

Costa, D., Mullan, B. A., Kothe, E. J., & Butow, P. (2010). A web-based formative assessment tool for Masters students: A pilot study. *Computers & Education*, 54(4), 1248-1253.

<http://dx.doi.org/10.1016/j.compedu.2009.11.011>

A web-based formative assessment tool for Masters students: A pilot study

Daniel S.J. Costa¹, Barbara A. Mullan², Emily J. Kothe*, Phyllis Butow³

School of Psychology, University of Sydney, Australia

The use of web-based learning and assessment tools is growing in tertiary institutions around the world. To date, very few papers have reported the development and evaluation of a web-based formative assessment tool for postgraduate students. The aim of the present paper was to report on the development and evaluation of an online formative assessment tool for this student group. The web-based formative assessment tool was evaluated by a sample of undergraduate students, postgraduate students and academic staff within a psychology department in order to determine the suitability and sensitivity of the tool. The results of this pilot test suggest that the development of such a tool is both appropriate and feasible for Masters students studying psychology.

Keywords: Evaluation of CAL systems; Interactive learning environments; Post-secondary education; Simulations; Teaching/learning strategies

1. Introduction

The number of students enrolled in tertiary education is increasing in many developed countries around the world (OECD, 2009). In some cases, for example as in Australia, student–teacher ratios have risen dramatically over the same period (DEST, 2004). At the same time there has been an increase in demand for alternatives to on-campus teaching (Krause, Hartley, James, & McInnis, 2005). Together larger class sizes and changes in student needs have prompted a shift towards alternative teaching methodologies in many institutions. Much of this shift has occurred in the form of computer-assisted learning and assessment, and in particular through the use of web-based delivery of course content and/or assessment (Anderson, Johnson, & Saha, 2002).

The use of technology in educational enrichment and assessment has been the subject of much discussion in recent years. The use of computers in education can be traced back to the 1960s (Harding, 2001) and with constant advances in computer power and accessibility, teaching and assessment via information and communication technology (ICT) is emerging as a useful supplement to more traditional methods. Computer-assisted learning and assessment provides a means through which assessment and associated feedback can be efficiently administered to students in the tertiary setting (Lyall and McNamara, 2000, Peat and Franklin, 2002 and Ring, 1993).

1.1. Online formative assessment

While it is now common practice for teaching staff to make course material available electronically (Anderson et al., 2002 and Peat and Franklin, 2002), use of structured programmes of formative assessment is a relatively recent development. The importance of formative assessment was highlighted by Gipps (2005), who noted that learning is not

enhanced by feedback in the form of marks or grades alone, but by comments on the good and bad aspects of performance. Gipps (2005) argued that only the latter constitutes true formative assessment. Online formative assessment, while requiring the training of staff, has the added advantage of allowing greater independence and self-direction on the part of the student, thus enabling staff to focus attention on students in greater need of assistance (O'Reilly and Patterson, 1998). Velan, Kumar, Dziegielewski, and Wakefield (2002) added that online self-assessment allows flexibility with regard to the time and place in which assessment is undertaken, rapid delivery of individualised feedback and opportunity for repetition.

Specific online formative assessment tools have been analysed in terms of usage and correlation with academic performance, with varying results (e.g. Clariana, 1993, Henly, 2003, Velan et al., 2002 and Zakrzewski and Bull, 1998). For example, Velan et al. (2002) found that in their online formative assessments for undergraduate pathology students, which consisted of questions of a variety styles that were able to be attempted on multiple occasions, nearly all of the targeted students completed the assessments at least once. Of those that made multiple attempts, performance on the last attempt tended to be better than that on the first. In contrast, Henly (2003) found that both performance and usage of an online self-assessment tool, designed for undergraduate dentistry students and featuring multiple-choice, short-answer and extended matching questions, declined over the course of the semester. Henly (2003) suggested that the material may have become progressively more difficult, or that the manner of usage changed, such that students initially revised and prepared for the self-assessment task, but subsequently used it as a guide to revision for summative assessment. Despite the decline in performance and usage, the top 10% of students (as determined by overall summative assessment) used the tool significantly more

than the bottom 10%, although the author conceded that it is not known whether using the tool improved performance or whether the top students were simply more conscientious. The issue of conscientiousness was addressed by Buchanan (2000), who found that online self-assessment usage amongst introductory psychology students was a good predictor of examination performance, even when class attendance, as a reflection of conscientiousness, was held constant. Results suggest that, at least within the discipline of psychology, online formative assessment tasks have the potential to improve student outcomes.

