Journal of Computer Assisted Learning (2000) **16**, 193-200

The efficacy of a World-Wide Web mediated formative assessment

T. Buchanan
Division of Psychology, University of Sunderland

Abstract Two studies evaluated the effectiveness of a WWW-based formative assessment package used in undergraduate psychology courses. Students taking on-line multiple-choice tests received instant feedback on areas of weakness and how to address them. In Study 1, students used the package as an integral part of their course syllabus. Level of use correlated with performance in the end-of-course summative assessment. In Study 2, the package was used as an 'optional extra'. Both studies found that students who used the package performed better than those who did not. Such systems may be useful learning tools which students may use to enhance performance.

Keywords: Assessment; Formative; Internet; Multiple-choice; Psychology; Summative; Undergraduate

Introduction

The Internet is increasingly being employed in education. In recent years, a number of articles have been published outlining the potential role of the Internet in teaching at all levels and reporting various ways in which it has been employed (e.g. Krantz, 1995; Duchastel, 1997). Quite apart from the much-hyped emergence of 'Virtual Universities' offering portfolios of on-line courses in which teaching and assessment are conducted remotely via the WWW and associated technologies, Internet-mediated teaching and assessment is increasingly being used to supplement traditional courses (e.g. Stockburger, 1998; Gibbs *et al*., 1998). Given such developments, there is a need to assess the extent to which the methods used are educationally sound. This paper describes the evaluation of a World-Wide Web based formative assessment package developed for use in undergraduate psychology courses at the University of Sunderland and seeks to answer the question: did students using it actually benefit from the experience?

Formative assessment

The value of formative assessment — that is, assessment which is used to

Accepted: 1 November 1999

Correspondence: Tom Buchanan, Department of Psychology, University of Westminster, 309 Regent Street, London W1R 8AL, Email: tb99@iname.com

provide feedback to students rather than to evaluate them for course grades — is well known (e.g. Brown & Knight, 1994; Wiliam & Black, 1996). Students, having received feedback on their performance, may then take steps indicated by that feedback to remedy whatever weaknesses the assessment has exposed. The function of formative assessment is essentially to assist learners in '*closing the gap between actual and desired levels of performance*' (Wiliam & Black, 1996; p. 543).

For a formative assessment to be useful, however, there are certain criteria it must meet. For example, there is little point in giving a student feedback on knowledge or skills when there is no time left for them to act upon it: feedback should be provided at an appropriate point in the learning process (Brown & Knight, 1994). It also should '*have embedded within it some degree of prescription about what must be done*' (Wiliam & Black, 1996; p. 543) in order for students to improve their performance. When large numbers of students or lengthy pieces of work are involved, practical constraints (such as time or workload pressure) may make it difficult for tutors to provide feedback which is both timely and useful to the student. It was this desire — to provide swift and useful feedback — which led to the development of the system described in this paper.

Computerisation of formative assessment

One solution to this problem is to conduct formative assessment by means of multiple-choice tests, which can be an efficient way of providing feedback (Ramsden, 1992), at least in knowledge-based disciplines (although their use in assessing skills or competences such as report writing or laboratory techniques is limited). Such tests are widely used as both formative and summative assessment instruments. Of all assessments, they are probably the easiest to automate: they can easily be administered and scored by computer programs which can automatically provide instant feedback on performance (e.g. Callear & King, 1997).

As well as ease of administration, computerised tests may have other advantages over their pencil and paper antecedents (Bostow *et al.*, 1995). For example, students may repeatedly take the same test in order to assess extent of improvement in performance after study. If the tests can be used unsupervised, students may work at their own pace (Clariana (1997) has shown that some individuals learn faster than others from computer-based materials, implying that people will differ in the amount of computer time they require) or at times and places of their own choosing. Individuals may thus tailor their use of the assessments to their own learning styles.

Such assessments are increasingly being implemented via the World-Wide Web[*] (e.g. Burgess *et al.*, 1998; Charman & Elmes, 1998; Stockburger, 1998). Students may access course and assessment materials — perhaps within constraints laid down by course tutors — using any Internet-connected computer with widely available browser software. This confers

[*] Of course, they are only one of many tools available for on-line education. On-line tests may vary in format, goal, nature of feedback given, and the type of knowledge or skill addressed: for example, the exercises used in Study 1 focused on factual knowledge, while those described in Study 2 targeted a variety of learning outcomes.

extra freedom on the amount of access students may have to learning materials (and offers the possibility of linking formative assessments with other course related materials).

PsyCAL

Buchanan (1998) has described the development of a WWW-mediated formative assessment package (PsyCAL — the acronym stands for Psychology Computer Assisted Learning) and its implementation in an introductory psychology course. Students were able to access PsyCAL exercises on-line through the Psychology Division's web pages. In each exercise, a series of multiple choice questions was presented as a form on the computer screen. Students completed the test and submitted the answers for processing by a program on the WWW server which instantly generated a feedback page.

