & Pachler (1999), and in terms of assessment more specifically (Timmis et al., 2012))

One area of work that seems particularly active in the domain of Computer Science education is the automated marking of programming (Ihantola et al., 2010) which includes the comparison of the design and implementation components (Hayes et al., 2007), the automated presentation and marking of scripting problems in the Linux Operating System (Solomon et al., 2006) and individualised interactive exercises on the topic of algorithms that provide automatically-generated feedback (Nikander et al., 2004). An example from beyond Computer Science education relates to the automated marking of free-text responses (Butcher & Jordan, 2010).

These examples may be identified as supporting individual interactions (Pyper et al., 2011); indeed the value of practice has been seen as a motivation for automating elements of the assessment in the literature, especially programming (Ihantola et al., 2010). However, whilst the focus of this project is on individual interactions, these occur in a social educational context and it is important to note that work on assessing cooperative and collaborative interactions is also well represented in the literature, for example in terms of peer assessment (Sitthiworachart & Joy, 2003) and the analysis of social learning activity (Rabbany et al., 2012).

In conclusion of this section, the range of activity is broad and diverse. However, one area that remains a substantial part of e-assessment provision in the sector is objective testing (Timmis et al., 2012) and this will discussed next.

3.2 Objective testing

Objective tests are made up of questions, or items as they are referred to commonly in the literature and as such in this programme of research. The items have predetermined answers. Examinees make a selection from a list of options available to them to provide their answer to the question. As such, the marking of such tests is considered to be objective; markers do not make a judgement about the examinee's response to the question itself. The design of the test however remains subject to the judgement of experts (Ward, 1981) and may not be considered objective.

There is a range of formats of objective items, including true/false items, matching items in lists, matching labels to parts of a diagram, Multiple Choice Questions (MCQs) (see Figure 3-2) and Multiple Response Questions (MRQs) (see Figure 3-2)(Bull & McKenna, 2003). MCQs and MRQs and are the most commonly used forms of objective test, and for this programme of research, the focus has been on MCQs.

Items in an MCQ are made up of a stem and options. The stem is the stimulus for the item – it may be a statement or a question. The options include distractors and a key. (For MRQs there is more than one key, as shown in Figure 3-2.)

Stem	This is the stimulus for the item
Distractor	Distractors are plausible alternatives to the key.
Key	The key is the correct response to the item
Distractor	Another plausible alternative to the key
Distractor	Another plausible alternative to the key

FIGURE 3-1: THE STRUCTURE OF A MULTIPLE CHOICE ITEM

Stem	This is the stimulus for the item
Distractor	Distractors are plausible alternatives to the key.
Key	The key is part of the correct response to the item
Key	The key is part of the correct response to the item
Distractor	Another plausible alternative to the key

FIGURE 3-2: THE STRUCTURE OF A MULTIPLE RESPONSE ITEM

Usually either four or five options are used, with four options providing a good balance between the feasibility of creating a sufficient numbers of items for a given test with creating items that have plausible distractors (Ward, 1981). The creation of items is already non-trivial – requiring five options (i.e. four distractors plus the key), places additional demands on the test designers. Requiring three distractors plus the key makes the creation of plausible distractors more feasible and still offsets the impact of examinees guessing in the test (Ward, 1981).

A more detailed discussion of the composition of good items is beyond the scope of this dissertation; however, an important characteristic of an item is its level of difficulty. Items may vary in difficulty; this may be expressed in terms of Bloom's taxonomy of educational objectives in the cognitive domain (Bloom, 1956). The hierarchy has six levels with lower levels being subsumed by the levels above them. It is constructed as follows: (Bull & McKenna, 2003), (Brown, Bull, & Race, 1999):

> **Knowledge.** Primarily tests recall or recognition of information including facts, theories, concepts and procedures. The knowledge level underpins the other levels, but whilst in those other levels it is a contributory factor, here it is the major factor.
>
> **Comprehension.** Relates to the understanding of facts, for example through the interpretation of information and extrapolation from it. The translation of a communication, for example from one language into another or into different terms from the original communication is also considered to be a skill at the level of comprehension.
>
> **Application.** Requires a demonstration of skill of going beyond comprehension, and applying what is comprehended to novel contexts. Here the skill is not to demonstrate comprehension of how to use a specified technique or approach but to apply the appropriate technique or approach without it being specified. As such, at the application level, the student is demonstrating that they can apply what they understand to novel situations without guidance about the appropriate technique or approach.
>
> **Analysis.** More than comprehension and application, the level of analysis requires the ability to break down the parts of a given topic and show how they

interrelate. This might involve demonstrating an understanding of how ideas interrelate, or being able to test a hypothesis by identifying and applying relevant information.

At this level, students are expected to demonstrate that they can also infer assumptions and underlying themes from a given topic or task.

Synthesis. The ability to take information from different sources of existing knowledge and combine and recombine them into a new form, for example a new insight or idea. Unlike comprehension, application and analysis, where whole ideas or complete accounts may be presented, at the level of synthesis the student is expected to create that whole idea or complete account themselves.

Evaluation. The ability to make informed judgements about a given issue and to justify those judgements based on a critical evaluation of pertinent information or evidence. The evaluation may also involve the use of existing standards. Evaluation subsumes elements of knowledge, comprehension, application, analysis and synthesis and also may have value elements associated with it, so the student may also be making value judgements about a given topic, or may be defending a position with evidence derived from evaluation work.

When constructing items or other educational tasks, words associated with each of these levels may be used to describe what is required. This provides an indication of the level of difficulty of what is being asked. For example, Figure 3-3 represents an item at the knowledge level (examinees are being asked to recall and recognise information), and as such might be considered to be relatively low down in the hierarchy and be a relatively easy item for examinees within the anticipated test population. The item shown in Figure 3-4 however represents an item at the application level. In this example, the examinee is being asked to apply what they understand about control structures (specifically an If…Then…Else statement) and apply it to a novel problem.

Stem	What is the smallest unit of data that can be represented inside the computer?
Distractor	(a) One byte
Key	(b) One bit
Distractor	(c) One word
Distractor	(d) One nybble

FIGURE 3-3: EXAMPLE KNOWLEDGE LEVEL MCQ ITEM

Stem	Consider the code excerpt below:
	`Dim a, b, c As Integer`
	`a = 10`
	`b = 2`
	`c = a / b`
	`Label1.Text = "orange"`
	`If c < 0 Then`
	`  Label1.Text = "apple"`
	`Else`
	`  Label1.Text = "banana"`
	`End If`
	Which of the following will result from compiling this code, assuming all other code for your web page works properly?

Distractor	(a) Label1.Text is set to 0
Distractor	(b) Label1.Text is set to apple
Key	(c) Label1.Text is set to banana
Distractor	(d) Label1.Text is set to orange

FIGURE 3-4: EXAMPLE APPLICATION LEVEL MCQ ITEM

The value of applying an approach such as Bloom's taxonomy is in the creation and review of items. The taxonomy provides a basis upon which test designers can design items at different levels of difficulty using appropriate terms in the questions, for example Thompson et al. (2008).

There is debate about the extent to which objective tests can test the higher cognitive skills (for example Biggs, 1999). Whilst this debate is beyond the scope of this work, it is worth noting that the objective tests that are the subject of the research reported here remain part of a wider educational context. This is consistent with the observation that other forms of assessment may be more appropriate for assessing higher-level skills (Brown, Bull & Race, 1999). As such, objective tests could form part of an overarching learning ecosystem.

Additionally, within this ecosystem, embedding objective tests can provide the basis for assessment experiences that go beyond the question–response foundation of this approach, which is consistent with an assertion that Nicol (2007) makes within the auspices of his framework for using MCQs. Further, as he identifies, the way in which MCQs can be used to support feedback is also relevant to the extent to which objective testing may be applied to the higher cognitive skills; Davenport et al. (2009) showed that the nature of the feedback provided to answers to objective questions can engage students in using such higher-order skills.

Further pedagogical advantages of using MCQ tests include the ability to target specific areas of knowledge and understanding and to focus on the abilities being tested without the influence of other skills that may not be subject to the assessment.

Efficiency is also a key driver for the use of MCQ tests. Objective testing overall is a format that lends itself to a technical implementation: scoring examinees' responses using technology is a much easier task to perform when there are only a limited range of possible responses. Also, even though the task of creating objective questions is non-trivial, once they have been created, they can be deployed to a technical solution so that both the delivery and marking of the items is automated.

It is perhaps not surprising then that objective testing has been and continues to be strongly represented in assessment regimes in higher education, (Ward (1981), (Timmis et al., 2012)).

In contemporary settings there is a diverse range of technical implementations of objective testing. The plethora of technical solutions that was mentioned in the previous discussion about the use of technology in different assessment contexts remains true of objective testing. Largely as a result of the technology available, a broad range of implementations and formats are available (McAlpine & Hesketh, 2003). In addition to the presentation of such tests on screens, the use of Electronic Voting Systems (EVS), or clickers, has become a significant part of this provision.

EVS makes use of objective testing approaches and has been used to provide a range of feedback, for example in providing normative yet anonymous feedback live in lecture theatres (Davenport et al., 2009). This approach takes advantage of the potential value of normative feedback whilst managing the potential risks; students are able to obtain feedback on how they are doing in relation to their peers without their performance being made public.

Additionally, the provision of immediate feedback means that it fulfils an important property of effective feedback, in that it is timely (Charman (1999), Shute (2008)). This has added to the value of using EVS by others (Cliffe et al., 2010), and also within the University of Hertfordshire (Cubric & Jefferies, 2012).

Barker & Bennett (2012) used EVS in conjunction with peer review to provide an assessment experience that engaged higher-order skills and found that the students engaged more effectively with their practical work than previous cohorts of students who had not used the EVS and peer review approach.

As previously noted, a significant proportion of e-assessment relates to objective testing using Multiple Choice Questions (MCQs) that are presented on computer screens (Timmis et al., 2012) and indeed, MCQ tests for individual students form the focus of this work. In this area alone, the impact of technology along both the efficiency and innovation dimensions and the role of computers in supporting objective testing have been substantial.

There is a strong heritage of technology being used to go beyond issues of efficiency and to explore its potential to enhance assessment. The use of CAT is a good example of this. CAT approaches allow tests to be adapted to the performance of examinees. This has some important implications. After initial questions, examinees are presented with items that are selected based on their performance, and fewer items that are too easy or too difficult for the examinees are presented. To put it another way, this means that items that do not provide much information about the proficiency of an examinee in a given test are not presented. The ways in which this may be done varies with CAT approaches, and this will be discussed later in this chapter. However a common issue for both CAT and CBT is the provision of item pools that the items may be selected from.

In conventional testing, the dilemma for test designers is how to provide a sufficiently precise measurement of examinee performance across the range of examinee ability in a given group of examinees without requiring the presentation of a potentially overwhelming number of items. Two main approaches may be adopted, the peaked conventional test and the rectangular conventional test. These can be seen in Figure 3-5.

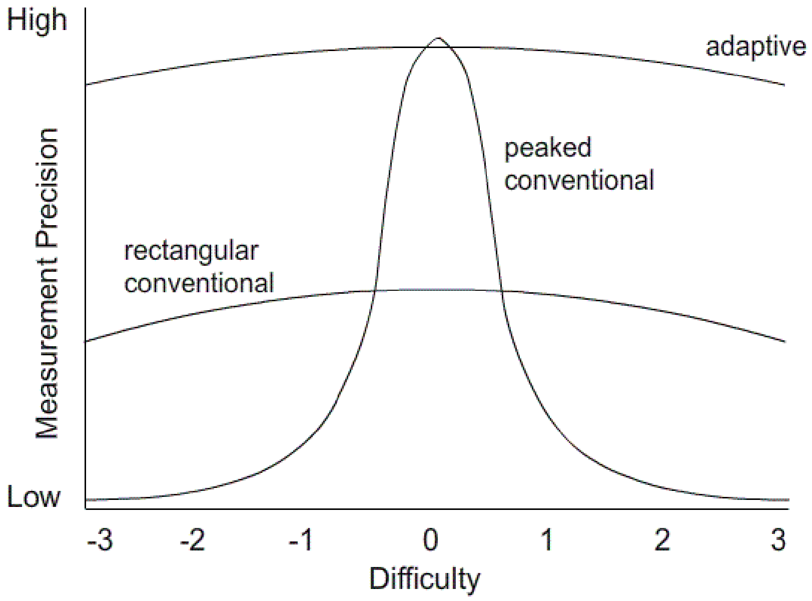


FIGURE 3-5: MEASUREMENT PRECISION FOR CONVENTIONAL TESTING AND ADAPTIVE TESTING (BASED ON WEISS, 1985, P775)

A peaked conventional test has a relatively large number of items within the average proficiency range. This provides high fidelity (or measurement precision) within the average range of proficiency (Flaugher, 1990). The rationale for this approach is that the proficiency of most examinees lies in the average range (Wainer et al., 1990) and that by populating the average range with most items, the test provides precise measurement for the majority of examinees. However, for those examinees outside the average range, both the accuracy and reliability with which their proficiency can be estimated is impaired. (Weiss (1985), Betz & Weiss (1975), Weiss (2004)). Additional implications of insufficient item coverage may be seen in issues associated with the affective impact of items that are outside the range of a given examinee's proficiency and their response to them. When confronted with items that are too difficult for them examinees may become disheartened. On the other hand, examinees confronted with items that are too easy for them may become disengaged (Weiss (1974), Carlson (1994), Wainer (2000)).

To address this issue, a rectangular conventional test provides wide bandwidth across the range of difficulty. This provides greater precision than the peaked conventional test outside the peak, and equal precision across the difficulty range, but the overall level of precision at any given point is relatively low. As previously noted, the dilemma here is that given a fixed length conventional test, it is not feasible to provide a test that provides both wide bandwidth and high fidelity since it would require many test items to be presented to examinees. It has been suggested that this would impair the measurement provided by the test, for example through examinee fatigue (Wise & Kingsbury, 2000).

Providing a measure that has a sufficiently high fidelity and bandwidth of measurement is critical because it has a direct bearing on the reliability of a given test; precise measurements reduce the amount of error in the measurement (Betz & Weiss, 1975).

This is not a binary choice however, and approaches adopting elements of both the peaked and rectangular conventional tests are used. Indeed, it can be seen in the adaptive item coverage in Figure 3-5 and also by inference from the point made above that a way of achieving measurement precision across the full range of proficiency is to provide sufficient item coverage across all areas of the proficiency range, but to only present items that are within a given examinee's proficiency.

One of the advantages of CAT over conventional CBT is that it can do this whilst balancing bandwidth and fidelity effectively. Additionally, it can potentially do so with shorter tests (Weiss (1985), Weiss & Kingsbury (1984)). Items that do not provide much information about an examinee's level of proficiency are not administered to that examinee; as a result, fewer items may be used than in a conventional CBT. This is an important consideration given that the task of devising test items is non-trivial;

the size of item pool for CAT is relatively large and the calibration of the items can be particularly resource intensive.

Indeed, even with the presentation of fewer items, the size and nature of the item pool is one of the issues that may have limited the uptake of CAT; however, whilst some forms of CAT require the use of large calibrated item pools, the resource requirements can vary substantially. Also, the potential value of being able to provide tailored MCQ tests in a broad range of real educational contexts that CAT seems to offer was a key motivation for the programme of research reported here. The Flexilevel test approach was selected initially as it seemed to offer the best balance of resource requirements and adaptivity. To provide a context for this decision, the next section considers CBT and variants of CAT, IRT and fixed-branch approaches.

The structure for the discussion of these issues is derived from the structure used by Wise & Kingsbury (2000) for their analyses of practical issues associated with CAT, specifically:

- Item pools
- Starting the test
- Continuing the test
- Stopping the test.

3.3 Computer Based Testing (CBT)

Conventional CBT tests present tests of the same level of difficulty to all examinees, often presenting the same items to all examinees. (Triantafillou et al., 2008). Variants of CBT tests include the random generation of items and also the random selection of items (Thelwall, 1999). Typically the items are peaked in terms of facility around the anticipated average proficiency of a given test population. Tests may be arranged by item type and topic (Bull, Brown & Pendlebury, 1999).

Item pools. The item pool for CBT tests is typically equal to the number of items to be administered, although random item selection from a larger pool of items could increase the pool required substantially.

Starting the test. The first item of the initial topic or test type section is presented to the examinee.

Continuing the test. The next item within a topic or test type section is selected according to the test design.

Items may be ordered by increasing difficulty to support students' confidence in the test, and also to prevent low-performing students from spending too much time on the more difficult questions (Bull & McKenna, 2003), although it has also been suggested that ordering by type of item then sub-topic is the preferred approach (Brown, Bull & Race, 1999). One issue to consider is the response of higher performing examinees; it has been suggested that the presentation of items that are too easy for them leads to frustration and potential attenuation of concentration (Carlson (1994), Wainer (2000)). As such, it may be argued that presenting easy items first may lead the higher performing students to become bored or frustrated, this in turn may affect their performance.

Stopping the test. The stopping conditions for the CBT are typically when all items have been answered by an examinee or the time limit for the test has been reached.

Tests are typically scored in terms of the number of items the examinee answered correctly.

3.4 Item Response Theory–based Computerised Adaptive Testing (IRT-CAT)

Item Response Theory (IRT) represents a way of showing the correspondence between a latent trait and its manifestation (De Ayala, 2009). In IRT a trait is conceptualised as a continuum upon which individuals are placed in terms of their level of the trait.

For the purposes of this work, the trait of interest is the proficiency of an examinee within a given topic domain and items are located on this continuum in terms of their difficulty. (This is a minimum – other parameters are introduced below that allow IRT to make use of more of the information that is available about the items.) The latent trait (proficiency), whilst not directly observable, may be estimated by measuring the behaviour (response to items) of the examinee.

In IRT, the main models that are used may be differentiated in terms of the parameters they allow to be varied. Specifically:

> **Difficulty**: the location of an item on the trait continuum;
>
> **Discrimination**: this parameter describes how well an item can discriminate between examinees who have an estimated proficiency above, and those who have an estimated proficiency below, the difficulty of the item;
>
> **Pseudochance**: this parameter sets the value that an examinee answers an item correctly by chance.

The 1 Parameter Logisitic (PL) model uses difficulty, the 2PL model uses both difficulty and discrimination, and the 3PL model uses difficulty, discrimination and pseudochance parameters.

Item pool: IRT-CAT requires the use of a large calibrated item pool, for example 3 to 4n where *n* is the number of questions to be administered during the test (Carlson, 1994), or 8 to 12n (De Ayala, 2009).

Additionally, to calibrate this pool large numbers of examinees are required to provide sufficient data for statistical analysis of the items (Lord, 1970), with each item requiring 200–1000 examinees responding to it depending on the approach used (Carlson, 1994). This is an issue that is exacerbated by the use of more parameters in the IRT model.

Additionally two assumptions of IRT concerning items are that they are unidimensional and also independent of other items within the pool. As such, an item only tests one trait (for the purposes of this research, proficiency in a given domain) and items do not provide information to examinees about other items. In other words, one item does not provide a clue to the answer to another item. These assumptions of the model place further constraints on the designers of items.

Starting the test: Commonly with IRT-CAT, the test starts with the presentation of an item of medium difficulty. A potential problem with this is that all examinees could be presented with the same initial item thereby increasing the exposure of that item. This has implications for the security of the item - if many examinees have seen the item, then the chance that knowledge of that item being disseminated through a given examinee cohort may be relatively high (Thissen & Mislevy, 2000). Item exposure applies also to CBT, and more generally across the item pool, for example when all items are presented to examinees.

However, in the case of IRT-CAT, the alternative option of using existing information about the proficiency of the examinees may also be available. An initial proficiency estimate can be made using information held about the examinee's proficiency in previous relevant testing. This would potentially provide a more efficient starting point, but care needs to be taken to ensure that this does not disadvantage examinees (Thissen & Mislevy, 2000). Another option would be to assume a range of proficiency and select items that fell within that range.

Continuing the test: In an IRT-CAT, items are selected on the basis that they are the most likely to improve the estimation of an examinee's proficiency. Initially, in the early stage of the test, this estimation may look like a fixed-branch approach in which easier or harder items are administered depending on examinee response. However, more accurate measures of proficiency are often found after the initial items have been presented and after this point the CAT is improving the precision of the estimate rather than jumping between points on the continuum.

The probability that a given item will be answered correctly increases with the level of proficiency. This is represented in an Item Response Function (IRF) (or Item Characteristic Curve (ICC)), which uses the item's parameters in its calculation. The IRF can be shown graphically as in the example shown in Figure 3-6.

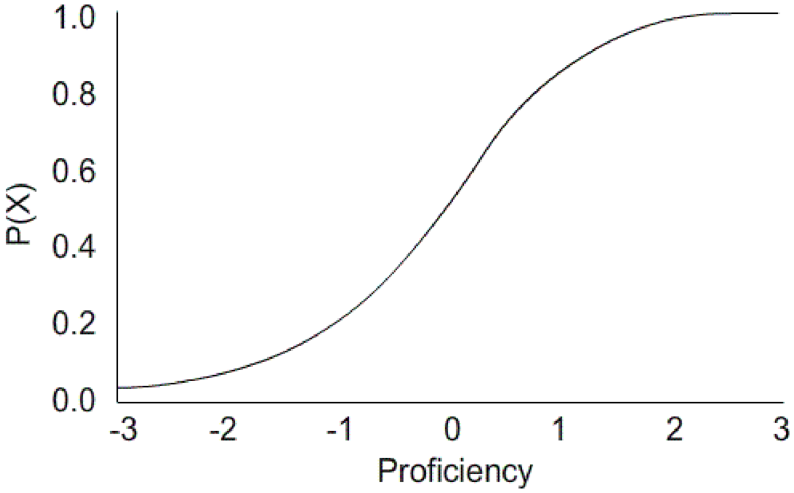


FIGURE 3-6: AN EXAMPLE IRF (FROM THOMPSON, 2009)

Each item has a complementary curve showing the probability that a given item will be answered incorrectly across the proficiency continuum. If an examinee answers correctly, they effectively are awarded the IRF and if they answer incorrectly they are given the complementary IRF (as shown in Figure 3-7).

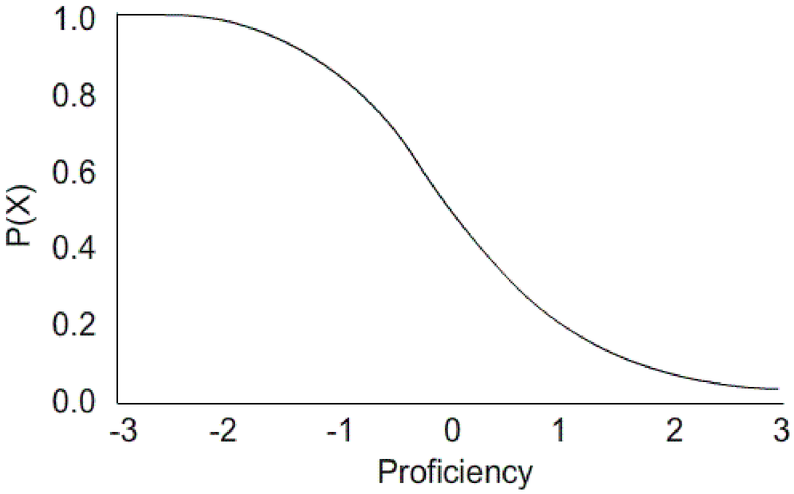


FIGURE 3-7: THE COMPLEMENT OF THE IRF SHOWN IN FIGURE 3-6 (FROM THOMPSON, 2009)

The IRFs for the items that an examinee has encountered can be multiplied together to produce a likelihood function (see Figure 3-8)

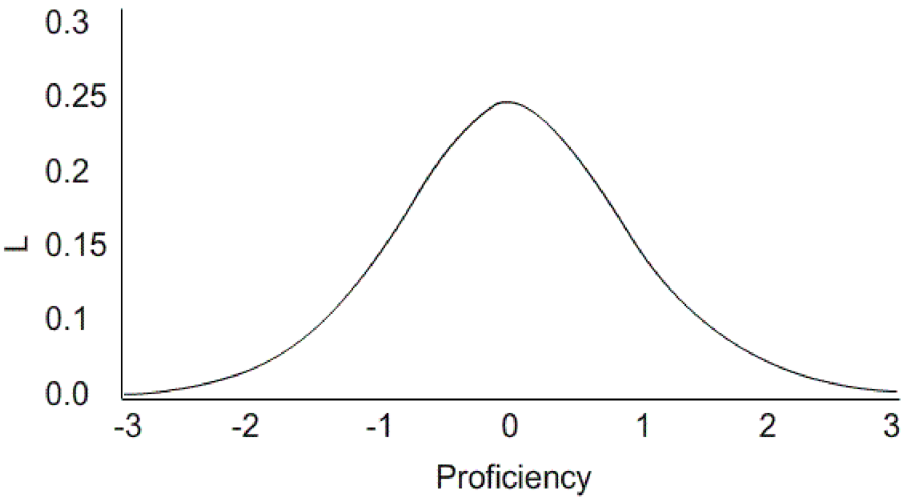


FIGURE 3-8: AN EXAMPLE LIKELIHOOD FUNCTION (BASED ON THOMPSON, 2009)

This gives the likelihood of the pattern of responses observed occurring at each level of proficiency across the continuum. The peak of the curve represents the maximum likelihood estimate and provides the estimate of proficiency for an examinee[1].

The precision of the estimate can be quantified using the Standard Error of Measurement (SEM) and, as previously noted, a target precision can be used as a stopping condition. As such, the exam designer can set a level at which they are confident in the precision of the estimate. Until this (or some other) stopping condition is met, the IRT-CAT will select items that improve the precision of the estimate, usually on the basis of the amount of information they provide given an estimated proficiency.

It is worth noting here that there is an inverse relationship between standard error and information. Each item provides some information about the proficiency of an examinee (the Item Information Function, or IIF). The more information they provide, the lower the standard error they will have. As such, an item that has high discrimination around a given level of proficiency would provide a lot of information

[1] It should be noted however that if an examinee has answered all items presented to them correctly or all items presented to them incorrectly, then the maximum likelihood estimate will be positively or negatively infinite. There are techniques to address this issue, but they are beyond the scope of this dissertation. It may also be argued that this would not be expected to persist for many items given that the item pool is assumed to contain a large number of high quality items across the range of difficulty.

about the proficiency of examinees at that level and would be more likely to be selected next.

As well as providing a basis for item selection, IIFs are important in that they can be used to provide information about a test. Given the assumption that items are locally independent, they are considered to be additive so their information functions can be summed to generate a Test Information Function (TIF). This shows where the test is providing most information, but also when inverted where the standard error is highest.

There are two implications of this. Firstly, exam designers can specify a target TIF (De Ayala, 2009) so that the exam design provides the most information at the point(s) required. As an example, a selection exam may require a TIF that provides a lot of information around the cut point for that selection (or to put it another way, the location on the continuum at which the TIF curve would be peaked). Secondly, to be able to design an exam in this way, such that the observed TIF is comparable to the target TIF, a minimal requirement is an item pool that contains sufficient high quality items (De Ayala, 2009).

Stopping the test

IRT-CAT is typically used as a variable length test allowing examinees to finish after having answered different numbers of items. In these circumstances, the stopping condition may be determined by the level of confidence of the proficiency estimate of a given examinee. So when a certain level of confidence is reached, the test is terminated.

Another stopping condition is the application of a time limit to the test, thereby providing a specified duration as a stopping condition. This may be combined with other stopping conditions, so a test may be stopped either when all items have been presented or the specified duration has been reached. These stopping conditions may be used in IRT-CAT, but could limit the approach’s capacity to shorten tests. Such conditions also limit the examiners’ ability to specify the level of confidence in proficiency estimates and, as noted, this has an impact on the reliability of the test (De Ayala, 2009).

However, as well as the practical difficulties of allowing tests without time limits, it should be noted that another important factor in determining stopping conditions is the attitude of students. Lilley et al. (2004) identified concerns that students held about the fairness of variable length tests, particularly in summative contexts.

As noted, IRT-CAT is perhaps the predominant approach to CAT, particularly in large scale CAT, but alternative approaches in the form of fixed-branch approaches are also in use contemporary educational contexts. The approaches considered next are the two-stage test, multi-stage test, Stradaptive test and the Flexilevel test.

3.5 Two stage tests

The two stage test (Betz & Weiss (1973), Larkin & Weiss (1975)) represented an early attempt to adapt tests and took the approach of providing two or more stages that route examinees through the test.

The routing test can be made up of items that peak the test around an average ability for the group taking the test, as shown in Figure 3-9. However, the range of facility in the routing test can also be broader (Weiss, 1974) where the average level of ability is not known. For example in the case of a wider test population that takes in a broader range of abilities. The testlets that appear in the second stage are peaked at levels of difficulty across the range of difficulty of the test.

The adaptation occurs as the route examinees are given through the test varies with their performance.

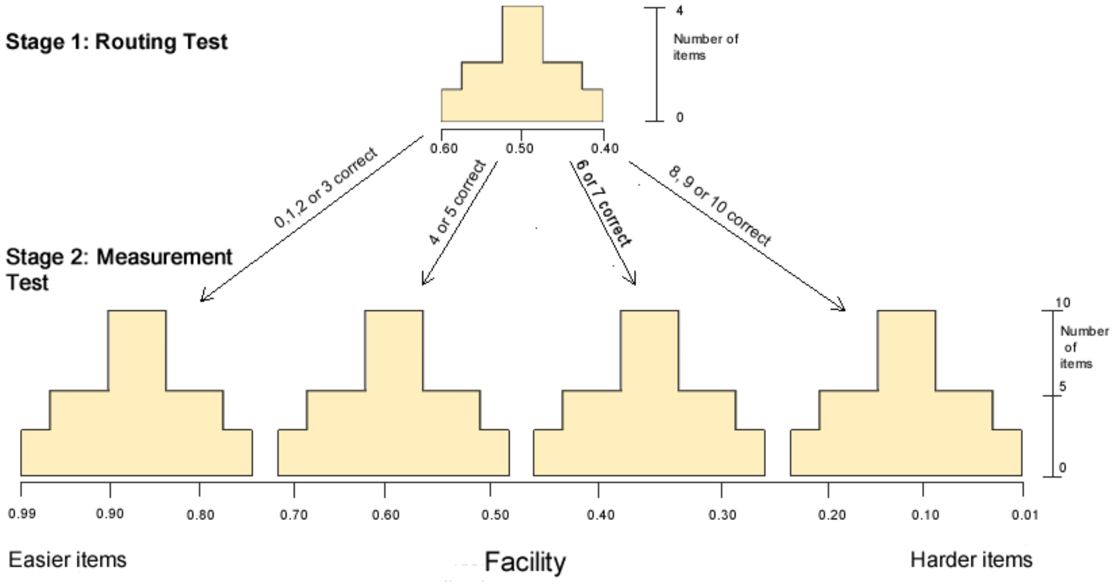


FIGURE 3-9: THE TWO-STAGE TEST (ADAPTED FROM (WEISS 1974, P.4))

As shown in the example in

Figure 3-9, a routing test that is peaked at the average level of facility for a given group of examinees is given first. Depending on their performance in the routing stage, examinees are branched to one of four measuring stages. These are peaked around different points in the facility scale in order to provide greater fidelity of measurement at different levels of facility.

Examinees performing well would be expected to be routed to one of the measurement tests that present items with a facility of around the 0.35 or 0.1 marks. Examinees who are not performing so well would be expected to be routed to one of the measurement tests that present items with a facility of around 0.6 or 0.85.

The format of two stage tests varies. This may in part be due to the difficulty of identifying optimum approaches (Lord, 1971b).

Item Pools. As compared with other objective testing approaches relatively large numbers of items are needed in the item pool. In the example above (Weiss, 1974), the routing test contains ten items and the measurement tests 30 each: an item pool of 130 items for a 40-item test.

Starting the test. The test begins with an initial routing test stage in which examinees take a test of average difficulty that is used to estimate their level of proficiency and place them in the correct testlet in the second, measuring stage. For multi-stage tests, there may be more than one routing stage and more than one measuring stage.

Continuing the test. The test continues with the second stage in which testlets aimed at low, medium and high performing examinees are presented. In this sense, examinees are presented with tests that are tailored to their level of proficiency.

Stopping the test. The test is stopped once examinees have answered all items within the measuring test stage or stages.

Scoring methods may vary, but one possible scoring method is to report the average difficulty of the items answered correctly by the examinee in the measuring test stage (Weiss, 1974).

3.6 Multi-stage tests

There is a risk that if there are any issues with the accuracy of the initial routing test, the accuracy of the estimate of proficiency may be reduced and this error may be propagated to the measuring test stage if an examinee is routed to the wrong testlet within the measuring test stage (Sands et al., 1997). To offset the potential implications of issues with the routing stage, more stages have been added to tests, giving multi-stage tests.

An example (from Armstrong et al., 2004, p.150) is shown in Figure 3-10. The test comprises six stages and 14 test bins, each showing the range of proficiencies they are targeting.

Stage					
1	**2**	**3**	**4**	**5**	**6**
					14:[87–100]
			7:[75–100]	10:[75–100]	
		4:[50–100]			13:[50–87]
1: [0–100]	2:[0–100]		6:[25–75]	9:[25-75]	

		3:[0–50]			12:[13–50]
			5:[0–25]	8:[0–25]	
					11:[0–13]

FIGURE 3-10: AN EXAMPLE OF A MULTI-STAGE TEST (FROM ARMSTRONG ET AL. (2004) P.150

Rather than having one routing test, other stages are added that further refine the routing of examinees through a given test. It is possible to apply rules to the test such that routing of examinees is restricted between stages, so no adaptation is made between those stages. This can be seen in the progression from Stage 4 to Stage 5 in Figure 3-10, whereby those in bins 5, 6 and 7 will remain within the top 25%, middle 50% and bottom 25% going into stage 5. There would not be any further differentiation between the examinees' performance and the groups as a whole would be routed to the next stage intact. This may preserve one of the benefits of fixed-branch approaches, that test experts can see the possible routes through a test and verify that they are appropriate. This becomes unfeasible when much adaptation is permitted in a test because of the increased numbers of possible forms the test may take.

Starting the test. Examinees are presented with one or more routing test(s), and based on their performance in this test are routed to measurement test bins. The number of bins will vary depending on the test design.

Continuing the test. Examinees respond to the items in the test bin and based on their score are routed to the next available bin.

Stopping the test. Once there are no further items to answer the test is stopped. Scoring methods may vary, but one possible scoring method is to use maximum likelihood estimates.

3.7 Pyramidal Test

Tests may be created in which routing is done on an item-by-item basis as in the Pyramidal Test.

Figure 3-11 shows a pyramidal test with four stages. A high performing examinee would be expected to take a route down the right side of the pyramid whereas a low performing student might be expected to follow the left side of the pyramid. For example, answering all items correctly would, in this example lead to the following sequence: Items 1, 3, 6, 10. Answering all items incorrectly would lead to the following sequence: Items 1, 2, 4, 7. An examinee performing averagely would be routed left and right down the test. In the example test, examinees would be presented with four items from a pool of 10.

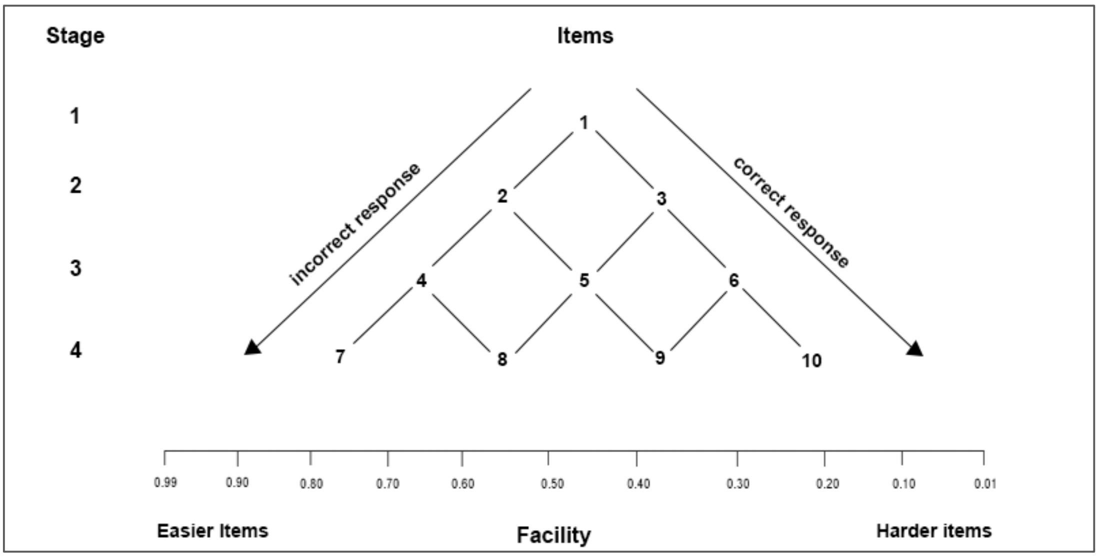


FIGURE 3-11: A DIAGRAM SHOWING THE PYRAMIDAL APPROACH (ADAPTED FROM LARKING AND WEISS (1975))

Item Pools: Item pools for pyramidal tests are relatively large as sufficient numbers of items to cover the range of difficulty in the test across the different stages are required. Longer tests clearly place considerable demands on the size of the item pool being used. Also, variants of the pyramidal approach that use more than one item per stage would require a still larger item pool.

Starting the test: The test starts with the presentation of the item of medium difficulty.

Continuing the test: Given an examinee’s response to the initial item, they are either branched right following a correct answer to a more difficult item or left following an incorrect answer to an easier item. Examinees are routed in this way down the pyramid.

Stopping the test: Pyramidal tests stop when the examinee has answered an item at each stage of the test.

There is a range of methods for scoring pyramidal tests, for example providing an average of the facility of all items answered correctly.

3.8 Stratified Adaptive (Stradaptive) Test

The Stradaptive test (Weiss, 1973) is composed of different strata at different levels of difficulty. Within those strata, individual items are presented to examinees depending on their proficiency. One of the outcomes of a Stradaptive test is to identify the basal and ceiling strata for an examinee. The basal stratum is the most difficult stratum in which examinees answer all or nearly all items correctly; the ceiling stratum is the stratum at which examinees answer items at or below the level of chance.

Figure 3-12 shows an example Stradaptive test with nine strata. It can be seen that each stratum is made up of a peaked testlet. The idea is to find the peaked test that is most appropriate for the proficiency of a given examinee.

As an example, an examinee is presented with an item in a stratum that corresponds to their estimated proficiency. For example, an examinee may start the test in the fifth stratum. They are presented with an item in the fifth stratum, answer incorrectly and so are presented with an item from the fourth stratum. If they answer this correctly, they are routed back to the fifth stratum. If they answer incorrectly, they are routed to stratum three. The test continues until the point at which they are answering at chance in a given stratum. So, assuming items have four options and each stratum has ten items, if the examinee in the example is returned to stratum five and answers only two items out of the ten correctly, then the test terminates.

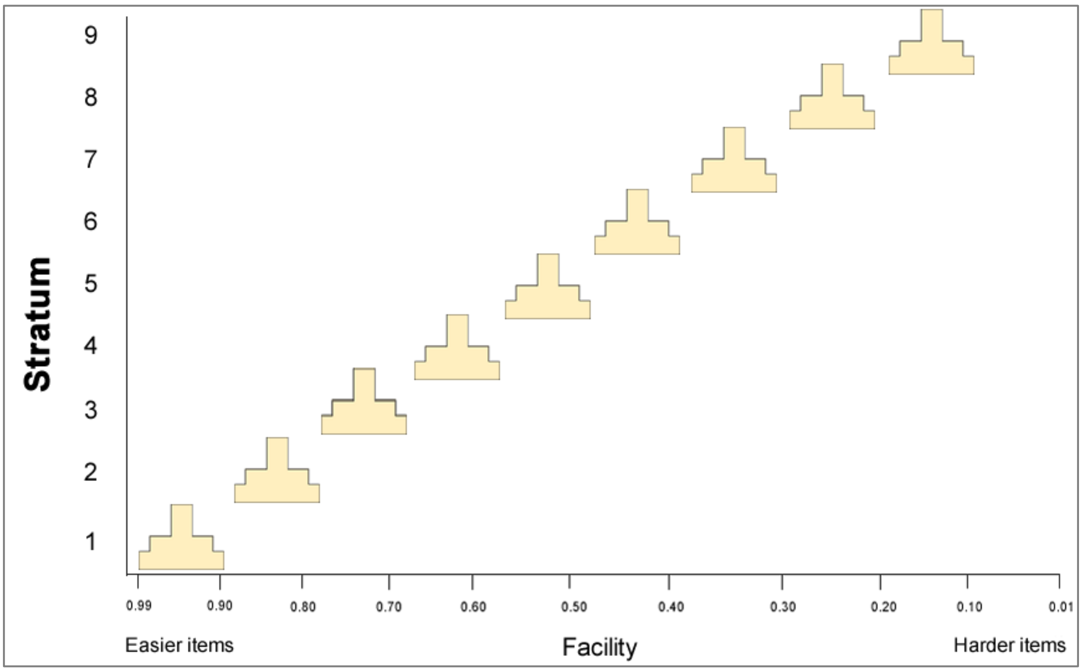


FIGURE 3-12: AN EXAMPLE STRADAPTIVE TEST (BASED ON WATERS (1977))

Item Pools. The item pool required by the Stradaptive test is relatively large and requires calibratiwon with a large number of representative participants. They are organised into strata ordered in terms of difficulty with each stratum containing a peaked test that is peaked around the difficulty of the strata (see Figure 3-12). There should be no overlap in terms of difficulty between the strata.

Starting the test. The test begins with the presentation of an item that is pitched at an examinee's estimated proficiency. The estimate is based on prior knowledge of the examinee's proficiency that may be derived from performance in related assessments or from entry questions asked prior to the test. In the situation where there is no prior estimate of proficiency, the test may be started with an item of median difficulty.

Continuing the test. There are various approaches to routing the examinee through the test, but the classical approach is as follows: If an examinee answers an item at a given stratum correctly, then they are presented with the next available item from a more difficult stratum. If they answer the item incorrectly, then they are presented with the next available item in a less difficult stratum.

Stopping the test. The test stops when an examinee answers items at the level of chance in a given stratum. This provides the ceiling stratum of their ability, and this may be identified at different numbers of items for different examinees. As such, the Stradaptive test has a variable stopping condition.

In cases where examinees are performing extremely well or extremely poorly and there are no harder (in the case of examinees performing extremely well) or easier (in the case of examinees performing extremely poorly) strata available, then they are presented with items from the stratum they are currently on. It is possible that the test will end without an accurate estimate of proficiency being found for them.

There are a variety of ways to score Stradaptive tests, for example averaging the facility of all items answered correctly or averaging the facility of all items answered between the examinee's basal and ceiling strata. Maximum likelihood estimates can also be used.

3.9 The Flexilevel test

The Flexilevel test (Lord (1970), Lord (1971a)) provides a basis for a "minimalist adaptive test" (Thissen & Mislevy, 2000, p.102). There has recently been renewed interest in the potential of the approach to support flexible learning (Gordon, 2014). Figure 3-13 shows the item selection algorithm for the Flexilevel test. It uses a fixed-branching algorithm that presents items based on whether or not an examinee has answered the previous item correctly. If the examinee has answered the item correctly, they are routed to the next more difficult available item; if the examinee has answered the item incorrectly they are routed to the next available easier item. This continues until there are no further items available.

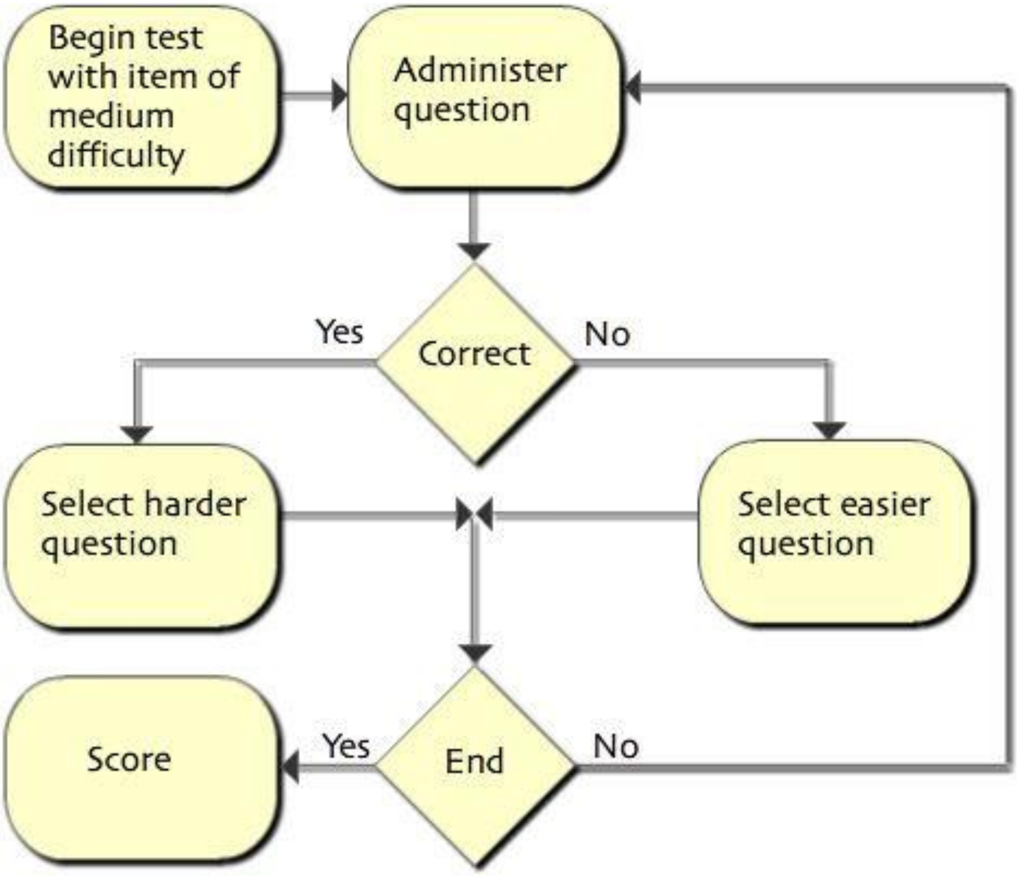


FIGURE 3-13: THE FLEXILEVEL ALGORITHM

Figure 3-14 shows an item pool for an example test of five items derived from Lord (1980). The difficulty of items in the item pool ranges from 0.15 (easiest item) to 0.85 (hardest item).

[I5, difficulty = 0.5]	
Red 1 [I4, difficulty =0.4]	Blue 1 [I6, difficulty =0.6]
Red 2 [I3, difficulty =0.3]	Blue 2 [I7, difficulty =0.7]
Red 3 [I2, difficulty =0.2]	Blue 3 [I8, difficulty =0.8]
Red 4 [I1, difficulty =0.15]	Blue 4 [I9, difficulty =0.85]

FIGURE 3-14: THE RANKING OF ITEMS FOR THE FLEXILEVEL TEST: EASIER ITEMS IN RED, MORE DIFFICULT ITEMS IN BLUE ADAPTED FROM LORD (1980).

In this hypothetical test (Lord, 1980), the examinee would start the Flexilevel test by answering I5 (i.e. item of medium difficulty). Each time a correct answer is given, the item to be answered next is the lowest numbered "blue" item not previously answered. Each time a wrong answer is given, the item to be answered next is the lowest numbered "red" item not previously answered.

A student who answers all items correctly will answer the following sequence of items: I5, I6, I7, I8 and I9. A student who provides wrong answers for all items will answer the following sequence of items: I5, I4, I3, I2 and I1. A student who answers the first item incorrectly, and all the following items correctly will answer the following sequence of items: I5, I4, I6, I7 and I8.

A somewhat more abstract example is shown in Figure 3-15, the test beginning with item 10 (medium difficulty).

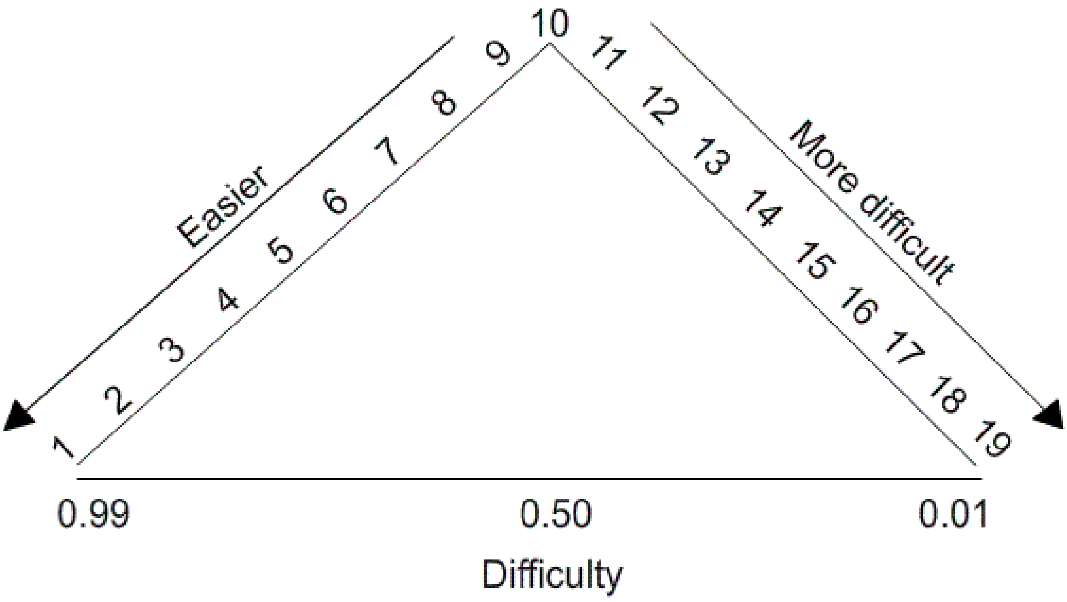


FIGURE 3-15: THE FLEXILEVEL TEST STRUCTURE (ADAPTED FROM BETZ AND WEISS, 1975)

If the examinee answers the item correctly, they are presented with the next available difficult item, item 11. If the examinee answers the item incorrectly, they are presented with the next available easier item, item 9.

