The Potential Use of the Flexilevel Test in Providing Personalised Mobile E-Assessments

Andrew Pyper(✉), Mariana Lilley, Paul Wernick, and Amanda Jefferies

School of Computer Science, University of Hertfordshire, Hatfield, UK
a.r.pyper@herts.ac.uk

Abstract. Sixteen students took a test that included a Flexilevel stage and a standard Computer Based Test (CBT) stage. The results were analysed using a Spearman's Rank Order correlation and showed a significant positive correlation (rs = 0.58, p <=0.05). This was taken to provide support for the notion that it is possible to provide shorter Flexilevel objective tests that are as efficacious as CBTs. Implications that this finding may have for the use of the Flexilevel Test in mobile learning contexts is discussed.

Keywords: Flexilevel · E-assessment · Mobile assessment · Computerised adaptive testing · Mobile and/or ubiquitous learning · Personalization · Technology enhanced learning

1 Introduction

There has been increased interest in the use of the Flexilevel Test [6, 7] in Higher Education contexts [2, 3, 5] due to its potential to personalize educational experiences.

This study is part of a programme of research aimed at understanding how the Flexilevel Test may be applied in genuine educational contexts and an increasingly important part of our educational context is mobile learning and assessment. This 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 [8].

However, the case for supporting mobile learning and assessment is compelling, both from pedagogical and practical perspective [2, 17], for example learners are often under significant time pressure and, given appropriate opportunities, can make use of short periods of time to engage in their studies [17].

In principle, one of the possible benefits of the Flexilevel test is the ability to present fewer items in a test than a standard Computer Based Test (CBT) approach yet still obtain a comparably accurate measurement of a test-taker's proficiency [20]. This is something that the authors have also found in a simulation study [14]. 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 [2, 17]. Further, it may provide these opportunities whilst mitigating some of the challenges to deploying formative assessments in a mobile context. As such, it is of interest to

P. Zaphiris and A. Ioannou (Eds.): LCT 2015, LNCS 9192, pp. 271–278, 2015.
DOI: 10.1007/978-3-319-20609-7_26

investigate whether or not this effect can be reproduced in a genuine educational context. This study is intended to investigate the extent to which shorter Flexilevel tests may provide comparably accurate measures of a test-taker's proficiency as a CBT in a summative assessment.

2 The Flexilevel Test

The Flexilevel Test was first proposed by Lord [6] as a paper-based test. One of its aims was to tailor the test difficulty level to the proficiency level of individual test-takers. This is an important difference between the Flexilevel Test and traditional CBTs as in the case of the latter, all test-takers are presented with the same set of test items.

The Flexilevel Test is a fixed branched test that supports the personalization of objective tests by presenting items to test-takers depending on their performance. It represents an approach to adapting assessment that is less resource intensive than other forms of adaptive testing, for example approaches based on other pyramidal approaches [20] or Item Response Theory (IRT) [6, 7].

A Flexilevel Test requires a set of $2n$-1 items where n is the number of items to be presented in the test. Test items are ranked in order of difficulty. There are different approaches to ranking the items; for example the use of expert calibration or calibration from an existing set of test-taker responses from preceding tests. In the work reported here, an existing set of responses from a previous test and the formula shown below, adapted from Ward [19], were used to calculate the difficulty of individual items in which n_p is the number of test-takers who answered the item correctly and n_r is the total number of test-takers answering the item.

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

After the difficulty of each item was calculated, items were ranked from the easiest to the most difficult.

A Flexilevel Test usually begins with the presentation of the item of median difficulty, typically an item with a difficulty of 0.5. If test-takers answer the item incorrectly, they are presented with the next available easier item. If they answer the item correctly, they are presented with the next available more difficult item (as shown in Fig. 1).

Once the true ability of a test-taker is reached, subsequent patterns will show an increasingly large range of difficulty between the items selected, as illustrated in Fig. 2.

Typical stopping conditions for the test are when the number of items to be administered have all been presented to the test-taker or the duration of the test has been reached, whichever happens first.

3 The Study

In this study, a test of 40 items was presented to 18 online distance learning students on an undergraduate Computer Science programme of study as a formative assessment.