An important feature of the system is that the correct answers are not given (this was deemed likely to encourage rote memorisation — a characteristic of CAL packages criticised by Ramsden, 1992). Instead, for each question answered incorrectly, a reference (to the appropriate section of one or more text books) was given which students could consult to find out the correct answer. This required students to engage more fully with the course materials. They are advised to repeat the test after doing the suggested reading in a test-learn-retest cycle which continues until the subject matter is mastered. Students are required to do exercises at set times during the semester and general revision exercises are also available for them to use.

Buchanan (1998) found that the package was widely used and popular with students. Feedback forms returned by students indicated that PsyCAL was perceived as equal in usefulness to other teaching and learning methods (lectures, seminars and textbooks) and that students would appreciate access to other such materials.

Purpose of current research

The fact that the assessment was popular, however, does not mean that it was effective. The purpose of the research described in this paper was to evaluate the effectiveness of the PsyCAL package: does it facilitate learning? Is usage associated with enhanced performance?

Study 1

Participants

The participants in this study were drawn from a cohort of 232 undergraduate students of mixed ability and background studying an introductory psychology module (PSY111) in 1997–98. All students were required to complete three set exercises, each comprising 11–15 multiple-choice questions mainly assessing factual knowledge of topics covered in the module. They also had access to two additional revision exercises, in each of which 10 questions randomly selected from a larger pool were presented. The exercises could be accessed from any networked computer, with a forms-capable browser, on or off campus. All students had access to extensive computing facilities in timetabled campus classrooms and open-access areas, and it was from these areas that most uses occurred. Access to each question set was only permitted once the relevant material had been covered in teaching sessions. For more details of the exercises, see Buchanan (1998).

Procedure

Students were asked to enter identification numbers when using the package. However, it was possible to complete the exercises without entering an identification number. Each time the package was used, a record was made of the identification number (if any) entered by the user. The variables measured in the study were the number of times each student used the package, and the score achieved by each student in the end-of-course summative assessment (an examination comprising multiple choice and essay elements, assessing performance on most of the topics covered by the PsyCAL exercises). Also noted was the number of timetabled seminar classes each student attended during the semester.

Results and discussion

During the period of the study, the package was used 1056 times by 161 identifiable students. There were also 1243 anonymous hits. Among the identifiable students, there were some with a very high number of uses (up to 32). As log files suggested that these may have been due to repeated clicking on the 'Submit' button rather than genuinely different attempts at the exercises, any student whose number of uses was more than three standard deviations above the mean was excluded from analysis (six in total). For the 155 students remaining in the sample, the average number of uses was 5.92 ($sd = 5.07$). Some students attended the course but did not sit the final exam. Due to this, exam marks were only available for 148 of the 155 students known to have used PsyCAL. For these 148, the number of uses was found to correlate positively with exam performance, $r_{(148)} = 0.24$, $p < 0.003$, although the proportion of variance accounted for is small (5.76%).

This does not necessarily mean that PsyCAL use leads to higher marks — a third factor might influence both variables. Students who are dedicated, conscientious or highly motivated to succeed might be expected to use the exercises extensively. They might also be expected to work hard in other ways — and it could be this hard work in other areas which influences their exam performance, not their PsyCAL use.

This may be unpacked by looking at the role of a third variable which should be strongly influenced by the same factors: attendance at classes. Should the findings outlined above simply reflect hard work in other areas, then if the effect of class attendance is controlled for in the analysis, PsyCAL use should not emerge as a significant predictor of performance.

The effects of PsyCAL use and class attendance upon exam performance were simultaneously assessed using multiple regression. Results indicated that performance was statistically significantly associated with both class attendance, $t_{(145)} = 4.33$, $p < 0.00005$, $\beta = 0.34$, and PsyCAL use, $t_{(145)} = 2.18$,

$p < 0.03$, $\beta = 0.17$: both variables had a unique and positive effect on exam mark, although the effect size was larger for class attendance.

Further information on the efficacy of the package comes from comparing the performance of those students who used it with those who did not. There were 148 students for whom a mark was recorded who were known to have definitely used PsyCAL. There were another 71 who may or may not have done so. Assuming that many of these 71 actually did not, one might hypothesise that the performance of the first group should be better than that of the second.

A one-way ANOVA with class attendance as a covariate indicated that this was the case: $F_{1,216} = 9.104$, $p < 0.003$. For this main effect, Ω^2 is 0.03 — again, a small effect size. However, as a test of the effectiveness of the package, this analysis is compromised by the fact that some of the 71 'nonusers' might actually have used the package anonymously — if anything this is a conservative estimate of the performance differential between actual users and nonusers. In real terms, the performance of the two groups on the final exam differed by almost 10% ($M = 39.75$, $sd = 1.30$ and $M = 48.42$, $sd = 12.32$, respectively).