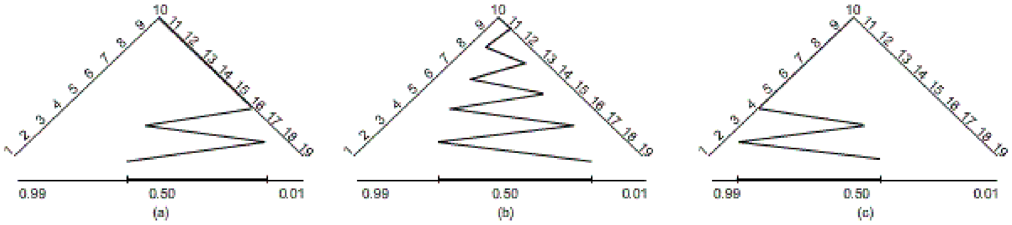


FIGURE 3-16: EXAMPLE PATTERNS FOR DIFFERENT PROFICIENCIES (BASED ON BETZ AND WEISS, 1975)

Figure 3-16 shows patterns of performance for examinees of different levels of proficiency. It is evident that the Flexilevel will present items that span an examinee's level of proficiency (in this case, the proficiencies are (a) high, (b) medium and (c) low). Scoring is straightforward with correct responses receiving a mark (Lord, 1970).

Item pools: The Flexilevel approach requires a database of 2n-1 items, where n is the number of items to be presented in a test. Calibration is also less troublesome than other forms of CAT. Indeed Lord (1980, p.117) suggests that "any rough approximation" of the difficulty of the questions will be adequate.

Starting the test: The item of medium difficulty is presented first. If item exposure is a concern, it is possible to vary the starting item, for example by selecting randomly from a set of items around median difficulty.

Continuing the test: With the Flexilevel test, the test is continued through the selection of harder or easier items depending on the response of the examinee as the example below depicts.

Stopping the test: Consistently with other fixed-branching techniques, the Flexilevel test is a fixed-length test that is concluded when a specified number of items, formally ½(N+1) where N is the number of test items, have been answered.

This may be combined with other stopping conditions, so a test may be stopped either when all items have been presented or the specified duration has been reached.

3.10 Rationale for selection of Flexilevel test

Understanding the proficiency of a given student is a key element of personalising the learning experience for that individual, something that could be valuable in the context of a rich and potentially complex learning ecosystem, for example through the reduction of transactional distance. An area that has shown real potential in achieving this is CAT.

This chapter has discussed how in CAT items are presented to the examinee based on their performance in a given test. The detail of how they do this may differ, but broadly examinees are presented with items that have a difficulty that is matched to their performance. By contrast, in a conventional CBT, examinees in a given test are presented with the same set of items, or a test of equivalent difficulty when items are selected randomly from a larger item pool, or generated automatically. The focus is on the presentation of tests to cohorts of examinees rather than items to individuals.

It has been suggested that an important part of the educational benefit of tailored tests is that individual examinees are less likely to be presented with items that are too difficult for them, nor too easy. When faced with items that are too easy for them, examinees may lose concentration or become frustrated. When items are presented that are too difficult for them, they may become unduly stressed or demoralised, (Weiss (1974), Carlson (1994), Wainer (2000)).

Assessment using the CAT approach has been shown not to disadvantage students when compared with conventional CBT approaches (for example see work conducted with IRT-CAT by Lilley & Barker 2004). It also provides potential procedural advantages over the conventional CBT approach (Weiss (1985), Carlson (1994)), including the presentation of fewer items to examinees.

It seems that IRT-CAT would be an ideal candidate for this work, but there are practical issues around it being embedded in real educational contexts. As an indication of the main issues, it seems that the practical uptake of the paradigm is limited to large-scale testing endeavours in which there are significant resources available for the design, development and maintenance of the components required for an IRT-CAT.

The issues at hand are largely practical (see for example Lord's list of possible requirements (Lord, 1970)). In the context of ubiquitous and embedded assessment where assessments are taken formatively on an ad hoc basis as and when students are engaged in their learning activities, some of the requirements for most forms of CAT would be too onerous. These requirements have been discussed above, and relate principally to the item pool. For IRT-CAT the calibration of the item pool and the numbers of items required in a test are prohibitively high. Bull & McKenna (2003) point out that this is one of the issues with IRT-CAT, but as has been seen, other CAT techniques also suffer from this issue.

Fixed-branch approaches have a long heritage - going back to Binet's IQ test in 1905 – in which the route examinees take through a test is determined by their performance on that test. Variants of fixed-branching approaches for CAT remain part of the assessment landscape. Indeed, it seems that the use of multi-form tests is increasingly common in contemporary assessment (for example, (Armstrong et al., 2004), Edwards & Thissen (2007)), and this seems to indicate the potential for branching algorithm-based tests to offer a viable approach even in a computationally rich learning ecosystem.

Whilst the computational capacity for relatively complex IRT-based forms of Computer Adaptive Testing is available, there remains much scope for branched alternatives to the complexities of IRT-CAT (Edwards & Thissen, 2007). One of the advantages of fixed-branch approaches is that test designers can see the various routes through a test and can ensure that such sequences are sound; this is not feasible in IRT-CAT.

Whilst perhaps less onerous than IRT-CAT the resource implications for multi-stage testing are not insignificant, in particular when considering using individual items instead of testlets or bins at the different stages, as in the pyramidal approach.

The requirements of potentially very large item pools in fixed branching approaches also means they would not be feasible for this for this work. Further, similarly to the IRT the scoring can be opaque to examinees.

Lastly the Flexilevel test (Lord (1970), Lord (1971a)) was considered. Whilst it shares a similar approach to routing examinees through a test, it is less resource-intensive than either IRT-CAT or the other fixed-branching approaches discussed here.

When compared to conventional testing, it was found to be more effective when testing over a wide range of examinee proficiency (Lord, 1970). Further, as Kocher (1974) observes, the Flexilevel test shows the capacity to match item difficulty to examinee proficiency. Consistent with this, De Ayala & Koch (1986) and De Ayala, et al. (1990) have shown that Flexilevel proficiency estimates are comparable to those obtained using IRT-based tests. As previously noted, the firm potential for the use of the Flexilevel test in supporting flexible learning has recently been reinforced in a Higher Education Academy (HEA) commissioned study by Gordon (2014).

In summary, the Flexilevel test seems to offer comparable efficacy for conventional testing (and also to other forms of CAT), but with lighter demands on resources, it presents an ideal candidate for this programme of research.

The next stage of the work is to establish the effectiveness of the Flexilevel test in a contemporary higher education setting. This is the focus of the next chapter.

4 Studies investigating the effectiveness of the Flexilevel test

This chapter presents the studies that were undertaken as part of this research that focused on the effectiveness of the Flexilevel approach in a Higher Education context. These include one simulation study and ten empirical studies involving examinees in real educational contexts.

The motivation for carrying out the studies was that whilst there is a basis in the literature for the effectiveness of the Flexilevel approach (see, for example, Lord (1980), De Ayala & Koch (1986), De Ayala et al. (1990)), it was necessary to establish the extent to which this applied in contemporary Higher Education assessment contexts, whether formative or summative.

Research into the effectiveness of approaches to e-assessment typically involves the comparison of scores from different assessment formats (see, for example, Clariana & Wallace (2002) and Jeong (2014)). McBride (1997) identifies four main sources of data that are particularly relevant for research into adaptive testing approaches such as the Flexilevel test:

1. Theoretical analysis
2. Computer simulation
3. Live testing
4. Real data simulation

Evaluation of educational technology is a complex undertaking, and one of the perennial issues that contributes to this is the extent to which an evaluation may be considered to be authentic (Oliver, 2000). Studies that provide closely controlled experimental conditions may be able to control for variables that may affect an outcome, but the extent to which the results can be extrapolated to real educational contexts may be limited (Oliver et al., 2006). Barker & Barker (2002) amongst others report on the value of evaluating educational software in real educational contexts. A substantial part of the motivation behind this programme of research is to gain an understanding of the extent to which the Flexilevel test can be applied in real educational contexts. In this respect, and in the context of McBride's (1997) sources of data, the studies conducted as part of this work focused predominantly on live testing, with also one study on real data simulation. Theoretical analysis and computer simulation were considered to be beyond the scope of this programme of research, given the importance outlined earlier of evaluating the Flexilevel test in real educational environments.

Live testing involves the administration of conventional tests to examinees and the processing of the data that this generates, particularly the calculation of some form of score. This can then be analysed to compare examinee performance in the different forms of testing. Live testing is important but also resource intensive, whatever form it takes. In practical terms, there are procedural and academic quality concerns – writing items and administering the test itself is a non-trivial task.

The live testing in real educational contexts conducted as part of this programme of research was organised into two phases. The first phase of live testing included summative assessments where the number of items in the test was the same for the Flexilevel and conventional CBT tests. The second phase of live testing included diagnostic, formative and summative assessments; this phase focused on investigating the extent to which shorter Flexilevel tests may be comparable to longer conventional CBT tests.

Given that a principal motivation for the work was to apply the Flexilevel approach in real educational contexts, live testing was the predominant methodology for understanding the educational implications of the work. Other key considerations are discussed later in this dissertation, including the attitudes of students and staff to the approach.

It should be noted that in addition to the live studies in real educational contexts, two exploratory pilot studies were conducted. As part of these two exploratory studies, examinees were presented with an assessment that comprised a Flexilevel test stage and a conventional CBT stage under supervised conditions following the same guidelines typically used for in-class summative assessments. As well as providing the opportunity to establish the methodology of the test, this methodology allowed the testing of the effectiveness of the approach before launching the assessments in a real educational context.

The contexts of use in this programme of study spanned formative and summative tests, tests supervised in computer labs, online tests that employed remote live invigilation and unsupervised formative tests.

This programme of research also made use of real data simulation. There is much value in using real data that has already been generated in live testing: the overhead of creating new items and tests is obviated whilst the data in use is still genuine data in the form of examinees' responses to items in real tests. Given a set of responses to items, this would provide the basis for simulating the selection of items based on whichever approaches are of interest. So real data simulation allows for the generation of a route through a test based on real data. Once these have been generated, it is possible to analyse the results as with live testing – for example in correlating the actual score achieved with the simulated score for a given approach.

As discussed in section 3.9, the Flexilevel approach requires items to be ranked in order of facility. In all datasets used in the studies introduced in this chapter, items (i.e. test questions) were calibrated using historical test data. The formula used to calculate the facility of individual items was adapted from Ward (1981):

$$F = \frac{n_p}{n_r}$$

EQUATION 1: ITEM FACILITY, ADAPTED FROM WARD (1981)

In Equation 1, n_p is the number of students who answered the item correctly, and n_r is the total number of students who answered the item.

In the next section, the two exploratory studies are described.

4.1 Exploratory Studies

The exploratory studies were conducted under controlled conditions. Participants were required to take part in the study in a computer lab, under supervised conditions. Standard exam conditions applied; for example, participants were not permitted to communicate during the test, use mobile devices or use unauthorised materials.

The advantage of conducting studies in controlled conditions is evident from both the perspective of the experimental methodology and the quality assurance associated with the outcome of the assessment. Controlling the environment in the way outlined above provides conditions that are both authentic in terms of the assessment experience of students and rigorous in terms of experimental controls.

4.1.1 Pilot Study

The pilot study was intended to compare participants' performance in a Flexilevel test with their performance in a conventional CBT. The pilot was also intended to test the methodological design and software application for the studies that would follow. Findings from this study were originally reported in Lilley & Pyper (2009).

Participants. Twenty-four Level 6 Computer Science undergraduates participated in the pilot study. They had previously studied the topic that was the subject of the test, but were briefed that the experiment was a test of the software application and not of them. The pilot study was not conducted in the students' usual educational context, hence the need to inform the subjects that we were testing the software application and not them.

Methodology. The participants took the test in a computer laboratory under supervised conditions using a software application developed as part of this research (please see Appendix A). The test was organised into two test stages: Flexilevel and CBT. The order in which the tests were presented to participants (i.e. Flexilevel or CBT presented first) was randomised. The items used for each test had the same range of difficulty. Participants were unaware of the test order. The test consisted of 20 items (ten Flexilevel items, ten CBT items) and had a time limit of 25 minutes.

The time taken for participants to complete the test was monitored by the software application. The application also enabled the recording of the performance of participants in the two test stages and the reporting of their results at the end of the test.

It was important to the experimenters that students engaged with the test and attempted to answer test items to the best of their ability. To encourage this, participants were informed that the participant with the highest overall score would be given a small prize.

Furthermore, once the participants had completed the test they were asked to fill in a questionnaire that had been placed next to their computers. This supported the collection of data associated with participants' attitudes towards the Flexilevel approach. This was a necessary aspect of the pilot study and clearly is important overall. Findings from the questionnaire are discussed in section 5.2.

Two members of the research team were present to observe participants' interaction with the software application, and to provide assistance in using the application where necessary.

Results. A summary of participants' performance is presented in Table 4-1; scores for each of the test stages could range from 0 (minimum) to 10 (maximum). Table 4-1 shows that the range of scores achieved by participants was relatively large for both test stages and that there is little apparent difference between the scores achieved in each of the test stages.

Test Stage	Minimum	Maximum	Mean	Std. Deviation
Flexilevel	2.0	8.5	4.89	1.91
Conventional CBT	1.0	9.0	4.62	2.10

TABLE 4-1: SUMMARY OF PARTICIPANT SCORES FOR EACH SECTION OF THE TEST (N=24)

A Pearson product-moment correlation coefficient was computed to assess the relationship between the Flexilevel test and conventional CBT scores shown in Table 4-1. The results show a significant correlation ($r=0.764$, $N=24$, $p \leq 0.01$). A paired-samples t-test shows no significant difference between the Flexilevel test and the conventional CBT scores, $t(23)=-0.954$, $p=0.350$. The scatterplot chart shown in Figure 4-1 illustrates the results.

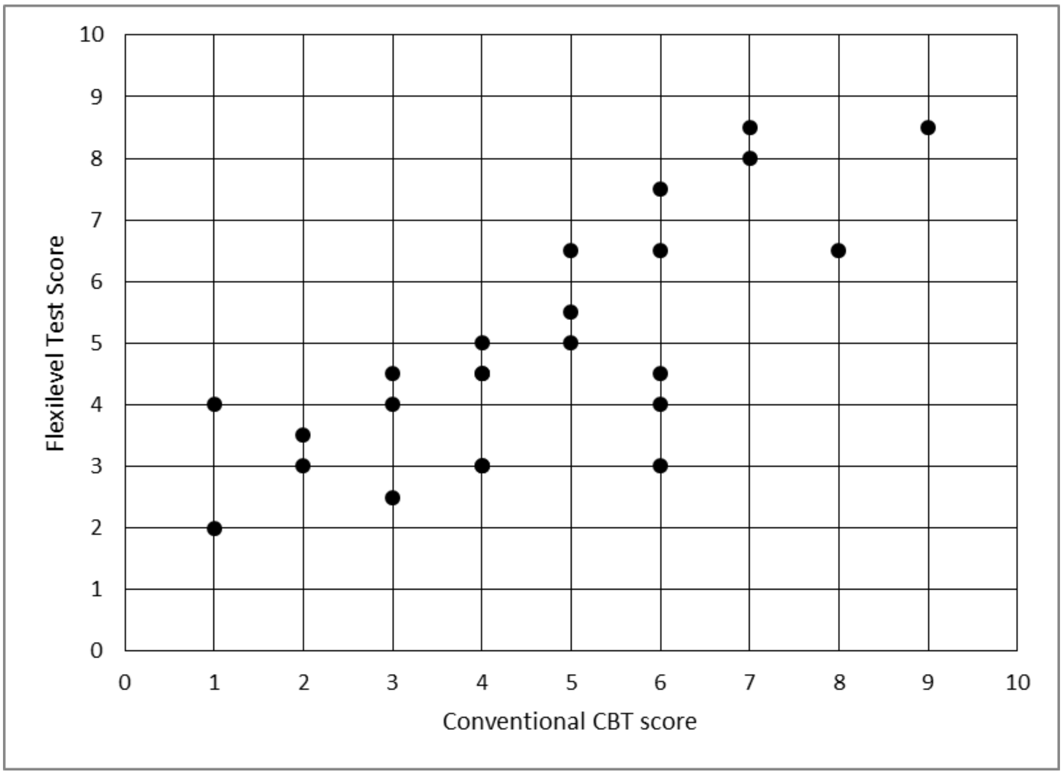


FIGURE 4-1: A SCATTERPLOT CHART SHOWING THE SCORES ACHIEVED BY PARTICIPANTS IN THE TWO TEST CONDITIONS (N=24)

Findings. The results of the pilot study were very encouraging. They showed a significant correlation between participants' performance on the two different types of test (r=0.764, N=24, p≤0.01). This was taken to indicate that participants were not disadvantaged by the Flexilevel approach. Moreover, there was a relatively wide distribution of scores in both tests, with levels of attainment achieved in each being comparable for each student. Therefore it was concluded that that performance of participants is consistent in the two tests across a range of attainment.

Direct observation did not uncover any usability problems. This was an important finding, particularly as participants had not used the application before.

4.1.2 POSTGRADUATE STUDY

It was of interest to the research to conduct a study involving postgraduate participants, given that they are exposed to curricula and educational experiences that may be expected to be qualitatively different from other educational levels SEEC (2010). The work reported here was published in Pyper et al. (2014b).

Participants. Twenty-nine participants took part in the study. They were all studying on an MSc. programme, and were enrolled on a web programming course. Participants had previously studied the topic that was the subject of the test, but were briefed that the test was about the software application and not about their performance.

Methodology. The test was conducted under supervised conditions as close as possible to those of a typical assessment session, both to provide an authentic test environment for the participants and to ensure the experiment was closely controlled.

The study was conducted using the software application designed and developed for this programme of research. The test comprised two stages: one Flexilevel test and one conventional CBT. The order of the presentation of the test stages was randomised and participants were unaware of the order. Each stage of the test consisted of ten objective items and had a time limit of 10 minutes. The time remaining in each stage was displayed to participants, as was their progress in the test. Once a participant had completed the whole test, their results for each stage were displayed.

Participants were informed that the top two scores would win a small prize. This incentive, together with the educational relevance of the test, was intended to control for motivational issues that may have otherwise affected the participants' performance.

Two members of the research team were present to observe participants' interaction with the software application, and to provide technical support in using the application where necessary. It should be noted that this was the participants' first interaction with the software application, and no previous training had been provided.

Results. A summary of the participants' performance can be found in Table 4-2; scores for each of the tests could range from 0 (minimum) to 10 (maximum).

Test Stage	Minimum	Maximum	Mean	Std. Deviation
Flexilevel	2.5	9.0	5.07	1.69
Conventional CBT	1.0	8.0	4.24	1.95

TABLE 4-2: SUMMARY OF PARTICIPANT SCORES FOR EACH SECTION OF THE TEST (N=29)

The Pearson product-moment correlation coefficient for the data presented in Table 4-2 showed that the scores for the Flexilevel and conventional CBT tests were significantly correlated (r=0.614, N=29, p≤0.01). Moreover, a paired t-test showed no significant difference between the scores for the Flexilevel and conventional CBT tests (t=-1.279, df=28, p≤0.212).

Figure 4-2 illustrates the scores obtained by participants in the Flexilevel and CBT stages.

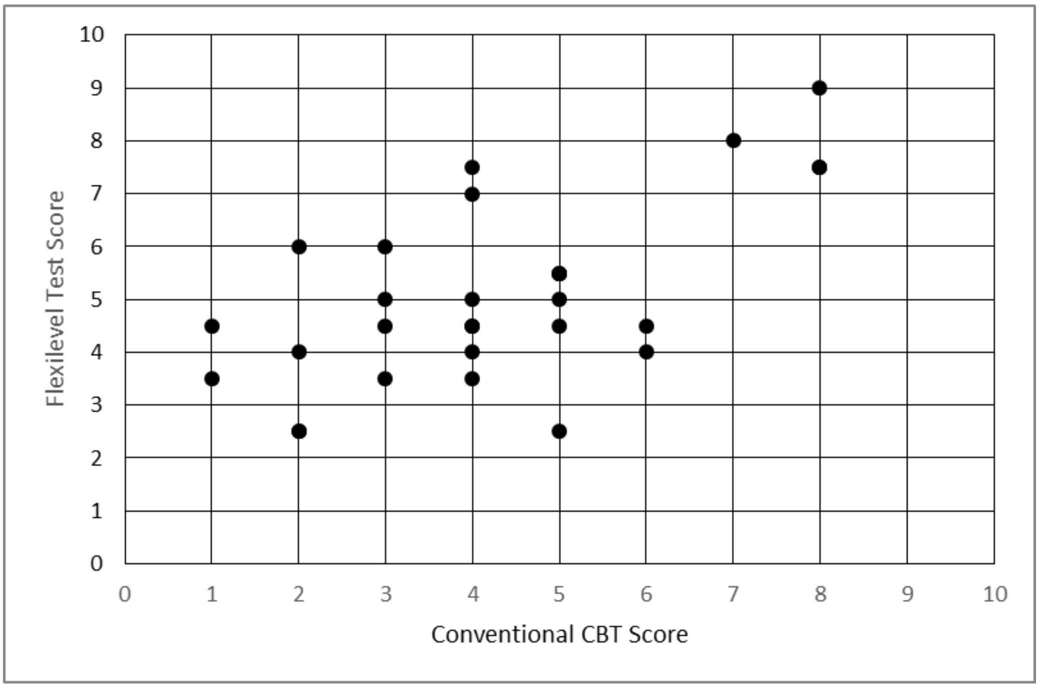


FIGURE 4-2: A SCATTERPLOT CHART TO SHOW THE SCORES ACHIEVED BY PARTICIPANTS IN EACH OF THE TWO TEST CONDITIONS (N=29)

Findings. As with the study reported earlier, there was a significant correlation between participants' scores in the Flexilevel and CBT tests (r=0.614, N=29, p≤0.01).

This was taken to indicate that participants were not disadvantaged by the Flexilevel approach. Furthermore, direct observation did not uncover any usability problems.

4.2 First Phase of Live Testing

The exploratory studies were important in ensuring that there were no usability problems with the software application that could impact on participants' performance in a detrimental way. Additionally, these studies provided a useful insight into the effectiveness of the Flexilevel test when compared with a conventional CBT test.

It was important to follow up on these encouraging findings and investigate the effectiveness of the Flexilevel test in authentic assessment contexts. This is because one of the potential limitations of the exploratory studies was that they were not conducted in genuine educational contexts, and this may to some extent have affected participants' attitudes and motivation towards the test.

4.2.1 Study A

The aim of this study was to investigate the effectiveness of the Flexilevel test when compared to a conventional CBT test in a summative assessment context.

Participants. A total of 131 first-year Computer Science undergraduates took the test as part of the assessment regime of the programming module they were studying.

Methodology. The supervised conditions implemented in the exploratory studies were repeated here. The test comprised two stages: one Flexilevel test and one conventional CBT test. The order of the presentation of the test stages was randomised and examinees were unaware of the order. The test consisted of 40 items (20 Flexilevel, 20 CBT) and had a 45 minute time limit. The time remaining in each stage was displayed to examinees, as was their progress in the test. Once an examinee had completed the entire test, their results for each stage were displayed. Additionally, examinees had the opportunity to take a practice test if they wished to do so.

Results. Table 4-3 provides an overview of examinees' performance in the test. Scores for each of the test stages range from 0 (minimum) to 20 (maximum).

Test Stage	Minimum	Maximum	Mean	Std. Deviation
Flexilevel	4.0	16	9.29	2.88
Conventional CBT	0.0	17.0	8.79	3.24

TABLE 4-3: SUMMARY OF EXAMINEES' SCORES FOR EACH SECTION OF THE TEST (N=131)

Figure 4-3 shows the scores achieved by examinees in the Flexilevel and conventional CBT tests.

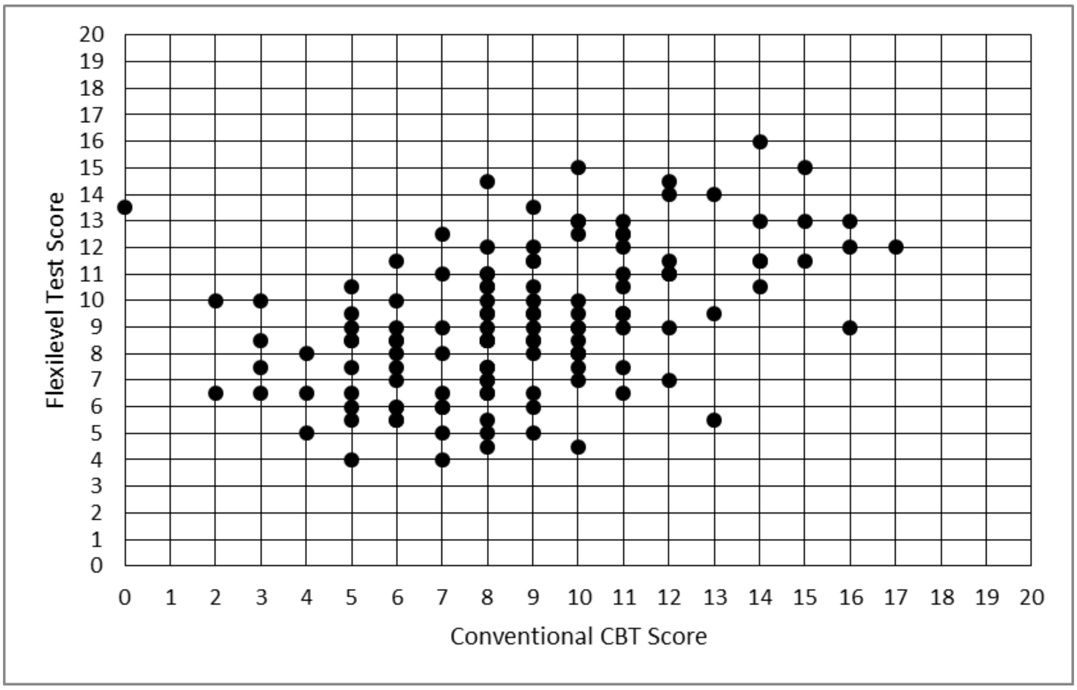


FIGURE 4-3: A SCATTERPLOT CHART TO SHOW THE SCORES ACHIEVED BY EXAMINEES IN EACH OF THE STAGES (N=131)

The Pearson product-moment correlation coefficient was calculated for the data in Table 4-3 and shows that the Flexilevel and conventional CBT test scores are significantly correlated (r=0.461, N=131, p≤0.01). Additionally, a paired t-test showed no significant difference between the scores for the Flexilevel and CBT test (t=-1.836, df=130, p=0.069).

Findings. The Pearson product-moment correlation and t-test results were interpreted as showing that the examinees' were not disadvantaged by the use of the Flexilevel test when compared with conventional CBT test. These findings provide additional support for the notion that the Flexilevel test approach does not disadvantage students.

4.2.2 STUDY B

This study was conducted in a summative assessment context. It should be noted that findings from this study were previously published in Pyper & Lilley (2010).

Participants. A total of 180 first year Computer Science undergraduates took the test as part of the assessment regime of the programming module they were studying.

Methodology. The test was conducted under supervised conditions in computer laboratories. Similarly to Study A (please refer to section 4.2.1), the test comprised 40 items, administered in two stages: one Flexilevel test (20 items) and one conventional CBT (20 items). The maximum duration of the test was 45 minutes and the order in which the stages were presented to each examinee (i.e. Flexilevel test followed by conventional CBT test or vice versa) was randomly selected. The examinees were unaware of the presentation order. Furthermore, examinees were able to take a practice test prior to interacting with the application in a summative assessment context.

Results. A summary of the examinees' performance can be found in Table 4-4. Scores for each of the test stages range from 0 (minimum) to 20 (maximum).

Test Stage	Minimum	Maximum	Mean	Std. Deviation
Flexilevel	2.5	18.5	9.56	2.97
Conventional CBT	3.0	19.0	9.85	3.20

TABLE 4-4: SUMMARY OF EXAMINEES' SCORES FOR EACH SECTION OF THE TEST (N=180)

The Pearson product-moment correlation coefficient was calculated for the data presented in Table 4-4 and shows that the scores for the Flexilevel and conventional CBT tests were highly and significantly correlated, (r=0.646, N=180, p≤0.01). This was interpreted as showing that examinees were not disadvantaged by the use of the Flexilevel test as compared to conventional CBTs. Moreover, a paired t-test showed no significant difference between the scores for the Flexilevel and conventional CBT tests, t= 1.488, df=179, p=0.139.

Figure 4-4 below illustrates the scores obtained by examinees in the two different tests:

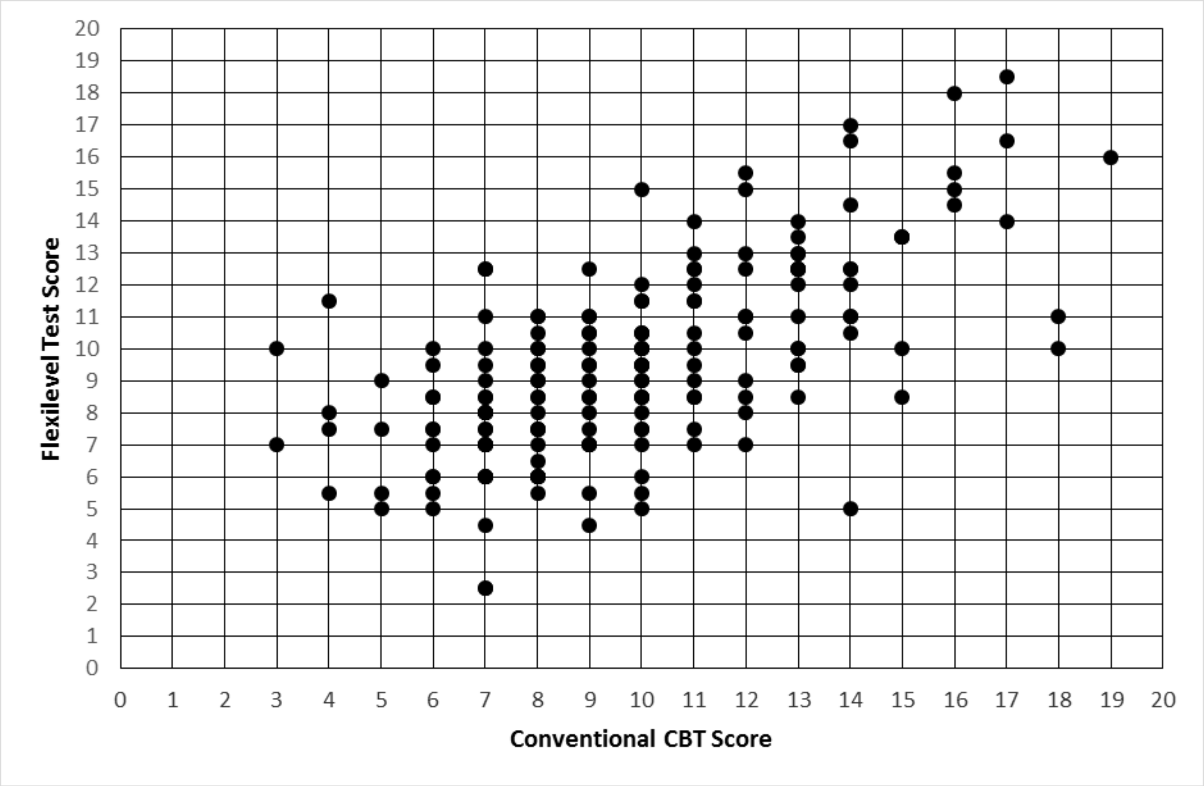


FIGURE 4-4: A SCATTERPLOT CHART TO SHOW THE SCORES ACHIEVED BY EXAMINEES IN EACH OF THE STAGES (N=180)

Findings. The findings from this study are in line with other studies conducted as part of this programme of research. They support the view that the correlation between Flexilevel and CBT scores is statistically significant, as well as the view that examinees are not disadvantaged by the Flexilevel approach.

4.3 Real Data Simulation

Previous studies conducted as part of this research showed that examinees were not disadvantaged by the Flexilevel approach when compared to conventional CBT tests. An anticipated benefit of the Flexilevel approach would be the ability to achieve comparable estimates of examinees’ performance using fewer items than those required in a conventional CBT test.

Shorter tests could increase the feasibility of having a greater range of formative assessment opportunities for students, particularly in mobile assessment contexts where challenges to the attention of students as they interact with any mobile assessment application may be limited (Oulasvirta et al., 2005).

It was therefore important to this programme of research to investigative the extent to which the scores generated by shorter Flexilevel tests are comparable to those generated in a conventional CBT test.

To this end, a simulation study was conducted. Simulation studies are often used to evaluate the efficacy of test methodologies based on fully simulated data (see, for example, Armstrong et al. (2004) and De Ayala et al. (1990)). However, the study reported here is based on datasets derived from the results of actual tests that were originally presented to examinees as a conventional CBT.

The approach adopted in this real data simulation study was to apply the Flexilevel test algorithm to two different existing sets of test data and then to compare the results obtained using the simulated Flexilevel test with those actually achieved by examinees in the original CBT test. It should be noted that this work was previously published in Pyper et al. (2014a).

Data sets. Examinee responses from two different conventional CBTs were collated for analysis. The tests in question were part of the summative assessment for one Level 4 Databases module and one Level 6 Databases module. Both tests consisted of objective items only.

Methodology. For each of the datasets, all items were ranked according to their facility. The facility of the items was calculated using Equation 1. The Flexilevel test algorithm as outlined previously was then applied to the two existing datasets in a simulated environment, as illustrated in Table 4-5.

The Databases Level 4 dataset contained responses from 11 examinees. The Databases Level 6 dataset contained responses from 65 examinees.

Dataset	Level	CBT (Total number of items)	Simulated Flexilevel (Total number of items)
1	4	22	11
2	6	17	9

TABLE 4-5: NUMBER OF ITEMS IN EACH TEST FOR THE TWO DATASETS USED

Dataset 2 is used below to illustrate how the simulation was conducted. The ordering of the items for this test, ranked in order of facility from the easiest (highest facility) to the most difficult (lowest facility), is: Q16, Q10, Q9, Q7, Q12, Q4, Q8, Q1, Q15, Q17, Q3, Q6, Q13, Q5, Q11, Q2, Q14.

The Flexilevel test typically starts with an item of medium difficulty because the proficiency level of the examinee is not known at the start of the test. In this example, the item of medium difficulty is Q15. An examinee's simulated test where the examinee answers all items correctly would result in the following item sequence: Q15, Q17, Q3, Q6, Q13, Q5, Q11, Q2, Q14. An examinee's simulated test where the examinee answers all items incorrectly would result in the following item sequence: Q15, Q1, Q8, Q4, Q12, Q7, Q9, Q10, Q16.

Table 4-6 and Table 4-7 are intended to illustrate how the simulation worked; the set of responses provided by an examinee (Examinee number 40) can be seen in Table 4-6. Here, a value of 1 indicates a correct response and 0 indicates an incorrect response; Examinee number 40 achieved a score of 12 out of 17 (70.6%).

Examinee	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17
40	1	1	1	1	1	1	0	1	1	1	0	1	0	0	0	1	1

TABLE 4-6: SET OF RESPONSES PROVIDED BY EXAMINEE NUMBER 40

Table 4-7 shows the outcome of the Flexilevel test simulation for the same examinee. The Flexilevel score would be 6.5 out of 9 (72.2%).

Examinee	Q15	Q1	Q17	Q3	Q6	Q13	Q8	Q5	Q11
40	0	1	1	1	1	0	1	1	0

TABLE 4-7: SET OF SIMULATED RESPONSES FOR EXAMINEE NUMBER 40, USING THE FLEXILEVEL TEST ALGORITHM

Results. Table 4-8 summarises actual (CBT) and simulated (Flexilevel) scores for all examinees for the first dataset (Level 4 Databases). Table 4-9 summarises actual (CBT) and simulated (Flexilevel) scores for all examinees for the second dataset (Level 6 Databases).

As can be seen from Table 4-5, for the Level 4 Databases Test, the scores could range from 0 to 22 for the CBT test and 0 to 11 for the Flexilevel test.

Test Stage	Minimum	Maximum	Mean	Std. Deviation
Simulated Flexilevel	3.5	9.5	6.95	1.85
Conventional CBT	9	20	14.82	3.74

TABLE 4-8: SUMMARY OF ACTUAL (CBT) AND SIMULATED (FLEXILEVEL) SCORES FOR THE LEVEL 4 DATABASES TEST (N=11)

For the Level 6 Databases Test, as can be seen from Table 4-5, the scores could range from 0 to 17 for the CBT Test and 0 to 9 for the Flexilevel test.

Test Stage	Minimum	Maximum	Mean	Std. Deviation
Simulated Flexilevel	3	9	6.73	1.52
Conventional CBT	3	17	12.27	3.47

TABLE 4-9: SUMMARY OF ACTUAL (CBT) AND SIMULATED (FLEXILEVEL) SCORES FOR THE LEVEL 6 DATABASES TEST (N=65)

Figure 4-5 provides a visualisation of the relationship between the CBT and the Flexilevel scores for the Level 4 Databases test.

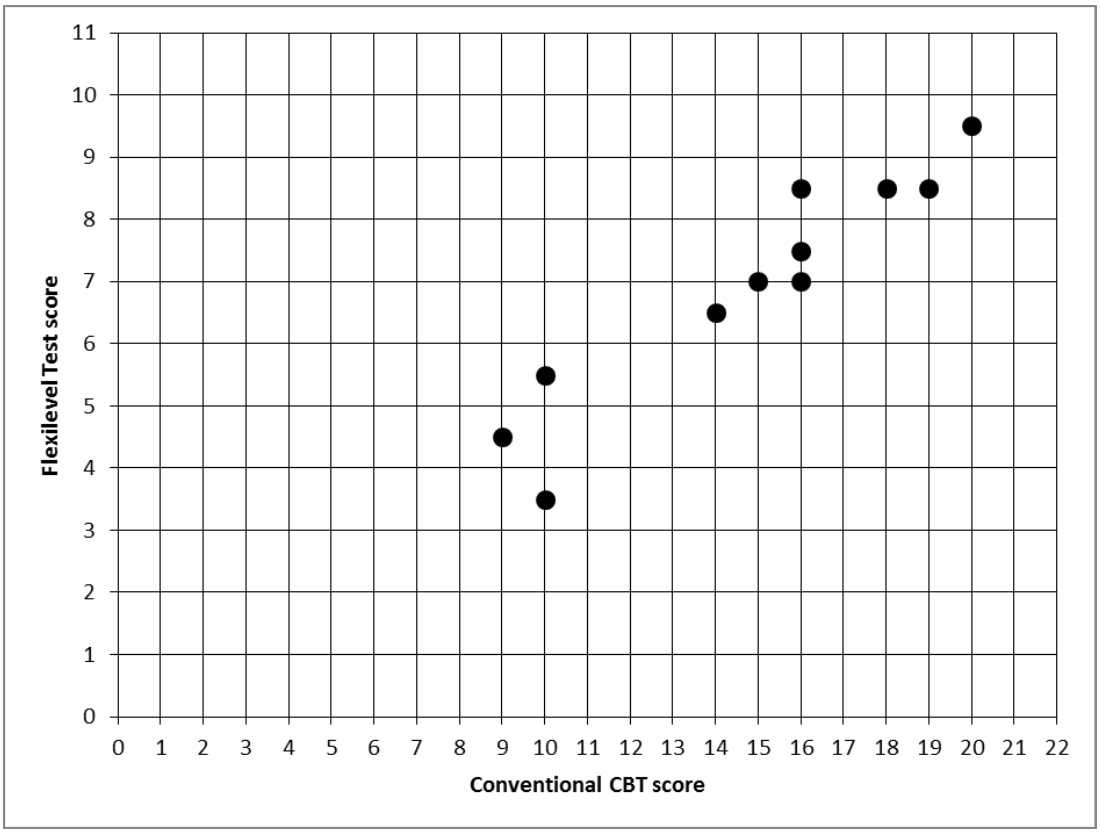


FIGURE 4-5: A SCATTERPLOT CHART TO SHOW THE ACTUAL (CBT) AND SIMULATED (FLEXILEVEL) SCORES FOR THE LEVEL 4 DATABASES MODULE (N=11)

A Pearson product-moment correlation coefficient was calculated for the Level 4 Databases test and showed a strong positive correlation between the scores calculated for the simulated Flexilevel test and those achieved by students in the actual conventional CBT test (r=0.946, N=11, p≤0.01).

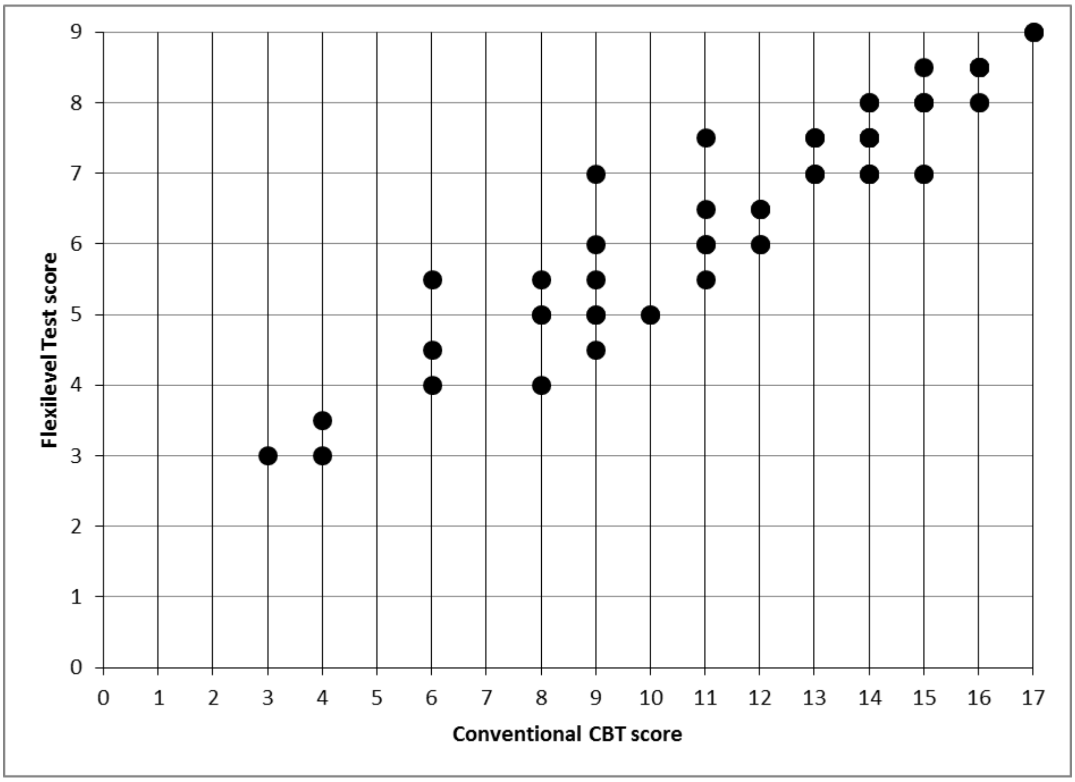


FIGURE 4-6: A SCATTERPLOT CHART TO SHOW THE ACTUAL (CBT) AND SIMULATED (FLEXILEVEL) SCORES FOR THE LEVEL 6 DATABASES MODULE (N=65)

A Pearson product-moment correlation coefficient was calculated for the Level 6 Databases test and showed a strong positive correlation between the scores calculated for the simulated Flexilevel test and those achieved by students in the actual conventional CBT test (r=0.941, N=65, p≤0.01).

Findings. This study was concerned with a simulation study in which two sets of data collected from an authentic summative assessment setting were analysed and used as the basis for a simulation of Flexilevel tests. The calculation of the Pearson correlation for each of the tests showed that there was a significant correlation between the simulated Flexilevel and actual CBT test scores for both datasets.

4.4 Second Phase of Live Testing

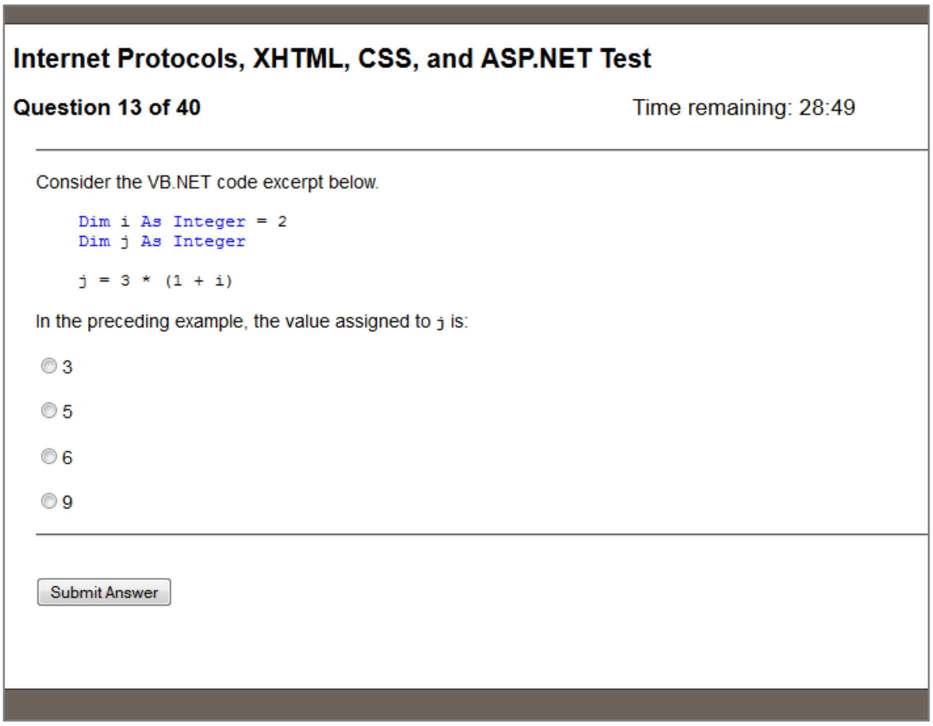
Following the encouraging results from the real data simulation, the second phase of live testing was concerned with evaluating the effectiveness of shorter Flexilevel tests on examinees’ scores in real educational contexts.

4.4.1 Study C

As with previous studies, participants engaged with a test that consisted of one Flexilevel and one conventional CBT test stage, where the number of items for the Flexilevel and CBT test stages was the same. The main difference when compared to previous studies is that in this case there were half as many Flexilevel items as CBT ones.

Participants. A total of 21 Level 6 examinees took part in a diagnostic test for a Level 6 Computer Science online distance learning module.

Methodology. The test was presented to examinees online via the web application developed as part of this programme of research (please see Appendix B). The development of a web application was necessary as the participants were geographically dispersed.



Internet Protocols, XHTML, CSS, and ASP.NET Test

Question 13 of 40 Time remaining: 28:49

Consider the VB.NET code excerpt below.

```
Dim i As Integer = 2
Dim j As Integer

j = 3 * (1 + i)
```

In the preceding example, the value assigned to j is:

○ 3

○ 5

○ 6

○ 9

FIGURE 4-7: SCREENSHOT OF THE USER INTERFACE OF THE WEB APPLICATION, SHOWING THE PRESENTATION OF AN ITEM

The e-assessment application supported two different item selection algorithms: a Flexilevel test and a conventional CBT test algorithm. Test items were selected from the database and presented individually. This is consistent with the need for the Flexilevel algorithm to select an appropriate item depending on the performance of individual examinees. Figure 4-7 shows how the user interface contained minimal information and gave no cues as to the approach (CBT or Flexilevel) being used.

The test consisted of three stages. The first, made up of ten items, was of relevance to the module, but not to the study. The last two stages were a Flexilevel test and a conventional CBT test, presented to the examinee in a random order unknown to the examinee. The Flexilevel test stage contained ten items and the conventional CBT test stage contained 20 items. The test had a total of 40 test items with a time limit of 40 minutes.

Results. A summary of examinees' performance can be found in Table 4-10. Scores could range from 0 (minimum) to 10 (maximum) for the Flexilevel test, and from 0 (minimum) to 20 (maximum) for the conventional CBT test.

Test Stage	Minimum	Maximum	Mean	Std. Deviation
Flexilevel (out of 10)	2.0	8.0	5.62	1.75
Conventional CBT (out of 20)	7.0	19.0	12.71	3.45

TABLE 4-10: SUMMARY OF EXAMINEES' SCORES FOR EACH SECTION OF THE FIRST TEST (N=21)

Figure 4-8 shows the results for the first test. Here it can be seen that there is a noticeable trend in the data such that examinees achieving relatively low scores in one stage also achieve relatively low scores in the other. Similarly, those scoring relatively high scores in one stage tend to achieve relatively high scores in the second.