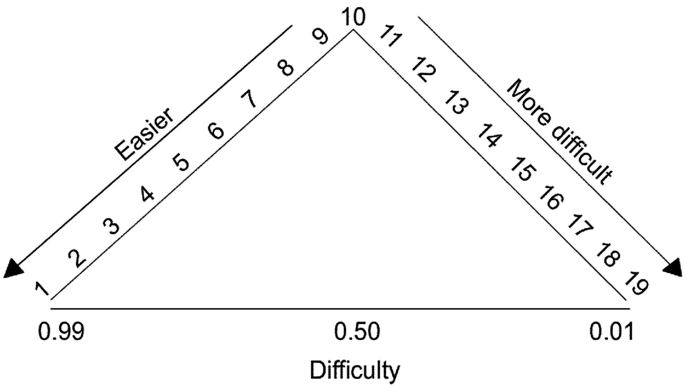

Fig. 1 The Flexilevel Test Structure (adapted from Betz and Weiss 1975)

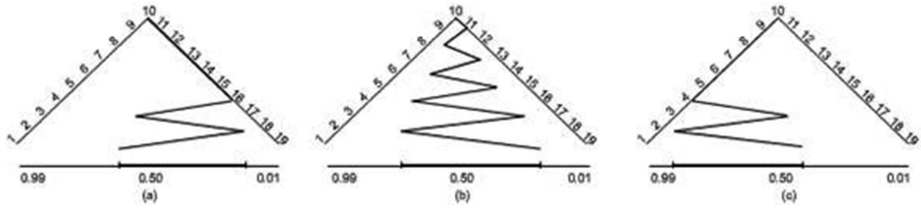

Fig. 2 Example patterns for different proficiencies (from Betz and Weiss, 1975)

3.1 The E-Assessment Application

It was necessary that the Flexilevel Test would be made available online, as the participants were geographically dispersed. A web application using HTML5, JavaScript and CSS on the client side and PHP and MySQL on the server side was purpose built for this reason.

The e-assessment application supported two different item selection algorithms: a traditional CBT and a Flexilevel Test. 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 test-takers. Figure 3 shows how the user interface contained minimal information and gave no cues to the approach, CBT or Flexilevel, being used.

3.2 Methodology

The test presented in this study was part of the formative assessment of a group of 18 first year online distance learning Computer Science students. The test related to their knowledge and understanding of internet technologies and, in particular, ASP.NET.

The test was presented to students online via the web application introduced earlier. It was invigilated by a remote live invigilation service to ensure students were not

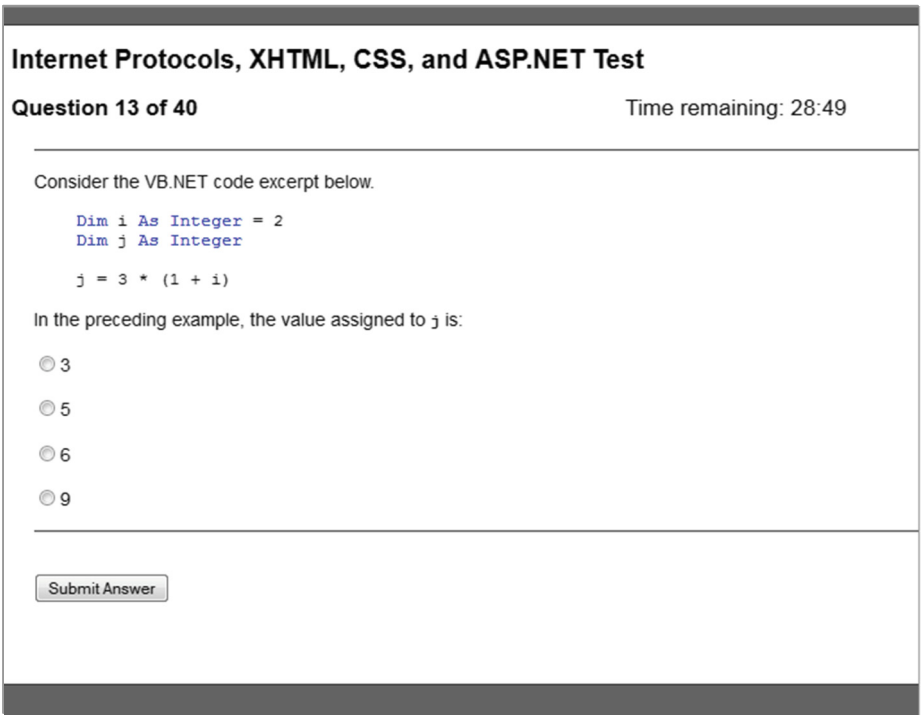

Fig. 3 Screenshot of the user interface of the application, showing the presentation of an item