1.2. The need for a formative assessment tool for Masters students

The aim of the present paper is to outline the development and preliminary evaluation of an online formative assessment module developed for use by students enrolled in a Masters level psychology course. A pre-requisite for entry into the course was either a 4-year psychology degree or a cognate degree and relevant work experience. As such, enrolling students were expected to have received some instruction regarding research methods and data analysis. However, given the differences in course content between universities and disciplines it was acknowledged that some students may have completed degrees that did not address research methods and statistics with the same level of detail as expected in psychology. Furthermore, regardless of the degree studied, a long period of time may have elapsed since some students enrolled in the course had formally studied these concepts. In light of this, an electronic self-directed learning tool, available through the University of Sydney eLearning environment, was deemed an appropriate way to ensure that students' knowledge of research and statistical concepts and issues was adequate.

Two particular characteristics of the Masters course resulted in the need for an online formative assessment tool that differed from those developed in the past. Firstly, the Masters

course is a postgraduate programme, while online formative assessment is typically used in undergraduate programmes. Two notable exceptions are the programmes reported by Crossouard, Pryor, and Torrance (2004), although the programme reported by those authors differed substantially to the one described here, in that students were provided with qualitative feedback, via discussion forum, for qualitative work, and that reported by Klecker (2007), who found that graduate education students randomly allocated to a condition in which they received weekly online formative tests produced higher final exam scores than those who were allocated to a control condition. Postgraduate students presumably take a more self-directed approach to study than do undergraduates (Gregory & Wohlmuth, 2002), and this is an important factor to consider in the development of a formative assessment tool. Secondly, the material in the online tool was not explicitly taught, and therefore not summatively assessed, in the coursework components of the Masters course. As a consequence, the online tool employed here was very much self-contained, meaning not only that the tool had to provide an adequate learning opportunity for students, as there was no corresponding class-work to consolidate learning, but that evaluation of the efficacy of the tool for Masters students was made difficult, as there would be no summative criterion against which it could be assessed.

The online self-assessment⁴ tool was designed with these particular learning needs of its target students in mind. For this reason, the tool included several features that had not been described previously in the literature. Most notably, each of six topic categories (research design, research ethics, qualitative research, descriptive statistics, inferential statistics and non-parametric statistics and regression) was divided into three levels of difficulty. Such categorisation of questions into levels of difficulty was employed by Peat and Franklin (2002) in their online tool developed for undergraduate Biology students. In their online tool,

students were able to choose the level of difficulty they wished to attempt, irrespective of their performance. In the present tool, progression to the next level of difficulty was contingent upon a certain level of achievement in the lower level of difficulty, according to some pre-defined criterion. In the event that this criterion was not achieved, students were prevented from accessing the quiz for a certain amount of time and were encouraged to pursue the given resources to ensure that their performance improved in their next attempt. In addition, upon completion of a quiz, correct answers were not revealed for incorrectly answered questions. Instead, a hyperlink to a relevant article, book chapter or web page was provided. The aim of providing these links rather than more direct feedback (e.g., the correct answer and the reason the selected answer was incorrect) was to encourage students to actively seek information rather than merely memorising the correct answers. A similar technique was employed by Buchanan, 1998 and Buchanan, 2000, although he provided references to useful resources, rather than the resources themselves. It was felt that providing the resources directly would increase the likelihood that they would be accessed by students.

1.3. Preliminary evaluation of the online formative assessment tool

Prior to being made available to the target students, the online tool was subjected to pilot testing by students (both undergraduate and postgraduate) and academic staff in the School of Psychology. The outcome of this pilot testing is reported in the present paper. The purpose of this study was to gather information that would allow the refinement of the tool in two ways:

- item statistics were gathered to determine which items were most appropriate for inclusion in the final version of the tool;
- feedback regarding the general usability (e.g., format, content, duration, difficulty) was collected.