However, there are scores that do not appear to conform to this trend. There are two instances where examinees only scored 2 or 2.5 in the Flexilevel test, which is around chance, whereas they did better in the CBT stage, in one case scoring 13. Further investigation on the examinees' submissions did not indicate a change in behaviour, for example rapid answers in the Flexilevel section, between the stages.

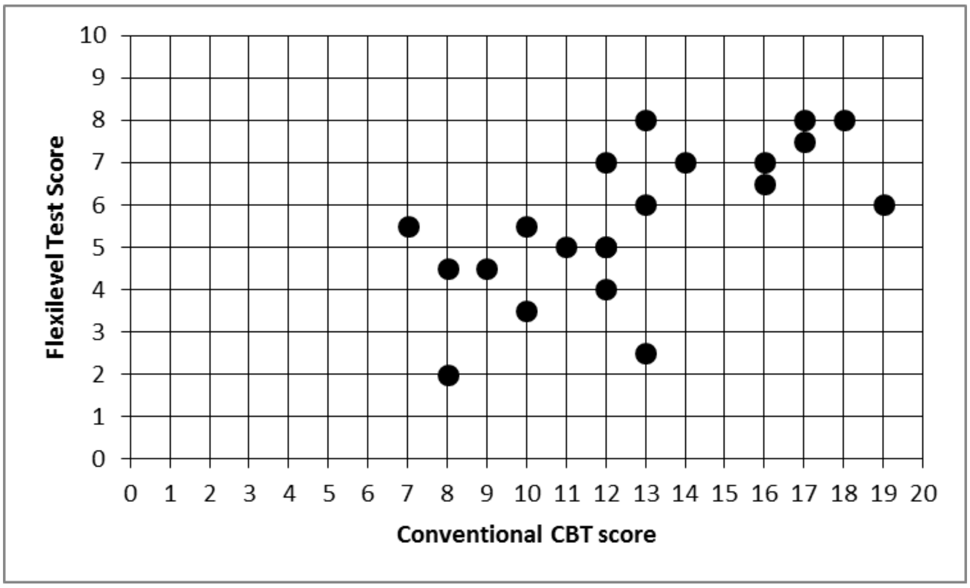


FIGURE 4-8: A SCATTERPLOT TO SHOW THE SCORES ACHIEVED BY EXAMINEES IN THE FLEXILEVEL AND CBT STAGES IN THE FIRST TEST (N=21)

The Pearson product-moment correlation was calculated for the data presented in Table 4-10 shows that the scores for the shorter Flexilevel test and the conventional CBT were significantly correlated (r= 0.645, N=21, p≤0.01).

Findings. The findings from this study support the view that scores obtained by examinees in a Flexilevel test and a conventional CBT test are comparable when the Flexilevel test has half the number of items that are present in the CBT test.

4.4.2 STUDY D

The approach used in Study C (please refer to section 4.4.1) was repeated here.

Participants. A total of 18 Level 6 examinees from the original Study C group took part in a low-stakes summative assessment (2%).

Methodology. Study D took place 21 days after Study C. The test was presented to examinees online via a web application developed as part of this programme of research. The test was not proctored, students took the test at a time and place of their choosing within a 48-hour window.

The test configuration was the same as in Study C. There were three stages. The first, with ten items, was part of the module, but was not pertinent to this study. The last two stages, which were of interest to the study, were a Flexilevel test and a conventional CBT test, presented to the examinee in a random order unknown to the examinee. The Flexilevel test stage contained ten items and the conventional CBT test stage contained 20 items. The test had a total of 40 test items with a time limit of 40 minutes.

Results. A summary of examinees' performance can be found in Table 4-11. Possible scores range from 0 (minimum) to 10 (maximum) for the Flexilevel test, and from 0 (minimum) to 20 (maximum) for the conventional CBT test.

Test Stage	Minimum	Maximum	Mean	Std. Deviation
Flexilevel (out of 10)	3.5	10.0	6.78	1.89
Conventional CBT (out of 20)	11.0	19.0	15.56	3.07

TABLE 4-11: SUMMARY OF EXAMINEES' SCORES FOR EACH SECTION OF THE SECOND TEST (N=18)

Figure 4-9 illustrates the performance of examinees in the two test stages.

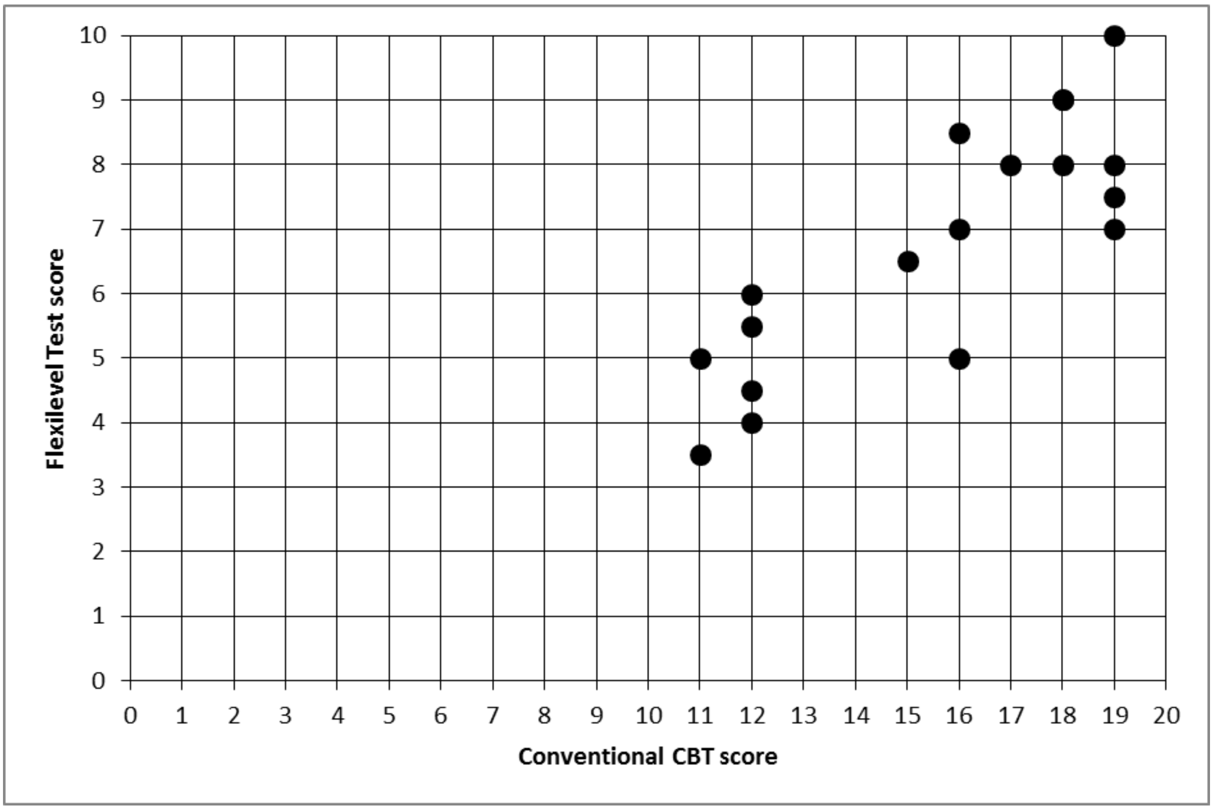


FIGURE 4-9: A SCATTERPLOT CHART TO SHOW THE SCORES ACHIEVED BY EXAMINEES IN EACH OF THE STAGES FOR THE SECOND TEST (N= 18)

The Pearson product-moment correlation coefficient was calculated for the data presented in Table 4-11 and shows that the scores for the Flexilvel test and conventional CBT were significantly correlated (r= 0.839, N=18, p≤0.01).

Additionally the performance of examinees across the two tests was calculated. The results are shown in Table 4-12. The correlations from within the first test are not included here for clarity; they were calculated on a larger group of scores than the correlations calculated below.

Component		Test 1 Flexilevel Test	Test 1 CBT Test	Test 2 Flexilevel Test	Test 2 CBT Test
Test 1 Flexilevel test (out of 10)	Pearson Correlation	*	*	0.696	*
	Sig. (2-tailed)	*	*	p≤0.01	*
Test 1 Conventional CBT test (out of 20)	Pearson Correlation	*	*	*	0.683
	Sig. (2-tailed)	*	*	*	p≤0.01
Test 2 Flexilevel test (out of 10)	Pearson Correlation	0.696	*	*	0.839
	Sig. (2-tailed)	p≤0.01	*	*	p≤0.01
Test 2 Conventional CBT test (out of 20)	Pearson Correlation	*	0.683	0.839	*
	Sig. (2-tailed)	*	p≤0.01	p≤0.01	*

TABLE 4-12: PEARSON'S PRODUCT-MOMENT CORRELATION RESULTS FOR THE TWO STUDIES (N=18)

Findings. The Pearson correlation between the Flexilevel test score and the conventional CBT test score found in Study D (r= 0.839, N=18, p≤0.01) is a somewhat stronger correlation than that found in Study C (r= 0.645, N=21, p≤0.01) which may be partly the result of there being fewer outliers in the Flexilevel test score when compared with performance on the CBT test stage. The differences are relatively small, but given the size of the test population, the potential influence of outliers is of interest.

Furthermore, the data in Table 4-10 and Table 4-11 shows that the ranges of scores for both test stages remain either the same (as in the case of the CBT test) or show only a slight shift (by one point for both the lower and upper bound in the case of the Flexilevel test stage).

4.4.3 Study E

This study is intended to investigate the extent to which shorter Flexilevel tests may provide comparable measures of an examinee's proficiency as a CBT test. It should be noted that elements of this work were discussed in Pyper et al. (2015a).

Participants. A group of 18 first year online distance learning Computer Science students took part in a test. The test was summative, and related to their knowledge and understanding of internet technologies and, in particular, ASP.NET.

Methodology. The test was presented to students online via the web application developed as part of this programme of research (please see Appendix B). It was invigilated by a remote live invigilation service to ensure students were not accessing materials that were not permitted during their test.

The remote online invigilation service uses live proctors; the service allows institutions to examine test-takers in their home environment and have them invigilated by means of a live proctor, using screen sharing technologies combined with live audio and video feeds. The service works in conjunction with any preferred assessment application but does not include assessment as part of its functionality. The service is available 24 hours a day, seven days a week, allowing participants to select a time to take part within a 48-hour window.

On the day of the test, participants were required to log in at the time of their pre-selected slot. An email reminder was sent 24 hours prior to the test. Once participants logged into the remote live invigilation service, they were prompted to download and run the software that would connect their webcam and desktop to a live proctor. After a connection had been established, the following authentication and environment checks were carried out:

1. An identity check was performed in which participants were asked to present a form of photo identification (e.g. their student ID). The proctor checked that

the photo and name on the card matched with the participant who had scheduled the slot and was present onscreen via webcam. Nothing was recorded or captured; it was simply logged as having been checked.

2. Participants had the option of uploading their photo in advance or having their photo taken by the proctor as part of the authentication process. If the photo was uploaded in advance, the proctor compared this to the individual participant onscreen.
3. Following authentication, participants were asked to pan over their work area using their web cam and hold a reflective surface to the camera to ensure there were no unauthorised materials or persons present. If all checks were completed to the satisfaction of the proctor, the participant was invited to log into the assessment application.

Once the test was running, the proctor monitored the participant’s desktop, their environment and also their conduct. None of the proctors reported any unusual examinee behaviour, technical or usability problems with the application.

The test was limited to 40 minutes and contained 40 items overall. The test items were subjected to calibration using data from the performance of an earlier cohort of students.

An initial stage outside of the scope of the study, but relevant to the coverage of topics in the module, was presented first for every student. This stage contained 10 items. Then, as with previous studies, the examinees were randomly assigned to one of two different groups. Half of the participants were assigned to Group 1, and the second half to Group 2. Group 1 was presented with the Flexilevel test followed by the CBT; and Group 2 was presented with the CBT followed by the Flexilevel test. For the students there was no distinction between these two test stages. The Flexilevel and CBT stages consisted of 10 and 20 items respectively.

Examinees also took part in a web scripting coursework as part of their studies that covered the same subject domain as was covered in the online tests introduced here.

Results. Table 4-13 shows the range of scores obtained and the mean score for each stage of the test. It can be seen that the scores for both of the stages varied relatively widely. The responses from three participants were not included in the analysis as they did not complete one or more components of the study.

Component	Minimum	Maximum	Mean	Std. Deviation
Flexilevel Test (out of 10)	5.0	9.5	7.83	1.27
Conventional CBT Test (out of 20)	6.0	19.0	12.46	3.37
Web Scripting coursework (out of 30)	10.0	28.0	23.33	5.70

TABLE 4-13: SUMMARY OF STUDENTS' PERFORMANCE FOR EACH SECTION OF THE TEST AND ON THE COURSEWORK (N=15)

Figure 4-10 illustrates the correlation between the Flexilevel and conventional CBT test scores.

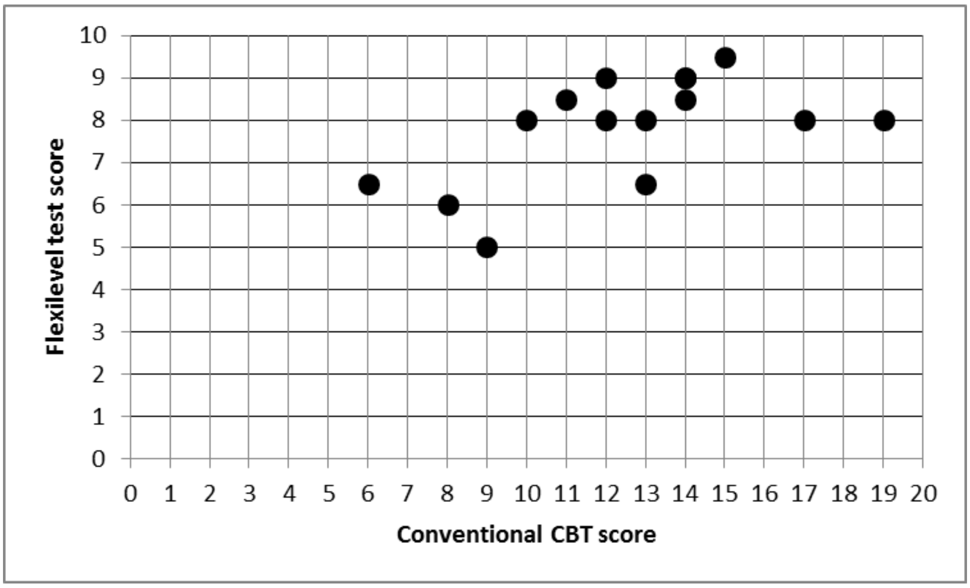


FIGURE 4-10: A SCATTERPLOT CHART TO SHOW THE SCORES ACHIEVED BY STUDENTS IN EACH OF THE STAGES (N=15)

The Pearson's product-moment correlation coefficient was calculated for the data presented in Table 4-13. This is reported in Table 4-14 below.

Component		Flexilevel Test	CBT Test	Web Scripting coursework
Flexilevel Test (out of 10)	Pearson Correlation	*	0.574	0.734
	Sig. (2-tailed)		p≤0.05	p≤0.01
Conventional CBT Test (out of 20)	Pearson Correlation	0.574	*	0.662
	Sig. (2-tailed)	p≤0.05		p≤0.01
Web Scripting coursework (out of 30)	Pearson Correlation	0.734	0.662	*
	Sig. (2-tailed)	p≤0.01	p≤0.01	

TABLE 4-14: PEARSON'S PRODUCT-MOMENT CORRELATION FOR ALL THREE ASSESSMENT CONDITIONS (N=15)

It can be seen from Table 4-14 that the scores achieved by examinees in the Flexilevel test and the conventional CBT test are significantly correlated. Furthermore, the correlations between the Flexilevel test and the Web Scripting coursework scores, and the conventional CBT test and the Web Scripting coursework scores are also statistically significant.

Findings. These results further support the notion that the Flexilevel test can provide comparable results to a standard CBT test whilst only presenting half the items. Interestingly, in this study the Flexilevel test was a better predictor of examinee performance in the practical component (i.e. the Web Scripting coursework) than was the conventional CBT test.

4.4.4 STUDY F

Following the encouraging findings from the previous studies in an online distance learning context, it was important for the research to investigate the comparability of shorter Flexilevel tests and conventional CBT tests in a campus-based context. Please note that this work is being published later this year (Pyper et al. 2015b)

Participants. A total of 58 Level 4 examinees took part in a summative test as part of the assessment regime for their module.

Methodology. The examinees took the test in a computer laboratory under supervised conditions using the web application developed as part of this research. The test was organised into two test stages: Flexilevel and CBT. The order in which the two tests were presented to examinees was randomised. The test items had been calibrated using existing test data and were ranked according to their facility. Furthermore, the items used for each test had the same range of difficulty. Examinees were unaware of the test order. The test consisted of 40 test items (15 Flexilevel items, 25 CBT items), and had a 45-minute time limit.

The examinees were not aware of the location of the web application nor what their login credentials would be prior to the test. Personalised login credentials and instructions were distributed at the start of the test. Once they had completed the test, examinees were presented with their score out of 40 directly through the web application and were free to leave.

Results. A summary of examinees' performance can be found in Table 4-15. Possible scores range from 0 (minimum) to 15 (maximum) for the Flexilevel test, and from 0 (minimum) to 25 (maximum) for the conventional CBT test.

Test Stage	Minimum	Maximum	Mean	Std. Deviation
Flexilevel (out of 15)	4.0	13.0	9.32	2.04
Conventional CBT (out of 25)	7.0	21.0	13.53	3.22

TABLE 4-15: SUMMARY OF STUDENTS' SCORES FOR EACH SECTION OF THE TEST (N=58)

Figure 4-11 illustrates the relationship between the performances of examinees on each of the test stages.

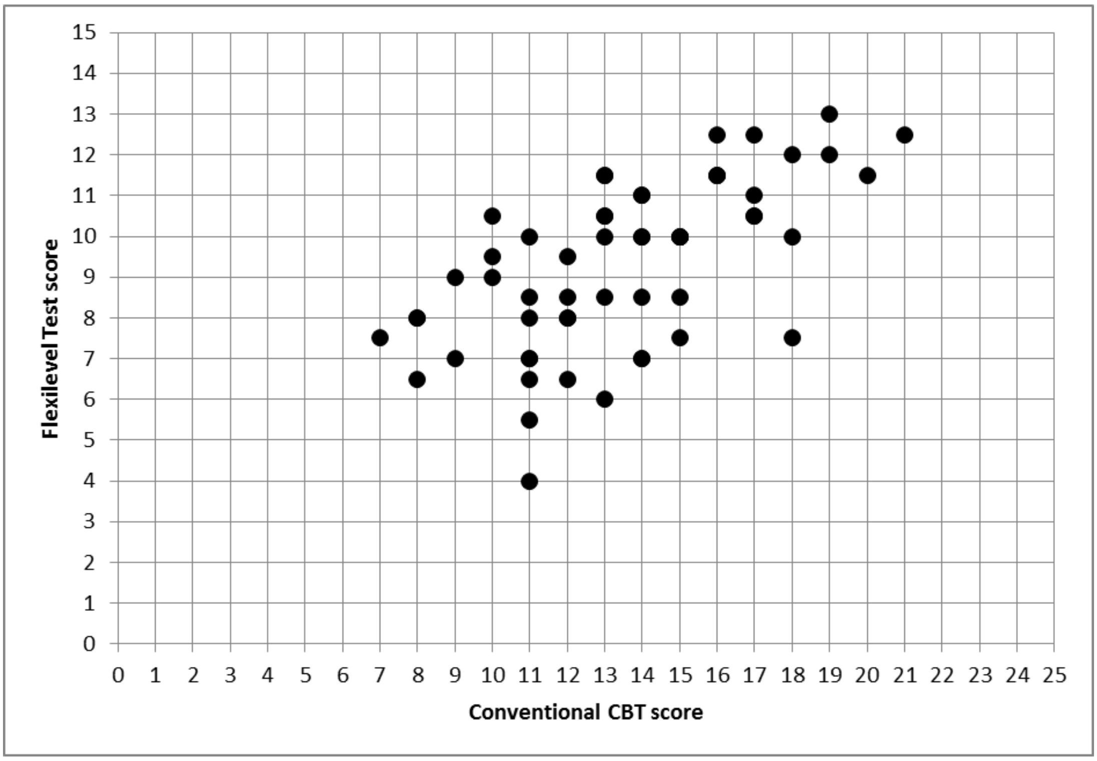


FIGURE 4-11: A SCATTERPLOT CHART TO SHOW THE SCORES ACHIEVED BY STUDENTS IN EACH OF THE STAGES (N=58)

A Pearson's product-moment correlation was performed on the data in Table 4-15 and showed a statistically significant correlation between the Flexilevel and conventional CBT scores (r=0.6232, N=58, p≤0.01).

Additionally, the performance of examinees in the Flexilevel section of the test and in the conventional CBT section of the test was compared with another component of the assessment regime for the module, a practical programming assignment. There are fewer examinees in this part of the analysis because only examinees who both took both the test and the practical programming assignment are included. The results are shown in Table 4-16.

Component		Flexilevel Test	CBT Test	Web Scripting coursework
Flexilevel Test (out of 15)	Pearson Correlation	*	*	0.394
	Sig. (2-tailed)			p≤0.01
CBT Test (out of 25)	Pearson Correlation	*	*	0.292
	Sig. (2-tailed)			p≤0.05
Programming practical test (out of 20)	Pearson Correlation	0.394	0.292	*
	Sig. (2-tailed)	p≤0.01	p≤0.05	

TABLE 4-16: PEARSON'S PRODUCT-MOMENT CORRELATION FOR ALL THREE ASSESSMENT CONDITIONS (N=48)

It can be seen that both the Flexilevel test score and the conventional CBT test score are significantly correlated with the scores examinees achieved for the programming practical. Further, the correlation for the Flexilevel test score is stronger than that for the conventional CBT score.

For clarity the correlation between the Flexilevel test and the CBT test is not included in Table 4-16. 58 examinees took the test whereas 48 took both the test and completed the coursework so the datasets for the two correlations are different.

Findings. In line with previous studies conducted as part of this research, the correlation between the Flexilevel test and the conventional CBT test was statistically significant. These results provide further support for the idea that shorter Flexilevel tests can provide examinees with assessment opportunities that are comparable in outcome with, but shorter than, conventional CBT tests, as well as being tailored to examinees' individual levels of ability within a subject domain.

In this study the Flexilevel test was also a better predictor of examinee performance in the practical component than the conventional CBT test.

Furthermore, no technical or usability problems with the application were raised or identified through direct observation by invigilators.

4.4.5 Study G

Given the results of studies using fewer Flexilevel test items, this study involved making a test available for self-assessment purposes with the option of using the test on a mobile device. A test that used the Flexilevel algorithm for all item selection was made available on a voluntary basis to students. It was based on a paper-based test (PBT) that participants had taken during a module they had just completed. The PBT was deemed to be a paper-based alternative to CBTs used elsewhere in the research; participants were presented with a set of objective items and were required to select the correct answer to be awarded a mark.

Participants. A total of 29 participants who were enrolled on a first year Computer Science undergraduate module took part in the study, each responding to at least one item. Of these participants, 21 completed the test and had also taken the PBT.

Methodology. The test was run as an open test in which participants were informed that they could take the test using whichever device they preferred and at whatever time they preferred within the test window. The test was run using a version of the web application (please see Appendix C) that also had some adaptations for mobile use.

The test consisted of 13 items and had a time limit of 20 minutes. All items were selected using the Flexilevel algorithm (i.e. the test consisted of one Flexilevel test stage).

Results. As outlined above, participants had the option to take part in the study at a time convenient to them, using a device of their choice. Fourteen out of 21 participants took part using a desktop computer or laptop. Five of the remaining seven used a smartphone (four Android and one iOS), and two a tablet (one Android and one iOS).

A summary of the scores obtained in each of the tests is shown in in Table 4-17. The possible range of scores is 0–13 (Flexilevel test) and 0–100 (PBT).

Test	Minimum	Maximum	Mean	Std. Deviation
Flexilevel Test (out of 13)	4.0	13.0	9.14	2.54
PBT Test (out of 100)	24.0	96.0	74.29	19.58

TABLE 4-17: SUMMARY OF STUDENTS' SCORES FOR EACH SECTION OF THE TEST (N=21)

Figure 4-2 shows the distribution of scores obtained in the Flexilevel and PBT tests.

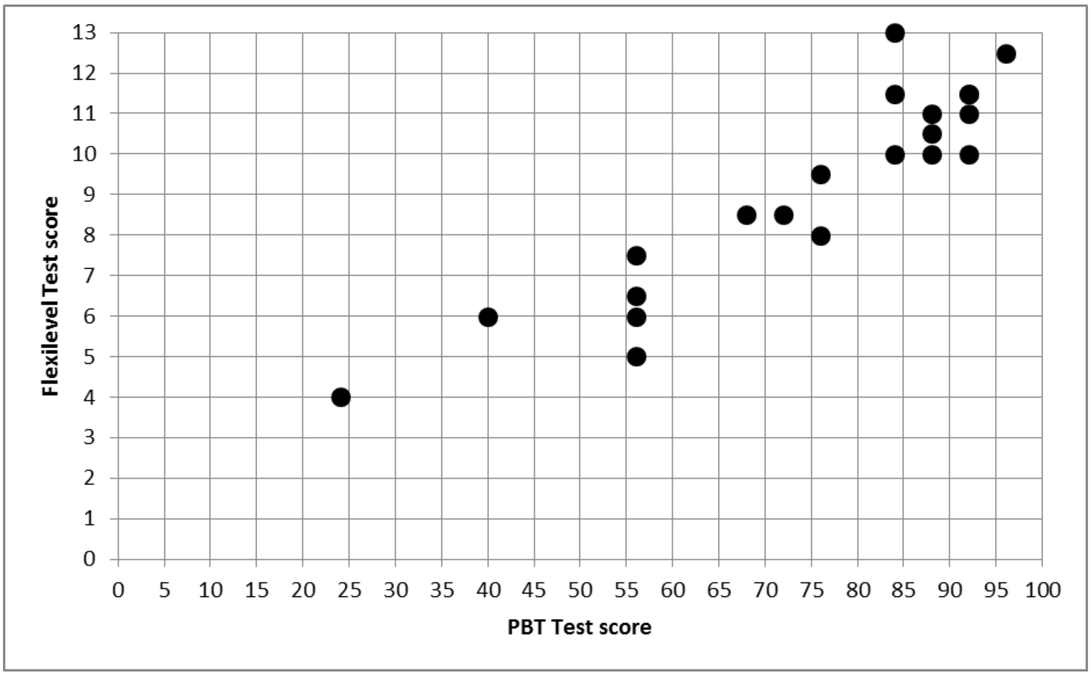


FIGURE 4-12: A SCATTERPLOT CHART TO SHOW THE SCORES ACHIEVED IN TWO TEST CONDITIONS (N=21)

A Pearson's product-moment correlation was performed on the data, showing a statistically significant correlation between the Flexilevel and PBT scores (r=0.916, N=21, p≤0.01).

It was of interest to the research to also analyse participants' responses according to the platform used for taking part in the study.

Table 4-18 provides a summary of performance for those participants who used a desktop or laptop computer to participate.

Test	Minimum	Maximum	Mean	Std. Deviation
Flexilevel Test (out of 13)	4	13	9.11	2.53
PBT Test (out of 100)	24	92	72.29	21.25

TABLE 4-18: SUMMARY OF PARTICIPANTS' PERFORMANCE WHEN USING DESKTOP OR LAPTOP COMPUTER (N=14)

A Pearson's product-moment correlation was performed on the data shown in Table 4-18, showing a statistically significant correlation between the Flexilevel and conventional PBT scores (r=0.916, N=14, p≤0.01) for desktop and laptop users.

Additionally, Table 4-19 provides a summary of performance for those participants who used a mobile device such as a smartphone or tablet to take part in the study.

Test	Minimum	Maximum	Mean	Std. Deviation
Flexilevel Test (out of 13)	5	12.5	9.21	2.75
PBT Test (out of 100)	56	96	78.29	16.46

TABLE 4-19: SUMMARY OF PARTICIPANTS' PERFORMANCE WHEN USING A SMARTPHONE OR TABLET (N=7)

A Pearson's product-moment correlation was performed on the data shown in Table 4-19, showing a statistically significant correlation between the Flexilevel and conventional PBT scores (r=0.980, N=7, p≤0.01).

Findings. The correlation between the Flexilevel test and the conventional PBT test scores was strong and statistically significant, regardless of the platform used by participants. These results provide further support for the idea that shorter Flexilevel tests are capable of providing examinees with assessment opportunities that are comparable to conventional objective tests. Furthermore, it provides support to the notion that the Flexilevel test would lend itself to use in mobile contexts.

4.4.6 Study H

This study was comprised of two Flexilevel tests that were presented as diagnostic and low-stakes summative tests separated by 14 days.

Both tests provided detailed feedback about work the students would benefit from doing based on their score. Thus a tailored test was used to provide tailored feedback to the students. Also, once a student had completed the second test, they were presented with a comparison of their scores achieved.

Methodology. Forty-three students took one or both of the tests on devices of their own choice. The test was a low-stakes summative test that students were permitted to take at a time of their choosing, within the limits of the test window. The tests consisted of 20 items from the same pool of 39 items.

Results. Results were excluded from the analysis if an examinee did not complete one of the tests. This meant that the analysis was performed on the results of 22 students. These can be seen in Table 4-20.

Flexilevel	Minimum	Maximum	Mean	Std. Deviation
Test 1	7.5	17.5	11.14	3.09
Test 2	8.0	18.0	11.62	3.50

TABLE 4-20: COMPARISON BETWEEN EXAMINEES’ SCORES ON A REPEATED TEST USING THE FLEXILEVEL ALGORITHM (N=22)

The relationship between examinee’s performance on the two Flexilevel tests is shown in Figure 4-13.

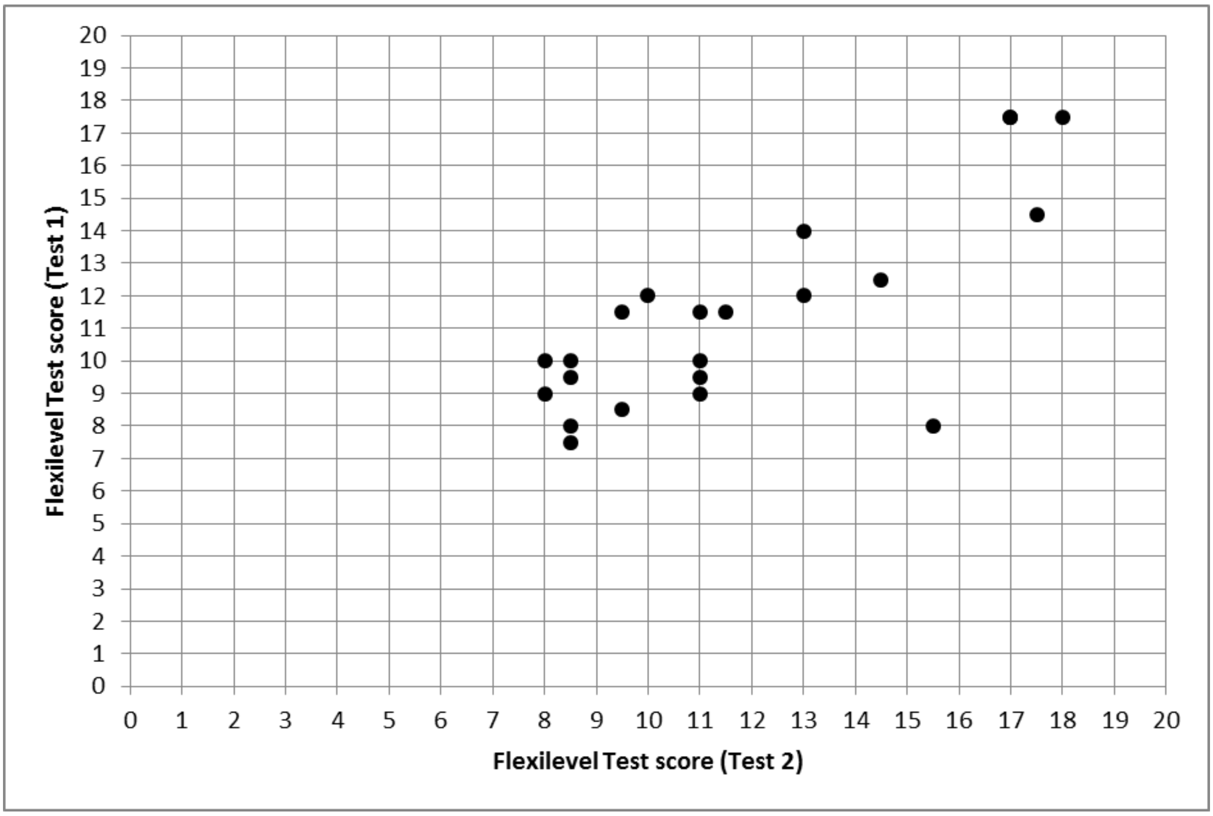


FIGURE 4-13: A SCATTERPLOT CHART THAT ILLUSTRATES SCORES ACHIEVED BY EXAMINEES ON THE TWO TESTS (N=22)

A paired t-test was calculated using the data shown in Table 4-20. This shows no significant difference between the first and second Flexilevel test scores, t(21) = -0.905, p=0.375).

Findings. Given that the subject matter being tested was stable over the 14 days between the two tests, the paired t-test was used to analyse the extent to which the Flexilevel test generates reliable results. The results of the t-test show that there is no statistical difference between the scores of examinees in the first and second Flexilevel tests. This finding supports the notion that the Flexilevel test is able to generate reliable scores.

4.5 Discussion

The empirical studies reported in this chapter are intended to address the question of the extent to which Flexilevel tests afford effective assessment opportunities in Higher Education contexts.

The key findings are:

1. **Scores for the Flexilevel test and conventional CBT test are significantly correlated when the number of items in both tests is the same.**

   This was taken to indicate that scores generated by Flexilevel tests and conventional CBT tests are comparable. Additionally, these findings were taken to indicate that examinees are not disadvantaged by the Flexilevel test. Finally, the results of the Flexilevel test are taken to be good predictors of scores for a conventional CBT test. This component of the research was explored in the Pilot Study involving undergraduate students (section 4.1.1), postgraduate study (section 4.1.2), Study A (section 4.2.1, first phase of live testing) and Study B (section 4.2.2, first phase of live testing).

2. **Scores for the Flexilevel test and conventional CBT test are significantly correlated when the number of items in both tests is the same, and this applies to different levels of study.**

   This was taken to indicate that the scores obtained through Flexilevel tests and conventional CBT tests are comparable for students at different levels of study. This component of the research was explored as follows:

   - **Level 4:** Study A (section 4.2.1, first phase of live testing) and Study B (section 4.2.2, first phase of live testing);
   - **Level 6:** Pilot Study involving undergraduate students (section 4.1.1);
   - **Level 7:** Pilot Study involving postgraduate students (section 4.1.2).

3. **Scores for the Flexilevel test and conventional CBT test are significantly correlated in the case of shorter Flexilevel tests (provided that the length of the Flexilevel test is at least half that of the conventional CBT).**
   This was an important finding, and taken to indicate that scores generated by Flexilevel tests and conventional CBT tests are comparable when the Flexilevel test has at least half the number of items of a conventional CBT test. Furthermore, these findings were taken to indicate that examinees are not disadvantaged by the Flexilevel test. Finally, scores obtained using a Flexilevel test with at least half the number of items of a conventional CBT test are good predictors of scores for a conventional CBT test. This component of the research was explored in the Real Data Simulation (section 4.3), Study C (section 4.4.1, second phase of live testing), Study D (section 4.4.2, second phase of live testing), Study E (section 4.4.3, second phase of live testing) and Study F (section 4.4.4, second phase of live testing).

   It was of interest to the research to explore whether the correlation between conventional MCQ scores and Flexilevel scores would also be significant where the MCQ test was paper-based rather than a CBT. As part of Study G (section 4.4.5, second phase of live testing), it was found that scores obtained through Flexilevel tests and those obtained through paper-based MCQ tests are comparable. Additionally, these findings were taken to indicate that examinees are not disadvantaged by the Flexilevel test.

4. **Scores for the Flexilevel test and conventional CBT test are significantly correlated, in the case of shorter Flexilevel tests (provided that the length of the Flexilevel test is at least half that of the conventional CBT), and this applies to different levels of study.**
   This was taken to indicate that scores generated by Flexilevel tests and conventional CBT tests are comparable for students at different levels of study when the Flexilevel test has at least half the number of items of a conventional CBT test. This component of the research was explored in the following studies:

- **Level 4:** Study E (section 4.4.3, second phase of live testing) and Study F (section 4.4.4, second phase of live testing);
- **Level 6:** Study C (section 4.4.1, second phase of live testing) and Study D (section 4.4.2, second phase of live testing).

5. **Scores for the Flexilevel test and those obtained through other forms of assessment are significantly correlated.**
   This was taken to indicate that the scores obtained through Flexilevel tests and those obtained through other forms of assessment within a given subject domain are comparable. Moreover, these findings were taken to indicate that examinees are not disadvantaged by the Flexilevel test. Finally, it is also a finding that Flexilevel scores are a good predictor of scores obtained through other forms of assessment within a subject domain. This component of the research was explored in Study E (section 4.4.3, second phase of live testing), Study F (section 4.4.4, second phase of live testing).

The empirical studies reported in this chapter are taken to demonstrate that the Flexilevel test affords effective assessment opportunities in Higher Education contexts.

5 EVALUATION WITH STUDENTS

From the start of this programme of research, a second theme of work was instituted to complement the studies concerned with the effectiveness of the Flexilevel test in real educational contexts. Barriers to the uptake of the approach may not be limited to technical or practical issues. Barriers may include issues of perceived usefulness, and it was important to establish the attitudes of students and staff to CAT in general and the Flexilevel test approach specifically. As part of the pilot study, students were asked about their views of the Flexilevel test and this study was followed up with a similar study involving staff.

In functional usability terms, the interest is in the context of use of the system. This can be highly constrained as it has been for much of the programme of research where invigilated tests have predominated. However when considering unmoderated online and formative contexts of use, the context of use for the application is more diverse. As such, gaining an understanding of how students are already using networked and mobile technology in their studies is an important foundation for work on designing and developing an online e-assessment application that is usable on mobile devices.

Previous work in this area provides additional background to this work. Lilley, Pyper & Attwood (2012) surveyed 110 students and interviewed six of those in order to produce personas that represented student groups studying on the BSc. online degree programmes.

The personas showed that they tended to be in employment, often balancing work, studies and personal lives. They were widely distributed geographically and demographically although consistently were studying in order to advance their careers. The use of mobile devices was evident, although most personas connected to the internet via broadband at home or at work via laptops or PCs. Finally, there was a growing emphasis on interaction with tutors and other students, which was something of a departure from previous data gathered about the student population through formal channels such as the Student Feedback Questionnaire for modules. Of most interest to students previously were individual transactions, i.e. studying alone, and the flexibility to study at their own pace, something that is consistent with other findings (Winter et al., 2010).

5.1 Contexts of use

An important part of this programme of research is aimed at understanding how the Flexilevel test might be applied in genuine educational contexts, and an increasingly important part of our educational context is mobile learning and assessment. It should be noted that work in this section has been previously published in Pyper et al. (2015a).

Mobile learning and assessment brings new challenges to the work, particularly in terms of supporting students in attending to cognitively demanding tasks such as formative assessments in a mobile context, which itself imposes significant cognitive load (Mayer, 2008).

However, the case for supporting mobile learning and assessment is compelling, from both a pedagogical and a practical perspective (Herrington et al., 2009). For example students are often under significant time pressure and, given appropriate opportunities, can make use of short periods of time to engage in their studies (Traxler, 2007).

In principle, one of the benefits of the Flexilevel test is the ability to present fewer items in a test than a standard Computer Based Test (CBT) approach while still obtaining a comparably accurate measurement of an examinee's proficiency (Weiss & Betz, 1973). This has also been identified in this programme of research (please see chapter 4). The potential to provide shorter tests is of interest, since it may provide an opportunity for educationally useful experiences in a broader range of contexts as noted above (Traxler (2007), Herrington et al. (2009)). Further, it may provide these opportunities whilst mitigating some of the challenges associated with deploying formative assessments in a mobile context. As such, it is of interest to investigate whether or not this effect can be reproduced in a genuine educational context.

There are practical issues to address given the application of the Flexilevel test in mobile contexts that have not applied to the desktop and classroom contexts of studies conducted earlier in this programme of research (for example Lilley & Pyper

(2009), Pyper & Lilley (2010)). Key to the differences between desktop and mobile use of Flexilevel testing is the greater competition for limited attentional resources that mobile environments may impose.

Whilst desktop computer contexts may vary, the difference in mobile usage contexts may vary quite substantially (Oulasvirta et al., 2005), both between devices and during a given interaction. In short, the contexts for mobile use tend to be more diverse and potentially distracting than those involving desktop computers.

Oulasvirta et al. (2005) used a range of tasks on a mobile device to show how different contexts impact on the attention of users. Factors that impacted upon users' attention to the tasks they had been set included the amount of social interaction that users needed to engage in (for example in managing their personal space whilst using an escalator) and the predictability of their context. Where there was much going on in the users' context, it required that users attended to their contexts for longer and more often, their attention on the task being interrupted.

These are temporary changes in the focus of attention away from a given task with a mobile device and it is worth noting that many interruptions are triggered by the user themselves (Tsiaousis & Giaglis, 2008). Overall, an indication of the impact of such interruptions may be found in differences in the length of interaction users may have between desktop and mobile devices whereby desktop interactions have been timed as lasting substantially longer than mobile interactions (Monsell, 2003). Furthermore, it seems that interruptions may have an additional cognitive load in terms of switching between tasks (Sandy et al., 2013).

Framing the analysis of this cognitive load provides a good basis for understanding how mobile contexts may impact upon users' usage of a mobile application. Cognitive load theory (Sweller, 2010) has also been influential in learning, particularly multimedia learning (Mayer, 2008) and it seems pertinent also to mobile learning and e-assessment when considering the impact of extraneous loads on learning (Oviatt, 2006).

Clearly context of use is an important contributor to extraneous load, and in providing silent, invigilated exam conditions, extraneous load is minimised, freeing cognitive resources for the intrinsic load that tests impose. This is the case in previous studies conducted as part of this programme of research. Whilst the contexts may vary in these desktop environments, the studies have previously involved environments that were controlled to some substantial extent. For example in both formative and summative assessments in computer laboratories (Lilley & Pyper (2009), (Pyper & Lilley (2010)) and with students taking tests remotely at their own computers, all were invigilated and were the sole focus of the students' attention. This was done for both educational and empirical reasons, but the contexts of use of the Flexilevel assessment have not varied greatly.

However if a Flexilevel test were deployed in a mobile context, then it could be competing for attention in a much more diverse context, with a relatively high potential for interruption. As noted, this is something that impacts upon the capacity of students to attend to a given task (Monsell (2003), Oulasvirta et al. (2005), Oviatt (2006)).

Effectively it is more likely that there will be a higher extraneous load in the completion of tasks in a mobile context. This becomes most disruptive to the completion of tasks when the distractions occupy the same channels as the task, something that is consistent with the influential account of working memory (Baddeley, 2001). For example, it would be expected that interruptions that require visual perception would impact more acutely on the performance of a task since the task itself involves visual perception, at least in this study.

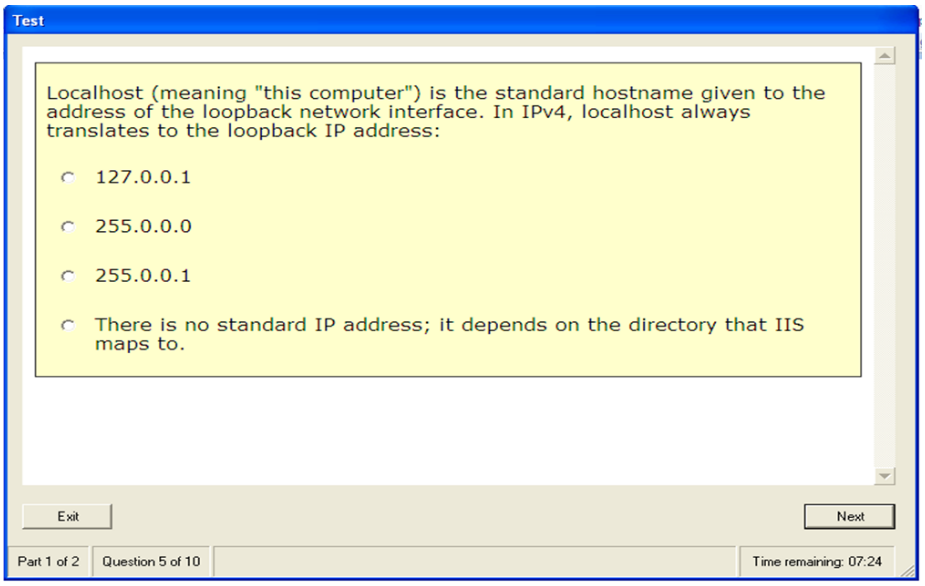
However, it may be expected that some tasks may be attended to with greater engagement and concentration than others, and an important question would be how mobile contexts may affect interactions that require greater engagement with the mobile task (Oulasvirta et al., 2005). This level of engagement is something that seems pertinent to even relatively short mobile e-assessments. A possible implication of this is that students may choose where and when they take formative assessments such that interruptions are less likely.

In respect of evaluation with students, there are two main strands of work reported in this chapter: the usability of software applications running the Flexilevel test in different contexts of use and the attitude students hold towards the approach overall and to its use in mobile contexts.

5.2 Pilot Study

The attitudes of students towards the Flexilevel test were part of the methodology from the beginning of this programme of study. The data reported here was gathered as part of the pilot study (please see section 4.1.1)

The methodology of the test has been noted previously, but here it is reported in terms of the examinee's experience of the test. The two different sets of items were presented by the application with different light background colours (blue for conventional and yellow for Flexilevel) on the question screens (Figure 2-1, please also see Appendix A). This was done to enable the experimenters and participants to refer to the different tests without naming them – participants were not aware what the colours related to, nor that there were two types of test being used. The colours were selected because they maintained a clear contrast between the text of the questions and the background of the screen. Figure 5-1 shows the layout of the question screen – in this case it is presented during the Flexilevel stage.



Localhost (meaning "this computer") is the standard hostname given to the address of the loopback network interface. In IPv4, localhost always translates to the loopback IP address:

○ 127.0.0.1

○ 255.0.0.0

○ 255.0.0.1

○ There is no standard IP address; it depends on the directory that IIS maps to.

FIGURE 5-1: SCREENSHOT OF THE FLEXILEVEL SOFTWARE APPLICATION (HERE USING THE FLEXILEVEL TEST ALGORITHM).

Finally, as can also be seen in Figure 5-1, the software allows for the timing of a test, and displays the time remaining and the number of questions answered throughout the test.

Student attitude to the Flexilevel test. Once the participants had completed the Flexilevel test, an account was provided about the way the Flexilevel test worked. The participants were then asked to fill in a questionnaire that covered their attitudes towards the test and its usability.

Results. As can be seen from Table 5-1, participants were positive when responding to statements about the perceived usefulness of the software application. This was particularly the case with respect to using the Flexilevel test for formative assessment (statement 4), areas that the participants might need to work on (statement 6), what they have learnt (statement 11) and whether participants would use Flexilevel tests if they were made widely available (statement 13).

Statement number	Statement	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree	Mode
4	I would find the Flexilevel approach useful for practice tests.	0	2	2	9	11	**5**
5	I would find the Flexilevel approach useful in summative tests (i.e. the test score counts towards my final grade).	0	1	11	8	4	**3**
6	The adaptivity supported by the Flexilevel approach would help me to identify the areas in which I need to work harder more quickly.	0	3	2	9	9	**4, 5**
7	The adaptivity supported by the Flexilevel approach would enhance my	1	1	6	13	3	**4**

	overall assessment experience.						
8	For practice tests, I would prefer using the Flexilevel approach to other forms of objective testing.	1	4	5	8	6	**4**
9	For summative tests, I would prefer using the Flexilevel approach to other forms of objective testing.	0	5	8	8	3	**3, 4**
10	The system used to score a Flexilevel test makes sense to me.	0	1	4	14	5	**4**
11	I would find the score provided by the Flexilevel approach useful at identifying how much I have learned.	0	4	2	12	6	**4**
12	I would find it useful if the level of difficulty of a test is tailored to my level of understanding.	0	4	3	14	3	**4**
13	Assuming the Flexilevel software was available to me for practice tests, I predict that I would use it on a regular basis.	0	1	7	8	8	**4, 5**
14	In practice tests, test questions that are too easy are less engaging than those questions that are tailored to my level of understanding.	0	1	6	12	5	**4**
15	In practice tests, test questions that are too difficult are less engaging than those questions that are tailored to my level of understanding.	0	6	9	6	3	**3**

TABLE 5-1: PERCEIVED USEFULNESS: MODE FOR THE RESPONSES (N=24)

Participants understood the scoring system (statement 10) with a mode of 4 and judged that it would be useful in identifying how much they had learned (statement 11), again with a mode of 4.

