The Application of the Flexilevel Approach for the Assessment of Computer Science Undergraduates

Mariana Lilley and Andrew Pyper

University of Hertfordshire, School of Computer Science, College Lane, Hertfordshire AL10 9AB, United Kingdom
{m.lilley,a.r.pyper}@herts.ac.uk

Abstract. This paper reports on the use of the flexilevel approach for the formative assessment of Computer Science undergraduates. A computerized version of the flexilevel was designed and developed, and its scores were compared with those of a traditional computer-based test. The results showed that the flexilevel and traditional scores were highly and significantly correlated ($p<0.01$). Findings from this study suggest that the flexilevel approach is a viable adaptive testing strategy, and may be a good candidate for smaller applications where IRT-based CATs may be too demanding in terms of resources.

Keywords: e-assessment, flexilevel, adaptive testing strategies.

1 Introduction

The past two decades have seen an increased use of e-assessment applications in higher education, to the extent that the use of computer technology in student assessment is rapidly becoming a common feature across the sector.

Most of these applications, however, are based on a "one-size-fits-all" approach of assessment, and tend to mimic traditional forms of static assessment (for example, paper-and-pencil objective tests). Examples of e-assessment applications that exploit the interactive nature of computers in order to adapt to the characteristics of individual students are few; indeed some argue that the full potential of the use of technology in higher education assessment has not yet materialized (see, for example, [2] and [6]).

An example of an e-assessment application that adapts to the characteristics of individual students is computerized adaptive testing. Adaptive testing differs from traditional testing in the way in which the questions to be administered during a given assessment session are selected. In traditional testing, all students are presented with the same set of questions, regardless of their proficiency levels within the subject domain. By contrast, in an adaptive test, the application's algorithm unobtrusively monitors the performance of students during the test, and then employs this information to dynamically adapt the sequence and level of difficulty of the questions (or tasks) to individual students.

Adaptive testing has been primarily associated with Item Response Theory (IRT) [8, 11]. Projects such as SIETTE [3, 4, 9, 10] have shown the efficacy of an IRT-based approach to adaptive testing in higher education.

J.A. Jacko (Ed.): Human-Computer Interaction, Part IV, HCII 2009, LNCS 5613, pp. 140–148, 2009.

Other studies on the use of adaptive testing in a higher education environment include research previously conducted by one of the authors [7]. In this work, it was shown that the application of an IRT-based approach to adaptive testing supported proficiency level estimates comparable to those obtained using traditional (i.e. static) tests. The IRT-based approach to adaptive testing was also shown to be effective at tailoring the level of difficulty of a test to the proficiency level of individual students. By tailoring the level of difficulty of the question or task to individual proficiency levels, students were challenged by test items at an appropriate level, rather than demotivated by items that were above or below their proficiency level.

Despite these positive findings, procedural factors impeded a greater uptake of the adaptive approach as proposed in the research [7]. The implementation of IRT-based adaptive testing applications is a complex task. Most importantly, the IRT model requires a large and calibrated database of questions. The calibration process usually requires a large item pool and specialist calibration programs, such as Xcalibre [1, 5].

This paper reports on a pilot study that investigated the implementation of computerized adaptive testing based on the flexilevel approach [8, 11] as an alternative to IRT-based algorithms. An overview of the flexilevel approach is presented in the next section of this paper.

2 The Flexilevel Approach

Like IRT, flexilevel algorithms attempt to match the level of difficulty of questions to the proficiency level of individual students. The Flexilevel approach, however, is based on a fixed branching strategy that is less complex to implement than IRT-based algorithms. Database calibration is also simpler; indeed Lord [8, p. 117] suggests that "any rough approximation" of the difficulty, of the questions will be adequate.

The difficulty of a question is determined by the following formula (adapted from Ward [12]):

$$D = 1 - \left(\frac{n_p \times 100}{n_r} \right) \tag{1}$$

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

In addition to simpler calibration, smaller question databases are required. The flexilevel approach requires a database of $2n$-1 question (where n is the number of questions to be administered during the test).

A flexilevel test typically starts with a question of medium difficulty. If the student answers the question correctly, a more difficult question is presented. Conversely, if the question is answered incorrectly, an easier question follows.