Due to various constraints, several of the characteristics unique to this online tool, described above, were not assessed in the present study. In an attempt to maximise completion rates,

progression to higher difficulty levels was not made contingent upon performance – this contingency upon achieving a criterion before progression was also deemed inappropriate because one aim of the study was to determine the appropriateness of the categorisation of items according to difficulty. In addition, although feedback links were made available to participants, correct answers were also given (unlike what is proposed for the tool to be implemented for Masters students), as part of the purpose of this pilot study was to receive feedback on question content, including whether the responses labelled correct were in fact correct. Usage of feedback links was not assessed directly, although participants were free to comment on this aspect of the tool.

It was predicted that in each of the six quizzes, the total score should decrease as participants progressed from the lowest to the highest level of difficulty. A further prediction was that performance would depend on whether participants were staff members, postgraduate students or undergraduate students. Although these three participant profiles were included in order to obtain a wide variety of feedback, and not for the purpose of comparing their performance, group differences were expected. Specifically, given the predominantly theory-based nature of the questions, performance was expected to be a function of how recently the material had been formally studied; that is, undergraduates were expected to perform best, followed by postgraduates and then staff.

2. Method

2.1. The eLearning module

A pool of 396 multiple-choice questions, each with five possible responses (only one of which was correct), were administered via the University eLearning environment. Questions were divided into six topic categories (research methods, ethics, qualitative research,

descriptive statistics, inferential statistics and non-parametric statistics/regression/correlation), and within each topic category questions were further categorised into three levels of difficulty, such that there were 22 questions per level of difficulty per topic category. Although progression from easier to more difficult questions is proposed to be contingent upon adequate performance in the tool that will be implemented for Masters students, in the present pilot study questions were presented in random order. Each participant was randomly allocated two topic categories (i.e., 132 questions), which could be completed in the participants' own time using any computer with Internet access. Upon completion of each quiz, participants were given detailed feedback about their performance, including a re-statement of questions and answer-choices, their response, the correct response and a hyperlink to a web page, electronic journal or article/book chapter that provided the information necessary to answer the particular question correctly. The total score for each quiz out of 66 was also reported.

2.2. Participant's ratings of the module

Participants were asked to complete an online survey relating to the online tool once they had completed both quizzes. Participants were asked nine 5-point Likert-scale items regarding the relevance, clarity, difficulty, appropriateness, and the ease of use of the tool. (e.g. "The online tool was easy to use – strongly agree, agree, neutral, disagree, strongly disagree"). All participants were also asked to respond to six open-ended questions regarding their experience with using the tool (e.g. "How well did you think you did before you got your results?").

3. Results

3.1. Participants

Forty-three individuals agreed to take part in the study, 14 academic staff, 15 postgraduate psychology students and 14 undergraduate psychology students. Overall, nine academic staff, 13 postgraduate students and five undergraduate students contributed to the study.

3.2. Likert-scale item feedback

The responses to the Likert-scale items were numerically coded (5 = strongly agree, 4 = agree, 3 = neutral, 2 = disagree, 1 = strongly disagree) and mean ratings for each item were calculated. The frequencies of each of these responses, along with item means, for each question are reported in Table 1. As shown in that table, 83% of participants reported that questions were relevant, and 79% agreed that questions were clearly expressed. Ninety-two percent of participants agreed that the online tool was easy to use, 92% agreed that the format was appropriate and only 8% reported that the instructions were not clear enough. With regard to the difficulty of the task, no participants reported that the questions were too easy, with 13% reporting that they were too difficult. Overall, 67% of participants reported obtaining some benefit from using the tool, with 58% of participants agreeing that they enjoyed using the online tool.

Table 1. Frequencies and mean ratings for Likert-scale evaluation items.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Mean
The questions were relevant	6	14	3	0	1	4
The questions were expressed clearly	3	16	4	0	1	3.833
The questions were too difficult	0	3	3	17	1	2.333
The questions were too easy	0	0	7	15	2	2.208
The online tool was easy to use	9	13	1	0	1	4.208
I obtained some benefit from using the online tool	2	14	8	0	0	3.75
The format of the online tool was appropriate	3	19	1	0	1	3.958
The instructions were not clear enough	0	2	0	11	11	1.708
I enjoyed using the online tool	2	12	8	2	0	3.583

3.3. Open-ended feedback

Twenty-four participants provided qualitative feedback. Some responses were concerned primarily with typographical, formatting and content errors, and the clarity of the instructions. However, the majority of the responses to the open-ended evaluation questions related to the difficulty (e.g., “Some of the parametric tests were really difficult”), length (e.g., “The questions were all relevant, however, there were too many items to sit through”) and repetitive nature (e.g., “sometimes I thought questions were quite repetitive – some questions seemed to answer other previous questions”) of the quizzes. Since one of the purposes of this study was to select from a larger pool of items the most appropriate items for the online tool to be made available to Masters students, such feedback was expected.