Participants were neutral overall about statement 15 that questions that are too difficult are less engaging than questions that are tailored to their ability (mode = 3). By contrast they mainly agreed with statement 14 that questions that were too easy for them were less engaging than those tailored to their level of ability (mode = 4).

Participants responded postively to the statements relating to the adaptivity offered by the Flexilevel test approach. They mainly agreed or strongly agreed that the adaptivity of the Flexilevel test would help them identify the areas they need to improve on (statement 6: mode = 4, 5) and agreed with statement 7 that the adaptivity of the Flexilevel test would enhance their overall assessment experience (mode = 4). They also agreed with statement 12 that they would find it useful if the difficulty of a test was tailored to their understanding (mode = 4).

These trends are also reflected in participants' responses to open questions.

Participants were positive about the formative use of the Flexilevel test approach. One participant commented that "providing multilevel questions after one another is good". As noted previously, participants believed it could support them in their academic progress: "…would really help to outline where problems in understanding lie and help students to address those areas". Also "…good approach to learning giving better students harder questions".

In terms of summative assessment, it was suggested that: "all students may not be tested equally" and "I think that there would be a smaller gap in marks between good and bad students than in normal tests".

For formative assessment, participants tended to agree or strongly agree with statement 8 that they would prefer it to other forms of assessment (mode = 4).

Usability of the Software Application. A consideration from the beginning of this work was the usability of the software application. In addition to eliciting the views of students to the Flexilevel test itself, another objective of this work was to test the software application that was used to mediate the test in order to gather participants' opinions on the interface of the application.

Initial usability work on the software artefact had involved expert reviews of the software application. These were supplemented by questionnaires that were presented to participants as part of the pilot study and the postgraduate study.

Results. Table 5-2 shows the participants' responses to the questionnaire presented to them in the pilot study. The statements relate to fundamental usability principles (for example as described in Dix, Finlay, Abowd & Beale, 2004), and are intended to identify any potential issues with the interface that may need further investigation. The results indicate that participants found the application easy to use and would be able to both learn how to use it and remember how to use it.

Table 5-1 below shows that the majority of participants agreed with positive statements concerning the ease of use of the software application.

Question Number	**Statement**	**Strongly disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**	**Mode**
1	Learning to use the Flexilevel software application would be easy for me.	1	0	4	15	4	**4**
2	I would find it easy to remember how to perform tasks (e.g. how to answer a question) using the Flexilevel software application.	0	2	3	16	3	**4**
3	I would find the Flexilevel software application easy to use.	0	0	5	16	3	**4**

TABLE 5-2: EASE OF USE RESPONSES (N=24)

Postgraduate participants. In the other exploratory study, the postgraduate students study (section 4.1.2) participants were asked to respond on a scale between 1 (strongly disagree) and 5 (strongly agree) to two statements:

1. I would think that most students would find the application easy to use
2. I found the application easy to use.

As shown in Table 5-3, participants were positive about the ease of use of the software artefact's user interface.

Question Number	Statement	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Mode
1	I would think that most students would find the application easy to use.	0	2	7	12	7	**4**
2	I found the application easy to use.	0	1	7	8	12	**5**

TABLE 5-3: POSTRGRADUATE PARTICIPANTS' RATINGS OF THE EASE OF USE OF DESKTOP APPLICATION'S USER INTERFACE (N=28)

It can be seen from the results that the postgraduate students also found the application easy to use.

Findings. Participants were positive about the Flexilevel test in formative assessment contexts, and valued the adaptivity it affords for assessment. They understood how the scoring system worked - the simplicity of the scoring is one of the advantages of the Flexilevel test over other approaches. Moreover they agreed that they would find it useful to identify how they are progressing in their studies.

Participants were more neutral about the use of the approach in summative contexts (Statement 13), with a mode of 3 and median of 3.5. Interestingly, this seems consistent with the attitudes of students to IRT-CAT approaches reported by Lilley et al. (2004), where the use of IRT-CAT in summative assessments is of concern to them. It is also consistent with the previously discussed idea that summative assessment is treated more conservatively than are other forms of assessment (Bull & McKenna (2003), Knight (2002)).

Another interesting finding was the participants' attitudes to the presentation of items that are not within their proficiency. It has previously been suggested that the presentation of items that are too easy or too difficult for an examinee has a demotivating effect (Carlson (1994), Wainer (2000)). However while participants in this study were not so interested in seeing the easier items, they seemed to want to see the more difficult items.

The usability of the interface was judged positively by the participants in both the pilot study and the postgraduate study.

5.3 Study G

Further information about context of use can be gleaned from the devices students are using to access a given system. This gives a useful indication of the extent to which a given population of users is using mobile devices to access a system, in this case the Flexilevel test application.

This study was concerned with how students engaged with the assessment application given an open choice of which of their own devices they could use to take a test. The test used for the study was authentic, and was derived from an existing test from a Level 4 module the students had recently studied. The development of the test involved the calibration of items based on test data generated during the module itself and used the Flexilevel algorithm to select items depending on performance on the test.

In terms of student engagement, the main point of interest was the extent to which students would engage with the test on mobile devices when given the opportunity. For the purposes of this work, mobile devices are considered to be smartphones and tablets.

Another part of the motivation for the choice of this area was that the items in the set theory test include diagrams and notation that may pose more of a challenge to mobile-delivered tests than other platforms. The items also had five options, which is a departure for the test methodology used up to now.

Methodology. For this study, a test was created based on a module the participants had recently studied. The test covered set theory, an area that was studied as part of the module, although the motivation for this study was to see how students interacted with the test application. They were told that they were able to use the application on any device (i.e. they could use it on smartphone or tablet and it did not need to be taken using a desktop computer). There are limited opportunities for mobile study within students formal studies, and it was felt that it needed to be flagged that the test application could be used on mobile devices if participants so wished.

From this recruitment, 29 participants took part, each answering at least one item in the test.

The test covered the area of set theory and was set up based on a calibrated set of test items that had been used in participants' academic studies during the module. The test was comprised of 13 items and participants were given a time limit of 20 minutes to complete the test.

When the user starts the test, the application logs the user agent string and screen size of the user's device. This information is provided by the user agent, the web browser in this case, itself. The user agent string provides some information about the device being used in terms of the operating system, browser and device. An example of a user agent string is given in Figure 5-2.

```
Mozilla/5 0 (iPad; CPU OS 8_3 like Mac OS X)
AppleWebKit/600 1 4 (KHTML, like Gecko) Version/8 0
Mobile/12F69 Safari/600 1 4
```

FIGURE 5-2: AN EXAMPLE USER AGENT STRING

The screen resolution of the device used was also logged to support the detection of the user agent type (mobile or laptop/desktop computer).

Results. Table 5-4 shows the total usage of mobile device as against laptop or desktop computer.

Type of device	N
Mobile devices	17
Laptop / desktop computer	22

TABLE 5-4: NUMBER OF DEVICES USED BY PARTICIPANTS (N=29)

Table 5-5 shows the number of items participants attempted in the test. Twenty-three participants took the full test whilst six others answered five or fewer items.

Total Items Attempted (out of 13)	Number of participants
13	23
5	1
4	1
3	1
2	1
1	2

TABLE 5-5: NUMBER OF ITEMS ATTEMPTED BY PARTICIPANTS (N=29)

Table 5-6 shows the progress made by test-takers using mobile devices and those using PCs/laptops. It can be seen that a larger proportion of PC/laptop test-takers than mobile test-takers completed the test (89% completion and 64% completion respectively).

Test progress	Mobile	PC / laptop
Started test	11	18
Completed test (13 items)	7	16

TABLE 5-6: PROGRESS OF TEST-TAKERS BY DEVICE TYPE

In terms of the mobile devices in Table 5-6, five of the participants that completed the test used smartphones and two used tablets.

Table 5-7 shows the descriptive statistics for the test where participants answered all 13 items; the average scores are high – in formal terms within the upper second to first class range.

Test descriptive statistics	
Mean	8.95
Standard deviation	2.49

TABLE 5-7: DESCRIPTIVE STATISTICS FOR TEST (N=23)

Findings. Twenty-nine participants attempted the test, of whom 23 completed all 13 items. The scores achieved by participants were high, supporting the notion that the students were treating the test seriously.

The dropout rate was higher for participants using mobile devices than for those using a PC/laptop although it was not possible to detect why participants dropped out given the limited data collected. There are at least two possible alternatives. First it is possible that given the complexity of some of the items mobile users were more inclined to drop out rather than try to complete the entire test. Secondly, there may have been something occurring in the mobile context participants may have been in at the time they took the test. The finding is consistent with the idea that mobile contexts are likely to be more distracting than more static contexts (Oulasvirta et al., 2005).

The extent to which the different devices can be accurately and reliably identified using the user agent string is somewhat limited. A full discussion of the use of user agent strings for user agent identification is beyond the scope of this report, but for the purposes of this study it was sufficient to distinguish between mobile devices and laptop/desktop PC platforms, this being possible using the user agent string and screen resolution. It would be of interest for a future study to attempt a more refined differentiation of devices used and perhaps map the definitions of mobile assessment with what this implies for the physical properties of the devices.

The study was intended to obtain a snapshot of how a group of students may interact with an assessment application in formative contexts. Eleven out of 29 participants chose to take the test on a mobile device, indicating that students do engage using mobile devices. Although this study had relatively few participants, its results were consistent with other evidence, anecdotal and in terms of the ubiquity of mobile devices in contemporary educational contexts. This in turn supports the idea that taking formative assessment in mobile contexts may be acceptable to students.

This points to the importance of determining how and when mobile devices can be best used in formative assessment if students are to be encouraged to use them in the learning ecosystem.

5.4 Study I

The personas inform the understanding of the BSc. online student population who may use the Flexilevel test application, but in order to understand students' views about the approach and its current implementation, a more specific study was designed. This investigated the extent to which participants would be willing to make use of the Flexilevel test in a formative mobile educational context. To understand how students may view a mobile Flexilevel test, interviews were carried out with 10 students.

Participants. Groups of students were identified on a module and level basis and were asked to take part in interviews about the Flexilevel test. Ten students (eight Level 7 (Masters) and two Level 4 (first year undergraduate) students) were recruited in this way.

All had prior experience of the Flexilevel test either through their studies or by taking a test online as part of another study. Two of the participants also asked to see the application in advance and were provided with sufficient login credentials to try the application out as many times as requested. Both had tried it out on mobile devices prior to the actual interview.

Methodology. All but one of the interviews took place online. The format of the interviews was to provide a briefing for the interviewee explaining the scope of the discussion; it was a semi-structured interview (Lazar et al., 2010) organised into three sections:

1. Participants' use, if at all, of networked and especially mobile technologies in their studies
2. Participants' views about the Flexilevel test approach
3. Participants' views about the Flexilevel test approach as delivered on a mobile device.

It was emphasised that they did not need to have any knowledge of the approach or, when demonstrating the test itself, the subject matter of the demonstration tests.

The interviews were recorded to facilitate the production of notes for each interview. The content of the interviews was then coded to identify key themes based on the questions detailed above.

Results. The results are reported below in terms of the participants' views of and use of mobile devices, their views of the Flexilevel test approach and of the application adapted to run Flexilevel tests on mobile devices.

Mobile use. Table 5-8 shows the devices owned by participants.

Device owned	**Number owned**
Smart phone	10
Laptop	10
Tablet	7
Desktop Computer	6
iPod	1

TABLE 5-8: COMPUTING DEVICES OWNED BY PARTICIPANTS (N=10)

All participants had a smartphone and most had a tablet. Participants were also asked about their use of laptops and other devices were included and it can be seen that all participants owned a laptop (one of which was a MacBook), six owned a desktop computer (five owned a PC and one a Mac). Participants had multiple devices as is evident from the total number of devices mentioned.

In the cases of both the smartphone and tablets, participants were asked to identify the brand and model in order to class the devices consistently. It was decided that a distinction beyond tablets and smartphones did not provide any further insight into the use of mobile devices in this study; there was no difference evident in the descriptions of the usage of somewhat larger and somewhat smaller smartphones. The main distinction within the mobile devices used was between smartphones and tablets.

Most participants had a preferred form of technology that they used, so when they reviewed work on their mobile device it tended to be a smartphone or tablet, but not both. This is reflected in the patterns of usage identified later in this section. The instances of use in Table 5-10 are largely made up of different participants rather than the same participant carrying out the same learning activity on different devices.

Participants were asked if they used mobile technology in their studies already, as can be seen in Table 5-9.

Mobile devices used in studies	**N**
Yes	8
No	2

TABLE 5-9: STUDENTS USING MOBILE DEVICES IN THEIR STUDIES (N=10)

Those that used mobile technology in their studies were asked for the context in which they used it. This is shown in Table 5-10.

Use of mobile device	Smartphone	Tablet	Total
Reading	4	4	8
Research (including internet searches, finding definitions for words)	5	1	6
Review (including reviewing web based learning materials, reading discussions)	3	2	5
Organisation (including planning, checking schedules, checking assignments details)	3	2	5
Checking emails	2	2	4
Light editing (including correcting typographical errors in essays)	2	1	3
Note-taking	1	2	3
Listening to audio books	1	0	1
Taking photos of hand-drawn diagrams	1	0	1
Ad hoc communication (student–student)	1	0	1
Online individual meetings with tutor	0	1	1
Online synchronous group study sessions	0	1	1
Essay writing	0	1	1

TABLE 5-10: LEARNING ACTIVITIES CARRIED OUT BY STUDENTS USING MOBILE DEVICES (N=8)

Reading is the most common learning activity engaged in on mobile devices, with eight instances reported. Although one participant asserted that smartphones were not suitable for reading, another participant indicated that they would like to do more reading using their smartphone, but that it was not feasible when e-books were produced in portable document format (pdf). Research is the next most common use of mobile devices with six instances of use reported. Review and Organisation each saw five instances.

Content creation activities are further down in terms of instances of use reported, with light editing and note-taking each having three instances of use.

Participants were also asked specifically about the benefits of the mobile technology they used. Table 5-11 shows the number of responses offered. The main benefits reported are portability and device(s) being close by, both with 3 instances.

Benefit of mobile technology	**N**
Portability – easy to carry about/can use it anywhere	3
Device(s) close by/handy	3
Always on/don't have to boot up	3
Being able to catch up	1
Always getting updates/podcasts	1

TABLE 5-11: WHAT ARE THE MAIN BENEFITS TO USING YOUR MOBILE DEVICE? (N=8)

One participant commented that a benefit of having the device close by is being able to post notes online “if something pops into your head”. Another found their mobile device enabled them to catch up quickly “stuff that involves quick checking, going online to see what’s going on, mobile devices are very helpful for that”.

Another participant valued “being able to do little bits and pieces”, “to do a little bit of studying just kind of in between things which I wouldn’t be able to do [on the laptop]”. They took their laptop with them sometimes, but the tablet was considered to be easier to take out as it was more mobile.

Participants were asked about any issues they had with use of their mobile technology or issues that prevented them using mobile technology if they did not. Table 5-12 shows the responses of the participants. Where they are specific to smartphones or tablets, this is noted. The small screen of the smartphones was the issue most often mentioned, with network issues being the next most common issue. The difficulty of interaction relates to both smartphones and tablets. For smartphones the issue surrounds the interaction in general, one participant describing it as “a bit small and a bit fiddly” and another suggesting that for going online to pick up bits of information, the “screen size is perfectly ok”, but if it involved downloading a document to edit, then screen size can become an issue.

It will be noted that laptops and PCs are not considered directly in this study but represent an important perimeter around the use of mobile devices; most participants used laptops for most of their studying (this is discussed in more detail in the patterns of use). However the disadvantages of their use can be seen in the statements above in terms of portability, device(s) close by/handy and always on/don’t have to boot up. As one participant responded, it was “so handy not having to boot up”.

Issue with mobile technology	**N**
Small screen (smartphone)	6
Network issues (coverage, speed)	5
Difficulty of interaction	4
Battery life	3
Existing environment	3
WiFi access	2
Privacy/security	1
Text to speech is poor	1

TABLE 5-12: ISSUES WITH MOBILE USE (N=10)

For some participants, the appropriateness of using a smartphone for studying at all was in question, one saying “I can’t imagine myself studying on the mobile actually” and another commented that their smartphone was a “personal device” rather than an academic one.

The existing environment refers to substantial portions of the participants’ learning ecosystem: issues with e-books and the Virtual Learning Environment when using both smartphones and tablets.

It should also be noted that whilst some participants identified the issues listed, others found them not to be problematic. Battery life was not an issue for two other participants, although one of these was systematic in keeping devices charged. Also, issues with network access varied between participants: one of the participants could find themselves without network access for weeks, whereas another would always be able to find good network signal.

Participants were also asked about the contexts in which they used their mobile devices to study. Table 5-13 shows participants’ responses and the number of times they were given.

Context	**N**
Commute/travelling (plane and train)	6
Home	2
Work	2
Waiting (for example to pick someone up)	2
Taking dogs for a walk (smartphone)	1
Cooking (smartphone)	1
Gym (smartphone)	1
Learning Resources Centre (smartphone)	1
Halls of residence (smartphone)	1

TABLE 5-13: CONTEXTS FOR MOBILE STUDY (N=8)

It can be seen that the context that was mentioned most often, six times, was the commute to work, including occasional commute by plane. Commutes were generally reported as being 45 minutes to an hour long, although for one participant this could stretch to a four- or five-hour bus ride if traffic was bad.

Further, even though using mobile devices, well established learning environments – the Learning Resources Centre and halls of residence – are represented, albeit by one participant. Similarly other locations where it seems likely there would be opportunities for studying using either PCs or laptops easily including home and work were each mentioned twice.

Also of interest in terms of the context of use was the extent to which participants had planned their studies in a given context. There is an almost even balance between planned and unplanned study reported by participants. Planned study is when a student knows they will have some time to study whilst mobile, although they may not be sure they would be able to (for example if there was no network signal). So planned study will usually also involve preparation, for example a student pre-downloading learning materials to a mobile device when they are on a good network connection in anticipation that they won't have one when they are ready to study. Participants reported that they were careful about when and where they study to make best use of their mobile devices. Planning for when they expect network coverage to be patchy was a well-represented example of this.

Also, when the option existed, participants not only planned what to study, but where. For one participant, studying at home is sometimes not an option so they go to their parents' house or a public library where they know they will be able to concentrate. In this case, the participant made the point that when they are able to plan their studies, they use their laptop. The interaction between planning and context may allow participants to make use of their laptops despite of the overheads already mentioned (for example booting up), rather than using mobile devices.

Unplanned study might involve taking advantage of an unexpected 30-minute break to do some reading or research whilst waiting for someone or, in one case, using a smartphone to correct typographical errors in an essay whilst walking the dogs.

Flexilevel test. Part of the discussion concerned the Flexilevel test. Participants were asked to identify positive points about the Flexilevel test approach and negative points.

Table 5-14 shows the positive points made about the Flexilevel test approach.

Positive points about the Flexilevel test approach	**N**
Flexilevel test concept is good overall	4
Adaptivity of the approach	2
Fixed length	2
Seeing only one question at a time	1
Not being able to go back	1
Personalisation of feedback	1

TABLE 5-14: PARTICIPANTS' POSITIVE VIEWS OF THE FLEXILEVEL APPROACH OVERALL (N=9)

Overall, participants were positive about the Flexilevel concept, with nine participants offering positive comments: "there are those who are good and they don't like having tests which are very very easy". Another participant thought it was a "very good approach" and liked the idea of "the next question depending on last". Another participant considered the presentation of one question at a time to be a positive feature: "I think it is good because it makes you more careful".

One participant who had been assessed with a variable length adaptive test for a professional qualification observed that it was a positive point, "there is a comfort with knowing how many questions".

The potential for personalised feedback was also considered a positive.

Table 5-15 shows responses to the request for negative points about the Flexilevel approach.

Negative points about the Flexilevel test approach	N
Miss out on difficult questions	6
Didn’t like that you couldn’t go back	2
Wouldn’t use it for summative assessment	2
Miss out on questions (either easy or difficult)	1

TABLE 5-15: PARTICIPANTS’ NEGATIVE VIEWS OF THE FLEXILEVEL APPROACH OVERALL (N=8)

Eight participants offered at least one negative point about the Flexilevel test. Interestingly most were concerned with missing out on difficult questions. The point about not being able to go back in the approach (nor in the demonstration application) seems to relate to the way in which conventional CBTs are often presented as sets of questions on a page. As one participant put it, being able to “flag a question and then go back to it” might be useful. However, this attracted only two comments and as shown in Table 5-14, not being able to go back was also considered to be a positive feature of the Flexilevel test approach.

Scoring. Another point of interest in terms of participants’ attitude towards the Flexilevel test was their view on how the test should be scored. As part of the demonstration, participants were shown the final criterion-referenced scoring in which the score achieved by examinees is shown as the number of questions correct, as well as a percentage.

Participants were asked what they thought of the scoring. This generated much discussion, something that is evident in the number of responses shown in Table 5-16. The score depended on the context of the test, summative or formative and also on other feedback that was provided with it. Weighting the scoring based on the difficulty of the items answered correctly was consistently mentioned as an option for scoring Flexilevel tests. Weighting of the scores would also allow for a more engaging, ‘gamification’ of the system (“If I'm able to go through the first level of the game, I would be so much encouraged to go through to the second level”) and one that the participant could play on their phone. Further, in terms of engaging students it would be something that offered more than just reading. Students would have something to do.

In this sense, rather than number-based scoring, the result was progress or no progress; references to level and the use of a golfing analogy ("you are right on par for what we would expect") exemplify this theme. For some participants, the emphasis was on the weighted score, for others it was on the potential for gamification, of which weighting the score was a part. These are represented separately in Table 5-16 to reflect this different emphasis.

Scoring	N
"Number right" score	7
Weighted score	5
Game-based score	4
Other forms of feedback more important than score	2

TABLE 5-16: PARTICIPANTS' RESPONSES TO QUESTIONS ABOUT THE SCORING SYSTEM (N-10)

Moreover, with reference to Table 5-15, if this was coupled with informing students of the difficult items they had not been shown it was considered that this would provide a real motivation for students to progress their studies.

Test length. Here the possibility of providing shorter tests was introduced to participants. Table 5-17 shows the responses.

Views on shorter tests	N
Positive	7
Neutral	2
Negative	1

TABLE 5-17: PARTICIPANTS' RESPONSES TO THE POSSIBILITY OF HAVING SHORTER TESTS (N=10)

Participants were positive about the possibility of having shorter tests. In the context of travelling, for example, "you will know that you can complete a test before you reach the end of your destination."

Web application. When participants were asked if they would use the Flexilevel web application in their studies, the responses in Table 5-18 show a strongly positive response. Nine participants stated they would use the application in their studies if it was available. One participant stated “oh yeah, definitely”.

Use of web application in studies	N
Would Use	9
Neutral	1
Would not use	0

TABLE 5-18: PARTICIPANTS ATTITUDE TO USING WEB APPLICATION IN THEIR STUDIES (N=10)

Useful to your studies. Participants were asked to identify how the Flexilevel web application would be useful to their studies. Table 5-19 shows that participants were positive about using the application in mobile assessment contexts. Indeed one participant observed that they could see how you could takes formative tests “pretty much anywhere”.

How would web application be useful in studies	N
Mobile assessment	9
Formative assessment	7
Summative assessment	4
Feedback	3
Improved engagement with studies	1

TABLE 5-19: USEFULNESS OF THE FLEXILEVEL E-ASSESSMENT APPLICATION (N=9)

One participant who had tried out the application on a smartphone reported that they were “able to use it on small mobile phone”

Four participants indicated they would be willing to take summative assessments using the application in mobile contexts, although one participant would not want to use it for summative assessment that was worth more than 25% of the module. Two participants stated they wouldn’t want to use it for summative assessments and another participant would “definitely use the application”, but would find a quiet spot to do it and would prefer not to use it for a scored test, but to obtain other forms of feedback.

It was suggested that the web application could be useful for the provision of quick feedback, including feedback that identifies areas of strength and weakness.

Context is considered to be of importance here. Where before participants might study in varied contexts, for assessments, there is an evident tendency to find an appropriate spot to do an assessment; as one participant put it they would do it on a mobile device, but "would give it my full attention and do it in a quiet room". Also, doing a summative assessment on a mobile device was not a popular option.

Not useful to studies. Overall, as shown in Table 5-20, there were fewer statements about how the web application would not be useful in the participants' studies than those made about how it would be useful.

Assessment web application not useful to your studies	Totals
If the course is more practical, more programming	1
If it is not aligned with other forms of assessment	1
If the test did not fit on the screen	1
If it is slow to load	1
If you can tell if questions are getting easier or harder	1
Only a score provided for feedback	1
The time limit being imposed	1
You only get one opportunity to take the test	1
If students are unaware of how the test works	1

TABLE 5-20: PARTICIPANTS' NEGATIVE VIEWS OF THE FLEXILEVEL E-ASSESSMENT APPLICATION (N=8)

Also the comments made about how the web application would not be useful to participants' studies were sometimes hypothetical (e.g. if a stated problem was included in the test, it would have a negative effect). Of the three statements that related directly to the web application, one related to the use of a timer for a formative test, the participant wondering "why it was necessary to give a time limit", another to the fact that the application only allows one test per participant (login) and finally, the feedback taking the form of a score.

Of value to the participants seems to be the opportunity to attempt the test as many times as they like with intervening feedback about what they got wrong and how they can improve their performance before taking the test again.

Findings. It seems that participants in this study make use of mobile devices for educational experiences that are relatively brief and between other activities; the commute to work is an example of this.

It also seems that participants carefully plan out the nature of the learning activities they will engage in with mobile devices and when they will revert to their laptops. Mobile devices are closely integrated with their use of other technologies but participants indicated that they do the bulk of their studies on fully capable computers, predominantly in the form of laptops. Desktop computers did not often feature in the discussions – although Table 5-8 shows six participants have access to desktop computers they are only being used on relatively rare occasions.

Most of the learning activities carried out on mobile devices do not include significant content creation. This seems to be because of the difficulty that participants had with typing both on smartphones and tablets.

Also, the constraints on their use of mobile devices are not limited to the devices themselves. There is also the issue of how educational technologies are implemented for mobile devices within Universities. Whilst screen size was the most often reported issue, it was pointed out that it is also the way the screen is used, something emphasised by another participant who suggested the problem is when applications don’t scale for mobile display.

This applies to both smartphones and tablets. Issues with support for mobile devices in current VLEs and existing infrastructure apply more broadly; one participant also reporting similar problems with their own VLE. The participant had attempted to access these systems on mobile devices but the experience had been poor so now they don’t even attempt to use the systems. This represents an important influence on their expectations for other systems on mobile devices. It also indicates the potentially damaging effect of poor mobile experiences on the expected uptake of something like a Flexilevel test delivered on a mobile device.

The combination of the learning activities, contexts, the devices used as well as the extent to which the whole learning experience is planned provides some patterns of study that are of interest for identifying how participants are integrating their mobile devices in their overall studies.

Patterns of study

1. Most studying is done using mobile technology, mainly tablet, but smartphone used for opportunistic learning. Laptop only used for specific tasks.
2. All studying is done using a laptop.
3. Before going to work do the necessary preparatory work using smartphone, when in the office switch to laptop, or work PC, to do the bulk of the studying.
4. When on train, do the preparatory work on smartphone, then transfer to laptop when at home or office to do bulk of the studying.
5. Mobile learning without mobile technology: read book on train, annotate with Post-it notes, take notes home do more online research, write extra notes and type them up on a laptop.
6. Work on smartphone up to the point of needing to do typing.
7. Check what studies need to be done using a mobile device in spare gaps to plan what studying to do when a more a more capable device is available.
8. Laptop on the go and tablet used together, laptop for typing, tablet for reading; 30 minutes quick study with tablet then go into serious work on laptop.

This study provides support for the notion that students would be willing to take Flexilevel tests in formative mobile contexts. Indeed it seems the tests would fit in well with the patterns of study participants reported that they engaged in.

5.5 Study J

The System Usability Scale (SUS) (Brooke, 1996) was devised to provide an efficient means of capturing users' perception of the usability of system. The SUS is a ten question survey that alternates positive and negative statements. Alternation was included in the SUS to encourage participants to really think about their answers (Hartson & Pyla, 2012), and was standard practice at the time of its publication to deal with acquiescence bias (Brooke, 2013). However there are two possible issues with alternating items: user selection error and coding error (Sauro & Lewis, 2011). Sauro & Lewis (2011) go on to propose a positive SUS that shows similar results as the classic SUS, to use the term used by Brooke (2013).

It produces a single score based on the responses of participants, something that is of value when communicating results (Hartson & Pyla, 2012). This is not straightforward given the context sensitivity of most usability studies (Brooke (1996), Sauro & Kindlund (2005)) but a score derived from questionnaires – in this case, SUS – seems to provide a means of doing this (Bangor et al. (2008), Sauro & Lewis (2012)). Additionally, the scores for individual items are not in themselves meaningful. They contribute to the overall score but should not be taken in isolation (Brooke, 1996), and only SUS scores should be reported (Bangor et al., 2008).

It should be noted however that there is some doubt that the SUS is unidimensional, for example Lewis & Sauro (2009) propose two factors – usability and learnability – that can be discerned. Nonetheless, they propose that this provides more information for the use of the existing SUS rather requiring a significant change to the SUS. Further, the SUS has been used extensively in usability practice and has been shown to be reliable across a range of usability evaluation contexts (Bangor et al., 2008), something that lends itself to usability studies of an application across different devices. In terms of this study, this is beneficial because although this is a test of the application as used on mobile devices, it could have the potential to offer a reliable insight into the varying user satisfaction with the application in different contexts, including desktop.

Additionally the individual score for the SUS has been used reliably and successfully for some time and the ease with which the score can be communicated to stakeholders remains an important motivation for using the survey (Hartson & Pyla, 2012).

As a means of enhancing the ease with which results of a SUS may be communicated, developments have been introduced including the use of appropriate adjectives to describe the results, for example, “good”, “awful” (Bangor, Kortum, & Miller, 2009) and in terms of university grades (Bangor et al., 2008).

Similarly to the SUS, the Usefulness, Satisfaction and Ease of use (USE) questionnaire provides an evaluation instrument that is applicable to a range of different contexts. The Usefulness, Satisfaction and Ease of use refer to the factors that this survey evaluates (Lund, 2001).

The questionnaire uses a seven-point Likert scale with positive statements. The scale ranges from 1 for ‘strongly disagree’ to 7 ‘strongly agree’.

Methodology. As part of their studies on a Level 7 Interaction Design module, ten students were tasked with taking a test in the domain of numbering systems using the e-assessment application on a mobile device and then to complete SUS and USE questionnaires. The use of mobile devices was verified through the collection of information about the user agents that were used; all participants ran the e-assessment application at least once on a mobile device.

The questionnaires were delivered using Bristol Online Survey (BOS).

For the SUS surveys the standard calculation was carried out. The item scores are converted from the 1–5 scale in the survey to a 0–4 scale. For items that have positively phrased statements, the item score is calculated by subtracting 1 from the option selected by the user. For items that have negatively phrased statements, the item score is calculated by subtracting the option selected from 5.

This gives a set of items sharing the same scale. The individual scores for each item are summed and then, to give a score out of 100 rather than 40, they are multiplied by 2.5. This is done for each participant in the study and the scores can then be averaged to provide a SUS score for the system.

For reporting the USE questionnaire, standard response frequency, mode and median scores may be used.

Results. Table 5-21 shows the SUS scores from each participant and the average SUS score that was calculated.

Participant	SUS Score
1	85
2	72.5
3	70
4	85
5	65
6	87.5
7	82.5
8	75
9	70
10	77.5
Average SUS score	**77**

TABLE 5-21: SUS SCORES GIVEN BY PARTICIPANTS (N=10)

The average SUS score produced in this study is 77, placing it above 68 which is often cited as an average SUS score (Sauro & Lewis, 2012). When mapped to the adjectives suggested by (Bangor, Kortum, & Miller, 2009) the SUS score of 77 places it in the good range. It also gives it a C rating (Bangor et al., 2008).

Table 5-22, Table 5-23: USE responses about the ease of use of the mobile web application (N=10), Table 5-24: USE responses about the ease of learning of the mobile web application (N=10) and Table 5-25: USE responses about participants' rating of their satisfaction with the mobile web application (N=10) shows the responses of participants to the USE questionnaire together with the mode responses for the different items in the USE questionnaire.

Table 5-22 shows the participants' responses to the usefulness items. It can be seen that participants were somewhat positive in their responses. The applicability of some statements may be greater than that of others, but there is support for the web application being useful to the participants. In particular getting things done and being more effective and productive attracted positive responses, as did the statement "It is useful".

USEFULNESS

3. Please rate each of the statements below.	**Strongly Disagree**						**Strongly Agree**	**mode**
	1	**2**	**3**	**4**	**5**	**6**	**7**	
3.1.a. It helps me be more effective.	1	1	0	2	4	1	1	5
3.2.a. It helps me be more productive.	1	1	0	2	4	1	1	5
3.3.a. It is useful.	0	0	0	1	6	1	2	5
3.4.a. It gives me more control over the activities in my life.	2	0	0	4	2	2	0	4
3.5.a. It makes the things I want to accomplish easier to get done.	2	0	0	1	3	4	0	6
3.6.a. It saves me time when I use it.	2	0	0	1	4	1	2	5
3.7.a. It meets my needs.	2	0	0	3	2	2	1	4
3.8.a. It does everything I would expect it to do.	0	1	0	1	4	3	1	5

TABLE 5-22: USE RESPONSES ABOUT THE USEFULNESS OF THE MOBILE WEB APPLICATION (N=10)

Table 5-23 shows the results for the ease of use section of the USE questionnaire. The items “It is easy to use”, “It is simple to use”, “I can use it without written instructions” and “I can use it successfully every time” showed positive responses.

EASE OF USE

4. **Please rate each of the statements below**.	**Strongly Disagree**						**Strongly Agree**	**mode**
	1	**2**	**3**	**4**	**5**	**6**	**7**	
4.1.a. It is easy to use.	0	0	1	0	2	4	3	6
4.2.a. It is simple to use.	0	0	0	0	1	4	5	7
4.3.a. It is user friendly.	0	0	1	3	1	4	1	6
4.4.a. It requires the fewest steps possible to accomplish what I want to do with it.	0	0	0	1	2	4	3	6
4.5.a. It is flexible.	1	0	1	3	4	1	0	5
4.6.a. Using it is effortless.	0	0	0	0	5	4	1	5
4.7.a. I can use it without written instructions.	0	0	1	1	1	4	3	6
4.8.a. I don't notice any inconsistencies as I use it.	1	0	1	4	1	1	2	4
4.9.a. Both occasional and regular users would like it.	0	1	1	0	4	2	2	5
4.10.a. I can recover from mistakes quickly and easily.	2	1	2	2	0	2	1	6
4.11.a. I can use it successfully every time.	0	0	1	2	2	1	4	7

TABLE 5-23: USE RESPONSES ABOUT THE EASE OF USE OF THE MOBILE WEB APPLICATION (N=10)

Table 5-24 shows the responses to ease of learning items, again there were consistently positive responses including “I learned to use it quickly”, “I easily remember how to use it”, “It is easy to learn to use it” also consistently attracted very positive responses with modes of 7.

EASE OF LEARNING								
5. **Please rate each of the statements below**.	**Strongly Disagree**						**Strongly Agree**	**mode**
	1	**2**	**3**	**4**	**5**	**6**	**7**	
5.1.a. I learned to use it quickly.	0	0	1	0	1	3	5	7
5.2.a. I easily remember how to use it.	0	0	1	0	1	3	5	7
5.3.a. It is easy to learn to use it.	0	0	1	0	0	4	5	7
5.4.a. I quickly became skilful with it.	0	0	1	0	2	3	4	7

TABLE 5-24: USE RESPONSES ABOUT THE EASE OF LEARNING OF THE MOBILE WEB APPLICATION (N=10)

Table 5-25 shows the participants’ responses to statements about their level of satisfaction with the application. It can be seen that participants were positive in terms of the statements “I am satisfied with it” and “it works the way I want it to work”. There is less support for statements about the system being ‘wonderful’ or ‘fun’.

SATISFACTION

6. **Please rate each of the statements below.**	**Strongly Disagree** **1**	**2**	**3**	**4**	**5**	**6**	**Strongly Agree** **7**	**mode**
6.1.a. I am satisfied with it.	0	1	1	0	4	3	1	5
6.2.a. I would recommend it to a friend.	1	0	1	4	2	2	0	4
6.3.a. It is fun to use.	1	0	2	5	1	1	0	4
6.4.a. It works the way I want it to work.	0	1	1	3	4	1	0	5
6.5.a. It is wonderful.	1	0	3	3	2	1	0	3, 4
6.6.a. I feel I need to have it.	1	0	4	4	1	0	0	3, 4
6.7.a. It is pleasant to use.	0	1	2	4	1	1	1	4

TABLE 5-25: USE RESPONSES ABOUT PARTICIPANTS' RATING OF THEIR SATISFACTION WITH THE MOBILE WEB APPLICATION (N=10)

Findings. The results are positive overall for the usability of this mobile Flexilevel test. The SUS score and results from the USE questionnaire indicate that the system is considered to be usable by participants.

The study provides a useful basis for usability work; the SUS and USE questionnaires are well established, especially in the case of the SUS, and have been applied in full to this study. It would be of interest to run this study with more participants, although it is worth noting that the participants in this study were Level 7 Interaction Design students so have greater expertise and insight into the usability of systems than other groups of participants. Also, the SUS can be effectively used with 10 participants.

The results are taken to support the notion that usability of the Flexilevel test running on mobile devices is acceptable to students and points to the potential for the Flexilevel test to be run on mobile devices.

5.6 Discussion

The studies reported in this chapter explore the question of where the potential applications for Flexilevel testing in HE contexts are and what the response is from the users. To address this question the attitudes of students to the Flexilevel test were investigated, additionally the usability of the software applications was also investigated.

1. **Students are positive about the use of the Flexilevel test as a method of assessment**

   This is an important finding because the acceptability of the approach by stakeholders is a key part of identifying potential applications of the Flexilevel test. In particular, students are positive about the Flexilevel test for formative assessment. They are also willing to take the Flexilevel test in summative assessments.

   This component of the research was investigated as part of the Pilot Study (section 5.2, Study I (section 5.4), and Study J (section 5.5).

2. **Students are positive about taking formative assessments in mobile contexts using the Flexilevel test**

   This is taken to confirm that the application of the Flexilevel test in mobile contexts would be acceptable to students.

   Additionally, the use of shorter tests was received positively by students. This would further enhance the potential for the Flexilevel test to be applied in mobile contexts.

   This component of the research was investigated in Study I (section 5.4).

3. **Students want to see the difficult items**

   Students consistently reported that they wanted to see those items that were more difficult than their level of performance. It seems that they may not be affected by being faced with harder questions in tests.

   Formatively students see the difficult items as an important benchmark of how they are performing. So even if they do not see them in a given test, students want to know that they didn’t see them and have the opportunity to take them, perhaps after further study.

   This is also an important finding as it relates to part of the educational rationale for the use of adaptive testing that the presentation of items that are too easy or too difficult is considered to be an issue in examinees’ test performance.

   This component of the research was investigated as part of the Pilot Study (section 5.2), Study I (section 5.4), and Study J (section 5.5).

4. **The usability of the applications used in this programme of research is acceptable to students**

   This is taken as an indication that students would be willing to use the applications to take assessments. Additionally it indicates that the applications themselves did not have an extraneous influence on the results of the studies conducted.

   This component of the research was explored in the Pilot Study (section 5.2), Study I (section 5.4), and Study J (section 5.5).

5. **Students use mobile devices to take formative Flexilevel tests**

   When students are given an opportunity to use a device of their choice to take a test, a substantial minority use their mobile devices. This is taken to indicate that students are willing to take formative Flexilevel tests on mobile devices.

This finding provides a useful addition to the positive attitude of students to the use of Flexilevel tests in formative mobile assessment contexts. Students both exhibit a positive attitude to the use of these tests and actually engage with them when the opportunity arises.

This component of the research was investigated as part of Study G (section 5.3)

6. **There was interest in applying the Flexilevel algorithm to the gamification of educational experiences.**

   A consistent point of discussion in interviews with students was the possibility of using the Flexilevel algorithm to support more game-like educational activities. This is an interesting finding that represents a range of possibilities for the application of the Flexilevel test algorithm, both in the selection of items in a test and also potentially the selection of more substantial educational experiences. As an example, the selection of an easier or more difficult level to route a student through a set of educational challenges (an example being coding challenges for Computer Science students).

   Students were engaged with the possibilities and this provides support for the idea that this is an area that the Flexilevel test could be applied to. This component of the research was explored in the Study I (section 5.4).

The results of these studies support the idea that students would accept the Flexilevel test as a formative assessment that could be taken in a range of contexts of use including mobile contexts of use.

Further, the results support the idea that students would accept the Flexilevel test as a means of enhancing existing educational experiences through the differentiation of challenges such that students are challenged to attain goals of different levels of difficulty.

Finally, the results of these studies point to the value of re-examining the idea that students' may be detrimentally affected by the presentation of items that are beyond their current proficiency in a given domain in both summative and formative assessment contexts.

6 EVALUATION WITH ACADEMIC STAFF

Tutors are primary stakeholders in any assessment approach, and as such it was important to gain an understanding of their attitudes towards Computer-Adaptive Testing (CAT) in general and the Flexilevel test approach in particular.

Study K (section 6.1) reports the results of a survey that is concerned with the attitudes to and usages of CAT in the Higher Education sector generally. This provides a background for more specific subsequent studies: Study L (section 6.2) and Study M (section 6.3) concerning the Flexilevel test. There is also a discussion about the possibilities for providing tailored feedback based on the Flexilevel test.

6.1 Study K

To understand the attitudes of academic staff to CAT generally, a survey was set up using an online survey application (SurveyMonkey, 2010). The survey (Appendix G) was made available to colleagues within the University of Hertfordshire and was publicised through the Higher Education Academy (HEA) and the Association for Learning and Technology (ALT).

In addition, targeted recruitment involved the identification and contact of colleagues in the sector who were prominent in the field. Results of this study were reported in (Lilley, Pyper & Wernick, 2011).

Participants. Sixty-nine participants responded to the survey. It can be seen from Table 6-1 that, given a choice of Higher Education Academy subject centres, most participants taught within the Information and Computer Sciences subject centres.

Subject Centre	N
Art Design Media	2
Bioscience	2
Business Management Accountancy and Finance	4
Education	5
Engineering	3
English	1
Geography, Earth and Environmental Sciences	1
Information and Computer Sciences	46
Maths, Stats & OR	3
Medicine, Dentistry & Veterinary Medicine	1
Social Policy and Social Work	1

TABLE 6-1: DISTRIBUTION OF PARTICIPANTS ACCORDING TO SUBJECT AREA (N=69)

Further, 67 of the participants also stated their Higher Education Institution (HEI) and this showed they were associated with 37 HEIs.

Methodology. Participants were routed through the survey depending on their use or non-use of CAT. Participants were presented with up to 23 questions. The questions covered:

- their current practice in terms of assessment,
- their attitude to the use of objective testing in general,
- their level of knowledge of CAT,
- their usage of CAT, and
- their attitude towards CAT.

Results. Participants were asked to select all options that applied to their current formative and summative assessment practice.

Assessment method	N
Objective tests (e.g. MCQ, MRQ, fill- in- the blanks)	44
Practical projects	37
Short answer questions	32
Essays	30
Group projects	22
Portfolios	17
Peer assessment	2
Presentations and/or seminars	2
Problem-Based based learning	2
In-class Socratic dialogue	1
Online discussion forums	1
Reflective journals	1
Self-assessment	1

TABLE 6-2: APPROACHES TO FORMATIVE ASSESSMENT USED BY PARTICIPANTS (N=69)

It can be seen from Table 6-2 that objective tests are the most selected option for formative assessment. In Table 6-3 it can be seen that practical projects, exams and essays were selected more often than objective tests for summative assessments.

Assessment method	N
Practical projects	50
Exams	47
Essays	41
Objective tests (e.g. multiple-choice, multiple-response, fill in the blanks)	29
Group projects	27
Vivas	26
Short answer questions	23
Portfolios	18
Practical tests	2
Presentations	2
Problem-based learning	2
Reports	2
Video reports and/or reflections	2

TABLE 6-3: APPROACHES TO SUMMATIVE ASSESSMENT USED BY PARTICIPANTS (N=69)

Participants were asked to respond to a general statement about their attitude towards objective testing in different assessment contexts. The results can be seen in Table 6-4.

Statement	Strongly disagree (1)	Disagree (2)	Neither agree nor disagree (3)	Agree (4)	Strongly agree (5)	Mode
I am happy to use objective testing in formative assessment.	6	6	8	26	23	4
I am happy to use objective testing in low-stakes summative assessment (i.e. less than 10% of the overall mark).	5	7	9	24	24	4, 5
I am happy to use objective testing in high-stakes summative assessment (i.e. greater than 10% of the overall mark).	12	13	15	18	11	4

TABLE 6-4: STAFF ATTITUDE TOWARDS OBJECTIVE TESTING (N=69)

The pattern of results indicates that although academic staff are positive about using objective testing in formative and low-stakes summative assessments, they are not so positive about using it for high-stakes summative assessment.

Also of interest is participants’ awareness of CAT. Table 6-5 shows that 52 (75%) of participants were aware of CAT.

Response	N
Yes	52
No	17

TABLE 6-5: PARTICIPANT IS AWARE OF CAT (N=69) [2]

[2] Full question: Computer-Adaptive Testing is a form of e-assessment in which the questions administered during an assessment session are tailored to the proficiency

The 52 participants who answered that they were aware of CAT were also asked more specifically about their level of knowledge. Table 6-6 shows the number of responses for each option.

Limited knowledge **(1)**	**Some knowledge** **(2)**	**Good knowledge** **(3)**	**Expert** **(4)**	**Mode**
14	25	9	4	2

TABLE 6-6: PARTICIPANTS' LEVEL OF KNOWLEDGE OF CAT (N=52) [3]

The mode value indicates that participants typically had 'some knowledge' of the approach. Overall, 39 participants rated their level of knowledge as being 'limited knowledge' or 'some knowledge', 13 participants rated their level of knowledge as 'Good knowledge' or 'Expert'.

In terms of using CAT in their practice, Table 6-7 shows that seven participants are using some form of CAT.

Response	**N**
Use CAT	7
Don't use CAT	45

TABLE 6-7: PARTICIPANTS' USE OF CAT IN THEIR PRACTICE (N=52) [4]

Participants were also asked about the assessment contexts in which they used their CAT approaches. The responses received are detailed in Table 6-8 and show a similar pattern to the general use of objective testing shown in Table 6-2 and Table 6-3

level of individual students. Are you aware of computer-adaptive testing as an assessment method?

[3] Full question: Please indicate your level of knowledge of computer-adaptive testing techniques.

[4] Full question: Do you use Computer-Adaptive Testing in your current assessment practice?

whereby it is most commonly used in formative and low-stakes summative assessment contexts.

Assessment context	Response count
Formative	4
Low-stakes summative assessment (i.e. less than 10% of the overall mark)	3
High-stakes summative assessment (i.e. greater than 10% of the overall mark)	1

TABLE 6-8: PARTICIPANTS USE OF CAT APPROACHES IN THEIR PRACTICE (N=7)

Those participants who stated that they did not use CAT were asked to provide more information about their reasons for not using CAT. These were provided in a set of predetermined options as well as a free-text response option. The results can be seen in Table 6-9.

Reason for not using CAT	Response count
Adaptive testing algorithms are more difficult to implement than conventional testing.	19
The items in the pool must be calibrated.	16
Students may perceive the test as being unfair, given that the set of items administered for each student is different.	14
The item pool is larger than that required by conventional testing.	13
Students may not fully understand how their final scores are calculated.	12
The technique is not suitable for the subject that I teach.	8
I think that the advantages of Computer-Adaptive Testing are outweighed by its disadvantages.	7
The software required is too expensive. Substantial start-up costs.	7
I see little or no merit in the idea of Computer-Adaptive Testing in an educational setting.	3
To date, I have not had an opportunity to make use of computer-adaptive testing.	3
I would not know how to set up a software for this.	2

It is not possible to integrate technique into existing VLE.	2
Institutional practices make this prohibitive in terms of the effort required to implement it.	2
I am computer-phobic.	1

TABLE 6-9: REASONS SELECTED BY PARTICIPANTS FOR THEIR NON-USE OF CAT BY TIMES SELECTED (N=45)

Those participants who stated that they did use CAT were asked about their experience of CAT in terms of its potential benefits (derived from Carlson (1994), Lord (1980) and Wainer (2000)) and its potential limitations. Participants' responses are shown in Table 6-10.

Statement	**Strongly disagree** (1)	**Disagree** (2)	**Neither agree nor disagree** (3)	**Agree** (4)	**Strongly agree** (5)	**Mode**
The ability to tailor tests to the proficiency level of individual learners.	0	0	1	2	4	5
The ability to administer tests that take less time.	0	0	5	2	0	3
The ability to administer tests with fewer items.	0	1	5	1	0	3
The ability to provide learners with tailored feedback.	0	0	1	3	3	4, 5
The potential to reduce cheating, given that learners sitting the same test will be presented with a different set of items.	0	0	2	5	0	4

TABLE 6-10: POTENTIAL BENEFITS OF CAT (N=7)

Participants agreed most strongly that the ability to provide tailored tests was a potential benefit and also tended to agree that tailored feedback and the potential to reduce cheating were also potential benefits. Statements related to efficiency did not attract such clear agreement.