The flexilevel approach was initially devised as a self-scoring paper-and-pencil test [8]. As an example, consider a paper-based test with 19 questions; the difficulty of the questions ranging from 0.05 (easiest) to 0.95 (hardest).

[Q10, difficulty = 0.50]	
Red 1 [Q9, difficulty = 0.45]	Blue 1 [Q11, difficulty = 0.55]
Red 2 [Q8, difficulty = 0.40]	Blue 2 [Q12, difficulty = 0.60]
Red 3 [Q7, difficulty = 0.35]	Blue 3 [Q13, difficulty = 0.65]
Red 4 [Q6, difficulty = 0.30]	Blue 4 [Q14, difficulty = 0.70]
Red 5 [Q5, difficulty = 0.25]	Blue 5 [Q15, difficulty = 0.75]
Red 6 [Q4, difficulty = 0.20]	Blue 6 [Q16, difficulty = 0.80]
Red 7 [Q3, difficulty = 0.15]	Blue 7 [Q17, difficulty = 0.85]
Red 8 [Q2, difficulty = 0.10]	Blue 8 [Q18, difficulty = 0.90]
Red 9 [Q1, difficulty = 0.05]	Blue 9 [Q19, difficulty = 0.95]

Fig. 1. Adapted from Lord [8]. Layout of a paper-based flexilevel test, in which *Q1* is the easiest question and *Q19* is the hardest question

In the example shown in Fig. 1, the student will start the test by answering Q10 (i.e. question of medium difficulty). The answer sheet will inform the student whether each response was correct or incorrect; for example, a red spot appears where the student has selected an incorrect answer and a blue spot appears where the student's response was correct. Each time a correct answer is given, the question to be answered next is the lowest numbered "blue" question not previously answered. In the event of an incorrect response, the next question to be answered is the lowest numbered "red" question not previously answered. A student who answers all questions correctly will answer the following sequence of questions: Q10, Q11, Q12, Q13, Q14, Q15, Q16, Q17, Q18, and Q19. A student who answers all questions incorrectly will answer the following sequence of questions: Q10, Q9, Q8, Q7, Q6, Q5, Q4, Q3, Q2, and Q1. A student who answers the first question incorrectly, and all the following questions correctly will answer the following sequence of questions: Q10, Q9, Q11, Q12, Q13, Q14, Q15, Q16, Q17, and Q18.

Following the directions for a paper-based flexilevel test can be an onerous and disengaging task. This issue can be avoided by developing a software application to select, administer and score the questions, and this is the focus of the study reported here.

3 Methodology

The experiment was designed to model a real formative assessment situation as far as possible, given the constraints of the experimental design. So, whilst participants were free to leave without consequence and they were not the subject of the test, they were encouraged to engage with the test as if it were a real formative assessment.

3.1 Participants

Twenty-four final year Computer Science undergraduates participated in the experiment. 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.

3.2 Software

The software prototype was developed using VB.NET with an Access database to store the questions and responses. There were 29 questions; these were separated into two pools, one to populate the flexilevel test (19 questions), and the other to populate the statically selected test (10 questions). The questions had previously been used in standard testing and calibrated based on the calibration formula described previously. The order of presentation of the two tests was randomised.

The two different sets of questions had different light background colours on the question screens. This was done in order to enable the experimenters and participants to refer to the different tests without naming them. The colours used were selected because they maintained a clear contrast between the text of the questions and the background of the screen. Fig. 1 shows the layout of the question screen.