accessing materials that were not permitted during their test. It was limited to 40 min and contained 40 items overall. The test items had gone through 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, the examinees were randomly assigned to one of two different groups. Half the participants were assigned to Group 1, and the second half to Group 2. Group 1 was presented with Flexilevel followed by CBT, and Group 2 was presented with CBT followed by Flexilevel. Both the Flexilevel and CBT stages covered ASP.NET and for the students there was no distinction between these two test stages. The Flexilevel and CBT stages consisted of 10 and 20 items respectively.

3.3 Results

Table 1 shows the range of the scores obtained and the mean score for each stage of the test. It can be seen that the scores for both of the stages ranged relatively widely. The responses from two participants were removed from the analysis as they did not complete one or more test stages.

Table 1 Summary of test performance (N = 16)

Test Stage	Minimum Score	Maximum Score	Mean Score
Flexilevel (out of 10)	5	9	7.625
CBT (out of 20)	5	19	12

A Spearman's Rank Order correlation was run to determine the relationship between the CBT and Flexilevel scores. It showed a significant positive correlation (rs = 0.58, p <=0.05) between the scores achieved by students in the Flexilevel and CBT test stages.

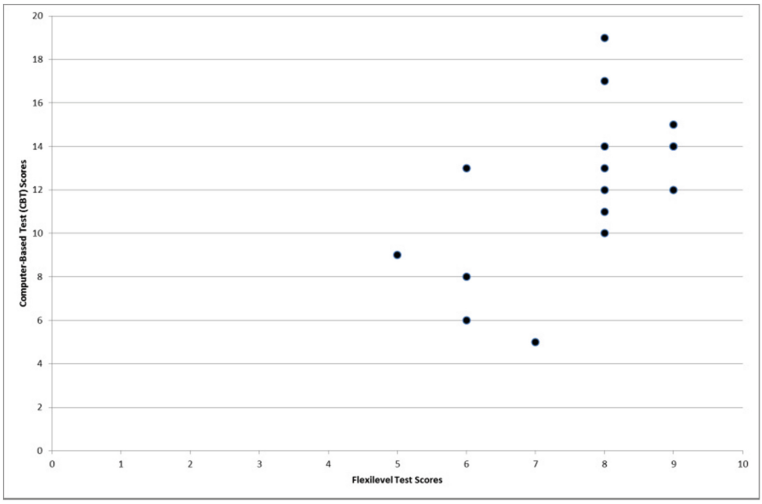

Fig. 4 Scatterplot showing relationship between scores

These results provide support for the notion that the Flexilevel approach to e-assessment can provide comparable results to a standard CBT test whilst only presenting half the items.

4 Discussion

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 by the authors so far (for example [5, 13]). 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 [11] and also during a given interaction. In short, the contexts for mobile use tend to be more diverse and potentially distracting than those using desktop computers.

Oulasvirta et al. [11] used a range of tasks on a mobile device to elucidate how different contexts may 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 to the task was 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 [18]. 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 [10]. Furthermore, it seems that interruptions may have an additional cognitive load in terms of switching between tasks [15].

Framing the analysis in terms of cognitive load provides a good basis to understand how mobile contexts may impact upon users' usage of a mobile application. Cognitive load theory [16] has also been influential in learning, particularly multimedia learning [8] and it seems pertinent also to mobile learning and e-assessment when considering the impact of extraneous loads on learning [12].

Clearly context of use is an important contributor to extraneous load, and in providing silent, invigilated exam conditions, extraneous load is minimized freeing cognitive resources for the intrinsic load that tests impose. This is the case for previous studies conducted by the authors. 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 [5, 13], or with students taking tests remotely at their own computers as in the study reported here; all were invigilated and were the 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 much.

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 learners to attend to a given task [10, 12].

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 [1]. 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 effect interactions that require greater engagement with the mobile task [11] something that seems pertinent to even relatively short mobile e-assessments. A possible implication of this is that learners may choose where and when they take formative assessments such that interruptions are less likely.