Statement	Strongly disagree (1)	Disagree (2)	Neither agree nor disagree (3)	Agree (4)	Strongly agree (5)	Mode
Students may not fully understand how their final scores are calculated.	0	1	1	4	1	4
Students may perceive the test as being unfair, given that the set of items administered for each student is different.	0	1	1	4	1	4
The item pool is larger than that required by conventional testing.	0	2	3	1	1	3
The items in the pool must be calibrated.	0	0	2	4	1	4
Adaptive testing algorithms are more difficult to implement than conventional testing.	1	2	0	4	0	4

TABLE 6-11: POTENTIAL LIMITATIONS OF THE CAT APPROACH (N=7)

Table 6-11 indicates that almost all options were considered to be potential limitations by the participants. There was less consensus about the limitations imposed by the item pool requirements of the approach with most responses disagreeing or 'neither agreeing or disagreeing'. This is seen as less of an issue than the need to calibrate the items.

Findings. Participants made use of objective testing in formative and low-stakes summative contexts, although less so in high stakes summative contexts. Fifty-two participants were aware of CAT (75% of the respondents), of these, 13 (25%) classed themselves as possessing a good knowledge or expertise in their knowledge and use of CAT.

Most participants indicated that they were not currently using CAT in their practice and that the main reasons they selected for this were: the resource requirement of creating a large, calibrated item pool; the difficulty of implementing CAT algorithms; and the possibility that students may perceive the test as being unfair or not understand how their final score is calculated.

Users of CAT also indicated that the perceived fairness of the test by students is a potential limitation of the CAT approach, but they also identified the value of tailoring tests and automated feedback for students as being key benefits of the CAT approach. Few of the non-users of CAT selected the options indicating that in their opinion CAT was flawed as an educational approach. So non-users of CAT did not exhibit an objection to the CAT approach itself.

Resourcing and concerns about student views about the CAT approach seemed therefore to be the main perceived barriers to greater uptake of CAT approaches rather than any perceived educational issues with the approach itself.

6.2 Study L

This study investigated the views of academic staff to the Flexilevel test overall and also about the software application being used to run the test (please see Appendix A)

Participants. A group of nine members of the University of Hertfordshire's academic staff from the School of Computer Science and the School of Engineering and Technology took part in the study.

The participants were all experienced tutors and were all actively involved in learning and teaching initiatives.

Methodology. The participants were given a short presentation about the Flexilevel test. The presentation was part of a series of Learning and Teaching lunches. Following the presentation, the participants were invited to take a test using the application designed and developed as part of this programme of research. They were then asked to complete a questionnaire about their views of the Flexilevel test and also the usability of the software application used to run the Flexilevel and conventional CBT tests.

Results. After taking the test, the participants were asked to rate 12 statements about the Flexilevel test using a five-point Likert scale. A summary of the responses provided by the participants can be found in Table 6-12.

The data indicates that the staff attitude towards the use of the Flexilevel test approach was positive in general.

Statement	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Mode
	(1)	(2)	(3)	(4)	(5)	
I would find the Flexilevel test approach useful in a formative assessment context.	0	0	1	6	2	4
I would find the Flexilevel test approach useful in a summative assessment context.	0	2	3	3	1	3, 4
The adaptivity supported by the Flexilevel test approach would help me in identifying those students who need greater support.	1	0	3	5	0	4
The adaptivity supported by the Flexilevel test approach would enhance the student assessment experience.	0	0	0	8	1	4
Students would find the Flexilevel test approach useful in a formative assessment context.	0	0	1	5	3	4
Students would find the Flexilevel test approach useful in a summative assessment context.	0	2	2	4	1	4
I would prefer to use the Flexilevel test approach than other forms of objective testing for formative assessment.	0	2	4	2	1	3
I would prefer to use the Flexilevel test approach than other forms of objective testing for summative assessment.	0	3	3	3	0	2, 3, 4
Assuming the Flexilevel software was available to me, I predict that I would use it on a regular basis to complement the existing range of assessment methods.	0	2	2	5	0	4

TABLE 6-12: PERCEIVED USEFULNESS (N=9)

In addition to responding to the statements listed in Table 6-12, participants were invited to list any barriers they identified to the implementation of the Flexilevel test as a supplement to existing assessment methods. Five out of nine participants listed at least one barrier, and a summary of these can be seen in Table 6-13.

Barrier	N
The Flexilevel test approach requires a calibrated item pool.	4
The Flexilevel test approach requires a larger item pool than that required by conventional testing.	2
In a Flexilevel test, students are unable to go to previous items.	1
Item management will be more complex than that required by conventional testing.	1
Inertia.	1
In a summative assessment context, students may not perceive the Flexilevel test as fair.	1
It may be difficult to explain to students how the Flexilevel test works.	1

TABLE 6-13: BARRIERS TO THE INTRODUCTION OF THE FLEXILEVEL TEST (N=5)

Participants were also asked to list potential benefits of the Flexilevel test approach. Four out of nine participants listed one benefit each; a summary of these can be seen in Table 6-14.

Benefit	N
The potential to enhance student motivation, as more able students will not be presented with items that are too easy for them.	1
The potential to enhance student motivation, given that students will be presented with items that are tailored to their ability.	1
The potential to provide students with tailored feedback.	1
The potential to create formative assessment opportunities that are more engaging for students.	1

TABLE 6-14: BENEFITS ASSOCIATED WITH THE INTRODUCTION OF THE FLEXILEVEL TEST (N=4)

Academic staff views of the usability of the software application. Staff were also asked to rate how they thought students would find the usability of the software artefact.

Statement	Strongly disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly agree (5)	Mode
Learning to use the Flexilevel software application would be easy for students.	0	0	0	6	3	4
Students would find it easy to remember how to perform tasks using the Flexilevel software application.	0	0	0	5	4	4
Students would find the Flexilevel software application easy to use.	0	0	0	4	5	5

TABLE 6-15: PERCEIVED EASE OF USE (N=9)

As can be seen from Table 6-15, participants believed that students would find the application easy to use. This was an important finding, as it suggests that they believe that students would not be disadvantaged by the user interface; it may be anticipated that such a perception may make it less likely that academics would use the application in their assessment practice.

Findings. The work reported here was aimed at gaining an understanding of how the Flexilevel test was viewed by academic members of staff. It also related to the software artefact designed and developed to deliver Flexilevel and CBT assessments. Although the extent to which the results can be extrapolated to any wider population is somewhat limited, they are nonetheless encouraging.

In the first section of the questionnaire, participants were asked about the usability of the software artefact. The responses of the participants, on behalf of their students, were strongly positive about the usability of the application. Although the length of the questionnaire was necessarily limited, the unequivocal support for the usability of the system does provide some confidence that it was not an interfering factor in participants' consideration of the utility of the software artefact.

The perceived barriers and benefits identified by participants are relatively low in number overall. Of the barriers, item pools relating to the calibration and number of items were the most often identified.

Participants indicated that they would be willing to use the Flexilevel test approach as a supplement to existing provision, but they indicated that they would be less willing to replace existing provision with the Flexilevel test approach.

The data suggests that participants were very positive about the statement that the adaptivity provided by the Flexilevel test approach would enhance the student assessment experience, and indeed the majority indicated that they would use it if it were made available to them.

There is also strong support for the use of the Flexilevel test approach in formative assessment, both from the perspective of the participants themselves and their perception of their students' attitudes. The response is less positive for the use of the Flexilevel test approach in summative assessment, both from the perspective of the participants and their perception of their students' attitudes.

Overall, the response of the participants to the Flexilevel test approach and the use of the application in their practice was positive.

6.3 Study M

Participants. Five tutors from the School of Computer Science were interviewed to discuss their views on the Flexilevel test and its use in mobile formative assessment contexts. All five participants were asked to take part in this study as they were experienced academic members of staff and had substantial experience of using technology in their assessment practice across a range of assessment contexts.

Methodology. The structure of the interviews was intended to elicit an understanding of participants' current practice, particularly in respect of objective testing. They were then asked about their use and opinions of mobile assessment. This was intended to provide a background for the discussion later in the interview about the use of the Flexilevel test application in mobile contexts. Before that part of the interview, the Flexilevel test application was demonstrated to participants.

Notes and recordings were taken of each interview and a content analysis was performed to elucidate the main points of interest. These are reported in the next section.

Results. As can be seen from Table 6-16, all five participants made use of objective testing in their practice.

Do you use objective testing?	N
Yes	5
No	0

TABLE 6-16: PARTICIPANTS' USE OF OBJECTIVE TESTING (N=5)

Table 6-17 sets out the motivations participants expressed for using objective testing.

Motivation use of objective testing	N
Coverage of module syllabus	3
Easy to mark	1
Students geographically dispersed (online)	1
Tests good for when you have many students	1
Students use their analytical skills to answer objective items	1
Test focuses in on what they know/understand	1

TABLE 6-17: PARTICIPANTS' MOTIVATION FOR USING OBJECTIVE TESTS (N=5)

Participants were asked about the issues they had encountered in their use of objective testing. These are reported in Table 6-18.

Issues encountered using objective testing	N
Setting up item pools	2
Creation of items	1
Selection of items	1
Item quality	1
How to act on the results generated by technology	1

TABLE 6-18: ISSUES ENCOUNTERED BY PARTICIPANTS IN THEIR USE OF OBJECTIVE TESTING (N=5)

Overall, the issues identified tended to relate to the creation of items and item pools. One of the issues identified, that of how to act on the results of assessments, relates to the meaning of the information generated by the technology and how to use it as a basis for feedback students – in short, what does the data mean and how can it inform the actions of tutors and students?

Use of technology currently. Table 6-19 shows the technologies that have been used by the participants. It should be noted that the focus is on objective testing, so the variety and number of technologies being used may seem low. However, there is much variety in the use of technology for other forms of assessment, such as Skype and Adobe Connect for vivas and feedback, plagiarism detection tools, concurrent versioning systems, invigilation services and video-based assessment and feedback, as well as the University of Hertfordshire's assignment submission system.

Technology usedfor objective testing	N
Questionmark Perception for objective testing	4
Questionmark Perception for short answers	2
Questionmark Secure	2
Institutional objective testing system	1

TABLE 6-19: ASSESSMENT TECHNOLOGY USED BY TUTORS FOR OBJECTIVE TESTING (N=5)

Participants were asked to identify issues they had encountered using the technology for assessment. Table 6-20 sets out those issues mentioned by participants.

Technical issue	N
Student problems with technology	2
Technical problems (infrastructural in student’s location)	2
Institutional support available (QMP)	2
Difficulty of using provision – support and technology	2
Technical problems unspecified	1
Ensuring students don't cheat	1
Formatting the questions	1

TABLE 6-20: TECHNICAL ISSUES ENCOUNTERED BY PARTICIPANTS (N=5)

Most responses relate to specific problems with students’ own technology (for example issues with installation of software) as well as more general problems (general infrastructural issues, such as reliability of electrical supply and network).

In this case, ensuring that students don’t cheat is shown from a technical perspective and ties in with the use of technologies to limit what students can access whilst taking the test and invigilation services.

Participants were asked about their use of mobile technology in their learning and teaching practice; none of the participants were using mobile technology directly in their learning and teaching. However the point of interest is the extent to which participants thought that mobile formative assessment in general, and mobile formative assessment using the Flexilevel algorithm for item selection in particular, would be of value to them and their students.

When asked if they thought whether or not assessments that were available on mobile devices would be beneficial to the students, all participants agreed that they would be.

Do you think formative assessment that is available on mobile devices would be beneficial to students?	**N**
Yes	5
No	0

TABLE 6-21: PERCEIVED USEFULNESS OF MOBILE FORMATIVE ASSESSMENT (N=5)

Participants were asked their views on formative assessment that is available on mobile devices. Table 6-22 shows the key themes that emerged from the interviews.

Views on formative assessment on mobile devices	**N**
Supports opportunistic learning	3
Supports informal communications with students	1
Students familiar with own devices	1
Gives students options	1

TABLE 6-22: VIEWS OF FORMATIVE ASSESSMENT ON MOBILE DEVICES (N=5)

The most value was placed on providing students with the opportunity to study opportunistically between other commitments.

Participants were asked for their views on the use of the Flexilevel approach. Table 6-23 shows that all five participants thought that it would be useful in their learning and teaching, although it should be noted that one participant did add the caveat that it depended on the extent to which it was integrated with the rest of the educational experiences for the module. If Flexilevel tests were offered as an optional extra, they would be of limited value.

Do you think the Flexilevel assessment application delivered on a mobile device would be useful in your learning and teaching?	**N**
Yes	5
No	0

TABLE 6-23: VIEWS ABOUT THE USEFULNESS OF THE FLEXILEVEL TEST APPLICATION (N=5)

This was followed up with a discussion about how it could be incorporated into participants' learning and teaching practice. Table 6-24 shows the participants' responses.

Possibilities for integrating the Flexilevel test in current learning and teaching	**N**
Formative assessment	5
Repeating tests would be useful	3
Gamification	2
Differentiating learning topics in terms of difficulty – easy to difficult	1
Summative assessment	1
Challenging students	1

TABLE 6-24: PARTICIPANTS' VIEWS ON THE POSSIBILITIES FOR INTEGRATING FLEXILEVEL TEST (N=5)

All participants saw the possibility of integrating the Flexilevel test application into their learning and teaching activities as a formative assessment tool. The ability to repeat tests was also perceived as being a good way of integrating the approach and gamification was also identified twice.

In terms of challenging students, one participant suggested that it would be more rewarding for students to take an adaptive test because they know that if they put more effort in they will see more difficult questions; their effort means something since it changes the test.

Participants were asked what they thought the main problems would be to adopting the Flexilevel test in their learning and teaching practice. Table 6-25 shows an aggregation of participants' responses.

Possible barriers to integrating the Flexilevel test in current learning and teaching	**N**
Estimating item difficulty	3
Creation of items	2

TABLE 6-25: PARTICIPANTS' VIEWS ON BARRIERS TO INTEGRATING THE FLEXILEVEL TEST (N=5)

It indicates that participants thought that the main perceived barrier would be with the creation and calibration of enough test items. However, this was not universal; one participant thought they would be able to write enough test items. Also, when the possibility of shorter tests was put to participants, they viewed it positively.

The approach in principle was supported: as one participant put it, the approach would be "really, really good" if the academic and technical resources were available. This view is also borne out in the responses participants gave to the request for open positive and negative comments about the Flexilevel test. These are shown in Table 6-26 and Table 6-27 respectively.

General positive comments about the Flexilevel test application	**N**
Increases student confidence	1
Shorter tests	1
Simple algorithm	1
Interesting concept	1
Visual representation of the performance of students	1
Item selection is very good	1

TABLE 6-26: POSITIVE OPEN COMMENTS ABOUT THE FLEXILEVEL TEST APPROACH (N=5)

One of the concerns as can be seen in Table 6-27 relates principally to the initial accuracy of the calibration of items. For one participant, calibrating an existing item pool that had already been used for testing would not be an issue.

General negative comments about the Flexilevel test application	**N**
Populating item pool	3
Accuracy of calibration	2
Sensitivity of items may be too high (routing on the basis of one item)	1

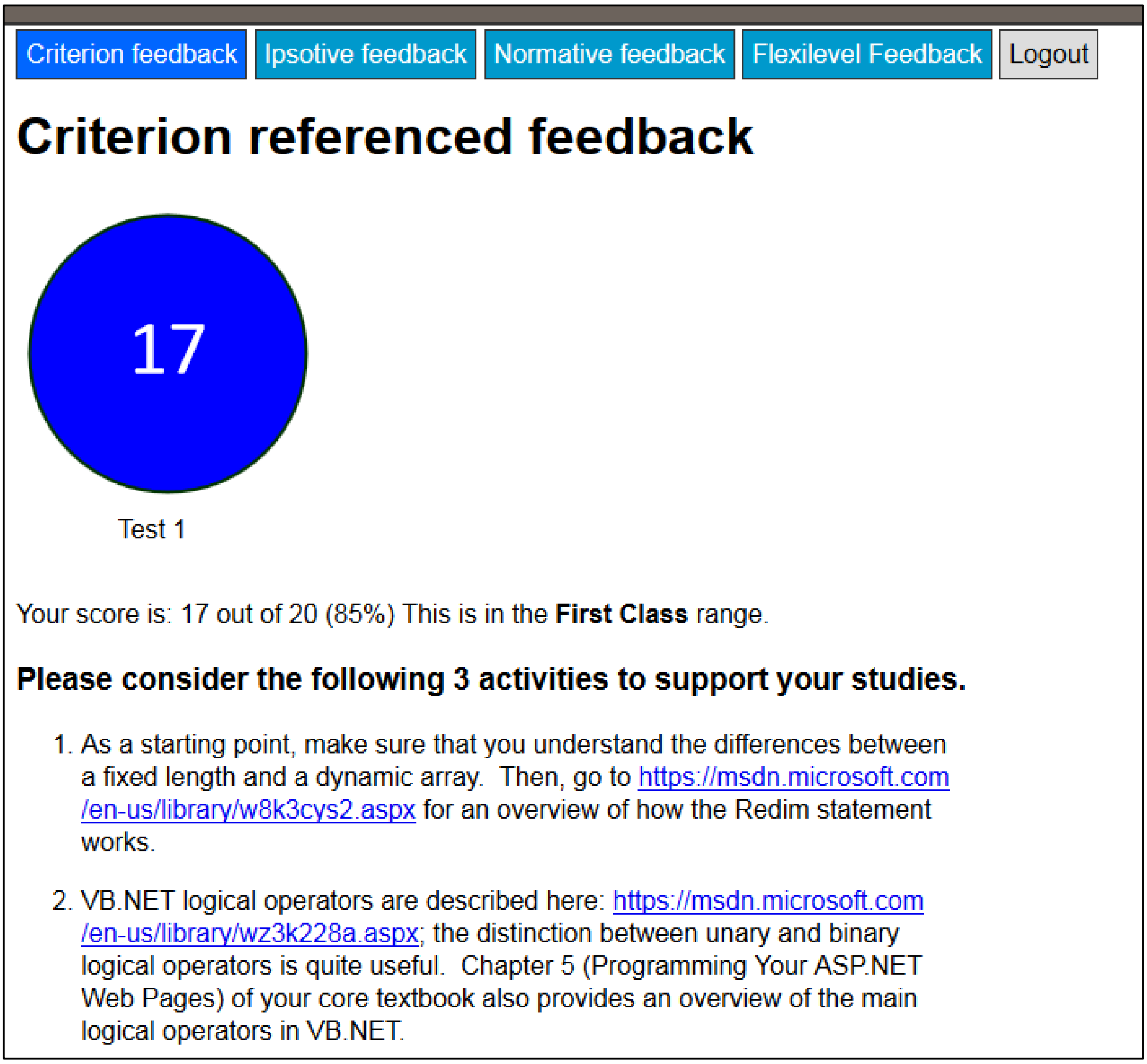
TABLE 6-27: NEGATIVE OPEN COMMENTS ABOUT THE FLEXILEVEL (N=5)

Feedback. The potential value of providing automated tailored feedback was also investigated as part of this study. The application was used to demonstrate the use of three forms of feedback: criterion-referenced, norm-referenced and ipsative feedback.

- Criterion-referenced feedback provides information about a student's performance based on pre-defined criteria, for example grades and feedback about why the student's work warrants a particular grade.
- Norm-referenced feedback is the comparison of students' scores, for example providing an average score achieved by a group of students.
- Ipsative feedback is the comparison of a student's performance with themselves, for example showing their progress through assessment points in a module.

The application also provides a representation of examinees' performance in the Flexilevel test. It should be noted that whilst the examples shown in the screenshots below (Figure 6-1, Figure 6-2, Figure 6-3 and Figure 6-4) are based on real test data only the criterion-referenced, or ipsative feedback was made available to students and was presented more generically as 'Feedback'. Participants in this study were shown the feedback and were asked about each of the main forms of feedback provided.

Criterion-referenced feedback was provided in the form of a score together with more detailed guidance on work that students could carry out to support their studies. This is shown in Figure 6-1. Table 6-28 shows participants' responses to the criterion-referenced feedback.



Criterion referenced feedback

Your score is: 17 out of 20 (85%) This is in the **First Class** range.

Please consider the following 3 activities to support your studies.

1. As a starting point, make sure that you understand the differences between a fixed length and a dynamic array. Then, go to https://msdn.microsoft.com/en-us/library/w8k3cys2.aspx for an overview of how the Redim statement works.
2. VB.NET logical operators are described here: https://msdn.microsoft.com/en-us/library/wz3k228a.aspx; the distinction between unary and binary logical operators is quite useful. Chapter 5 (Programming Your ASP.NET Web Pages) of your core textbook also provides an overview of the main logical operators in VB.NET.

FIGURE 6-1: CRITERION-REFERENCED FEEDBACK PROVIDED BY THE WEB APPLICATION

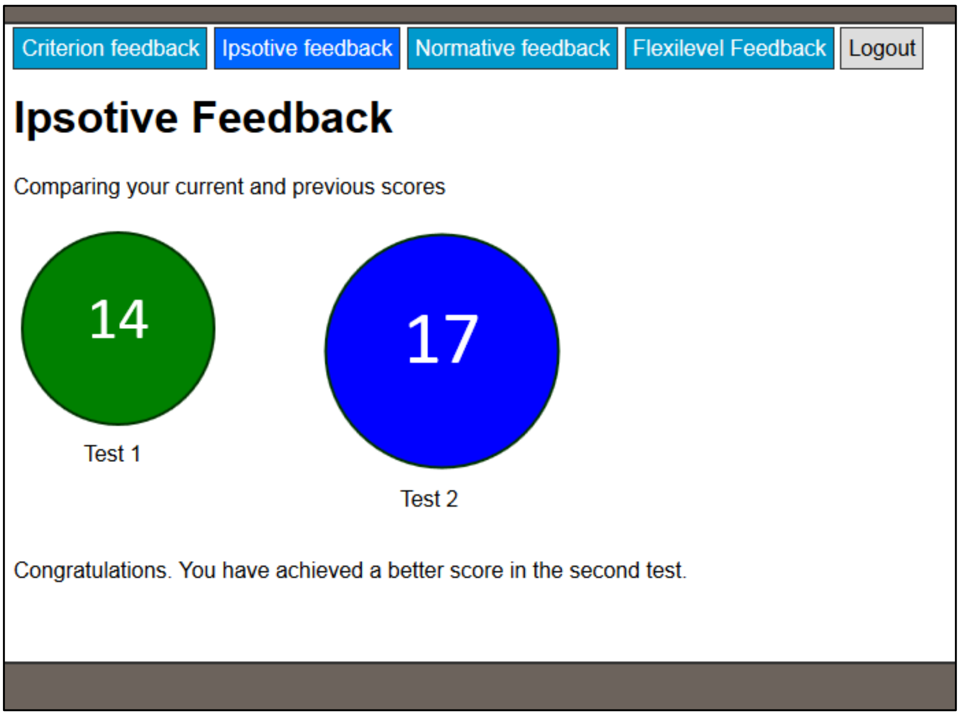
Participants were mainly positive about the criterion-referenced feedback. Mapping the score to degree classifications was considered positive and something that would be valued by students.

Criterion referenced feedback	**N**
Useful	4
Neutral	1
Not useful	0

TABLE 6-28: PERCEIVED USEFULNESS OF THE CRITERION-REFERENCED FEEDBACK (N=5)

Representing the score with a percentage when the items had different difficulties was considered a potential issue by one participant, but for most participants the scoring element of the feedback was deemed less important than other elements of the feedback.

The idea of using ipsative[5] feedback was introduced in the form of a basic example: a comparison of scores examinees achieved in a previous test with the score achieved in a follow-up test. This can be seen in Figure 6-2.



Ipsotive Feedback

Comparing your current and previous scores

Congratulations. You have achieved a better score in the second test.

FIGURE 6-2: IPSATIVE FEEDBACK PROVIDED BY THE WEB APPLICATION

As can be seen in Table 6-29, participants were mainly neutral about the usefulness of the ipsative feedback presented.

Ipsative feedback	N
Useful	1
Neutral	3

[5] Ipsative feedback is also spelled ipsotive as in the case of the application

Not useful	1

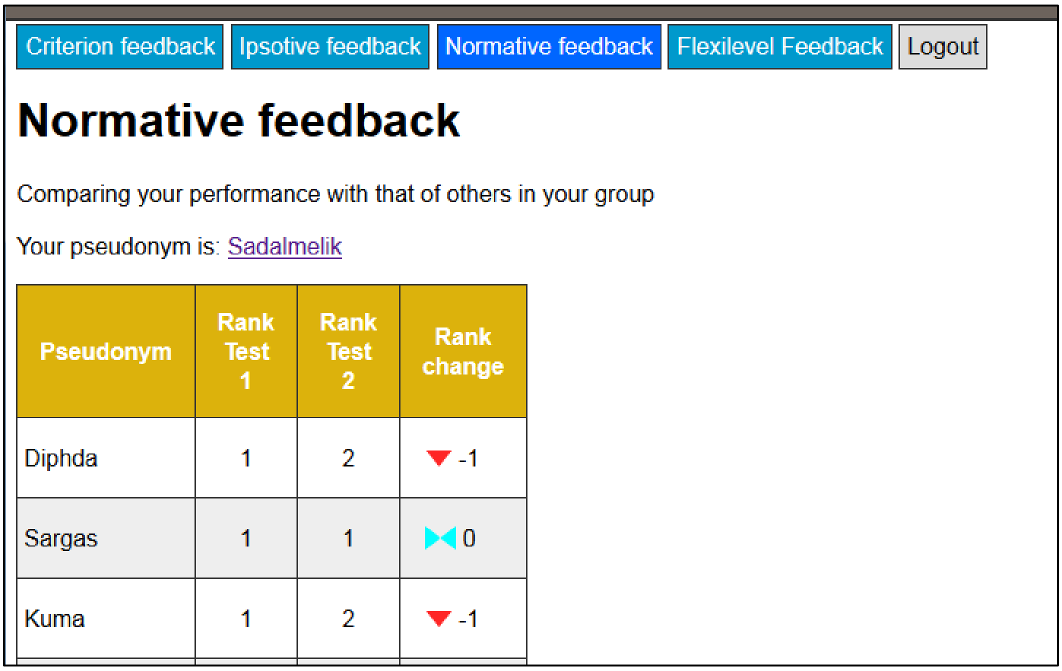
TABLE 6-29: PERCEIVED USEFULNESS OF IPSATIVE FEEDBACK PROVIDED BY THE FLEXILEVEL TEST APPLICATION (N=5)

It was felt that more information was needed and that the way in which feedback could be provided for faltering students could usefully be couched in terms of targets (for example in terms of development over time which lends itself to the gamification of the educational experience). This can be seen in Table 6-30.

Suggestions for ipsative feedback	N
Needs more information/feedback	2
Needs more specific feedback about work done in between tests	1
Gamification (for example attaining new levels of expertise)	1

TABLE 6-30: SUGGESTIONS FOR IPSATIVE FEEDBACK (N=4)

Normative feedback was presented in the form of a league table in which participants were given astronomy-based pseudonyms. They were shown their ranks in both tests as well as the direction and extent of any change to their ranking.



Normative feedback

Comparing your performance with that of others in your group

Your pseudonym is: Sadalmelik

Pseudonym	Rank Test 1	Rank Test 2	Rank change
Diphda	1	2	▼ -1
Sargas	1	1	⧓ 0
Kuma	1	2	▼ -1

FIGURE 6-3: SCREENSHOT OF THE NORMATIVE FEEDBACK PROVIDED BY THE FLEXILEVEL WEB APPLICATION

Normative feedback was viewed positively. The competitive element of the approach, as well as the perception that students find additional meaning in their scores from comparison with their peers, were cited as being positive features of the approach.

Normative feedback	N
Useful	3
Neutral	2
Not useful	0

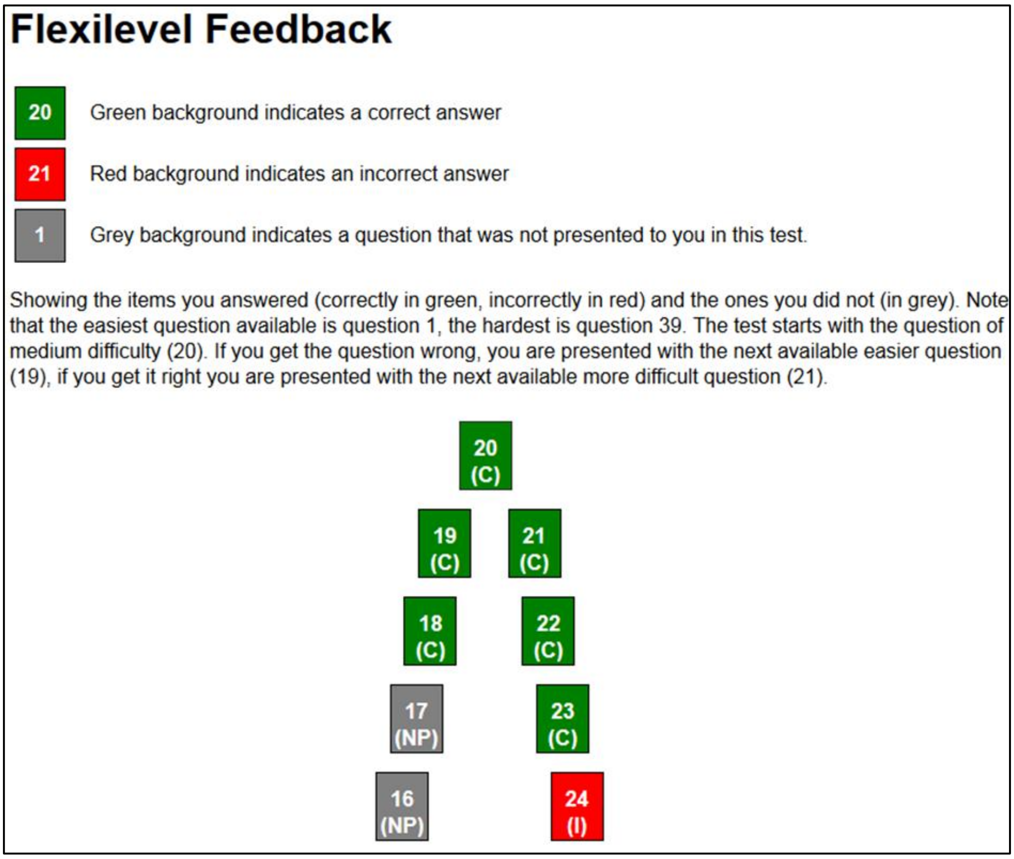
TABLE 6-31: PERCEIVED USEFULNESS OF NORMATIVE FEEDBACK (N=5)

As can be seen in Table 6-31, there were also some qualifications to the overall positive view of normative feedback. Concern with normative feedback on its own was that students may feel demotivated if they are listed in the bottom ranks of the table. Further suggestions for the provision of normative feedback were identified in Table 6-32.

Suggestions	**N**
Generic, 5 students got x out of n.	2
Spread of marks + distribution	1
You came *n* out of students	1
Combine with ipsative feedback	1

TABLE 6-32: PARTICIPANTS' SUGGESTIONS FOR PROVIDING NORMATIVE FEEDBACK (N=5)

The Flexilevel feedback was based on students' views that they wanted to know which items – in particular difficult items – they had missed. (This is discussed in more depth in Chapter 6.) The initial design shown in Figure 6-4 reflects a common representation of the Flexilevel test (for example, Figure 3-15) with the easier items being positioned down the left branch and the more difficult items being positioned down the right branch. Items are represented by their rank numbers. Those items that students answered correctly are shown in green and also contain a 'C' for 'correct', the items that students answered incorrectly are shown in red and have and also contain an ‘I’ for ‘incorrect’. Those items that the students did not see in the test are represented in grey with ‘NP’ for ‘Not presented’.



Flexilevel Feedback

🟩 20 Green background indicates a correct answer

🟥 21 Red background indicates an incorrect answer

▨ 1 Grey background indicates a question that was not presented to you in this test.

Showing the items you answered (correctly in green, incorrectly in red) and the ones you did not (in grey). Note that the easiest question available is question 1, the hardest is question 39. The test starts with the question of medium difficulty (20). If you get the question wrong, you are presented with the next available easier question (19), if you get it right you are presented with the next available more difficult question (21).

FIGURE 6-4: PARTIAL SCREENSHOT OF THE FLEXILEVEL FEEDBACK PROVIDED BY THE WEB APPLICATION

The Flexilevel feedback was viewed positively, as shown in Table 6-33.

Flexilevel feedback	N
Useful	3
Neutral	1
Not useful	1

TABLE 6-33: PERCEIVED USEFULNESS OF THE FLEXILEVEL FEEDBACK (N=5)

Table 6-34 shows suggestions for the Flexilevel feedback. Whilst it seems that the Flexilevel feedback would provide a basis for further explanation of the Flexilevel approach to the students, one reservation was that it was too academic and might not be easily communicated. However the combination of Flexilevel feedback with other forms of feedback was considered a potentially useful approach. For example, patterns of responses could also be mapped to degree classifications.

Suggestions	N
Combine with normative feedback	3
Present patterns in terms of degree classifications	1
Provide additional explanation of scoring	1

TABLE 6-34: SUGGESTIONS FOR THE FLEXILEVEL FEEDBACK (N=5)

Findings

The participants were experienced members of academic staff with extensive experience of e-assessment tools and techniques. All made use of objective testing and conventional CBT tests.

The Flexilevel test for formative assessment and the idea of implementing it in mobile contexts was well received. The possibility of embedding shorter Flexilevel assessments in between other learning and teaching activities was valued. The potential for including a gamification element in students' educational experiences generated positive discussion.

Perceived barriers to the use of adaptive testing included the creation of large item pools with calibrated items. In terms of the Flexilevel test, it seems that issues of calibration and item pool population would not present the same level of barrier in practice as is perceived for adaptive testing generally.

The calibration of items would not require much beyond what is done already in participants' reported practice. Indeed, where there are existing CBT item pools, there is the potential for the Flexilevel test to be applied without substantial additional preparation being required. Where they fitted in with the overall assessment design, shorter tests could make use of existing item pools in which case the additional work would be limited to the ranking of existing items. Also, once the item pool has been created further calibration could be achieved based on real test data.

Feedback and the nature of the feedback was considered to be key in terms of linking the assessments in with existing educational experiences, and providing meaning for the actions that both tutors and students need to take based on the results.

The feedback that was demonstrated was generally well received and there was interest in how the different forms of feedback could be usefully combined. There was support from the participants in this study for the Flexilevel feedback to be combined with other forms of feedback, particularly normative feedback. It seems there are promising possibilities to combine automated tailored feedback with other forms that provide information about how others have done and also what is needed to achieve more. This was also considered in terms of using the Flexilevel test for the gamification of educational experiences.

Overall the Flexilevel approach was perceived positively, as was the potential for embedding it in participants’ learning and teaching activities.

6.4 Discussion

The studies reported here were intended to gain a better understanding of the perception and usage of CAT and the attitudes of academic staff members to the Flexilevel approach in general and in their learning and teaching contexts.

The key findings are:

1. **The Flexilevel test is viewed positively by academic members of staff.**

   This was taken to indicate that academic members of staff would accept the Flexilevel test as a form of assessment in their learning and teaching practice.

   This was explored in Study L (section 6.2) and Study M (section 6.3)

2. **A perceived barrier to adaptive testing is the resource requirement associated with creating a calibrated item pool.**

   This was taken to indicate that the perception of CAT in general as requiring a large calibrated item pool. Although it seems from discussions with academic staff about the Flexilevel that this perception may also exist for the Flexilevel test, the resource implications for the Flexilevel test are lighter than for other forms of CAT.

   As has been noted, the size of the item pool, 2n-1 where n is the number of items in the test, is smaller than those required by other forms of CAT. Additionally, the calibration requirements are lighter than other forms of CAT where statistical methods are employed; these can require 200–1000 examinees' responses to an item for calibration. Indeed, a ranking of the items is sufficient for the Flexilevel test.

   These issues are discussed in more detail in Chapter 3, but point to the potential for the Flexilevel test to address the perceived issues as they have been expressed in these studies. This is particularly the case where objective testing is already actively used and item pools already exist. There is much scope to make more use of these item pools with minimal additional resource requirements – for example the use of shorter tests if this is consistent with the overall assessment design being used.

This was explored in Study K (section 5), Study L (section 6.2) and Study M (section 6.3)

3. **There is support for the use of the Flexilevel test in formative contexts, including for opportunistic or ad hoc educational experiences.**
   Opportunistic and ad hoc educational experiences refer to those experiences that occur when a student finds themselves unexpectedly in a position to engage with their studies. (Examples of unplanned studies are included in Chapter 5, but may include reading an article when someone you are waiting for is late.) This is taken to indicate that short Flexilevel tests – that students could take when the opportunity presented itself – would be acceptable to academic staff when properly aligned with the other learning and teaching activities in a module.

   This was explored in Study M (section 6.3).

4. **Academic staff were positive about the use of the Flexilevel test in mobile contexts.**
   This is taken to indicate that the Flexilevel test application would be acceptable to academic staff in their learning and teaching activities.
   This was explored Study M (section 6.3).

5. **There is a variety of possible ways the Flexilevel test could provide feedback to students.**
   There was much discussion about the possibilities afforded by the different forms of feedback provided, particularly in combination with each other and when constructively aligned with other learning and teaching activities. Examples included setting out where students were and where they needed to get to in terms of more difficult items or tasks, comparing their own performance in relation to others. This was taken to indicate that the Flexilevel test would be acceptable as a means of providing feedback and

supporting student engagement with the educational experiences available to them.

This was explored Study M (section 6.3).

The studies reported here show that academic members of staff are positive about using the Flexilevel test in formative assessment contexts and as the basis for feedback. Further, there is support for the embedding of Flexilevel tests in real educational contexts if perceived resource issues are addressed. The potential for the Flexilevel test to address these perceived issues as they have been expressed in these studies has been detailed both in this chapter and in Chapter 3, and this is encouraging for the embedding of the approach in real educational contexts.

7 SUMMARY, CONCLUSIONS AND FUTURE WORK

The motivation for this programme of research was to contribute to the existing body of knowledge concerned with embedding adaptive test algorithms in educational experiences in Higher Education (HE).

The studies conducted as part of this programme of research were concerned with the effectiveness of the Flexilevel test, the evaluation of its usability with students, the attitude of students to the Flexilevel test, and the views of academic staff about its usability and the attitudes of academic staff to the Flexilevel test itself. These strands of work continued throughout the programme of research and an overview of the work in terms of effectiveness, usability and attitude is set out below.

7.1.1 EFFECTIVENESS

Ten empirical studies and one real data simulation were conducted to investigate the effectiveness of the Flexilevel algorithm. These are described in detail in chapter 4 and an overview is provided below. This is structured as follows:

- Location of the study
- Assessment context
- Test Length
- Summary of the statistical analysis.

Exploratory studies (Section 4.1). Exploratory studies were carried out to test the effectiveness of the Flexilevel test without impacting on the formal learning and teaching activities the students were engaged in.

Two exploratory studies were carried out as follows:

Location of the study	PC laboratory
Assessment context	Not applicable
Test length	The Flexilevel test was the same length as that of the conventional CBT test

Summary of statistical analysis	Statistically significant correlations were found between the scores achieved by participants in the Flexilevel test and the conventional CBT

TABLE 7-1: SUMMARY OF THE EXPLORATORY STUDIES

Results from the exploratory studies indicated that the approach was appropriate for testing in real educational contexts.

First phase of live testing (section 4.2). The first phase of live testing involved the presentation of two test stages to participants in real educational contexts as follows:

Location of the study	PC laboratory
Assessment context	Summative
Test length	The Flexilevel test was the same length as that of the conventional CBT test
Summary of the statistical analysis	Statistically significant correlations were found between the scores achieved by participants in the Flexilevel test and the conventional CBT

TABLE 7-2: SUMMARY OF THE FIRST PHASE OF LIVE TESTING

The results again showed a statistically significant correlation between students' performance in the Flexilevel test condition and the standard CBT condition. This encouraged consideration of an additional possible benefit of the Flexilevel test: to provide shorter tests that are as effective conventional CBT tests.

Real data simulation study (section 4.3). Before using a shorter Flexilevel test in real educational contexts, a real data simulation study was run as follows:

Location of the study	Not applicable
Assessment context	The data used in this study had been generated in summative assessments
Test Length	The Flexilevel tests were at least half the length of the conventional CBT tests
Summary of the statistical analysis	Statistically significant correlations were found between the scores obtained by participants when items were selected for them using the Flexilevel algorithm in the simulation and the actual scores achieved by students in a conventional CBT test

TABLE 7-3: SUMMARY OF THE REAL DATA SIMULATION STUDY

The results of the real data simulation study indicated that shorter Flexilevel tests were fair and accurate. As such, the results were taken to provide a basis for a second phase of live testing.

Second phase of live testing (section 4.4). In the second phase of live testing tests containing a Flexilevel test stage that was at least half the length of the conventional CBT test stage were presented to students.

Location of the study	Online
Assessment context	Diagnostic, formative, summative
Test Length	The Flexilevel tests were at least half the length of the conventional CBT tests
Summary of the statistical analysis	Statistically significant correlations were found between the scores achieved by participants in the shorter Flexilevel test and conventional CBT tests

TABLE 7-4: SUMMARY OF THE SECOND PHASE OF LIVE TESTING

In addition to the effectiveness of the Flexilevel test approach it was necessary to establish whether the applications being used were having any extraneous effects on the performance of students. The next section provides an overview of work conducted to investigate the usability of the computer applications used.

7.1.2 Usability

Rogers et al. (2011) identify three main ways of gathering data: observation, questionnaires and interviews. All three were used in the usability evaluations carried out in this programme of research:

- **Direct observation**. This involves watching users interact with a system live. In the studies reported here, direct observation was part of the invigilation of the tests, whether online or in PC laboratories. For academic staff, it occurred when participants used the application in a study. The advantages to this approach are that the usage of the system is taking place in a real context, so the observations made can be argued to have ecological validity.
- **Questionnaires**. Questionnaires are useful in that they can provide subjective reactions to the system that participants have just used. Two main forms of questionnaire were used. Questionnaires were presented after participants had taken a Flexilevel test. This questionnaire comprised a section on attitude as well as questions about the ease of use of the system. This approach

enabled the collection of participants' views about the Flexilevel test at the point at which they had just finished trying it out for themselves.

Additionally, the System Usability Scale (SUS) (Brooke, 1996) and the Usefulness, Satisfaction, and Ease of use questionnaire (USE) (Lund, 2001) were used. Participants were asked to try out the system and then were asked to fill in the two questionnaires. The focus of this study was on the usability of the system.

- **Interviews.** The usability of the applications that students had used or had seen demonstrated was discussed as part of wider interviews about the Flexilevel test. This provided an opportunity for more in-depth discussion about the usability of the application being used and also elicited views about what makes an e-assessment application usable for them. The interviews were semi-structured, using a script to provide a structure for the interview (Lazar & Feng, 2010).

Overview of results

Students. Table 7-5 is a summary of the usability findings from studies with students (please see chapter 5 for details).

Form of data collection	Result
Direct observation	No usability issues were detected on the systems used as part of this programme of research. This applies to the desktop and web version of the application.
Bespoke questionnaire	Students were positive about the usability of the system they were presented with.
SUS and USE questionnaires	Students were positive about the usability of a Flexilevel test delivered on a mobile device. The SUS score was above average and the response to the USE questionnaire indicated that participants were positive about the usability of the application, particularly in terms of measures associated with ease of use.
Interview	After having used the system or seen it being presented, students were positive about the usability of the system.

TABLE 7-5: A SUMMARY OF USABILITY FINDINGS – STUDENTS

Academic staff. Table 7-6 provides a summary of the usability findings from studies with academic staff (please see chapter 6 for details).

Form of data collection	**Result**
Direct observation	No usability issues were detected on the systems used as part of this programme of research. This applies to the desktop version, the web version of the application used on a mobile device and on a laptop.
Questionnaire	Academic staff were positive about how usable they thought students would find the desktop application.

TABLE 7-6: A SUMMARY OF USABILITY FINDINGS – ACADEMIC STAFF

These results were taken to support the conclusion that there were no usability issues with the software applications used in this programme of research that affected the scores achieved by participants using them.

7.1.3 ATTITUDES

In tandem with this work, it was important to gain an understanding of the attitudes of both students and academic staff to the Flexilevel test as this would identify applications of the Flexilevel test that would be acceptable to students and academic staff.

Similarly to the usability studies the approach to gathering data about the attitudes of academic staff and students was through the use of questionnaires and semi-structured interviews.

Students. Questionnaire-based data: A five-point Likert scale questionnaire was used, as noted above, to capture immediate feedback from student participants of their views about the Flexilevel test straight after taking one. Participants were also asked to provide free-text responses to more general open questions.

Interview-based data: This provided a much more in-depth account of students' attitudes to the Flexilevel test. Findings relating to their learning contexts – their current use of technology provided a rich background to discussions about their perceived usefulness of the Flexilevel test approach.

This was a necessary part of the work to gain an understanding of the contexts in which participants would be accepting of the use of the Flexilevel test. The results show that the Flexilevel test would be acceptable to students as an embedded part of their educational experiences. This applies particularly to formative assessments, but also to the structuring of educational experiences that differ in levels of difficulty. Such variation in challenge lends itself to the introduction of game elements, for example achieving the next level in a sequence of tasks set at different levels of difficulty.

Interestingly, a consistent theme throughout this programme of research has been the students' desire to see difficult items even if they are outside their range of proficiency.

Academic staff. Questionnaire-based data: This was used to reach a wider population of academic staff in the UK HE sector and indicated that objective testing was widely used in formative contexts and to a lesser extent in summative contexts. There was an awareness of CAT reported, but few participants made use of CAT in their practice.

A similar pattern emerged when using questionnaires to gather information from colleagues at the University of Hertfordshire about their views of the Flexilevel test. Overall, attitudes were found to be positive for the use of Flexilevel testing in formative assessment, although a more conservative view was evident when considering the use of CAT and Flexilevel testing in summative assessment.

The potential value of tailored feedback was also explored with academic staff. They were positive about the potential for combining different forms of feedback to provide engaging tailored feedback. The feedback provided by the Flexilevel test was also considered to be a possible vehicle for enhancing the engagement of students in their educational experiences – representing feedback in terms of the levels of a game where greater attainment is achieved through answering more difficult items or completing more difficult tasks.

7.2 Contribution to knowledge

The principal contributions to knowledge of this programme of research are:

1. The Flexilevel test provides measures of examinee performance that are as fair and accurate as conventional CBT tests in a range of real Higher Education contexts including summative, formative, in-class and online assessment settings.
2. Academic staff and students exhibited a very positive attitude towards the use of the Flexilevel test in formative assessment contexts, online, in PC laboratories and on mobile devices.
3. Students consistently expressed the view that they wanted to be presented with difficult items. This does not follow the existing notion that examinees would suffer frustration or demotivation from the presentation of items that are too difficult for them and points to the potential value of reviewing this area of research.
4. The application designed and developed for this research which allowed the test to be run online in different contexts; no other such web application is known.

7.3 Research Questions

The research questions for this programme of research were:

1. Does the Flexilevel test afford assessment opportunities in real Higher Education contexts?
2. What, if any, are the potential applications of the Flexilevel to Higher Education contexts?

7.3.1 Does the Flexilevel test afford assessment opportunities in real Higher Education contexts?

The performance of students when taking a Flexilevel test have been demonstrated to be highly and significantly correlated with their performance on standard CBT tests in different assessment contexts and also at different levels of study. This demonstrates that the Flexilevel test affords assessment opportunities in Higher Education.

An important motivation for this programme of research was to evaluate the use of the Flexilevel test as a means of providing CAT with fewer of the resource requirements associated with other forms of CAT. There is support for the view that a CAT technique that was not resource intensive has the potential to be more widely adopted in real Higher Education contexts.

The programme of research reported here demonstrates that the Flexilevel algorithm could in practical terms be applied to both summative and formative assessment contexts. This includes standard invigilated exam conditions in a PC laboratory as well as open conditions including mobile assessment contexts.

The work carried out here has contributed to a number of peer-reviewed publications, and has been reported at conferences. These publications and dissemination work has contributed to a renewed interest in the Flexilevel test in contemporary Higher Education settings (Gordon, 2014).

7.3.2 What, if any, are the potential applications of the Flexilevel test to Higher Education contexts?

Stakeholder attitudes are key to understanding the potential applications to which the Flexilevel test can be applied in real educational contexts. Attitudes and expectations of staff and students were investigated as part of this programme of research. Both user groups were positive about the use of the Flexilevel test in formative assessment. They took a more conservative view of the use of the Flexilevel test in summative contexts.

The research reported here indicates that students would value further opportunities for online formative assessment. Such tests, mediated by the Flexilevel algorithm were viewed positively by participants. Additionally mobile assessment provision would be welcomed by both academic staff and students. Furthermore, both groups were positive about the Flexilevel test being used for formative assessments in this context. The possibility of providing shorter tests was also received positively by participants and would lend itself well to the provision of shorter tests suitable for mobile delivery.