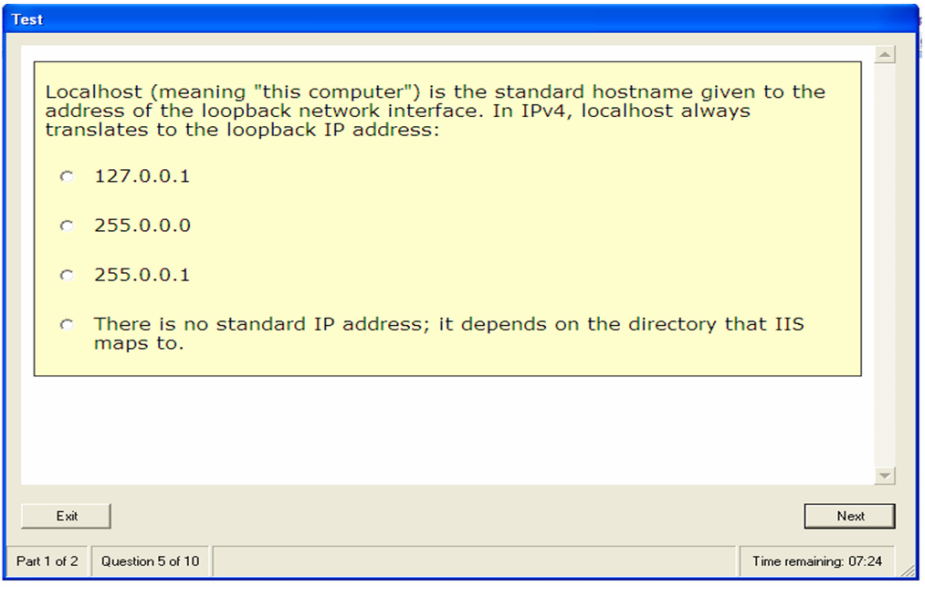

Fig. 2. Screenshot of the flexilevel software application

Finally, the software allows for the timing of a test, and displays time elapsed and number of questions answered throughout the test.

3.3 Questionnaire

The questionnaire document consisted of a briefing section setting out information about the flexilevel approach, test and questionnaire and 15 statements that participants could rate using a 5 point Likert scale from 1: Strongly Disagree to 5: Strongly Agree. The statements are shown in Tables 3 (ease of use) and 4 (perceived usefulness). The statements were designed to elicit participants' views on the ease of use of the software application and the perceived usefulness of the software application.

A further 3 questions that asked for free text responses were included, and are shown in Table 1 below.

Table 1. Open questions included in the questionnaire

	Statement
16	What problems, if any, can you see to the uptake of this adaptive testing approach?
17	Is there a question that you would like to have been asked? If so, what is it and how you would answer it?
18	Can you see any benefits of this adaptive testing approach?

3.4 Experimental Procedure

The participants were briefed that they were about to engage in an assessment using a new form of software assessment. They were told that the test was of the software and not of them. However, in order to encourage them to answer the questions to the best of their ability, participants were informed that the participant with the highest overall score would be given £20.00.

They were also told that they would be asked to fill in a questionnaire. All participants were offered £10.00 as a thank you for participating, but it was made clear that they could leave at any time and did not have to complete the test or the questionnaire in order to receive the money.

The participants took the test in a computer lab in test conditions. They started the test in their own time and had 25 minutes to complete the whole test; this was monitored by the software application. This enabled the collection of the performance of participants in the two tests and the reporting of their results at the end of the test.

Once the participants had completed the tests they were asked to fill in the questionnaires that had been placed by their computers. This enabled the collection of data associated with the views of participants.

3.5 Hypothesis

The hypothesis of the experiment was that there would be a significant correlation between students' results on the statically generated test and the flexilevel generated test.

4 Results

A summary of students' performance is presented in Table 2. Table 2 shows that the range of scores achieved by participants was relatively large for both test conditions and that there is little apparent difference between the conditions.

Table 2. Summary of student scores for each section of the test (N=24)

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

A Pearson product-moment correlation coefficient was computed to assess the relationship between the flexilevel and static scores shown in Table 2, r=0.764, n=24, p<0.01, Sig. (2-tailed)=.000. A paired-samples t-test showed no significant difference between the flexilevel and the static scores, t(23) = -.954, p=.350. The scatterplot shown in Fig. 2 illustrates the results.