5 Conclusion and Future Work

The study reported here provides further support for the idea that the Flexilevel approach can provide shorter tests in genuine educational contexts, although this must be treated with caution given the small scale of the study. It is intended that further studies will be conducted in order to establish if this effect can be demonstrated reliably.

It has also been noted that the approach of learners to their formative assessment activities in mobile contexts is of fundamental importance. As such, future work will focus on gaining an understanding of the attitude and approach of test-takers to their mobile formative assessment. How are test-takers using the application in mobile contexts, specifically in what kinds of contexts are they taking the tests? Also, what kind of learning activities are they willing to carry out on a mobile device, and does this include formative assessment? It has been suggested that a formative assessment may motivate learners to minimize interruptions themselves, but anecdotally learners have indicated that they learn opportunistically in mobile contexts, for example the short period of time between tutorials or lectures.

It seems that future studies will need to adopt two main strands of work; establishing the reliability of the correlation between the shorter Flexilevel test and full length CBT reported in this study and investigating how this may impact upon learner attitudes to formative assessment in mobile contexts.

References

1. Baddeley, A.: Is Working Memory Still Working? Am. Psychol. **56**, 849–864 (2001)
2. Betz, N.E., Weiss, D.J.; Empirical and simulation studies of flexilevel ability testing (Research Report 75-3) University of Minnesota, Department of Psychology, Psychometric Methods Program, Minneapolis (1975)
3. Gordon, N. (2014). "Flexible Pedagogies: technology-enhanced learning." Higher Education Academy, NIACE. https://www.heacademy.ac.uk/sites/default/files/resources/TEL_report_0.pdf [last accessed 06/03/2015]
4. 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. Strawberry Hills, NSW: Australian Learning and Teaching Council
5. Lilley, M., & Pyper, A. (2009). "The application of the flexilevel approach for the assessment of computer science undergraduates." In Human-Computer Interaction. Interacting in Various Application Domains (pp. 140-148). Springer Berlin Heidelberg
6. Lord, F.M.: The self-scoring flexilevel test. J. Educ. Meas. **8**, 147–151 (1971)
7. Lord, F.M.: Applications of item response theory to practical testing problems. Erlbaum, Hillsdale, NJ (1980)
8. Mayer, R.E.: Applying the science of learning: evidence-based principles for the design of multimedia instruction. Am. Psychol. **63**(8), 760 (2008)
9. Mendoza, A. (2013). "Mobile user experience: patterns to make sense of it all." Morgan Kaufmann Publishers Inc
10. Monsell, S.: Task switching. Trends in cognitive sciences **7**(3), 134–140 (2003)

11. Oulasvirta, A., Tamminen, S., Roto, V. & Kuorelahti, J. (2005) “Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI.” In Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2005
12. Oviatt, S. (2006). “Human-centered design meets cognitive load theory: designing interfaces that help people think.” In Proceedings of the 14th annual ACM international conference on Multimedia (pp. 871-880). ACM, 2006
13. Pyper, A., & Lilley, M. (2010). “A comparison between the flexilevel and conventional approaches to objective testing.” In Proceedings of CAA 2010 International Conference, Southampton
14. Pyper, A., Lilley, M., Wernick, P., Jefferies. A (2014) “A simulation of a Flexilevel test.” Paper presented at HEA STEM Annual Conference 2014, Edinburgh
15. Sandy J. J. Gould, Anna L. Cox, Duncan 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
16. Sweller, J. (2010). “Element interactivity and intrinsic, extraneous, and germane cognitive load.” Educational psychology review, 22(2), 123-138. Multimedia learning
17. Traxler, J.: 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) 2007
18. Dawood, Mohammad, Fieseler, Michael, Büther, Florian, Jiang, X., Schäfers, Klaus P.: A Multi-resolution Optical Flow Based Approach to Respiratory Motion Correction in 3D PET/CT Images. In: Zhang, David (ed.) ICMB 2008. LNCS, vol. 4901, pp. 314–322. Springer, Heidelberg (2007)
19. Ward, C.: Preparing and Using Objective Questions. Nelson Thornes Ltd, Cheltenham (1980)
20. Weiss, D.J., Betz, N.E.: Ability Measurement: Conventional or Adaptive? (Research Report 75-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program (1973)