The Flexilevel tests held the possibility of creating more engaging educational experiences, although interestingly, students' views consistently showed that they would not want to miss out on attempting the more difficult items. This is something of a departure from the view that students may become demotivated by the presentation of items that are too easy or too difficult for them (Carlson (1994), Wainer (2000)). In the studies conducted for this programme of research students consistently wanted to be presented with more difficult items.

The idea of students being aware of which difficult items they were not presented with – because they answered easier items incorrectly – and having the chance to attempt them lends itself to the gamification of educational experiences. Both academic staff and students identified gamification as an application that the Flexilevel test would be suited to.

For academic staff the main perceived barrier to greater uptake of CAT related to the creation of a large calibrated item pool, and it seems that the Flexilevel test was perceived to have similar requirements. However the extent to which these perceived barriers actually apply to the Flexilevel test are limited, especially when there is existing use of objective tests. This is taken to support the notion that embedding the Flexilevel test would be feasible, particularly where existing objective testing is already taking place.

The research began with the idea of investigating the Flexilevel test in contemporary higher education contexts, in particular embedding it in learners' educational experiences. Early research showed that students' performance in the Flexilevel test was comparable to their performance in a conventional Computer Based Test (CBT) and summative assessments were used for these studies, mainly to control for potential effects on student performance such as motivation. However, the main point of interest was in embedding the Flexilevel test in learner's educational experiences as formative assessment. This is in line with Gordon's subsequent analysis for the Higher Education Academy report into flexible pedagogies (Gordon (2014). Indeed an area of interest is in mobile assessment which seems to have a clearer role to play in formative objective assessment than in summative objective assessment.

The research data supports the use of the Flexilevel test in summative contexts; however, more research would be needed if large scale summative assessment using the Flexilevel test was planned.

Overall, participants exhibited a positive attitude to the use of the Flexilevel test in a range of assessment opportunities, and in formative assessment contexts in particular. As such the Flexilevel test can be argued to offer very good assessment opportunities embedded in real educational contexts in HE. It can also be argued that the Flexilevel algorithm shows real potential to be embedded in students' educational experiences more broadly than assessment, for example in providing tailored educational experiences that incorporate game elements.

7.4 Objectives

The extent to which the objectives set for this programme of research were achieved is set out below.

1. To compare student performance between a conventional CBT and Flexilevel test.

   Over 600 individual tests were analysed in the studies comparing student performance in conventional CBT and Flexilevel tests, with over 500 being carried out in real educational contexts. A core methodology was established in exploratory studies and applied in different assessment contexts: diagnostic, formative and summative. The assessments were conducted online as well as on campus as is common for conventional CBT.

A real data simulation study was also conducted to investigate the possibility that Flexilevel tests were as effective as conventional CBT. Here the data from real tests was used to simulate the scores that would have been achieved by students had the Flexilevel algorithm been used to select items in test.

Correlations between the two tests were found to be statistically significant, something that held true for tests held in different contexts (diagnostic, formative and summative), when delivered differently (on desktop PCs in a PC laboratory; online on desktop computers; and online using mobile devices) and at different levels of study (Levels 4, 6 and 7).

2. To gain an understanding of the attitudes of academic staff to tests using the Flexilevel algorithm for item selection.

Three studies were conducted involving over 80 academic members of staff to meet this objective. Study K (section 6.1) took a broad view and sought the attitudes of academic members of staff across the UK HE sector about CAT in general. The study was conducted via an online survey with the option for adding free-text responses. Participants were recruited through requests sent out to relevant mailing lists and also through direct contact with colleagues in the sector who are prominent in the field of e-assessment.

This provided a context for more specific studies with colleagues about the Flexilevel test approach: Study L (section 6.2) and Study M (section 6.3). Academic staff were positive about the use of the Flexilevel algorithm for item selection. Issues that emerged from Study K about CAT in general did seem to be perceived as issues for the Flexilevel test as well, so item pools and the calibration of items were considered potential barriers to the use of the Flexilevel approach. However, staff attitudes to the Flexilevel test were positive.

3. To gain an understanding of the attitudes of students to tests using the Flexilevel algorithm for item selection.

Over 70 students contributed to the evaluations of the Flexilevel Test as part of this programme of research, mainly through responses to surveys but also through participation in more in-depth interviews.

These evaluations identified a range of consistent attitudes that have been reported in the Pilot Study (section 5.2) and Study I (section 5.4).

An interesting finding of this programme of research was that students consistently reported that they wanted to see the difficult questions that may not have been presented to them. This is inconsistent with the idea that the presentation of items that are too easy or too difficult is an issue for the motivation and potentially the performance of students (Carlson (1994), Wainer (2000)).

4. To identify contexts of use in which the Flexilevel Tests may be applicable.

Studies were carried out in:

- Summative classroom-based assessment (Study A, Study B, Study F)
- Summative online proctored assessment (Study E)
- Diagnostic online assessment (Study C, Study H)
- Mobile online formative assessment (Study G)
- Low-stakes summative online assessment (Study D, Study H).

The Flexilevel test was as effective as conventional CBT in these contexts, both with the same number of items and fewer items than conventional CBT tests. Through direct observation and informal expert review, the applications were judged to be usable for these testing contexts.

Investigations into the attitudes of primary stakeholders demonstrate that the Flexilevel test is acceptable to them in diagnostic and formative contexts in the PC laboratories, online and also online using mobile devices. The attitude towards using the Flexilevel test for summative assessments is more conservative.

5. To evaluate the approach adopted in real educational contexts

Eight of the ten empirical studies carried out took place in real educational contexts. The two exploratory studies were closely modelled on real educational contexts, but the tests were not part of the programme of study of the participants.

Additionally, where tests were invigilated, it was possible through the use of direct observation techniques to observe students making use of the software applications used to run the tests in real educational contexts.

6. To apply Human Computer Interaction techniques to the development and design of e-assessment applications

HCI techniques were an important part of the evaluations with academic staff and students. Data gathering tools and techniques were employed, including the SUS and USE questionnaires. Evaluations of the desktop application involved measures of ease of use as well as perceived usefulness.

For the web application, discussions with colleagues already using web-based objective testing applications showed where the interface for the application could follow the conventions already established in the assessment provision available to students. Expert reviews of the interface were carried out. Whilst the focus of this programme of research was on the Flexilevel test rather than the software applications that mediated it, HCI techniques were used to establish that the applications were not having an uncontrolled impact on students' assessment experience.

This dissertation has set out the assessment opportunities afforded by the Flexilevel test and how Flexilevel tests can be applied in a range of assessment contexts. The evidence presented supports the notion that the Flexilevel test provides a CAT technique that is effective in different assessment contexts, one that is sufficiently lightweight to be able to be embedded in students' educational experiences. This provides for a number of interesting possible routes for future work and these are outlined next.

7.5 Future work

7.5.1 Mobile assessment using the Flexilevel test

It would be interesting to follow up on work that was conducted about the use of mobile devices and their use in contemporary learning ecosystems. In Study G, students were free to use whichever device they chose, and some chose mobile devices. The next stage would be to repeat this study as a formative assessment along with follow up interviews, as were conducted for Study I (section 5.4).

This is consistent with techniques already used in this programme of research, but poses the challenge of the participants being in a mobile environment. The observations may need to come through usage data gathered. Limited data was collected in Study G, but it points to the potential value for descriptive learning analytics to explain participants' usage of the system.

This would build on the work already done on mobile use of the Flexilevel test application.

7.5.2 Automated tailored feedback

Another interesting area for future work would be to see how the Flexilevel test could be used as a basis for the provision of tailored feedback. This work could build on the findings reported in Study M (section 6.3) as part of this programme of research in terms of the combinations of feedback that have emerged as possible routes to investigate.

The learning ecosystem is a complex environment requiring significant judgements on the part of the student as well as significant guidance from tutors and other experts. It would be interesting to investigate the extent to which feedback provided on the back of a tailored assessment would support students in navigating the ecosystem successfully.

7.5.3 Broadening the use of the Flexilevel algorithm

Tailoring assessment would provide two important functions in learning ecosystems. The first is to reduce the transactional distance of the educational experience since the assessment adapts to the student; and the second is to provide scaffolding for the learner through the provision of automated tailored feedback.

This work could investigate the extent to which the Flexilevel test could be applied more broadly to tailoring students' progress through educational experiences. Some interesting ideas came out of the research programme in this regard, including the potential for the Flexilevel algorithm to structure vivas, and to provide levels of challenges for students to try to successfully complete.

This latter idea also indicates another interesting path for future work: gamification. A consistent theme that emerged from this programme of research was the value of introducing game elements to students' educational experiences. This was identified both by students and by academic staff and the Flexilevel algorithm was viewed positively as a potential source of gamification features in their learning and teaching activities.

7.5.4 Broadening out to other Science, Technology, Engineering, Medicine and Mathematics subjects

The work here has focussed on Computer Science education. However, it would be of particular interest to verify that the same approach also works in other areas of study. The expectation is that for other STEMM subjects where objective testing is widely used, there will be an easy and seamless move to its adoption.

7.5.5 CERTAINTY-BASED MARKING

It would be of value to investigate the relationship between students' attitudes and their performance on tests. A consistent finding during this programme of research was that students reported a desire to see the difficult items that may not have been presented to them. One way of investigating this would be to create a test that used the Flexilevel algorithm for item selection and certainty-based marking (Gardner-Medwin & Curtin, 2007).

In a test using certainty-based marking, examinees answer an item, but also state how certain they are that they have answered the item correctly. The level of certainty ranges from 1 for 'uncertain' to 3 for 'certain'. If they answer correctly, but rated their level of certainty as 1 they score 1 mark. If they answer correctly and rate their level of certainty as 2, they are awarded 2 marks and if they rate their certainty as 3, they receive 3 marks. Conversely, if they answer incorrectly, the marks awarded are 0 for a certainty rating of 1, -2 marks for a certainty rating of 2 and -6 for a certainty rating of 3. This can be seen in Table 7-7

Level of Certainty	Correct answer	Incorrect answer
1 (uncertain)	1	0
2	2	-2
3 (certain)	3	-6

TABLE 7-7: SCORING FOR CERTAINTY BASED MARKING

This could be of real interest because apart from the additional value of providing information about the examinee's proficiency, it would be possible to gather information about how proficient they think they are. A feature of the Flexilevel algorithm is that it provides a tailored assessment experience and as such, it would be valuable to see how this may be captured in terms of certainty. It would be interesting to find out whether examinees report their certainty differently when presented with a test that is tailored to their level of proficiency than when presented with a conventional CBT test.

The thesis put forward for this programme of work is that the Flexilevel test would make a good choice of CAT for embedded assessment opportunities in Higher Education. This dissertation has supported this view. It has shown the effectiveness of Flexilevel testing in affording assessment experiences in Higher Education contexts and has also demonstrated the acceptability of the approach to primary stakeholders. A range of applications of the approach have been demonstrated and lines of future investigation have been outlined. It is hoped that the work reported here provides a good basis for future work and other avenues of research.

REFERENCES

Anderson, T., & Dron, J. (2010). Three generations of distance education pedagogy. The International Review of Research in Open and Distributed Learning, 12(3), 80-97.

Armstrong, R. D., Jones, D. H., Koppel, N. B., & Pashley, P. J. (2004). Computerized adaptive testing with multiple-form structures. Applied Psychological Measurement, 28(3), 147-164.

Attle, S., & Baker, B. (2007). Cooperative learning in a competitive environment: Classroom applications. International Journal of Teaching and Learning in Higher Education, 1, 77-83.

Baddeley, A. (2001). Is Working Memory Still Working? American Psychologist, 56, 849–864.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. International Journal of Human–Computer Interaction, 24(6), 574-594.

Bangor, A., Kortum, P., & Miller, J. (2009). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. The Journal of Usability Studies, 4(3), 114-123.

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.

Barker, T., & Bennett, S. (2012). The use of electronic voting and peer assessment to encourage the development of higher order thinking skills in learners. International Journal of e-Assessment, 2(1).

Benson, R., & Samarawickrema, G. (2009). Addressing the context of e-learning: using transactional distance theory to inform design. Distance Education, 30(1), 5-21.

Betz, N. E., & Weiss, D. J. (1973). An Empirical Study of Computer-Administered Two-stage Ability Testing. University of Minnesota, Minneapolis.

Betz, N. E., & Weiss, D. J. (1975). Empirical and simulation studies of flexilevel ability testing (Research Report 75-3). University of Minnesota, Department of Psychology. Minneapolis: Psychometric Methods Program.

Biggs, J. (1999). Teaching for Quality Learning at University. Buckingham, United Kingdom: Open University Press.

Biggs, J., & Tang, C. (2007). Teaching for Quality Learning at University (Society for Research Into Higher Education) (3rd ed.). United Kingdom: Open University Press.

Bloom, B. S. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.

Boud, D. (1995). Assessment and learning: contradictory or complementary? In P. Knight (Ed.), Assessment for Learning in Higher Education (pp. 35-48). London: Kogan Page.

Brooke, J. (1996). SUS-A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClleland (Eds.), Usability evaluation in industry (pp. 189-194). London, UK: Taylor & Francis.

Brooke, J. (2013). SUS: a retrospective. Journal of Usability Studies, 8(2), 29-40., 8(2), 29-40.

Brown, S., Bull, J., & Race, P. (1999). Computer-Assisted Assessment in Higher Education. Staff and Educational Development Series. Herndon, VA: Stylus Publishers Inc.

Brusilovsky, P., & Sosnovsky, S. (2005). Individualized exercises for self-assessment of programming knowledge: An evaluation of QuizPACK. Journal on Educational Resources in Computing (JERIC), 5(3).

Bull, J., & McKenna, C. (2003). A blueprint for computer-assisted assessment. Routledge.

Bull, J., Brown, G. A., & Pendlebury, M. (1999). Assessing student learning in higher education. London: Routledge.

Butcher, P. G., & Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. Computers & Education, 55(2), 489-499.

Carlson, R. D. (1994). Computer adaptive testing: A shift in the evaluation paradigm. Journal of Educational Technology Systems, 22(3), 213-224.

Change, V., & Guetl, C. (2007). E-learning Ecosystem (ELES)- A Holistic Approach for the Development of more Effective Learning Environment for Small-to-Medium Sized Enterprises (SMEs). In Proceedings of the Inaugural IEEE International Digital Ecosystems Technologies Conference (IEEE-DEST 2007).

Charman, D. (1999). Issues and impacts of using computer-based assessments (CBAs) for formative assessment. In S. Brown, J. Bull, & P. Race, Computer-Assisted

Assessment in Higher Education. Staff and Educational Development Series (pp. 85-93). London: Kogan Page.

Clariana, R., & Wallace, P. (2002). Paper–based versus computer–based assessment: key factors associated with the test mode effect. British Journal of Educational Technology, 33(5), 593-602.

Cliffe, E., Davenport, J., De Vos, M., Parmar, N. R., & Hayes, A. (2010). Using EVS And ResponseWare To Enhance Student Learning And Learning Experience. 11th Annual Conference of Higher Education Academy Subject Centre for Information and Computer Science. Bath.

Cubric, M., & Jefferies, A. (2012). EEVS Project. Retrieved 05 03, 2015, from Jisc Design Studio: http://jiscdesignstudio.pbworks.com/w/file/fetch/60025852/EEVSFinalReport%20reportNewFrontpage.pdf

Daly, C., & Horgan, J. (2004). An automated learning system for Java programming. IEEE Transactions on Education, 47(1), 10-17.

Daly, C., & Waldron, J. (2004). Assessing the assessment of programming ability. ACM SIGCSE Bulletin, 36(1), 210-213.

Davenport, J., Hayes, A., & Parmar, N. R. (2009). The use of an Electronic Voting System to enhance student feedback. Plymouth e-Learning Conference, University of Bath.

De Ayala, R. J. (2009). Theory and practice of item response theory. New York: The Guildford Press.

De Ayala, R. J., & Koch, W. R. (1986). A Computerized Implementation of a Flexilevel Test and Its Comparison with a Bayesian Computerized Adaptive Test. Paper presented at the Annual Meeting of the National Council on Measurement in Education. San Francisco.

De Ayala, R. J., Dodd, B. G., & Koch, W. R. (1990). A Simulation and Comparison of Flexilevel and Bayesian Computerized Adaptive Testing. Journal of Educational Measurement, 27(3), 227-239.

Dillenbourg, P. (1999). What do you mean by collaborative learning? In P. Dillenbourg, Collaborative-learning: Cognitive and Computational Approaches (pp. 1-19). Oxford: Elsevier.

Dix, A., Finlay, J., Abowd, G. D., & Beale, R. (2004). Human Computer Interaction (3rd ed.). Pearson Prentice Hall.

Edwards, M. C., & Thissen, D. (2007). Exploring potential designs for multi-form structure computerized adaptive tests with uniform item exposure. Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.

Elkjaer, B. (2009). Pragmatism: A learning theory for the future. In K. Illeris, Contemporary theories of learning: learning theorists... in their own words. (pp. 74-89). London & New York: Routledge.

Flaugher, R. (1990) Item Pools In: Wainer, H. (Ed) (1990) Computerized Adaptive Testing: A Primer Lawrence Erlbaum Associates.

Freire, P. (1970). Pedagogy of the Oppressed. (M. B. Ramos, Trans.) Penguin Books (1996).

Gardner-Medwin, T., & Curtin, N. (2007). Certainty-based marking (CBM) for reflective learning and proper knowledge assessment. In Proceedings of the REAP International Online Conference: Assessment Design for Learner Responsibility [Internet] (pp. 29-31).

Garrison, D. (2011). E-learning in the 21st century: A framework for research and practice. Taylor & Francis.

Garrison, D. R., & Anderson, T. (2003). E-Learning in the 21st Century: A framework for research and practice. RoutledgeFalmer. London & NY.

Gordon, N. (2014). Flexible Pedagogies: technology-enhanced learning. Higher Education Academy, NIACE.

Gorsky, P., & Caspi, A. (2005). A critical analysis of transactional distance theory. The Quarterly Review of Distance Education, 6(1), 1-11.

Gredler, M. E. (2004). Games and simulations and their relationship to learning. In D. H. Jonassen, Handbook of Research on Educational Communications and Technology (2nd ed., pp. 571-582). Mahwah, NJ: Lawrence Erlbaum Associates.

Hartson, R., & Pyla, P. S. (2012). The UX Book: Process and guidelines for ensuring a quality user experience. Waltham, MA 02451, USA: Morgan Kaufmann.

Hayes, A., Thomas, P., Smith, N., & Waugh, K. (2007). A Framework for the Automated Assessment of Consistency Between Code and Design. Proceedings of Informatics Education Conference II, (pp. 370-378).

Herrington, J., Herrington, A., Mantei, J., Olney, I., & Ferry, B. (2009). New technologies, new pedagogies: Using mobile technologies to develop new ways of teaching and learning. Final report to the Australian Learning and Teaching Council. NSW: Australian Learning and Teaching Council.

Higgins, C. A., & Bligh, B. (2006). Formative computer based assessment in diagram based domains. 11th annual SIGCSE conference on Innovation and technology in computer science education (ITICSE '06) (pp. 98-102). New York, NY, USA: ACM.

Ihantola, P., Ahoniemi, T., Karavirta, V., & Seppälä, O. (2010). Review of recent systems for automatic assessment of programming assignments. Proceedings of the 10th Koli Calling International Conference on Computing Education Research (pp. 86-93). ACM.

Irlbeck, S., & Mowat, J. (2007). Learning content management system (LCMS). In K. Harman, & A. Koohang, Learning Objects: Standards, Metadata, Repositories, and LCMS (pp. 157-184). Santa Rosa, California: Informing Science Press.

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. Behaviour & Information Technology, 33(4), 410-422.

JISC. (2014). Assessment for Learning. Retrieved 04 28, 2015, from The Design Studio: http://jiscdesignstudio.pbworks.com/w/page/52947115/Assessment%20for%20Learning

Jonassen, D. H. (1997). Instructional design models for well-structured and III-structured problem-solving learning outcomes. Educational Technology Research and Development, 45(1), 65-94.

Jones, C., Cook, J., Jones, A. and De Laat, M. (2007) Collaboration In: Conole, G. and Oliver, M (2007) Contemporary perspectives in E-learning research : themes, methods and impact on practice. Routledge

Knight, P. (2001). A Briefing on Key Concepts: Formative and summative, criterion and norm-referenced assessment. Learning and Teaching Support Network.

Knight, P. T. (2002). Summative assessment in higher education: practices in disarray. Studies in higher Education, 27(3), 275-286.

Kocher A., T. (1974) An empirical investigation of the stability and accuracy of flexilevel tests Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, Illinois, April 1974)

Larkin, K. C., & Weiss, D. J. (1975). An Empirical Comparison of Two-Stage and Pyramidal Adaptive Ability Testing.

Laurillard, D. (2001). Rethinking University Teaching: A Conversational Framework for the Effective Use of Learning Technologies (2nd ed.). London.: Routledge Falmer.

Lazar, J., Feng, J. H., & Hochheiser, H. (2010). *Research methods in human-computer interaction*. Chichester, West Sussex, U.K: Wiley.

Lazarinis, F., Green, S., & Pearson, E. (2010). Creating personalized assessments based on learner knowledge and objectives in a hypermedia Web testing application. Computers & Education, 55(4), 1732-1743.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. Human Centered Design (pp. 94-103). San Diego: Springer Berlin Heidelberg.

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.

Lilley, M., Barker, T. & Britton, C. (2004) The development and evaluation of a software prototype for computer-adaptive testing. Computers & Education 43(1-2): 109-123.

Lilley, M., & Pyper, A. (2009). The Application of the Flexilevel Approach for the Assessment of Computer Science Undergraduates. (J. A. Jacko, Ed.) Lecture Notes in Computer Science, 5613, 140-148.

Lilley, M., Pyper, A., & Attwood, S. (2012). Understanding the Student Experience through the Use of Personas. Innovation in Teaching and Learning in Information and Computer Sciences, 11(1), 4-13.

Lilley, M., Pyper, A., & Wernick, P. (2011). Attitudes to and Usage of CAT in Assessment in Higher Education. Innovation in Teaching and Learning in Information and Computer Sciences, 10(3).

Lord, F. M. (1970). The self scoring flexilevel test (Research Report RB·70- 43) . Princeton, N.J.: Educational Testing Service.

Lord, F. M. (1971a). The self-scoring flexilevel test. Journal of Educational Measurement, 8, 147-151.

Lord, F. M. (1971b). A theoretical study of two-stage testing. Psychometrika, 36(3), 227-242.

Lord, F. M. (1980). Applications of item response to theory to practical testing problems. Lawrence Erlbaum.

Lowenthal, P. R., & Dunlap, J. C. (2010). From pixel on a screen to real person in your students' lives: Establishing social presence using digital storytelling. Internet and Higher Education, 13, 70-72.

Lund, A. M. (2001). Measuring Usability with the USE Questionnaire. TC Usability SIG Newsletter, 8(2).

Mayer, R. E. (2008). Applying the science of learning: evidence-based principles for the design of multimedia instruction. American Psychologist, 63(8), 760.

McAlpine, M., & Hesketh, I. (2003). Multiple response questions–allowing for chance in authentic assessments. 7th International CAA Conference.

McBride, J. R. (1997). Research antecendents of Applied Adaptive Testing. In W. A. Sands, B. K. Waters, & J. R. McBride, Computerized Adaptive Testing from inquiry to operation (pp. 47-59). Washington DC: American Psychological Association.

Monsell, S. (2003). Task switching. Trends in cognitive sciences, 7(3), 134-140.

Moore, M. (1993). Theory of transactional distance. Theoretical principles of distance education, 22.

Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. Journal of Further and Higher Education, 31(1), 53-64.

Nikander, J., Korhonen, A., Seppälä, O., Karavirta, V., Silvasti, P., & Malmi, L. (2004). Visual algorithm simulation exercise system with automatic assessment: TRAKLA2. Informatics in Education-An International Journal, 3(2), 267-288.

Noss, R., & Pachler, N. (1999). The challenge of new technologies: doing old things in a new way, or doing new things. In P. Mortimore (Ed.), Understanding Pedagogy and its impact on learning (pp. 195-211). London: Paul Chapman Publishing.

Oliver, M. (2000). An introduction to the Evaluation of Learning Technology. Educational Technology & Society, 3(4), 20-30.

Oliver, M., Harvey, J., Conole, G., & Jones, A. (2006). Evaluation. In G. Conole, & M. Oliver, Contemporary Perspectives in E-Learning Research : Themes, Methods and Impact on Practice (pp. 203-217). Taylor and Francis.

Oulasvirta, A., Tamminen, S., Roto, V., & Kuorelahti, J. (2005). Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI. Proceedings of the SIGCHI conference on Human factors in computing systems. ACM.

Oviatt, S. (2006). Human-centered design meets cognitive load theory: designing interfaces that help people think. Proceedings of the 14th annual ACM international conference on Multimedia (pp. 871-880). ACM.

Pachler, N., Daly, C., Mor, Y., & Mellar, H. (2010). Formative e-assessment: Practitioner cases. Computers & Education, 54(3), 715-721.

Paulus, T. M. (2005). Collaboration or Cooperation? Analyzing Small Group Interactions in Educational Environments. In T. S. Roberts, Computer-Supported Collaborative Learning in Higher Education (pp. 100-124). Hershey, PA: Idea Group Publishing.

Plowman, L., Lucklin, R., Laurillard, D., Stratfold, M., & Taylor, J. (1999). Designing Multimedia for Learning: Narrative Guidance and Narrative Construction. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (pp. 310-317). ACM.

Polsani, P. R. (2006). Use and abuse of reusable learning objects. Journal of Digital information, 3(4).

Prpic, J. (2005). Managing academic change through reflexive practice: A quest for new views. Research and Development in Higher Education, 28, 399-406.

Pyper, A. (2011). Designing and developing groupware for different forms of groupwork; a case study in Human Computer Interaction. In D. Graham (Ed.), ICS HEA e-Learning and Teaching Workshop, University of Greenwich (pp. 11-14). ICS HEA.

Pyper, A., & Lilley, M. (2007). The impact of content type on educational dialogues. ICS HEA e-Learning and Teaching Workshop. University of Greenwich, United Kingdom. .

Pyper, A., & Lilley, M. (2008a). Student Attitudes to Discussion Forums as a Publishing Medium. Proceedings of the International Conference on Technology, Communication and Education, 7th-9th April, 2008 Gulf University for Science and Technology. Mishref, Kuwait .

Pyper, A., & Lilley, M. (2008b). Producing E-Learning Resources; By Design or By-Product? Online Educa Berlin 2008, the 14th International Conference on Technology Supported Learning & Training, December 3 – 5, 2008. Berlin.

Pyper, A., & Lilley, M. (2010). A comparison between the flexilevel and conventional approaches to objective testing. Proceedings of CAA 2010 International Conference. Southampton.

Pyper, A., Lilley, M., & Hewitt, J. (2009). A framework to support students in their individual studies. CSEDU 2009 - International Conference on Computer Supported Education. Lisbon.

Pyper, A., Lilley, M., Hewitt, J., & Wernick, P. (2011). A framework to support individual learners in learning ecosystems. Edmedia. Lisbon.

Pyper, A., Lilley, M., Wernick, P., & Jefferies, A. (2014a). A simulation of a Flexilevel test. HEA STEM Annual Conference. Edinburgh: Higher Education Academy.

Pyper, A., Lilley, M., Wernick, P., & Jefferies, A. (2014b). Applying the flexilevel test algorithm to postgraduate formative assessment: implications for the design of educational experiences. EDULEARN14 Proceedings, (pp. 3867-3871). Barcelona.

Pyper, A., Lilley, M., Wernick, P., & Jefferies, A. (2015a). The Potential Use of the Flexilevel Test in Providing Personalised Mobile E-Assessments. (P. Zaphiris, & A. Ioannou, Eds.) Lecture Notes in Computer Science 9192 Learning and Collaboration, 271-278.

Pyper, A., Lilley, M., Wernick, P., & Jefferies, A. (2015b). Comparing Shorter Flexilevel Tests with full Length Standard Computer Based Tests. The 13th European Conference on e-Learning . Hatfield.

Pyper, A., Meere, J., & Lilley, M. (2007). A Framework to Support Teaching and Learning Online. The 6th European Conference on e-Learning. Copenhagen.

Rabbany, R. K., Takaffoli, M., & Zaïane, O. R. (2012). Social network analysis and mining to support the assessment of on-line student participation. ACM SIGKDD Explorations Newsletter, 13(2), pp. 20-29.

Rogers, Y., Sharp, H., & Preece, J. (2011). *Interaction design: Beyond human-computer interaction*. Chichester: Wiley.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). Computerized adaptive testing: From inquiry to operation. Washington DC: American Psychological Association.

Sandy, J. J., Gould, A. L., Cox, D. P., & Brumby. (2013). Frequency and Duration of Self-Initiated Task-Switching in an Online Investigation of Interrupted Performance Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts AAAI Technical Report CR-13-01.

Sauro, J., & Kindlund, E. (2005). A method to standardize usability metrics into a single score. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 401-409). Portland: ACM.

Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2215-2224). ACM.

SEEC. (2010). SEEC Credit Level Descriptors for Higher Education. Southern England Consortium for Credit Accumulation and Transfer. Retrieved August 07, 2015, from http://www.seec.org.uk/wp-content/uploads/2013/seec-files/SEEC%20Level%20Descriptors%202010.pdf

Shute, V. J. (2008). Focus on formative feedback. Review of educational research, 78(1), 153-189.

Siemens, G. (2007). Connectivism: Creating a learning ecology in distributed environments. In T. Hug, Didactics of microlearning: Concepts, discourses and examples. Munster, Germany: Waxmann Verlag.

Sitthiworachart, J., & Joy, M. (2003). Web-based peer assessment in learning computer programming. The 3rd IEEE International Conference on Advanced Learning Technologies (pp. 180-184). IEEE.

Solomon, A., Santamaria, D., & Lister, R. (2006). Automated testing of unix command-line and scripting skills. Information Technology Based Higher Education and Training (pp. 120-125). Sydney: IEEE.

Stacey, E. (2005) A Constructivist Framework for Online Collaborative Learning: Adult Learning and Collaborative Learning Theory In: Roberts, T. S. (Ed) (2005) Computer-Supported Collaborative Learning in Higher Education. Idea Group Publishing

Surveymonkey (2010) [Online] Available from: www.surveymonkey.com [Accessed: 07/08/2015]

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. Educational psychology review, 22(2), 123-138.

Thelwall, M. (1999). Open access randomly generated tests: Assessment to drive learning. In S. Brown, P. Race, & J. Bull, Computer-assisted assessment in higher education (pp. 62-78). London: Kogan Page.

Thissen, D., & Mislevy, R. J. (2000). Testing Algorithms. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, & R. J. Mislevy, Computerized adaptive testing: A primer (pp. 101-133). Mahwah, New Jersey, USA: Routledge.

Thompson, N. A. (2009) Ability Estimation with Item Response Theory. White Paper. Assessment Systems Corporation

Thompson, E., Luxton-Reilly, A., Whalley, J. L., Hu, M., & Robbins, P. (2008). Bloom's taxonomy for CS assessment. Proceedings of the tenth conference on Australasian computing education. 78, pp. 155-161. Australian Computer Society, Inc.

Timmis, P. B., Oldfield, A., & Sutherland, R. (2012). Where is the cutting edge of research in e-Assessment? Exploring the landscape and potential for wider transformation. Proceedings of the 2010 International Computer Assisted Assessment (CAA) Conference. School of Electronics and Computer Science, University of Southampton, United Kingdom. .

Traxler, J. (2007). Defining, Discussing and Evaluating Mobile Learning: The moving finger writes and having writ.... The International Review of Research in Open and Distributed Learning, 8(2).

Triantafillou, E., Georgiadou, E., & Economides, A. A. (2008). The design and evaluation of a computerized adaptive test on mobile devices. *Computers & Education, 50*(4), 1319-1330

Tsiaousis, A. S., & Giaglis, G. M. (2008). Evaluating the Effects of the Environmental Context-of-Use on Mobile Website Usability. International Conference on Mobile Business, 2008. (pp. 314-322). IEEE.

Wainer, H. (2000). Introduction and history. In H. Wainer, Computerized Adaptive Testing: A primer. Lawrence Erlbaum Associates.

Wainer, H., Dorans, N.J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L. and Thiessen, D. (1990) Computerized Adaptive Testing: A Primer Lawrence Erlbaum Associates.

Warburton, B. (2013). CAA–Whither and Whence? The last decade and the next decade. In D. Whitelock, W. Warburton, G. Wills, & L. Gilbert (Ed.), CAA 2013 International Conference (pp. 1-13). University of Southampton.

Warbuton, B., & Conole, G. (2005). Whither E-Assessment? In Proceedings for 9th Computer-Assisted Assessment Conference. Loughborough: Loughborough University. Retrieved May 1, 2015, from https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/2010/1/WarburtonB_ConoleG.pdf

Ward, C. (1981). Preparing and using objective questions (Vol. 3). Hyperion Books.

Waters, B. K. (1977) An empirical investigation of the stratified adaptive testing model. Applied Psychological Measurement Vol. 1 No. 1. pp141-152. West Publishing Co.

Weiss, D. J. (1973). The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota.

Weiss, D. J. (1974). Strategies of adaptive ability measurement (Research Report 74-5). University of Minnesota, Minnesota.

Weiss, D. J. (1985). *Adaptive testing by computer*. Journal of Consulting & Clinical Psychology, 53, 774-789.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. Measurement and Evaluation in Counseling and Development, 37, 70–84.

Weiss, D., & Betz, N. E. (1973). Ability Measurement: Conventional or Adaptive? (Research Report 75-3). University of Minnesota, Department of Psychology. Minneapolis: Psychometric Methods Program.

Weiss, D. J. & Kingsbury, G. G. (1984) Application of Computerized Adaptive Testing to Educational Problems. Journal of Educational Measurement, Vol. 21, No. 4, 361-375

Wheeler, S. (2007). The Influence of Communication Technologies and Approaches to Study on Transactional Distance in Blended Learning. ALT-J: Research in Learning Technology, 15(2), 103-117.

Williams, R., Karousou, R., & Mackness, J. (2011). Emergent Learning and Learning Ecologies in Web 2.0. The International Review of Research in Open and Distributed Learning, 12(3), 39-59.

Wise, S. L. & Kingsbury, G. G (2000) Practical Issues in Developing and Maintaining a Computerized Adaptive Testing Program Psicologica (2000) 21, 135-155

Winter, J., Cotton, D., Gavin, J., & Yorke, J. D. (2010). Effective e-learning? Multi-tasking, distractions and boundary management by graduate students in an online environment. Research in Learning Technology, 18(1), 71-83.

APPENDIX A: SCREENSHOTS OF THE DESKTOP SOFTWARE APPLICATION

Appendix A

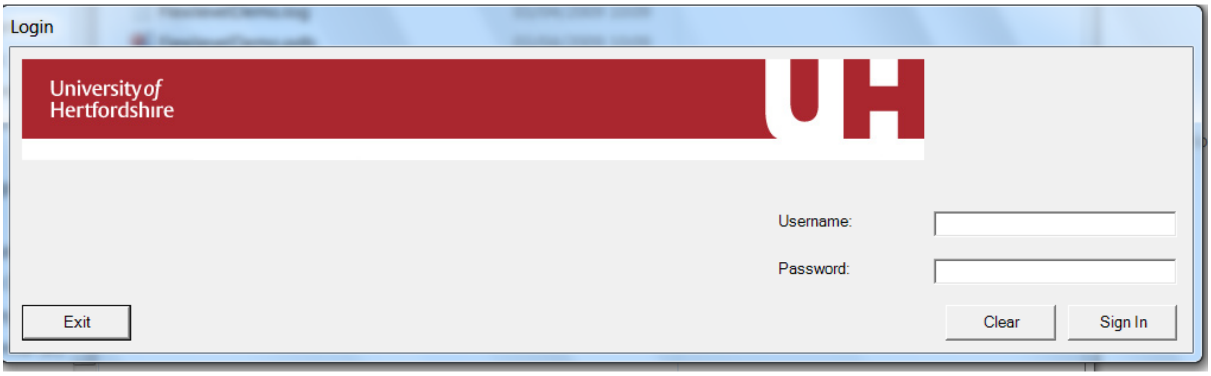
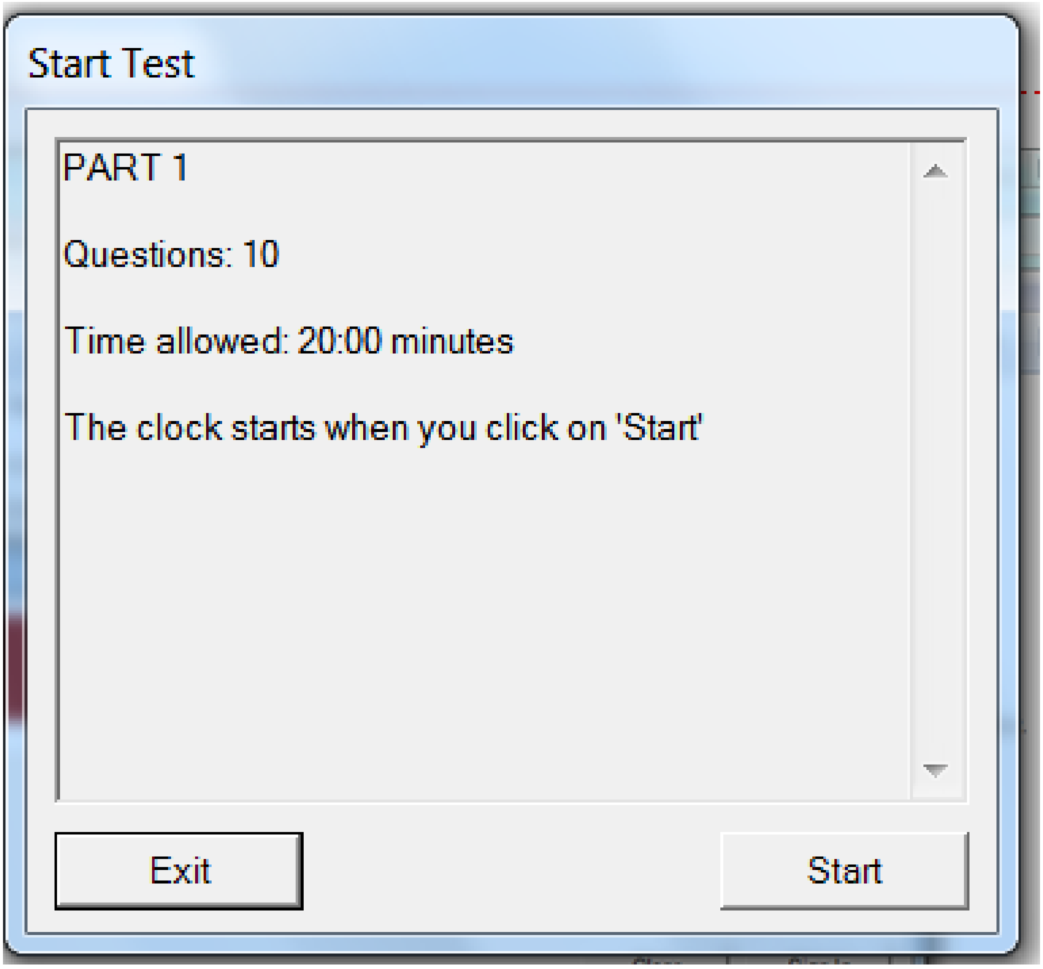


FIGURE A 1: LOGIN SCREEN



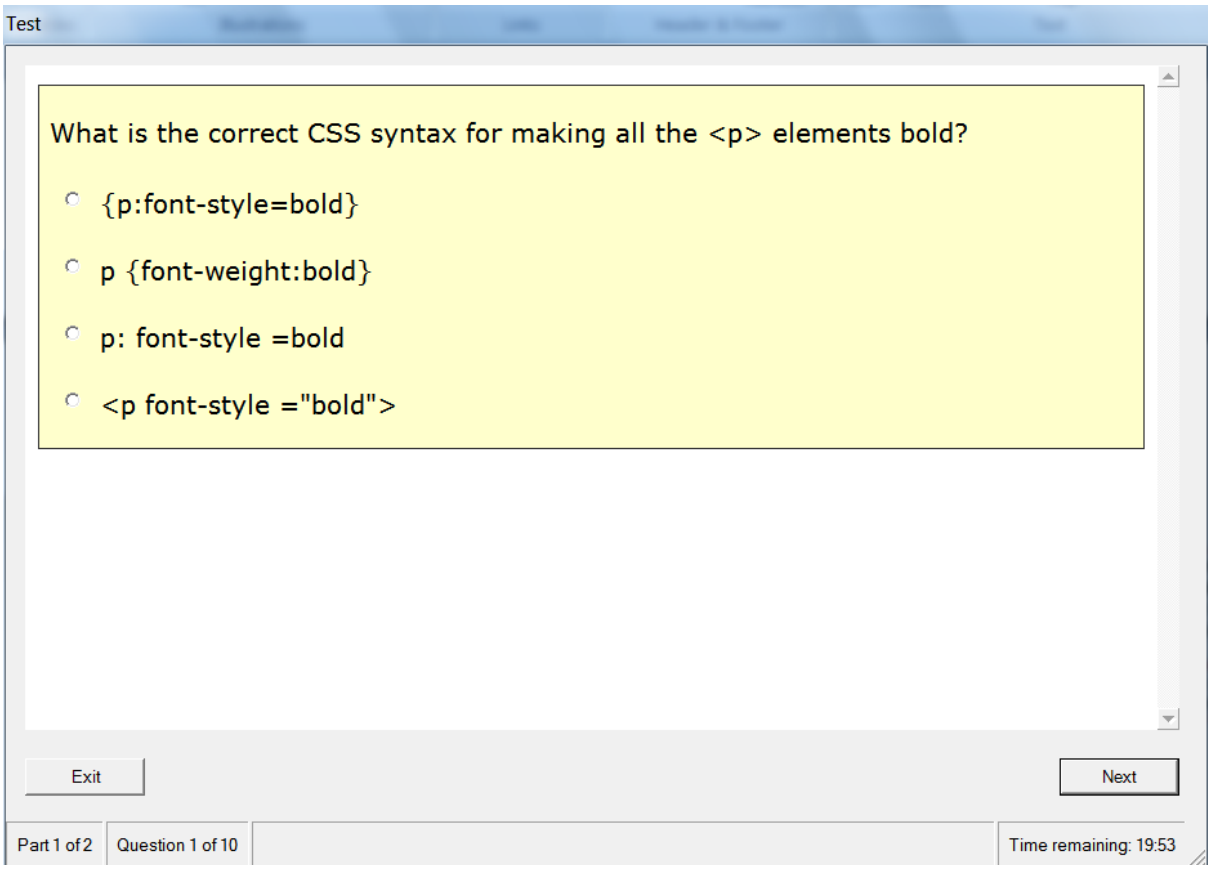
PART 1

Questions: 10

Time allowed: 20:00 minutes

The clock starts when you click on 'Start'

FIGURE A 2: INSTRUCTIONS FOR PART 1



What is the correct CSS syntax for making all the <p> elements bold?

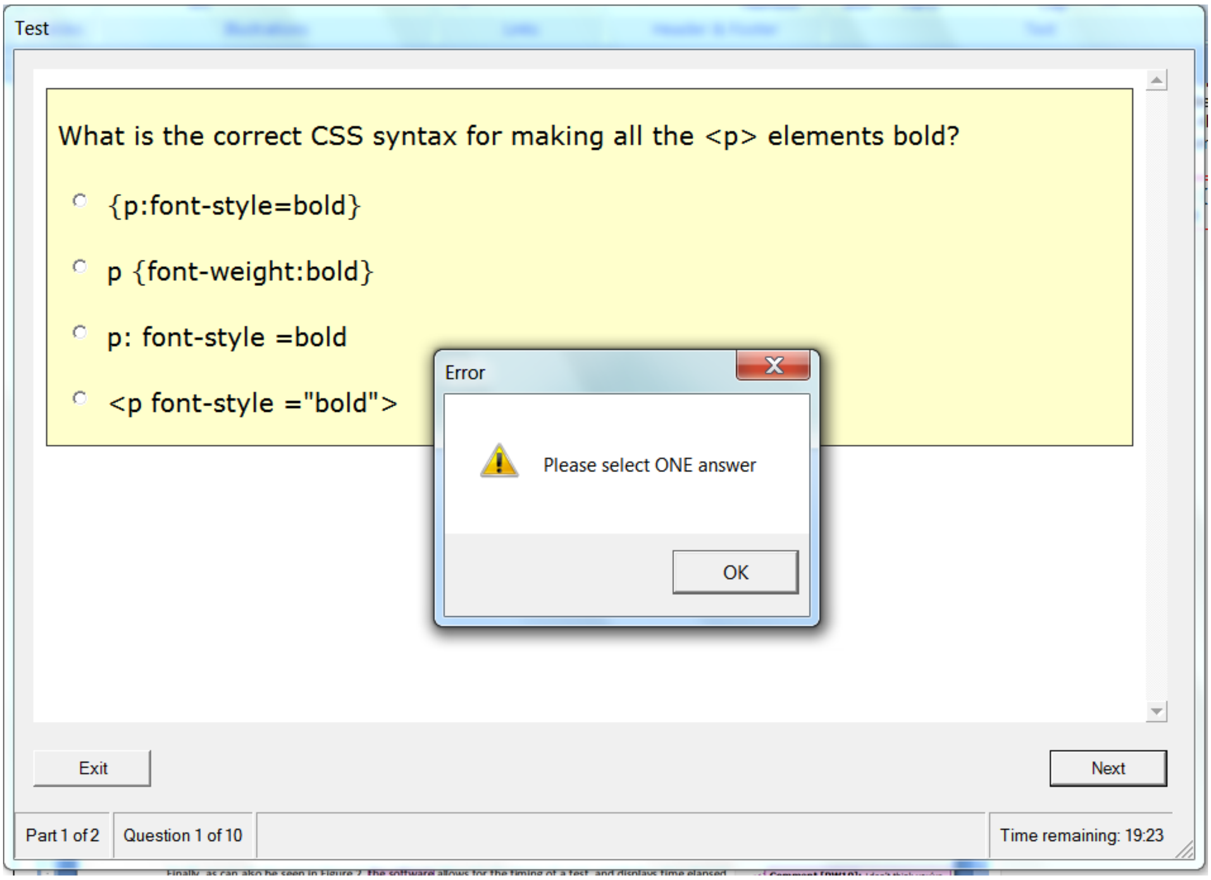
○ {p:font-style=bold}

○ p {font-weight:bold}

○ p: font-style =bold

○ <p font-style ="bold">

FIGURE A 3: QUESTION INTERFACE



What is the correct CSS syntax for making all the <p> elements bold?