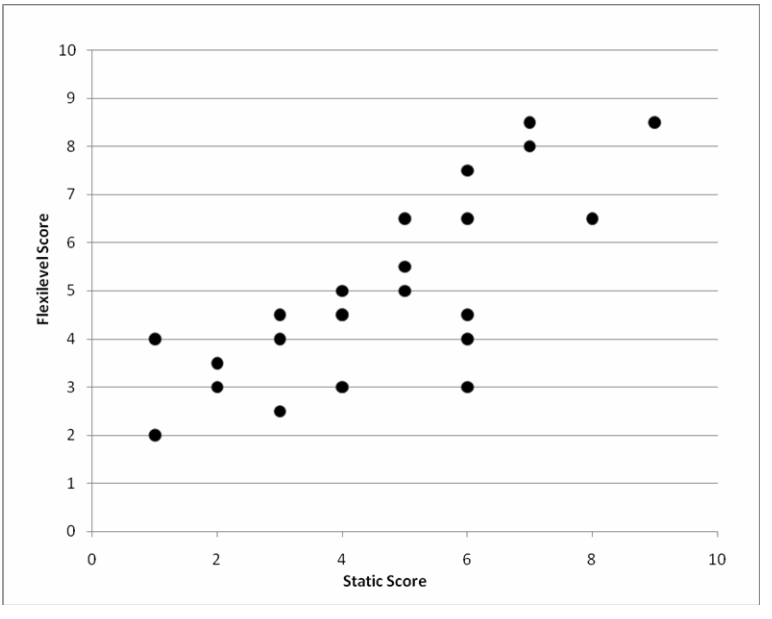

Fig. 2. Scatterplot diagram showing the scores achieved by participants in the two test conditions

Student responses to the questionnaires were also analysed. Table 3 below shows that the majority of participants agreed with statements concerning the ease of use of the software application.

As can be seen from Table 4, participants were also positive when responding to statements about the perceived usefulness of the software application. This was particularly the case in terms of using the flexilevel for formative assessment (statement 4), areas that the participants might need to work on (statement 6), or not (statement 11) and that if flexilevel tests were made widely available that they would use them (statement 13).

Table 3. Ease of use. Mode and median for the responses.

Statement	Median	Mode
1. Learning to use the Flexilevel software application would be easy for me.	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.	4	4
3. I would find the Flexilevel software application easy to use.	4	4

Table 4. Perceived usefulness. Mode and median for the responses.

Statement	Median	Mode
4. I would find the Flexilevel approach useful for practice tests.	4	5
5. I would find the Flexilevel approach useful in summative tests (i.e. the test score counts towards my final grade).	3.5	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.	4	5
7 The adaptivity supported by the Flexilevel approach would enhance my overall assessment experience.	4	4
8 For practice tests, I would prefer using the Flexilevel approach to other forms of objective testing.	4	4
9. For summative tests, I would prefer using the Flexilevel approach to other forms of objective testing.	3	3
10. The system used to score a Flexilevel test makes sense to me.	4	4
11. I would find the score provided by the Flexilevel approach useful at identifying how much I have learned.	4	4
12. I would find it useful if the level of difficulty of a test is tailored to my level of understanding.	4	4
13. Assuming the Flexilevel software was available to me for practice tests, I predict that I would use it on a regular basis.	5	4
14. In practice tests, test questions that are too easy are less engaging than those questions that are tailored to my level of understanding.	4	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.	3	3

Participants were less positive in their responses to the use of the flexilevel approach in summative assessments (statement 5) and tended not to agree that more difficult questions were less engaging (statement 15).

The open questions were included to enhance the richness of the data collected, but were too few to be subjected to content analysis. As such, they will be used to inform the discussion of the results that follows, rather than being reported in this section.

5 Discussion

The results of this pilot study are very encouraging, there is a high and significant correlation between participants' performance on the two different types of test and participants were positive about their use of the software. Additionally, participants were positive about the use of the flexilevel approach in formative assessment.

Moreover, there was a reasonably wide distribution of scores. So, it seems that performance of participants in the two tests is consistent across a range of attainment.

Nonetheless, this is a pilot study, and as such there are aspects of this study that future experiments will need to replicate. The participants are all final year undergraduate students. It may be expected that such a sample would have no problems with using the software application. The measures of ease of use are important in controlling for potential extraneous variables, and also to identify any usability issues in these early stages of development.

In this sense a sample of Computer Science undergraduates was a good choice for the pilot study. Expert analysis of the interface indicated no serious usability issues that might be expected to trip up even users with a high degree of technical literacy. The results of the study bear this out. Clearly the downside to using Computer Science students as participants is that this sample of participants may not be representative of a wider student population. Whilst the participants in this study reported no problems in using the software application, it seems likely that extrapolating this to a wider population would be unwise. Clearly an important part of future experiments would be to involve participants from a wider population of students.