Study 2

Study 1 seems to have demonstrated that use of a WWW-based formative assessment, when an integral part of a course syllabus, seems to be associated with superior performance. Study 2 was conducted in order to establish whether using the package brought any 'added value' to students on a course where its use was not compulsory.

Participants

Participants in this study came from a cohort of 214 undergraduate students studying a Level 2 psychology research methods and statistics module (PSY223) in 1997–98. These students were familiar with PsyCAL, having used it as part of their Level 1 course (they were drawn from the cohort reported on by Buchanan, 1998).

Procedure

As in Study 1, a set of exercises was made available to students. In this case, there were five exercises, each focussed on a particular topic (e.g. measures of association) and comprising 10 multiple choice questions. These were pitched at a higher level than those in Study 1: as well as factual knowledge, they addressed application of theories, selection of appropriate techniques; interpretation of sample statistical output and experimental results. In this course, use of the package was optional rather than being one of the core learning activities of the module: students were simply advised that the package was available and encouraged to use it. Before doing an exercise, students were required to enter an identification number. For the purposes of this study, the software was modified so that it could not be used unless this was done*.

* However, some students successfully entered fraudulent ID numbers: this demonstrates the difficulty in tracking access to on-line assessments and ascertaining whether a particular

The outcome variable in this phase was again performance in the module's summative assessment. This comprised two elements: a set of fairly mechanical statistical exercises completed using SPSS statistical software, and a group project which required some methodological and statistical sophistication. Each student's mark on each of these elements was recorded.

Results and discussion

During the teaching period of the module, the PSY223 exercises were used a total of 103 times by 27 individuals, of whom only 16 were identifiable (the others did not enter valid identification numbers). The mean number of uses was 4.50 ($sd = 4.23$).

When the performance of those 16 identifiable students who had used PsyCAL at least once was compared with the performance of those who had apparently never used it, Mann–Whitney *U*-tests (used because the exercises were marked on a non-interval grade point scale) revealed that on the Project element, PsyCAL users performed significantly better ($U = 1045.0$, $p < 0.04$) than nonusers. No difference was found on the SPSS element ($U = 1406$, $p > 0.4$). This seems to demonstrate that users of PsyCAL did appear to perform better, at least on the element of coursework where application of knowledge — rather than just following textbook procedures — was required.

The problem with these findings is, of course, the small sample size — why did so few of the students (many of whom had previously indicated that they would like to do so — Buchanan, 1998) on this course choose to use PsyCAL? One possibility is that only the brighter or more motivated students took advantage of the exercises — as it was an optional activity, use could be seen as reflecting greater engagement with the course. Superior performance of PsyCAL users could thus be due to the characteristics of those users, rather than an effect of the activity.

Other possibilities are that the exercises were insufficiently publicised or that the requirement to enter an identification number put students off. This latter suggestion is supported by informal feedback from students, and also by fact that log files generated in Study 1 suggested that some students seemed to use the package anonymously until they were sure of the correct answers, then identify themselves and submit a version of the test which would score full marks. This is a phenomenon worthy of future examination.

Another point is that due to the self-selecting nature of this sample of experienced users, students who did these exercises were likely to be those who felt that they benefited from their use in the past. If nothing else, this study does demonstrate that at least some subset of students considered PsyCAL useful enough to do these optional exercises.

Discussion

The results of these two studies suggest that use of this package is associated

exercise has been completed by one person rather than another (e.g. Snow *et al.* 1996; Charman & Elmes, 1998).

with superior exam performance. The fact that level of use was associated with performance in Study 1 suggests that it is not just the case that more able students choose to use the package. Along with the positive evaluations reported by Buchanan (1998), this supports its use as a learning tool, and the fact that usage statistically significantly predicted performance, even when class attendance was controlled for, suggests it has something additional to offer.

These conclusions may be qualified somewhat by the fact that this research was practice-based, rather than rigorously experimental. This was a situation where a system was actually being used and an attempt was made to evaluate it on the basis of observational measures. Experimental extensions to this work, controlling for possible confounds, could be implemented — for example a counter-balanced design where for part of the course half the class had access to the system while the other half did not. In a later section this would be reversed and assessments at the end of each section should reveal any differences. While this would increase confidence in the findings, there may be practical problems with conducting such experiments (e.g. the short time periods which would be involved — other studies have used time periods as short as one week) and perhaps ethical issues (e.g. would informed consent of participants be possible in such a design?).