○ {p:font-style=bold}

○ p {font-weight:bold}

○ p: font-style =bold

○ <p font-style ="bold">

Error

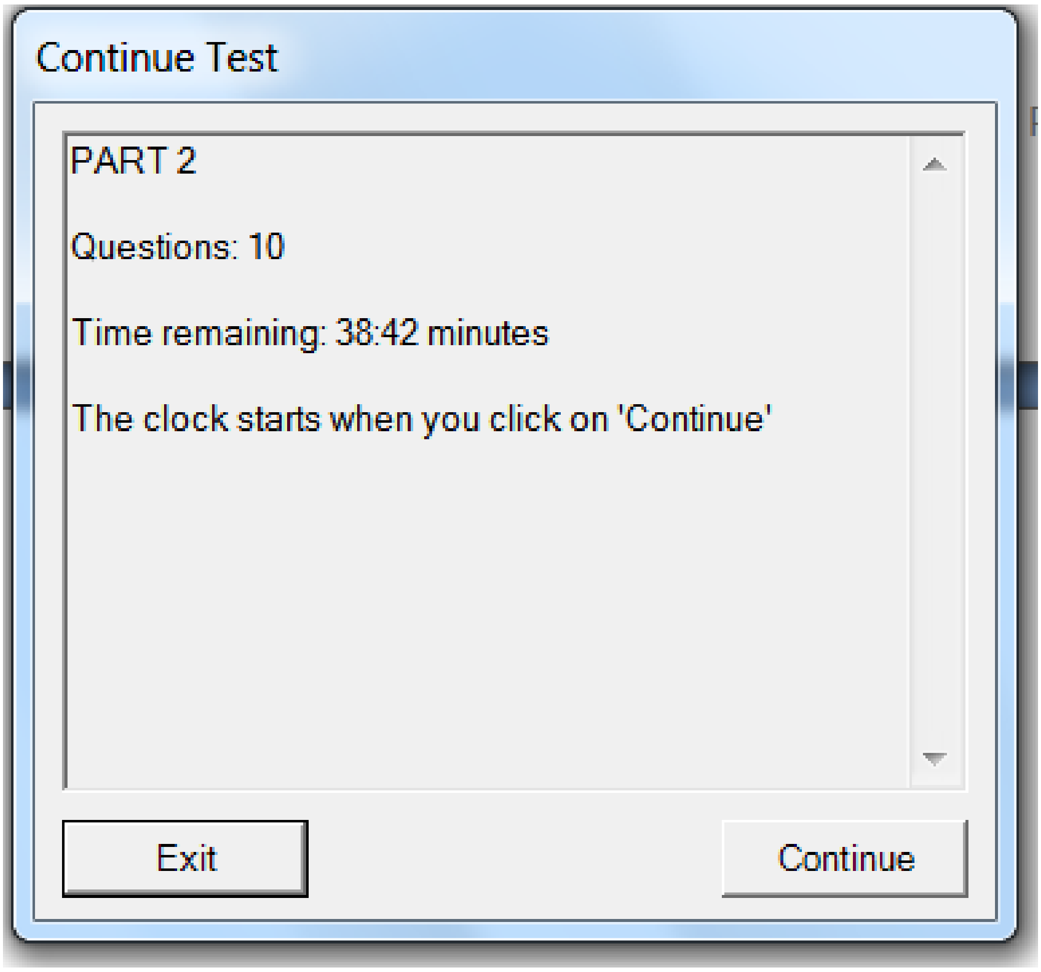
Please select ONE answer

Part 1 of 2

Question 1 of 10

Time remaining: 19:23

FIGURE A 4: ERROR SCREEN



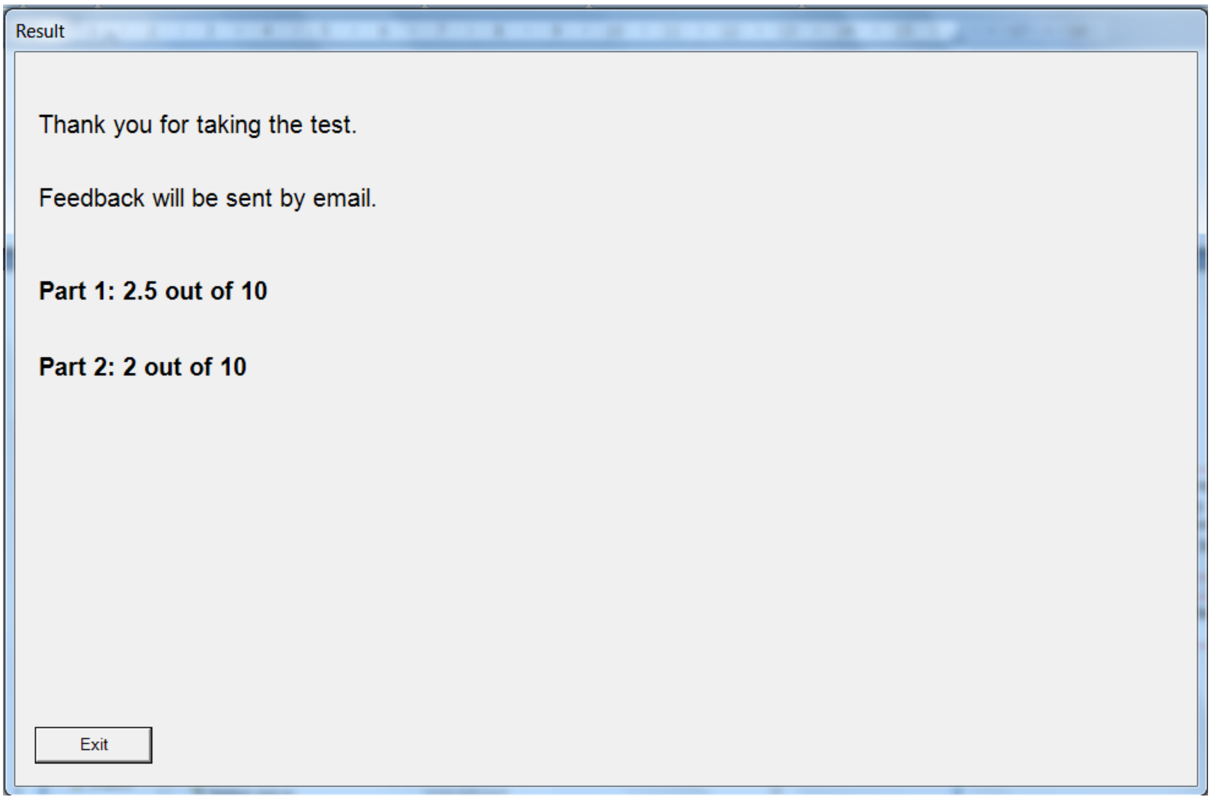
PART 2

Questions: 10

Time remaining: 38:42 minutes

The clock starts when you click on 'Continue'

FIGURE A 5: PART 2 QUESTIONS INSTRUCTIONS



Thank you for taking the test.

Feedback will be sent by email.

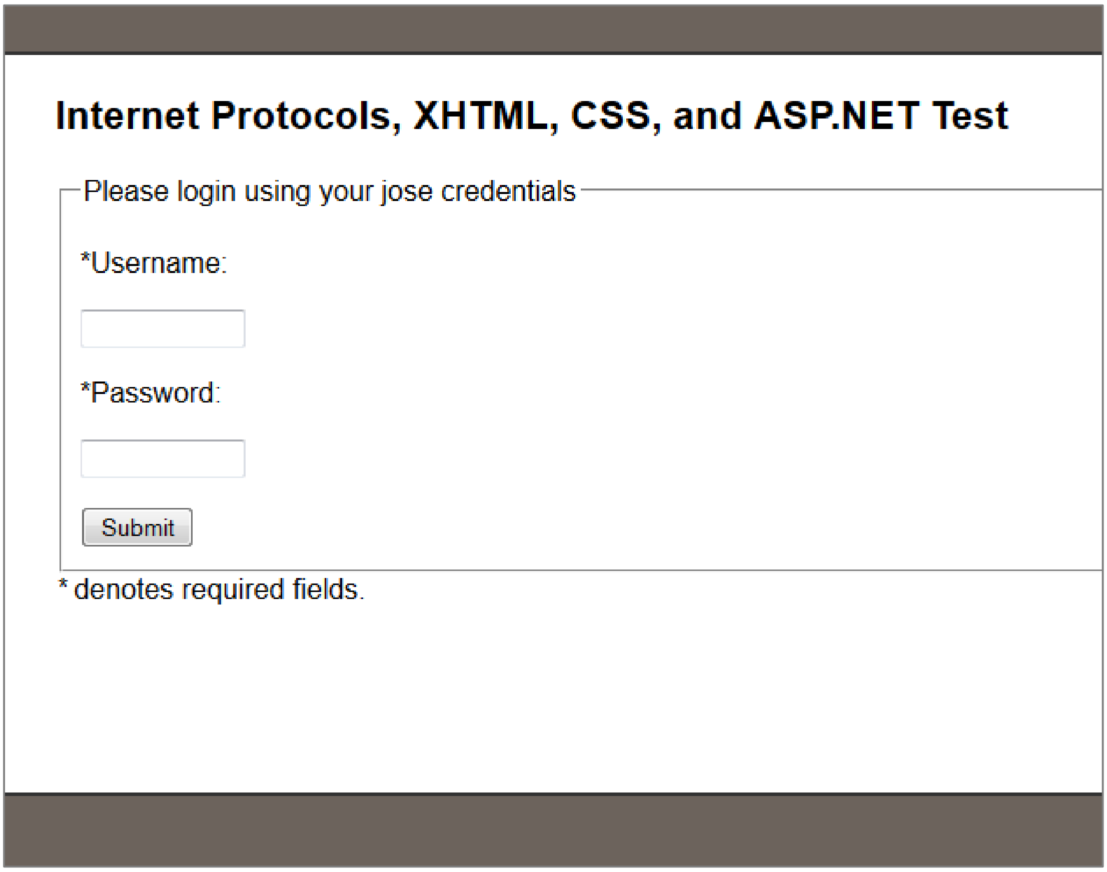
Part 1: 2.5 out of 10

Part 2: 2 out of 10

FIGURE A 6: FEEDBACK SCREEN FOR THE DESKTOP FLEXILEVEL TEST APPLICATION

APPENDIX B: SCREENSHOTS OF THE WEB APPLICATION

Appendix B



Internet Protocols, XHTML, CSS, and ASP.NET Test

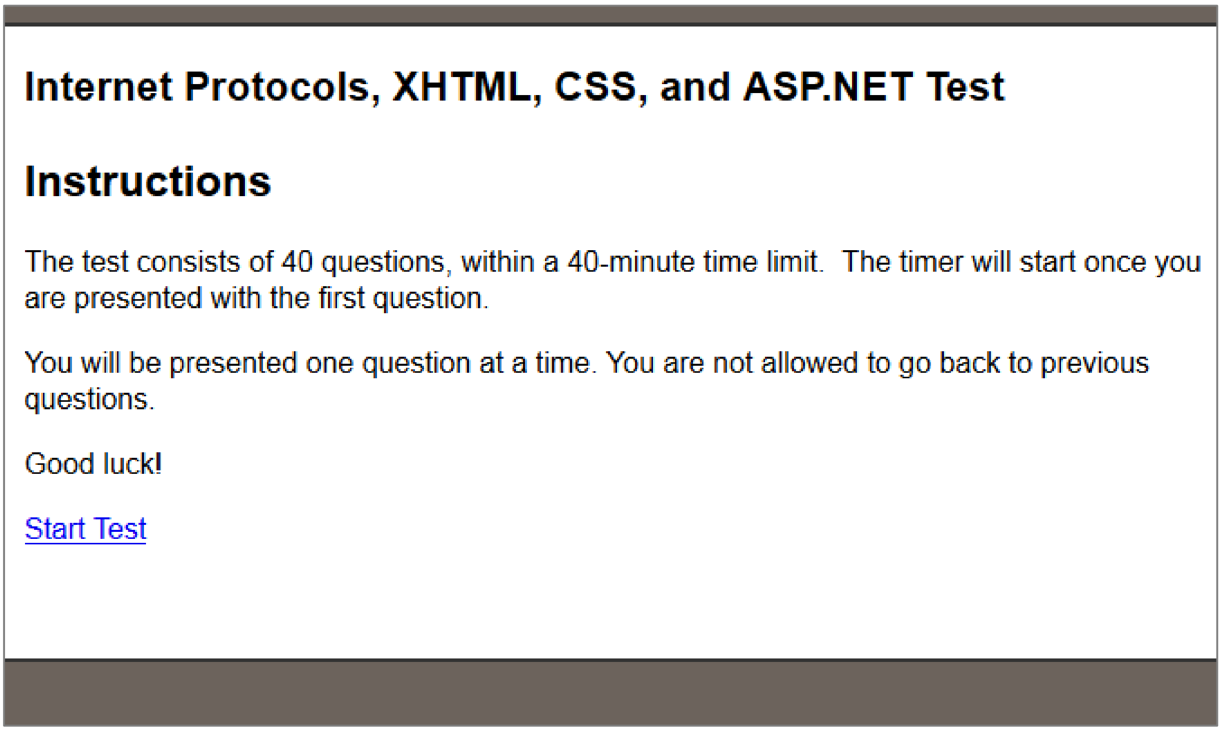
Please login using your jose credentials

*Username:

*Password:

* denotes required fields.

FIGURE B 1: LOGIN PAGE



Internet Protocols, XHTML, CSS, and ASP.NET Test

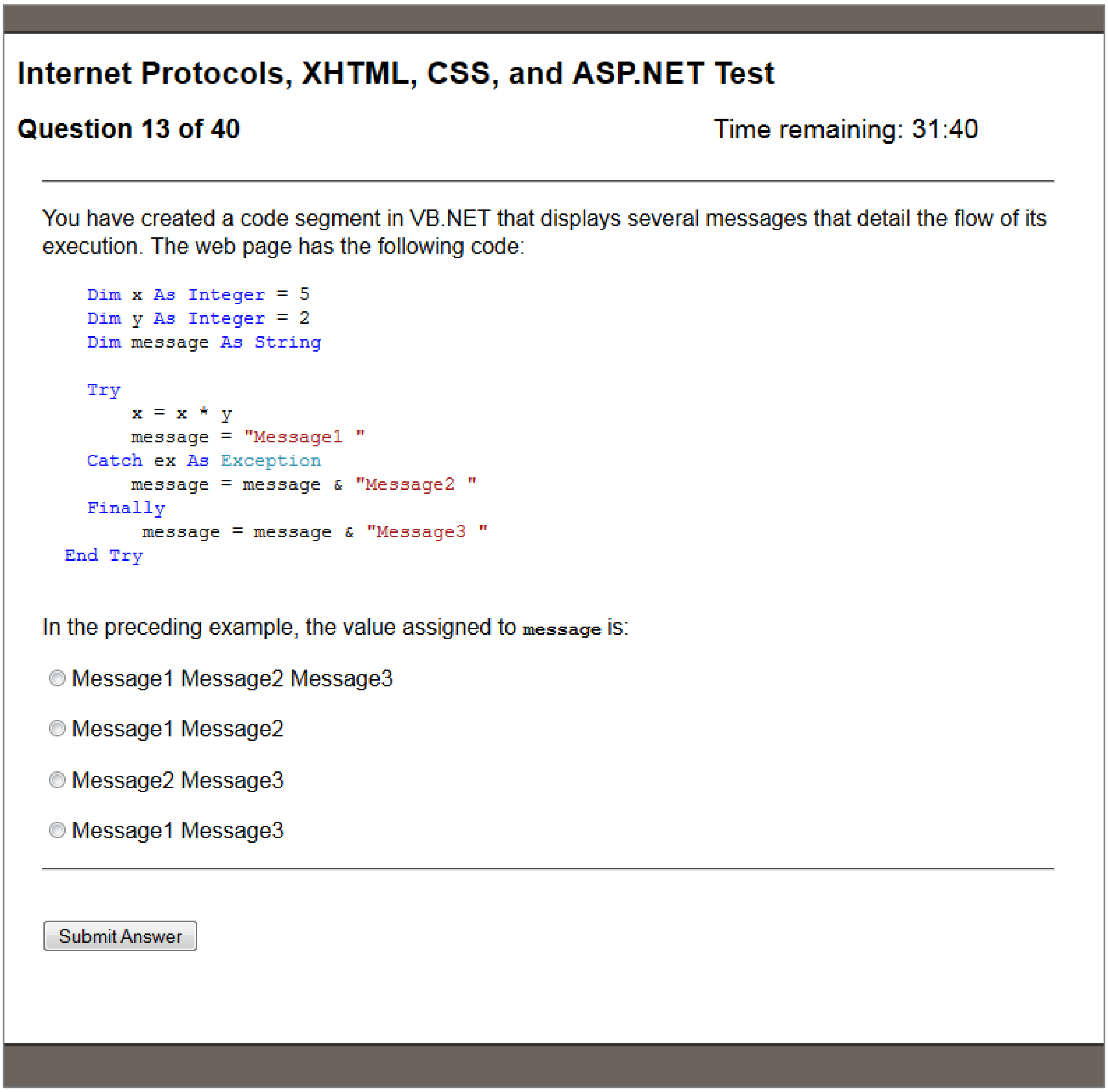
Instructions

The test consists of 40 questions, within a 40-minute time limit. The timer will start once you are presented with the first question.

You will be presented one question at a time. You are not allowed to go back to previous questions.

Good luck!

FIGURE B 2: WEB APPLICATION INSTRUCTIONS SCREEN



Internet Protocols, XHTML, CSS, and ASP.NET Test

Question 13 of 40

Time remaining: 31:40

You have created a code segment in VB.NET that displays several messages that detail the flow of its execution. The web page has the following code:

```
    Dim x As Integer = 5
    Dim y As Integer = 2
    Dim message As String

    Try
        x = x * y
        message = "Message1 "
    Catch ex As Exception
        message = message & "Message2 "
    Finally
         message = message & "Message3 "
  End Try
```

In the preceding example, the value assigned to `message` is:

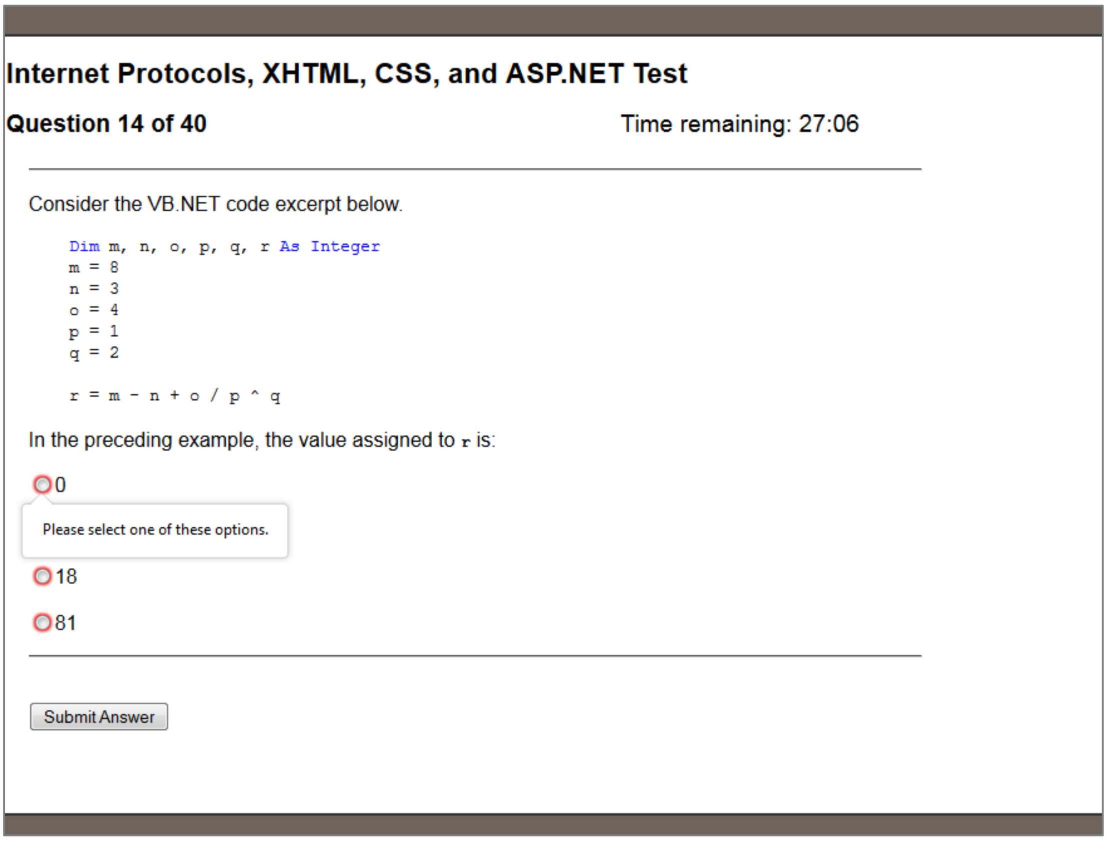
○ Message1 Message2 Message3

○ Message1 Message2

○ Message2 Message3

○ Message1 Message3

FIGURE B 3: TEST ITEM INTERFACE



Internet Protocols, XHTML, CSS, and ASP.NET Test

Question 14 of 40 Time remaining: 27:06

Consider the VB.NET code excerpt below.

```
Dim m, n, o, p, q, r As Integer
m = 8
n = 3
o = 4
p = 1
q = 2

r = m - n + o / p ^ q
```

In the preceding example, the value assigned to r is:

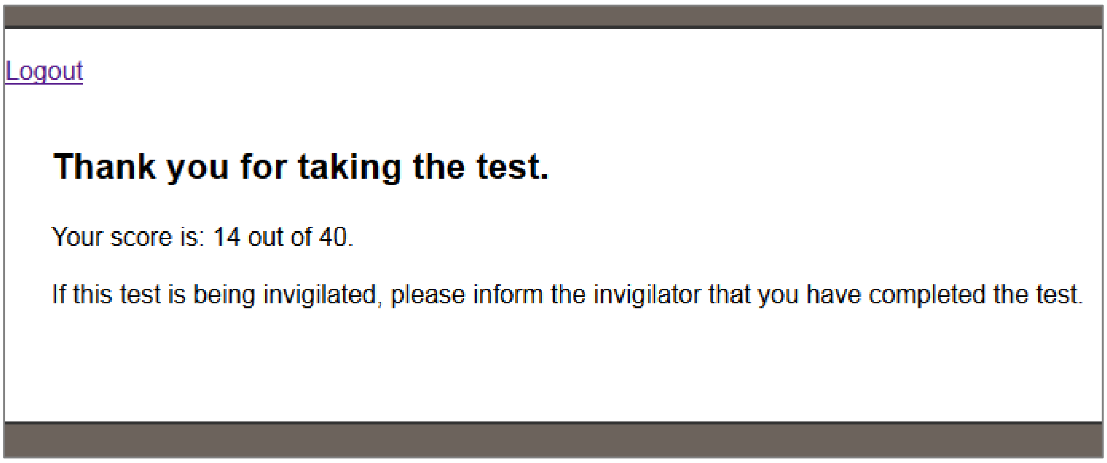
○ 0

Please select one of these options.

○ 18

○ 81

FIGURE B 4: ERROR MESSAGE



Thank you for taking the test.

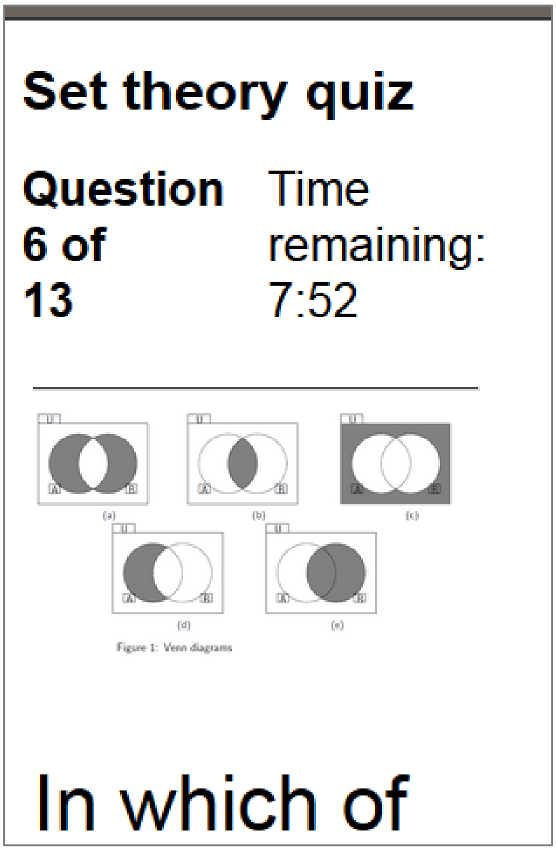
Your score is: 14 out of 40.

If this test is being invigilated, please inform the invigilator that you have completed the test.

FIGURE B 5: FEEDBACK SCREEN

APPENDIX C: SCREENSHOTS OF THE WEB APPLICATION SHOWN AT DIFFERENT MOBILE DEVICE RESOLUTIONS

Appendix C



Set theory quiz

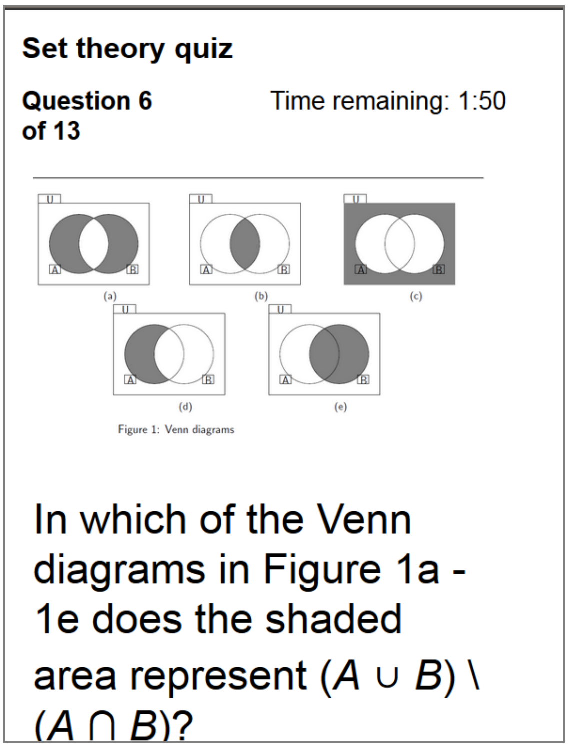
Question 6 of 13

Time remaining: 7:52

Figure 1: Venn diagrams

In which of

FIGURE B 6: AN IMAGE BASED ITEM AS SHOWN ON A MOBILE PHONE (320 X 480)



Set theory quiz

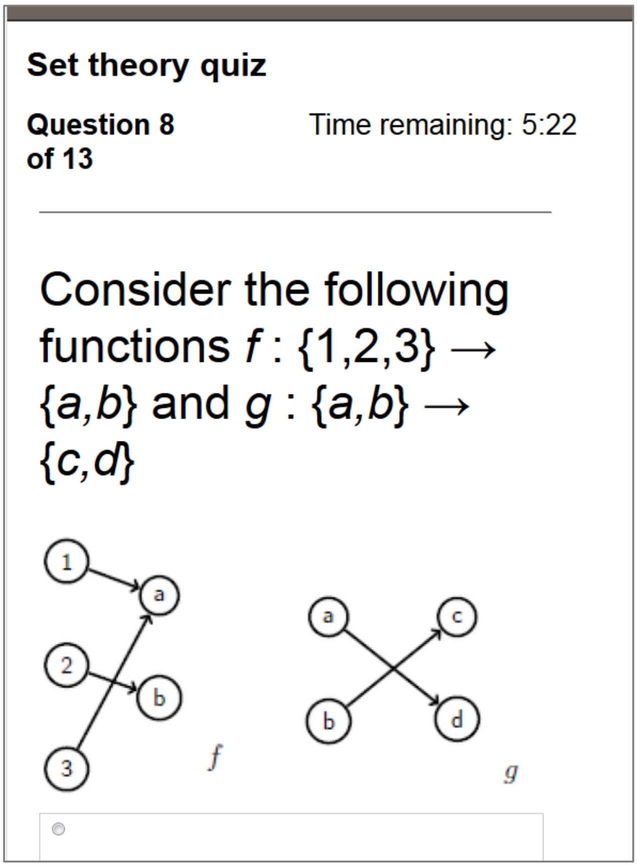
Question 6 of 13

Time remaining: 1:50

Figure 1: Venn diagrams

In which of the Venn diagrams in Figure 1a - 1e does the shaded area represent $(A \cup B) \setminus (A \cap B)$?

FIGURE B 7: AN IMAGE BASED ITEM AS SHOWN ON A SMALL TABLET IN PORTRAIT (600 X 800)



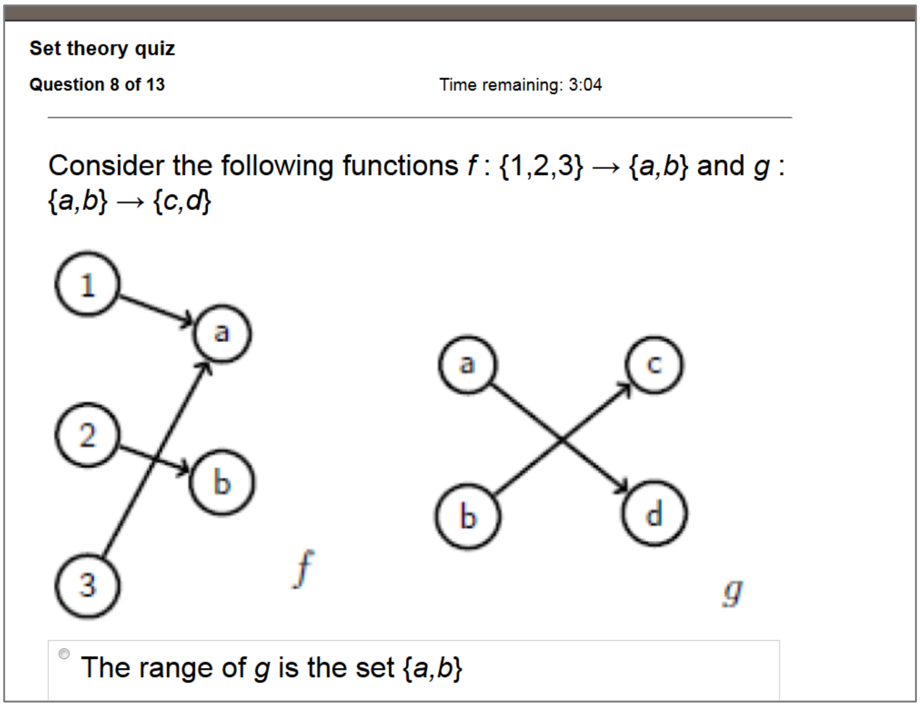
Set theory quiz

Question 8 of 13

Time remaining: 5:22

Consider the following functions $f : \{1,2,3\} \rightarrow \{a,b\}$ and $g : \{a,b\} \rightarrow \{c,d\}$

FIGURE B 8: AN IMAGE BASED ITEM AS SHOWN ON A SMALL TABLET (600 X 800)



Set theory quiz

Question 8 of 13

Time remaining: 3:04

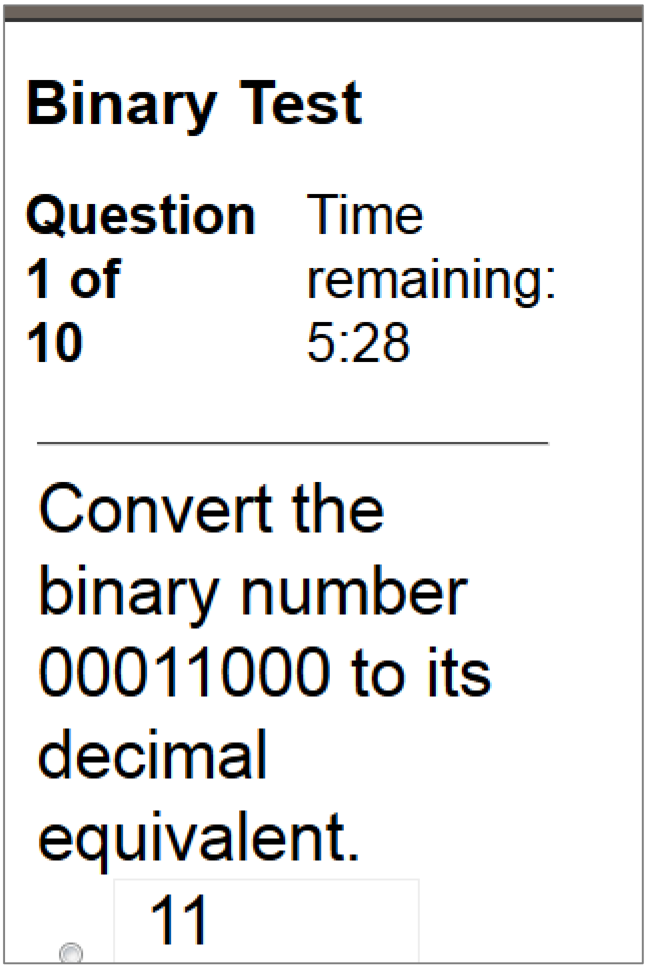
Consider the following functions $f : \{1,2,3\} \rightarrow \{a,b\}$ and $g : \{a,b\} \rightarrow \{c,d\}$

○ The range of g is the set $\{a,b\}$

FIGURE B 9: AN IMAGE BASED ITEM AS SHOWN ON A TABLET IN LANDSCAPE MODE (1024 X 768)

APPENDIX D: SCREENSHOTS OF THE WEB APPLICATION SHOWN AT DIFFERENT MOBILE DEVICE RESOLUTIONS

Appendix D



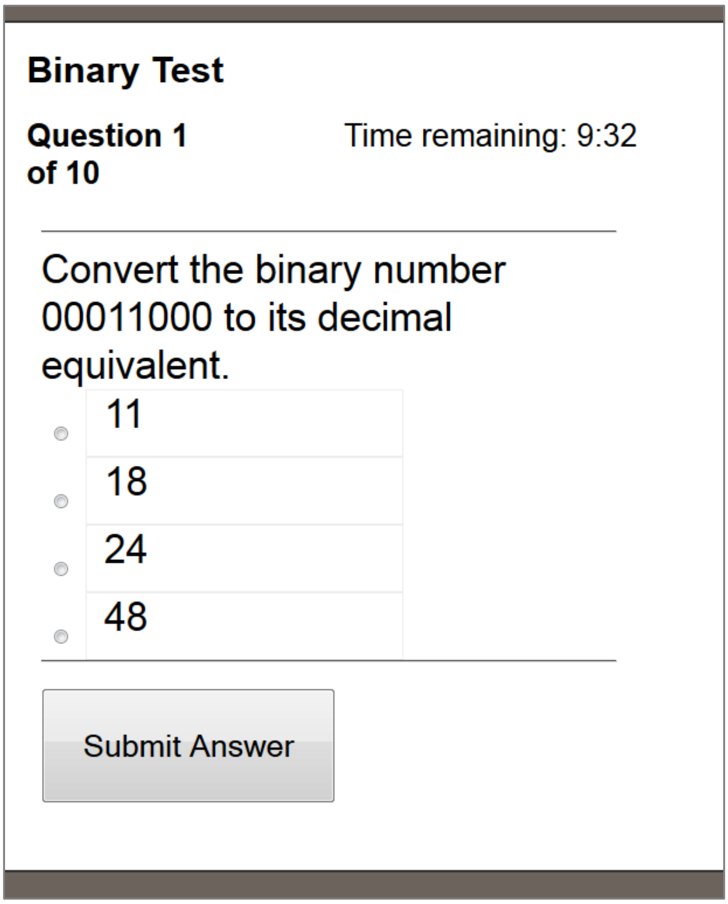
Binary Test

Question 1 of 10 Time remaining: 5:28

Convert the binary number 00011000 to its decimal equivalent.

○ 11

FIGURE D 1: A SIMPLE FORMAT ITEM AS SHOWN ON A MOBILE PHONE (320 X 480)


Binary Test

Question 1 of 10

Time remaining: 9:32

Convert the binary number 00011000 to its decimal equivalent.

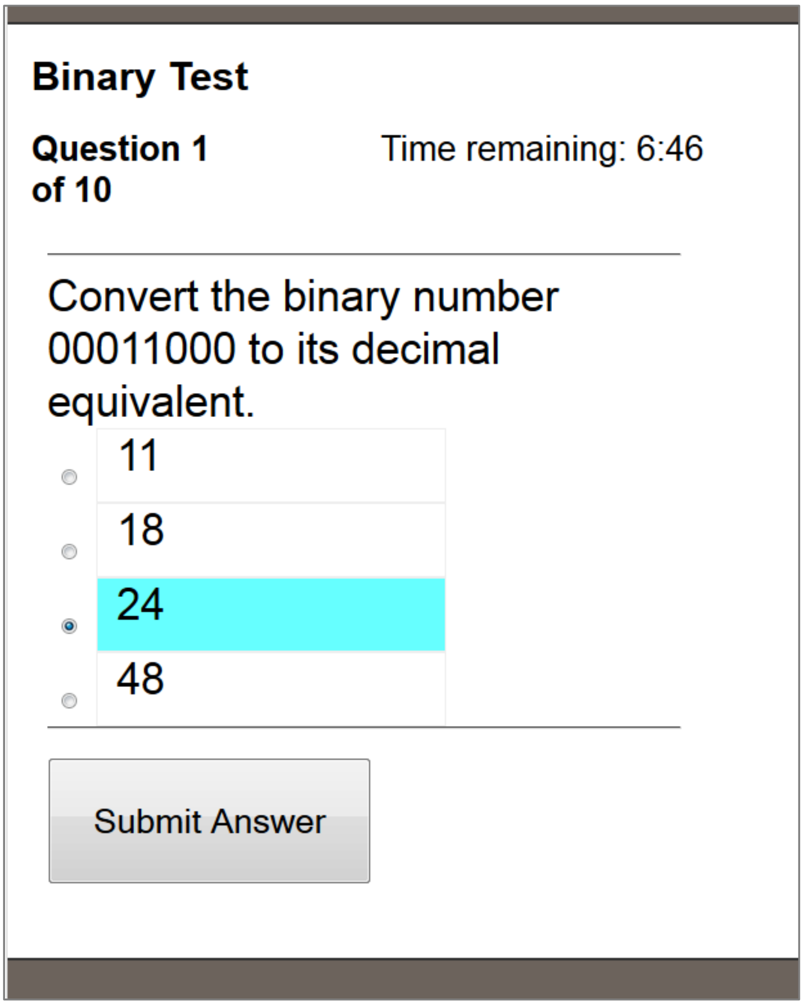
○ 11

○ 18

○ 24

○ 48

FIGURE D 2: A SIMPLE FORMAT ITEM SHOWN AS DISPLAYED ON A SMALL TABLET (600X 800)



Binary Test

Question 1 of 10

Time remaining: 6:46

Convert the binary number 00011000 to its decimal equivalent.

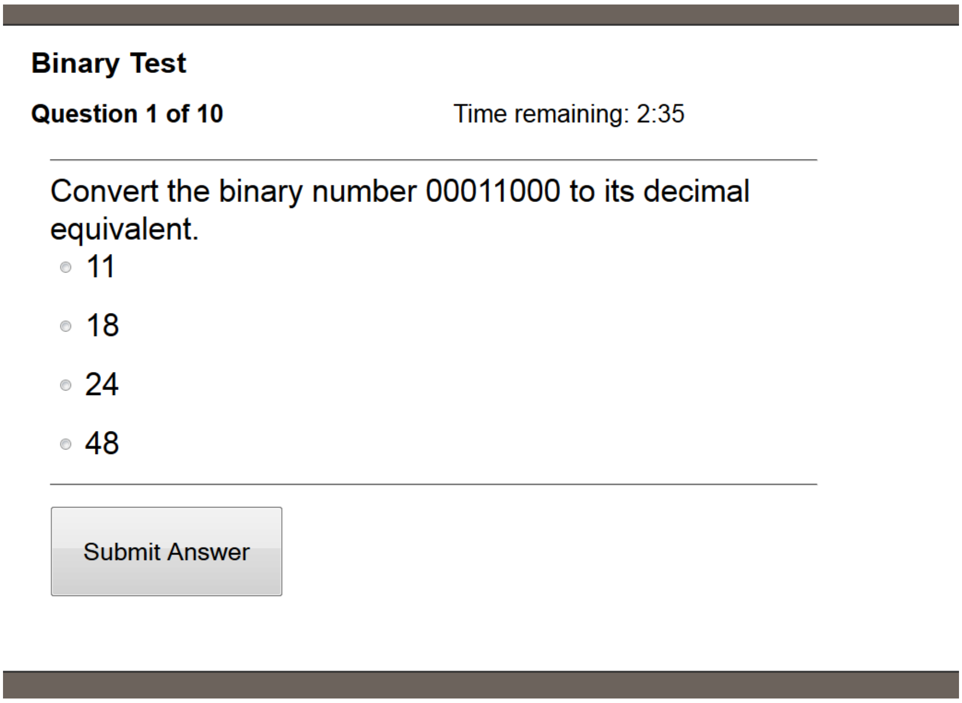
○ 11

○ 18

● 24

○ 48

FIGURE D 3: A SIMPLER ITEM WITH AN OPTION SELECTED SHOWN AS DISPLAYED ON A SMALL TABLET (600X 800)



Binary Test

Question 1 of 10 Time remaining: 2:35

Convert the binary number 00011000 to its decimal equivalent.

○ 11

○ 18

○ 24

○ 48

FIGURE D 4: A SIMPLER ITEM SHOWN AS DISPLAYED ON A TABLET (768 X 1024)

APPENDIX E: QUESTIONNAIRE USED IN STUDY L

Appendix E

Academic staff briefing

The application that you have just seen is an adaptive testing application based on the Flexilevel approach, as proposed by Lord (1980). In a traditional computer-based test, all test-takers are presented with the same set of questions regardless of their proficiency levels within the subject domain.

In a Flexilevel test, the test starts with a question of medium difficulty. In the event of a correct response, a more difficult question will be administered next. An incorrect response will cause an easier question to follow. This process will be repeated until a predetermined number of questions has been administered, or the time limit has elapsed (whichever happens first).

The following 12 statements refer to the application's ease of use and perceived usefulness. Please rate each statement using the scale provided. You will also be asked to provide your views on three points regarding this study.

Ease of Use

1. Learning to use the Flexilevel software application would be easy for students.

□	□	□	□	□
1	2	3	4	5 Strongly
Strongly disagree	Disagree	Neutral	Agree	agree

2. Students would find it easy to remember how to perform tasks using the Flexilevel software application.

□	□	□	□	□
1	2	3	4	5 Strongly

Strongly disagree	Disagree	Neutral	Agree	agree

3. Students would find the Flexilevel software application easy to use.

□	□	□	□	□
1	2	3	4	5 Strongly
Strongly disagree	Disagree	Neutral	Agree	agree

Perceived usefulness

4. I would find the Flexilevel approach useful in a formative assessment context.

□	□	□	□	□
1	2	3	4	5 Strongly
Strongly disagree	Disagree	Neutral	Agree	agree

5. I would find the Flexilevel approach useful in a summative assessment context.

□	□	□	□	□
1	2	3	4	5 Strongly
Strongly disagree	Disagree	Neutral	Agree	agree

6. The adaptivity supported by the Flexilevel approach would help me in identifying those students who need greater support.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

7. The adaptivity supported by the Flexilevel approach would enhance the student assessment experience.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

8. Students would find the Flexilevel approach useful in a formative assessment context.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

9. Students would find the adaptive Flexilevel approach useful in a summative assessment context.

□	□	□	□	□
1	2	3	4	5

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

10. I would prefer to use the Flexilevel approach than other forms of objective testing for formative assessment.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

11. I would prefer to use the Flexilevel approach than other forms of objective testing for summative assessment.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

12. Assuming the Flexilevel software was available to me, I predict that I would use it on a regular basis to complement the existing range of assessment methods.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Your views

What barriers, if any, can you see to the uptake of this adaptive testing approach?

What benefits, if any, can you see this of this adaptive testing approach?

Is there a question that you would like to have been asked? If so, what is it and how you would answer it.

APPENDIX F: QUESTIONNAIRE USED IN PILOT STUDY

Appendix F

Students briefing

The test that you have just taken is part of a project that we are working on.

The test was divided into two sections: the blue section and the yellow section.

The blue section was a traditional computer-based test or, in other words, a test in which all test-takers are presented with the same set of questions regardless of their levels of understanding. The yellow section was an adaptive test, based on the Flexilevel approach.

In a Flexilevel test, the test starts with a question of medium difficulty. In the event of a correct response, a more difficult question will be administered next. An incorrect response will cause an easier question to follow.

The scoring system is somewhat similar to that used in traditional computer-based tests, in that you receive 1 mark for a correct response, and no mark for an incorrect response. The main difference between the Flexilevel and traditional scoring methods is that, in the Flexilevel approach, you will be awarded 0.5 marks if your response to the last question of the test is incorrect. This is because it is assumed that test-takers will probably answer the next (easier) question correctly if they had the chance.

The following 15 statements refer to the application's ease of use and perceived usefulness. Please rate each statement using the scale provided.

Ease of use

1. Learning to use the Flexilevel software application would be easy for me.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

2. I would find it easy to remember how to perform tasks (e.g. how to answer a question) using the Flexilevel software application.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

3. I would find the Flexilevel software application easy to use.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Perceived usefulness

4. I would find the Flexilevel approach useful for practice tests.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

5. I would find the Flexilevel approach useful in summative tests (i.e. the test score counts towards my final grade).

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

6. The adaptivity supported by the Flexilevel approach would help me to identify the areas in which I need to work harder more quickly.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

7. The adaptivity supported by the Flexilevel approach would enhance my overall assessment experience.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

8. For practice tests, I would prefer using the Flexilevel approach to other forms of objective testing.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

9. For summative tests, I would prefer using the Flexilevel approach to other forms of objective testing.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

10. The system used to score a Flexilevel test makes sense to me.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

11. I would find the score provided by the Flexilevel approach useful at identifying how much I have learned.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

12. I would find it useful if the level of difficulty of a test is tailored to my level of understanding.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

13. Assuming the Flexilevel software was available to me for practice tests, I predict that I would use it on a regular basis.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly Agree

14. In practice tests, test questions that are too easy are less engaging than those questions that are tailored to my level of understanding.

□	□	□	□	□
1	2	3	4	5
	Disagree	Neutral	Agree	Strongly

Strongly disagree				agree

15. In practice tests, test questions that are too difficult are less engaging than those questions that are tailored to my level of understanding.

□	□	□	□	□
1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Your views

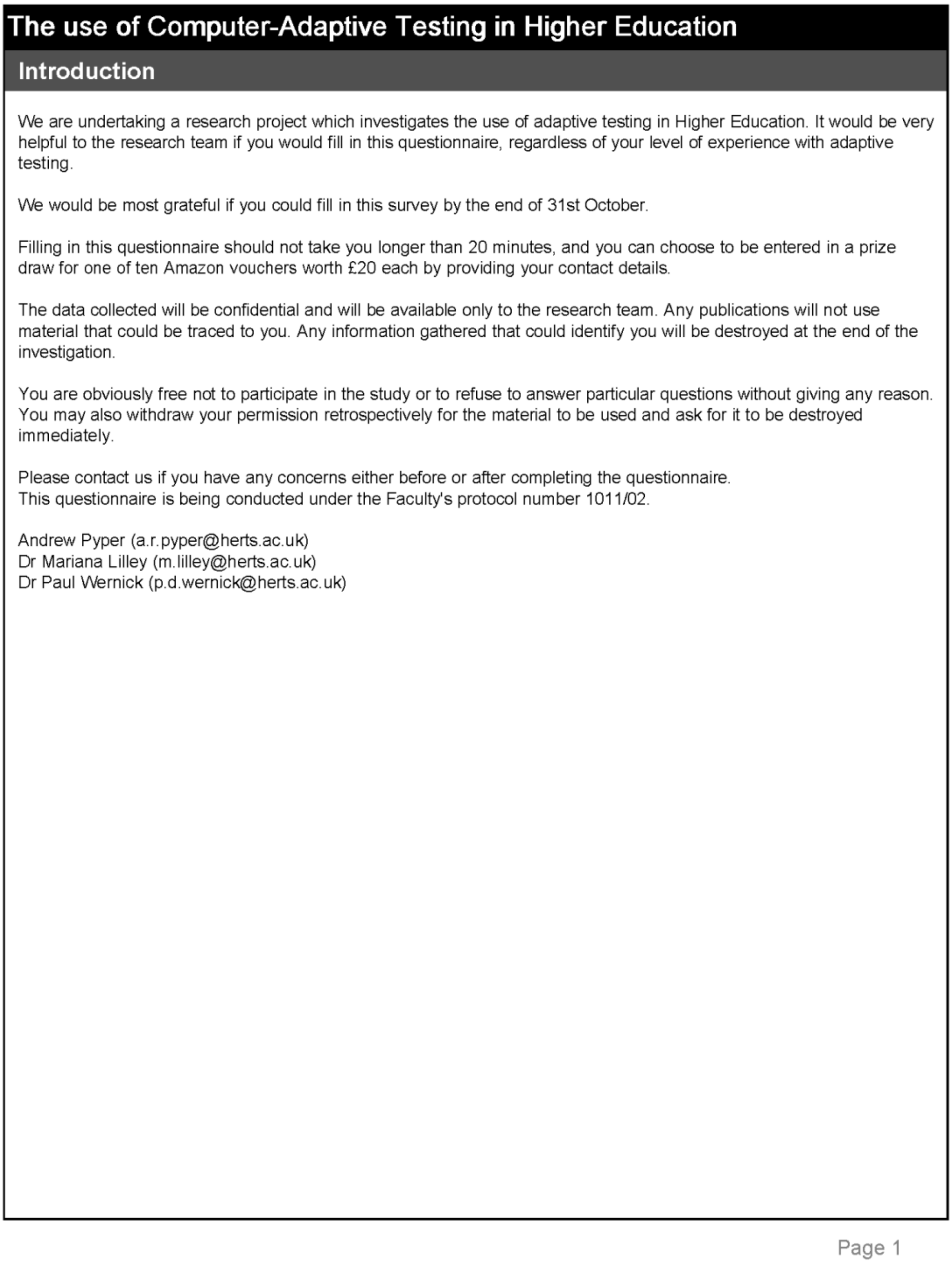
What problems, if any, can you see to the uptake of this adaptive testing approach?

Can you see any benefits of this adaptive testing approach?

Is there a question that you would like to have been asked? If so, what is it and how you would answer it.

APPENDIX G: QUESTIONNAIRE USED IN STUDY K

Appendix G



The use of Computer-Adaptive Testing in Higher Education

Introduction

We are undertaking a research project which investigates the use of adaptive testing in Higher Education. It would be very helpful to the research team if you would fill in this questionnaire, regardless of your level of experience with adaptive testing.

We would be most grateful if you could fill in this survey by the end of 31st October.

Filling in this questionnaire should not take you longer than 20 minutes, and you can choose to be entered in a prize draw for one of ten Amazon vouchers worth £20 each by providing your contact details.