One reservation raised about the software application was that "You can't go back and correct a previous answer". This is a necessary feature of the system, however, because if participants were able to go back and review the questions, they would soon be able to enhance their marks simply by going back and attempting the questions until they found the correct response.

In terms of the perceived usefulness of the application, there seems to be a difference in the way participants perceived the flexilevel approach as a tool for formative and summative assessment.

Participants were positive about the formative use of the flexilevel 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."

This attitude did not extend to the use of the flexilevel approach in summative contexts: "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 which would not be good for assessment".

Participants did suggest that the flexilevel approach would enhance their overall assessment experience (statements 7 and 12), but it seems that this enhancement would mostly be realised in the formative assessment they engage in.

An informal observation has been that students were uncomfortable using a system in which the scoring system was not clear to them; for example, maximum likelihood estimates, used in IRT-based adaptive testing may not be obvious to all. In this study, participants indicated that they did indeed understand how the scoring system worked

(statement 10). This could be taken to be support for the use of a simpler adaptive system than those that had been previously employed. However, this bears further scrutiny in future work given that participants felt that high-performing students may be disadvantaged by this scoring in summative assessments.

6 Conclusion

This pilot study was concerned with the efficacy of the flexilevel approach to adaptive testing when compared to standard computer based testing. The results show a highly significant correlation between scores achieved by participants on the two tests. This provides a basis upon which to replicate and extend this research to include larger groups of students and those in studying in different academic disciplines. This will be the subject of future work.

References

1. Assessment Systems Corporation, XCALIBRE Marginal Maximum-Likelihood Estimation, http://assess.com/xcart/product.php?productid=270&cat=0&page=1
2. Challis, D.: Committing to quality learning through adaptive online assessment. Assessment & Evaluation in Higher Education 30(5), 519–527 (2005)
3. Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-de-la-Cruz, J.L., Ríos, A.: SIETTE: A Web–Based Tool for Adaptive Testing. International Journal of Artificial Intelligence in Education 14, 1–33 (2004)
4. Conejo, R., Millán, E., Pérez-de-la-Cruz, J.-L., Trella, M.: An empirical approach to online learning in SIETTE. In: Gauthier, G., VanLehn, K., Frasson, C. (eds.) ITS 2000. LNCS, vol. 1839, pp. 604–614. Springer, Heidelberg (2000)
5. Gierl, M.J., Ackerman, T.: Software Review: XCALIBRE — Marginal Maximum-Likelihood Estimation Program, Windows Version 1.10. Applied Psychological Measurement 20(3), 303–307 (1996)
6. Joy, M., Muzykantskii, B., Evans, M.: An Infrastructure for Web-Based Computer-Assisted Learning. ACM Journal of Educational Resources 2(4), 1–19 (2002)
7. Lilley, M.: The Development and Application of Computer-Adaptive Testing in a Higher Education Environment, Unpublished PhD thesis, School of Computer Science, University of Hertfordshire, Hertfordshire (2007)
8. Lord, F.M.: Applications of Item Response Theory to Practical Testing. Lawrence Erlbaum Associates Inc., Mahwah (1980)
9. Ríos, A., Conejo, R., Trella, M., Millán, E., Pérez-de-la-Cruz, J.L.: Aprendizaje automático de las curvas características de las preguntas en un sistema de generación automática de tests. In: Actas de la Conferencia Española para la Inteligencia Artificial-CAEPIA 1999 (1999)
10. Ríos, A., Millán, E., Trella, M., Pérez-de-la-Cruz, J.L., Conejo, R.: Internet Based Evaluation System. In Open Learning Environments: New Computational Technologies to Support Learning, Exploration and Collaboration. In: Proceedings of the 9th World Conference of Artificial Intelligence and Education AIED 1999, pp. 387–394. IOS Press, Amsterdam (1999)
11. Wainer, H.: Computerized Adaptive Testing: A Primer. Lawrence Erlbaum Associates Inc., Mahwah (2000)
12. Ward, C.: Preparing and Using Objective Questions. Nelson Thornes Ltd., Cheltenham (1980)