There is nothing unique about the software described here. While the program used was purpose-written for use in the courses mentioned, its functionality can easily be replicated with modern test-authoring packages (e.g. Roberts, 1996). What does matter is probably the fact that it provides timely feedback which meets the requirements for formative use (Brown & Knight, 1994; Wiliam & Black, 1996). Its value is probably also enhanced by the fact that it does not supply correct answers but encourages the students to do further work on their own — a suggestion which it would be interesting to test experimentally by comparing a system such as this with one which provides correct answers (e.g. Burgess *et al*., 1998).

Does this system provide anything other than a study guide mounted on the WWW? While that in itself can be valuable, it is possible that it does provide something more: this system is an example of the meaningful interaction between student and instructional materials which Bostow *et al.* (1995) argue is an essential component of successful pedagogy — a component which can be provided through technology.

In practical terms, whatever this system provides, the important point seems to be that it 'works'. This is reassuring given the number of educators embracing such teaching methods. A system such as this may not be the ultimate teaching tool (the findings of Study 1 indicated that class attendance had a much stronger effect upon exam performance), and using it will not automatically lead to enhanced learning. There are areas (e.g. presentation skills, learning to work in teams) difficult to develop or assess through multiple choice tests; and there are types and sources of feedback (for instance, on essay writing skills from human tutors who have read the student's work) which may be more valuable.

However, such a system does offer the opportunity to provide some kinds of individualised feedback in a flexible and cost-effective manner. As part of a balanced curriculum utilising an appropriate selection of teaching, learning and assessment methods, it offers another form of learning opportunity, and these findings suggest that students who take up that opportunity are likely to benefit from the experience.

The question which remains is why they benefit. While these results indicate that the intervention had an effect, they tell us nothing about how or why. To answer these questions, more work is required: at the very least, an investigation of how students actually used the system and how they used the feedback they were given. For example, previous findings (Buchanan, 1998) indicate that there are individual differences in how students use this system. There is also an indication that at least some students (e.g. those who falsified their identities in Study 2) were more comfortable using the package in anonymity. Following up these findings could provide insights into the way students use such on-line learning resources and thus inform future practice.

Acknowledgements

Some of the results reported in this paper were presented at the Society for Computers in Psychology meeting in Dallas, Texas, November 1998. The assistance of Clive Lewis with development of materials for Study 2 is gratefully acknowledged.

References

Bostow, D.E., Kritch, K.M. & Tomkins, B.F. (1995) Computers and pedagogy: Replacing telling with interactive computer-programmed instruction. *Behavior Research Methods, Instruments & Computers*, **27**, 297–300.

Brown, S. & Knight, P. (1994) *Assessing Learners in Higher Education* Kogan Page, London.

Buchanan, T. (1998) Using the World Wide Web for formative assessment. *Journal of Educational Technology Systems*, **27**, 71–79.

Burgess, C., Lund, K., Keeney, M. & Audet, C. (1998) *A Generic Multiple Choice Test Program for Web-Based Applications* Paper presented at the November meeting of the Society for Computers in Psychology, Dallas, Texas.

Callear, D. & King, T. (1997) Using computer-based tests for Information Science. *ALT-Journal*, **5**, 27–31.

Charman, D. & Elmes, A. (1998) *Computer Based Assessment (Vol 1): a Guide to Good Practice* SEED Publications. University of Plymouth.

Clariana, R. (1997) Pace in mastery-based computer-assisted learning. *British Journal of Educational Technology*, **28**, 135–137.

Duchastel, P. (1997) A Web-based model for University instruction. *Journal of Educational Technology Systems*, **25**, 221–228.

Gibbs, G., Skinner, C. & Teal. A. (1998) *coMentor*: A collaborative learning environment on the WWW. In *Cip98 Conference Proceedings* (eds A. Trapp, N. Hammond & C. Manning), p. 32. CTI Centre for Psychology, York.

Krantz, J.H. (1995) Linked Gopher and World-Wide Web services for the American Psychological Society and Hanover College Psychology Department. *Behavior Research Methods, Instruments & Computers*, **27**, 193–197.

Ramsden, P. (1992) *Learning to Teach in Higher Education*. Routledge. London.

Roberts, A. (1996) Question Mark Designer for the Web (QM Web). *Active Learning*, **5**, 57.

Snow, H., Monk, A. & Thompson, P. (1996) Guidelines for the use of multiple choice and computer presented tests for University assessment. *Psychology Software News*, **7**, 4–8.

Stockburger, D.W. (1998) Automated grading of homework assignments and tests in introductory and intermediate statistics courses using active server pages. Paper presented at the November meeting of the *Society for Computers in Psychology.* Dallas, Texas.

Wiliam, D. & Black, P. (1996) Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal*, **22**, 537–548.