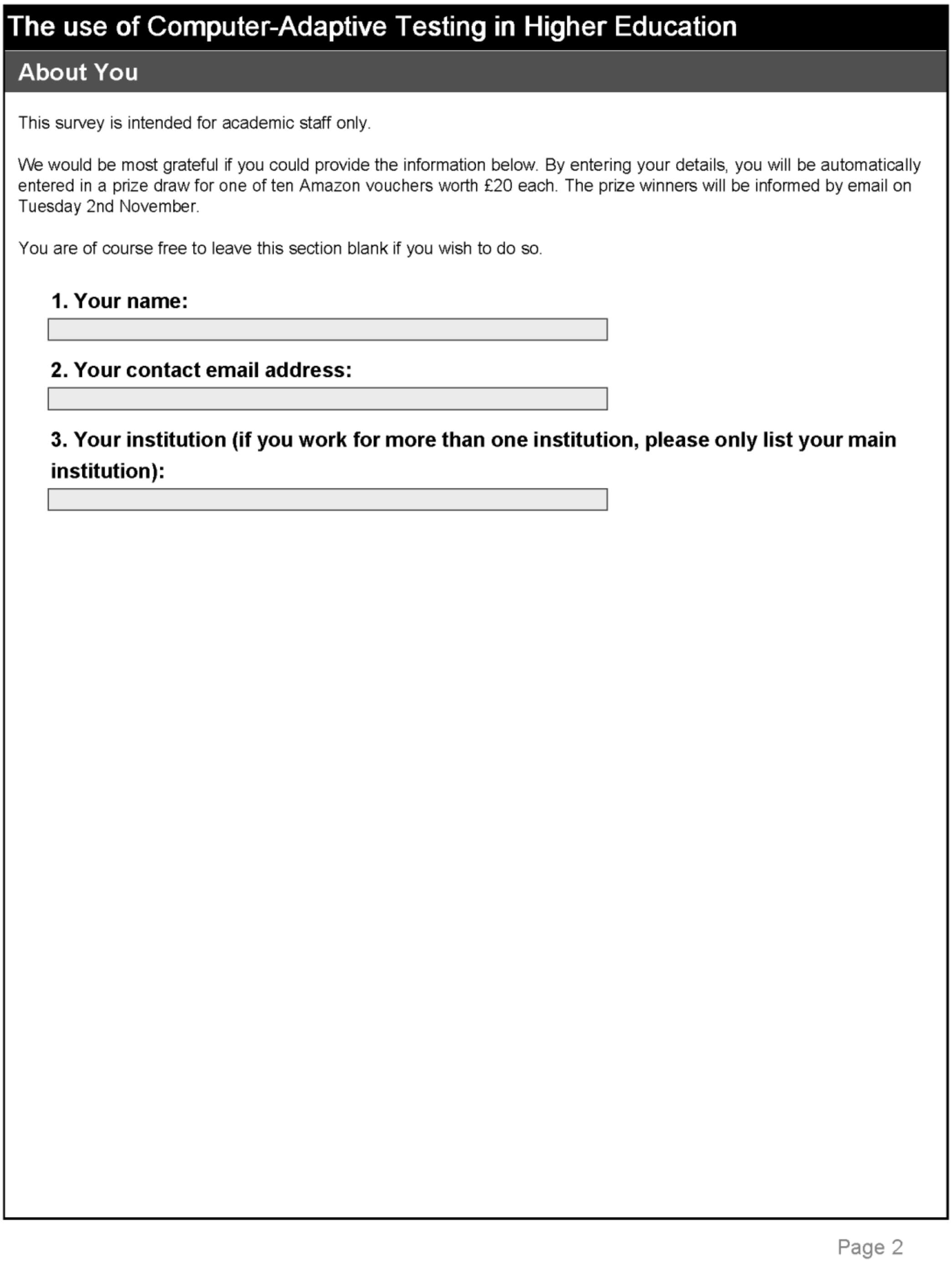
The data collected will be confidential and will be available only to the research team. Any publications will not use material that could be traced to you. Any information gathered that could identify you will be destroyed at the end of the investigation.

You are obviously free not to participate in the study or to refuse to answer particular questions without giving any reason. You may also withdraw your permission retrospectively for the material to be used and ask for it to be destroyed immediately.

Please contact us if you have any concerns either before or after completing the questionnaire.
This questionnaire is being conducted under the Faculty's protocol number 1011/02.

Andrew Pyper (a.r.pyper@herts.ac.uk)
Dr Mariana Lilley (m.lilley@herts.ac.uk)
Dr Paul Wernick (p.d.wernick@herts.ac.uk)

Page 1



The use of Computer-Adaptive Testing in Higher Education

About You

This survey is intended for academic staff only.

We would be most grateful if you could provide the information below. By entering your details, you will be automatically entered in a prize draw for one of ten Amazon vouchers worth £20 each. The prize winners will be informed by email on Tuesday 2nd November.

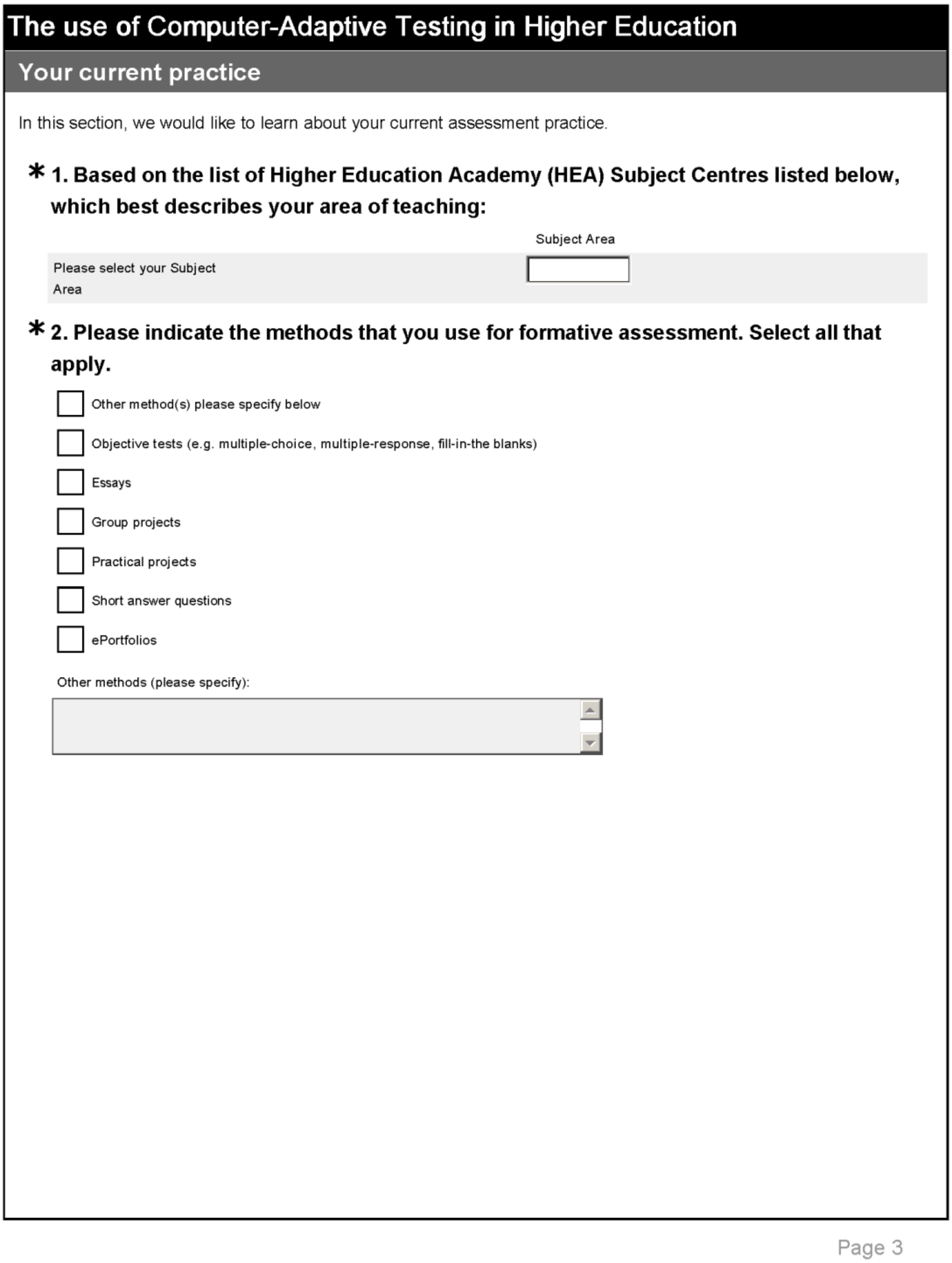
You are of course free to leave this section blank if you wish to do so.

1. Your name:

2. Your contact email address:

3. Your institution (if you work for more than one institution, please only list your main institution):

Page 2



The use of Computer-Adaptive Testing in Higher Education

Your current practice

In this section, we would like to learn about your current assessment practice.

*** 1. Based on the list of Higher Education Academy (HEA) Subject Centres listed below, which best describes your area of teaching:**

	Subject Area
Please select your Subject Area	□

*** 2. Please indicate the methods that you use for formative assessment. Select all that apply.**

□ Other method(s) please specify below

□ Objective tests (e.g. multiple-choice, multiple-response, fill-in-the blanks)

□ Essays

□ Group projects

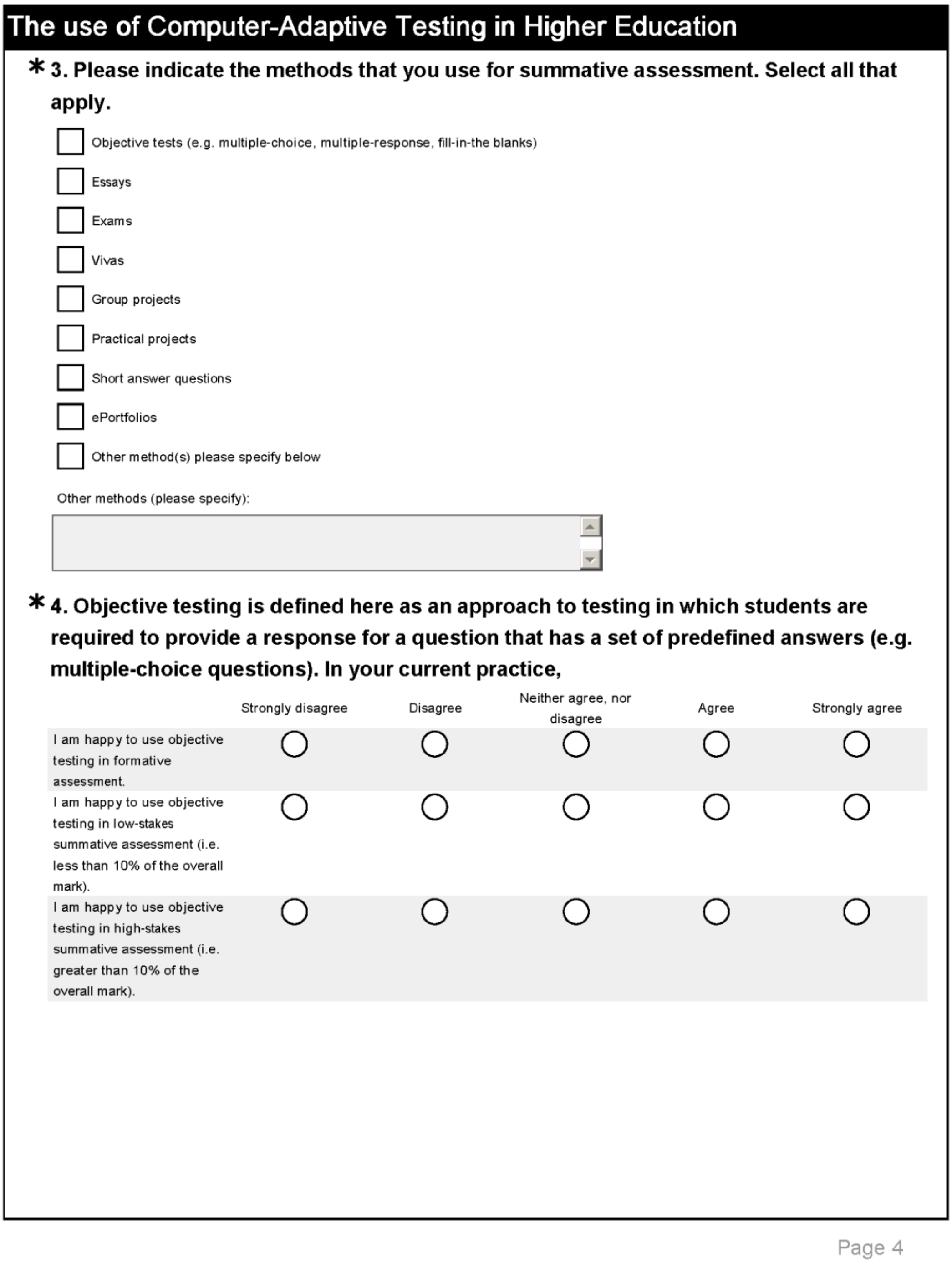
□ Practical projects

□ Short answer questions

□ ePortfolios

Other methods (please specify):

Page 3



The use of Computer-Adaptive Testing in Higher Education

*** 3. Please indicate the methods that you use for summative assessment. Select all that apply.**

□ Objective tests (e.g. multiple-choice, multiple-response, fill-in-the blanks)

□ Essays

□ Exams

□ Vivas

□ Group projects

□ Practical projects

□ Short answer questions

□ ePortfolios

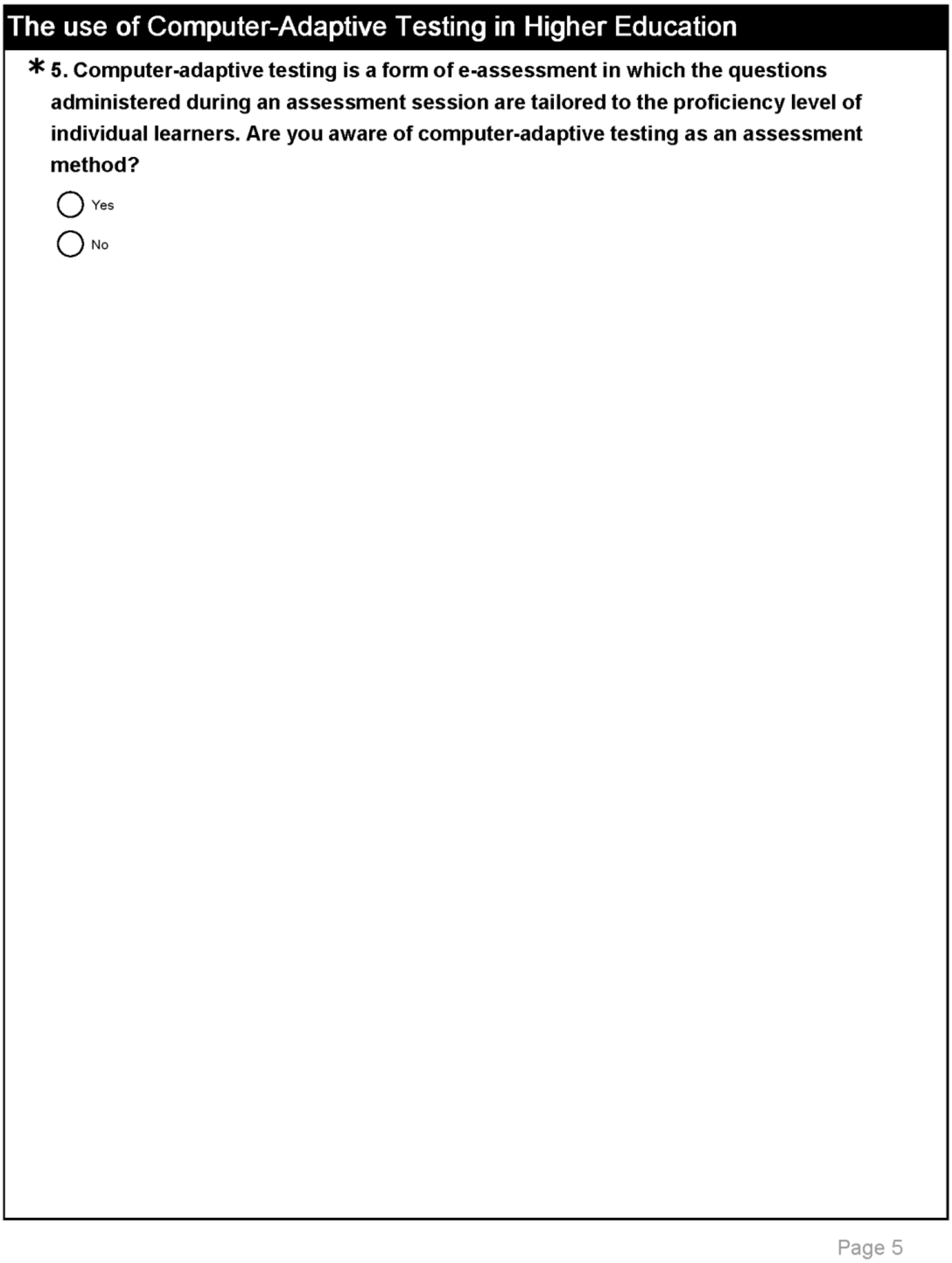
□ Other method(s) please specify below

Other methods (please specify):

*** 4. Objective testing is defined here as an approach to testing in which students are required to provide a response for a question that has a set of predefined answers (e.g. multiple-choice questions). In your current practice,**

	Strongly disagree	Disagree	Neither agree, nor disagree	Agree	Strongly agree
I am happy to use objective testing in formative assessment.	○	○	○	○	○
I am happy to use objective testing in low-stakes summative assessment (i.e. less than 10% of the overall mark).	○	○	○	○	○
I am happy to use objective testing in high-stakes summative assessment (i.e. greater than 10% of the overall mark).	○	○	○	○	○

Page 4



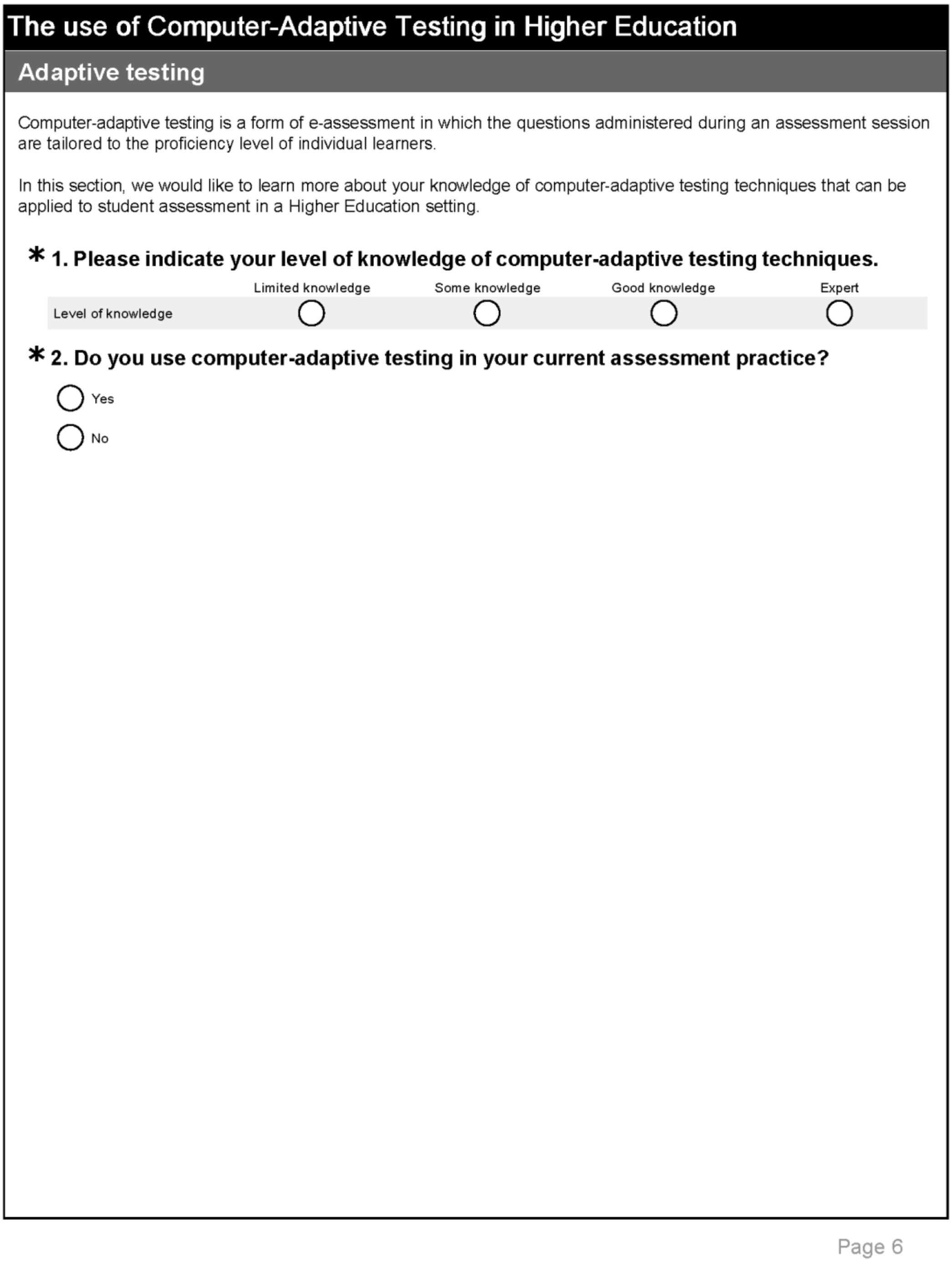
The use of Computer-Adaptive Testing in Higher Education

*** 5. Computer-adaptive testing is a form of e-assessment in which the questions administered during an assessment session are tailored to the proficiency level of individual learners. Are you aware of computer-adaptive testing as an assessment method?**

○ Yes

○ No

Page 5



The use of Computer-Adaptive Testing in Higher Education

Adaptive testing

Computer-adaptive testing is a form of e-assessment in which the questions administered during an assessment session are tailored to the proficiency level of individual learners.

In this section, we would like to learn more about your knowledge of computer-adaptive testing techniques that can be applied to student assessment in a Higher Education setting.

*** 1. Please indicate your level of knowledge of computer-adaptive testing techniques.**

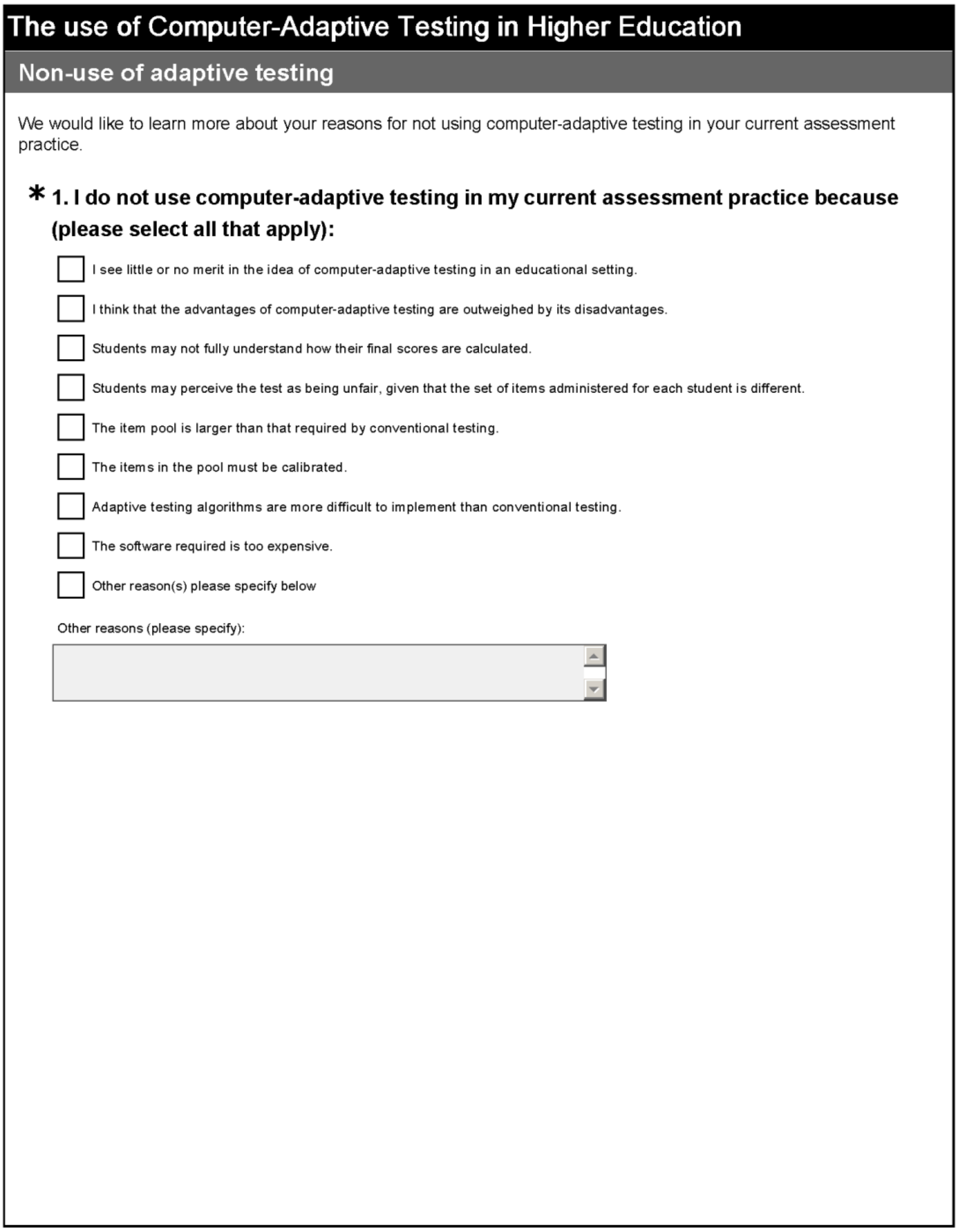
	Limited knowledge	Some knowledge	Good knowledge	Expert
Level of knowledge	○	○	○	○

*** 2. Do you use computer-adaptive testing in your current assessment practice?**

○ Yes

○ No

Page 6



The use of Computer-Adaptive Testing in Higher Education

Non-use of adaptive testing

We would like to learn more about your reasons for not using computer-adaptive testing in your current assessment practice.

*** 1. I do not use computer-adaptive testing in my current assessment practice because (please select all that apply):**

□ I see little or no merit in the idea of computer-adaptive testing in an educational setting.

□ I think that the advantages of computer-adaptive testing are outweighed by its disadvantages.

□ Students may not fully understand how their final scores are calculated.

□ Students may perceive the test as being unfair, given that the set of items administered for each student is different.

□ The item pool is larger than that required by conventional testing.

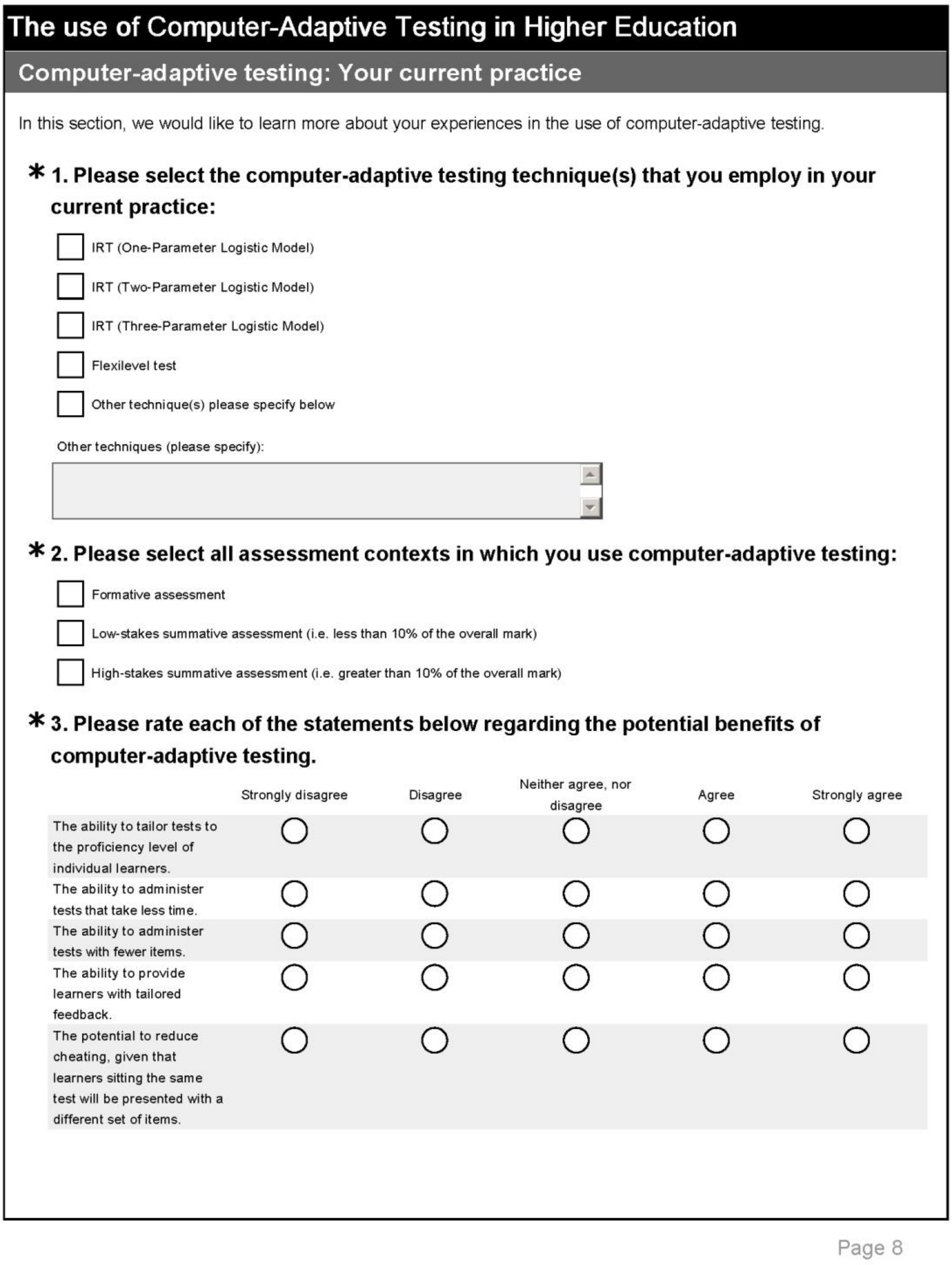
□ The items in the pool must be calibrated.

□ Adaptive testing algorithms are more difficult to implement than conventional testing.

□ The software required is too expensive.

□ Other reason(s) please specify below

Other reasons (please specify):



The use of Computer-Adaptive Testing in Higher Education

Computer-adaptive testing: Your current practice

In this section, we would like to learn more about your experiences in the use of computer-adaptive testing.

*** 1. Please select the computer-adaptive testing technique(s) that you employ in your current practice:**

□ IRT (One-Parameter Logistic Model)

□ IRT (Two-Parameter Logistic Model)

□ IRT (Three-Parameter Logistic Model)

□ Flexilevel test

□ Other technique(s) please specify below

Other techniques (please specify):

*** 2. Please select all assessment contexts in which you use computer-adaptive testing:**

□ Formative assessment

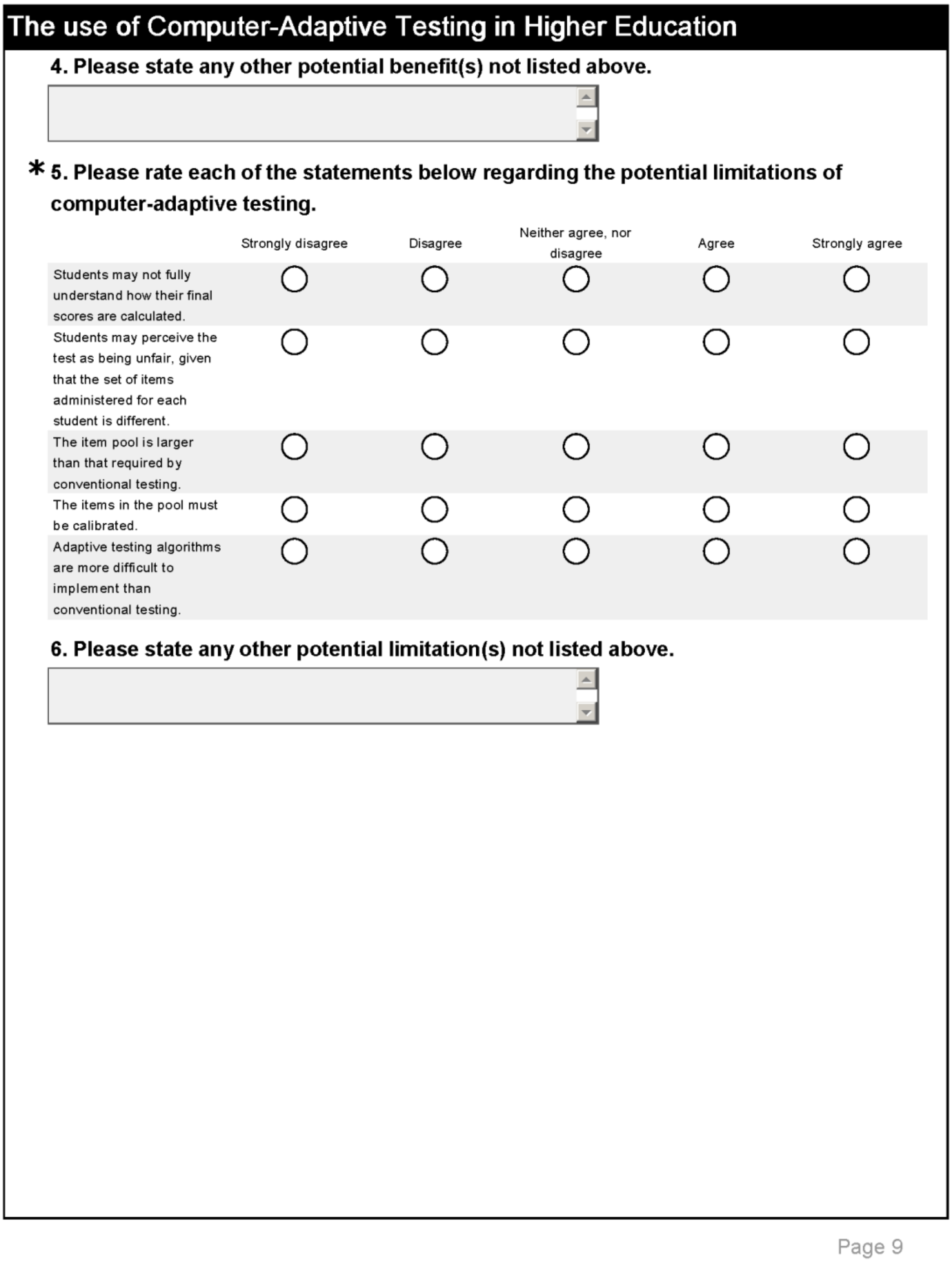
□ Low-stakes summative assessment (i.e. less than 10% of the overall mark)

□ High-stakes summative assessment (i.e. greater than 10% of the overall mark)

*** 3. Please rate each of the statements below regarding the potential benefits of computer-adaptive testing.**

	Strongly disagree	Disagree	Neither agree, nor disagree	Agree	Strongly agree
The ability to tailor tests to the proficiency level of individual learners.	○	○	○	○	○
The ability to administer tests that take less time.	○	○	○	○	○
The ability to administer tests with fewer items.	○	○	○	○	○
The ability to provide learners with tailored feedback.	○	○	○	○	○
The potential to reduce cheating, given that learners sitting the same test will be presented with a different set of items.	○	○	○	○	○

Page 8



The use of Computer-Adaptive Testing in Higher Education

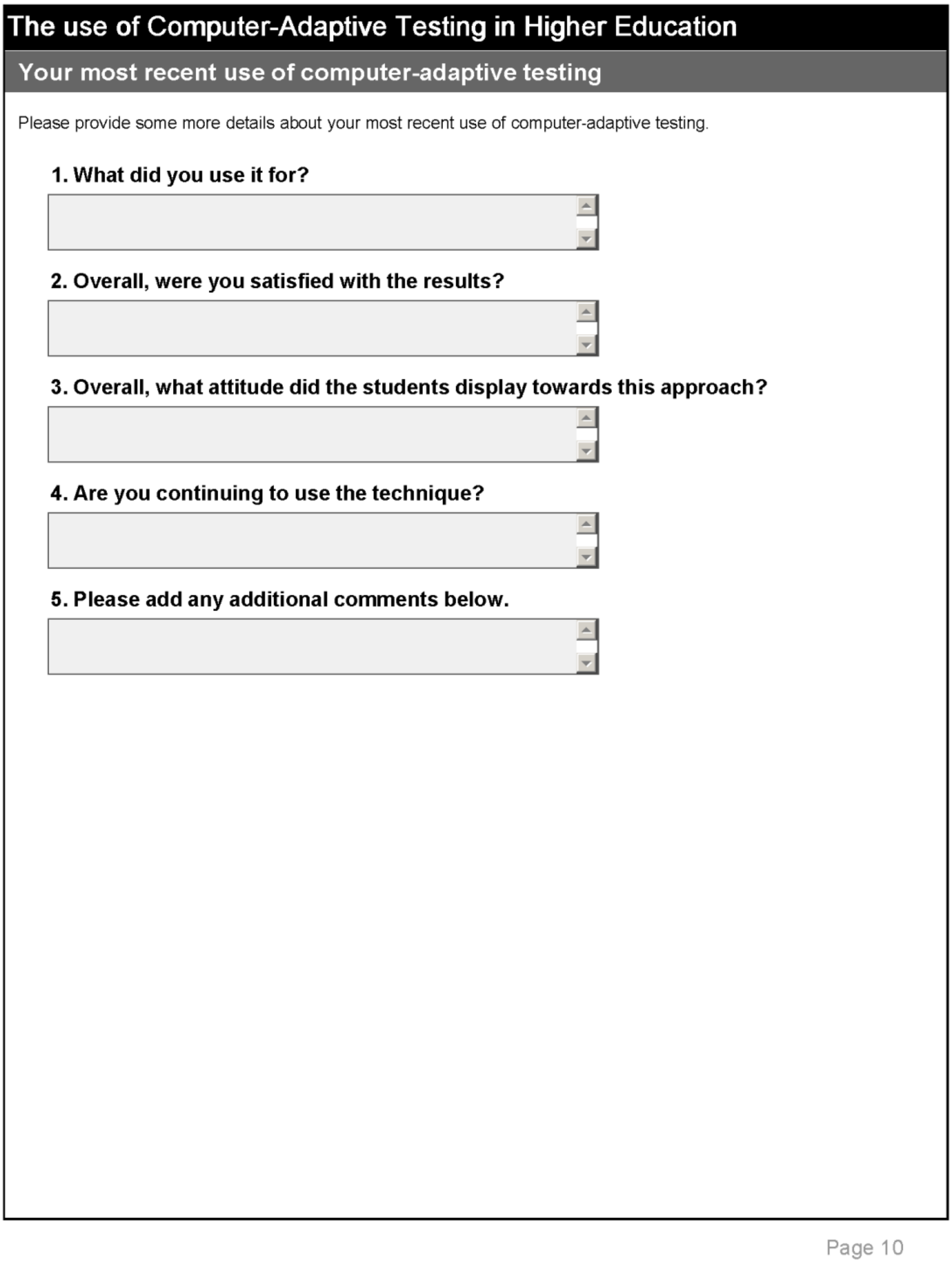
4. Please state any other potential benefit(s) not listed above.

*** 5. Please rate each of the statements below regarding the potential limitations of computer-adaptive testing.**

	Strongly disagree	Disagree	Neither agree, nor disagree	Agree	Strongly agree
Students may not fully understand how their final scores are calculated.	○	○	○	○	○
Students may perceive the test as being unfair, given that the set of items administered for each student is different.	○	○	○	○	○
The item pool is larger than that required by conventional testing.	○	○	○	○	○
The items in the pool must be calibrated.	○	○	○	○	○
Adaptive testing algorithms are more difficult to implement than conventional testing.	○	○	○	○	○

6. Please state any other potential limitation(s) not listed above.

Page 9



The use of Computer-Adaptive Testing in Higher Education

Your most recent use of computer-adaptive testing

Please provide some more details about your most recent use of computer-adaptive testing.

1. What did you use it for?

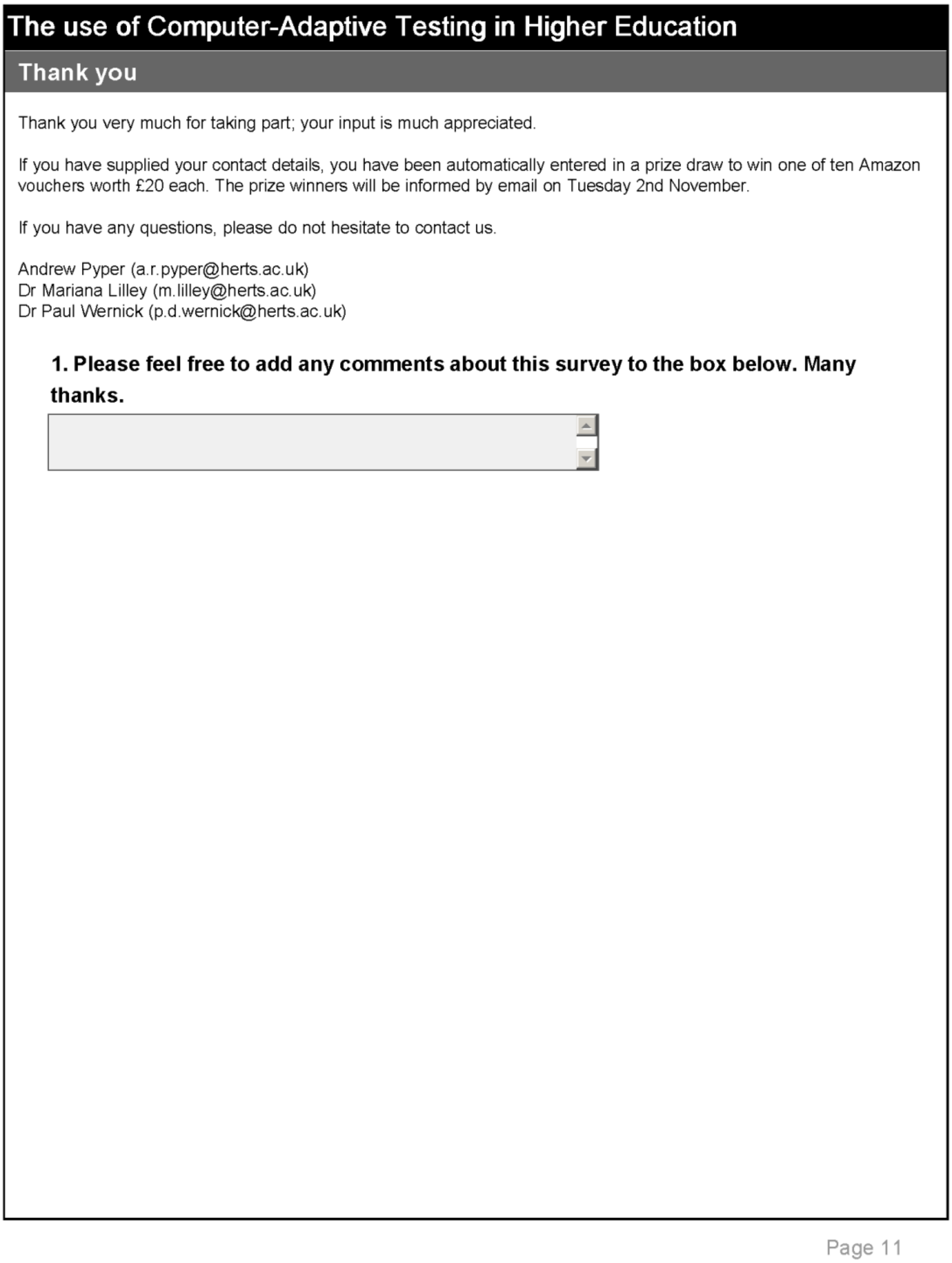
2. Overall, were you satisfied with the results?

3. Overall, what attitude did the students display towards this approach?

4. Are you continuing to use the technique?

5. Please add any additional comments below.

Page 10



The use of Computer-Adaptive Testing in Higher Education

Thank you

Thank you very much for taking part; your input is much appreciated.

If you have supplied your contact details, you have been automatically entered in a prize draw to win one of ten Amazon vouchers worth £20 each. The prize winners will be informed by email on Tuesday 2nd November.

If you have any questions, please do not hesitate to contact us.

Andrew Pyper (a.r.pyper@herts.ac.uk)
Dr Mariana Lilley (m.lilley@herts.ac.uk)
Dr Paul Wernick (p.d.wernick@herts.ac.uk)

1. Please feel free to add any comments about this survey to the box below. Many thanks.

APPENDIX H: SOFTWARE TESTING

Appendix H

Software testing was carried out for all new software development work using the tests below (the example below was used for Study F (please see section 4.4.4)

Section 1: Generic

Test case	Success Criteria	Pass / Fail	If fail, description of the error OR any other observations.
1. User provides valid username and password	User is forwarded to test instructions page. Timer is NOT active.		
2. User provides valid username, invalid password	User is re-directed to login page. User is provided with instructions on how to proceed Error message should indicate invalid combination of username and password but not give away information (e.g. “invalid username and/or password” is better).		

3. User provides invalid username, valid password	User is re-directed to login page. User is provided with instructions on how to proceed		
4. User provides invalid username, invalid password	User is re-directed to login page. User is provided with instructions on how to proceed		
5. User reads test instructions, and clicks proceed/next to start the test.	User is shown first question. Test countdown timer is activated at 40:00 (40 minutes).		
6. Start test.	The test consists of two consecutive modes, CBT and Flexilevel. First mode will be determined based on username.		
7. Number of questions per mode.	Twenty five (25) questions to be presented in CBT mode.		

	Fifteen (15) questions to be presented in CBT mode.		
8. Countdown timer is shown at all times, and it decrements at 1 second intervals.	Timer is shown at all times, and it decrements at 1 second intervals. When timer reaches 0:00 it no longer decrements. The timer updates periodically without user action. E.g. timer does not update only when user clicks a button or moves mouse. The timer applies to the overall test, rather than specific test sections.		
9. Questions are correctly displayed on the page.	Question stem and 4 options are presented on the page. There are no visual cues to indicate what the correct answer is.		
10. Question numbers are incremented correctly.	Question numbers are consecutive integers.		

11. User answers question and clicks proceed/next.	User answer is recorded in database. User is shown following question, question number is correct (e.g. incremented from ‘3’ to ‘4’) if the countdown timer is not yet 0:00. If the question answered was the final question in first mode, start second mode. If the question answered was the final question in second mode, terminate test.		
12. User must be prevented from resubmitting an answer for a previously answered question	If an answer already exists for the question then display error. Provide link to ‘current question’ to navigate the user to the question they should be answering.		
13. User clicks proceed/next without answering question.	Error message is provided; user is required to provide an answer before being presented next question.		

14. User accidentally closes browser window.	Test continues as before once re-logged in (timer still decrements)		
15. Internet connection is lost.	Test continues as before once network connection recovered		
16. Time limit has expired.	Do not record any answer to current question. Terminate test. Re-direct user to page with appropriate instructions		
17. Three users simultaneously take the test, using the following response patterns: User 1: all correct answers.	User answers are correctly stored on the database. Score is calculated correctly, i.e.: User 1: 40		

User 2: all incorrect answers. User 3: Right/Wrong/Right … alternate between right and wrong answers, starting with right answer.	User 2: 0 User 3: 15 + 5 (if response to last Flexilevel question is right) OR 15 + 15 + 5.5 (if response to last Flexilevel question is wrong)		
18. Terminate test	The test ends after 40 minutes, OR after 40 questions have been answered whichever happens first. Re-direct user to page with appropriate instructions		

Section 2: CBT

Test case	Success Criteria	Pass / Fail	If fail, description of the error OR any other observations.
19. User takes the test and provides right answers for all 25 CBT questions.	User answers are correctly stored on the database. Score is calculated correctly (i.e. 30).		
20. User takes the test and provides wrong answers for all 25 CBT questions.	User answers are correctly stored on the database. Score is calculated correctly (i.e. 0).		
21. User answers 25th CBT question and clicks proceed/next.	If CBT is first mode, then Flexilevel test starts. First Flexilevel question is shown. The question number is 26. If CBT is second mode, then terminate test.		

Section 3: Flexilevel

Test case	Success Criteria	Pass / Fail	If fail, description of the error OR any other observations.
23. User takes the test and provides right answers for all 15 questions.	User answers are correctly stored on the database. Score is calculated correctly (i.e. 15).		
24. User takes the test and provides wrong answers for all 15 questions.	User answers are correctly stored on the database. Score is calculated correctly (i.e. 0).		
25. User takes test and provides the response pattern below: Q1: W Q2: R	User answers are correctly stored on the database. Score is calculated correctly (i.e. 7.5).		

Q3: W Q4: R Q5: W Q6: R Q7: W Q8: R Q9: W Q10: R Q11: W Q12: R Q13: W Q14: R Q15: W			
26. User takes test and provides the response pattern below: Q1: W Q2: R	User answers are correctly stored on the database. Score is calculated correctly (i.e. 8.5).		

Q3: R Q4: R Q5: R Q6: R Q7: R Q8: R Q9: R Q10: R Q11: R Q12: R Q13: R Q14: R Q15: W			
27. User answers 15th Flexilevel question and clicks proceed/next.	If Flexilevel is first mode, then CBT test starts. First CBT question is shown. The question number is 16. If Flexilevel is second mode, then terminate test.		