3.4. Test performance

Overall, 57 quizzes were completed. Twenty-three participants completed 2 quizzes, while one participant completed 6 and another completed 3 (both postgraduates). Two participants (one staff, one undergraduate) completed only one quiz. The number of each topic completed is included in Fig. 2. To examine differences in performance between undergraduate students, postgraduate students and staff, the percentage of questions answered correctly was calculated for each participant who had completed at least one quiz, and subjected to regression analysis with dummy variables representing comparisons between postgraduates and each of the other two groups. In an attempt to reduce the impact of small sample size on the distributions of test scores, *p*-values based on bootstrapped confidence intervals (with 1000 bootstrap samples) are reported (cf. Efron & Tibshirani, 1993). The mean percentage of correctly answered items was significantly higher for both staff members (74.33%, SD = 13.03, *p* = .039) and for undergraduates (74.85%, SD = 11.02, *p* = .013) than for postgraduates (60.96%, SD = 17.41). Individual scores and group means are presented in Fig.

1. Note that one individual in the postgraduate condition produced a particularly low score. Although there is no reason to believe that this was not a serious attempt, the analysis was re-run with the participant excluded. The mean for postgraduates increased to 64.60 and the differences between this group and staff ($p = .096$) and undergraduates ($p = .059$) became only marginally significant.

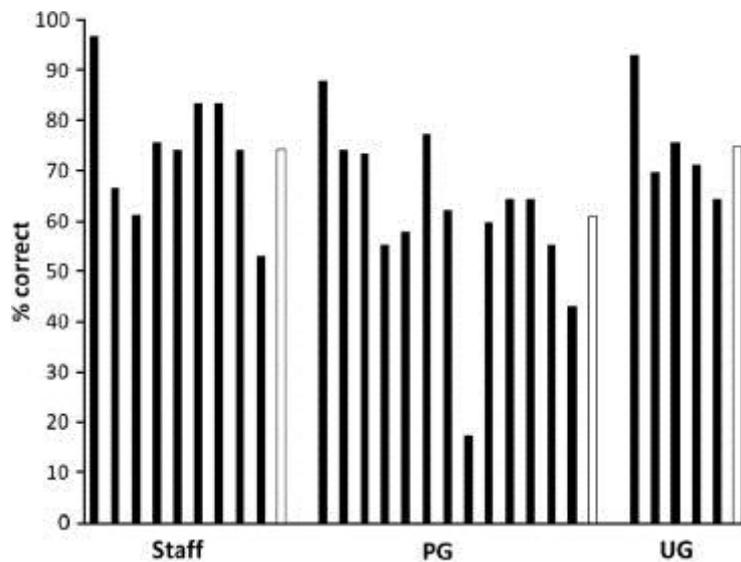


Fig. 1. Percentage correct for all quizzes attempted by individual staff, postgraduate students and undergraduate students (black), as well as the means of these three groups (white).

To determine whether the questions selected in each level of difficulty were appropriate to that level, mean performance was compared across the three categories. Since the online tool produced difficulty-level data for each item, and not each participant, the analysis of item difficulty was based on a between-group regression analysis (low, moderate, high) on individual items, not participants. The results of these analyses are displayed in Table 2 and are represented graphically in Fig. 2.

Table 2. Regression analysis results for performance predicted by difficulty level. The regression coefficients (*b*) reflect the mean change in percentage correct as difficulty increases by one level, and *p*-values were obtained via bootstrapping (1000 samples).

Quiz topic	<i>b</i>	<i>F</i> (1, 63)	<i>p</i>
Descriptive statistics	-6.06	5.78	.002
Ethics	-6.57	1.67	.108
Inferential statistics	-6.14	1.58	.069
Non-parametric tests, correlation and regression	-1.36	0.61	.761
Qualitative research	-10.39	3.20	.005
Research methods	-5.87	1.33	.055

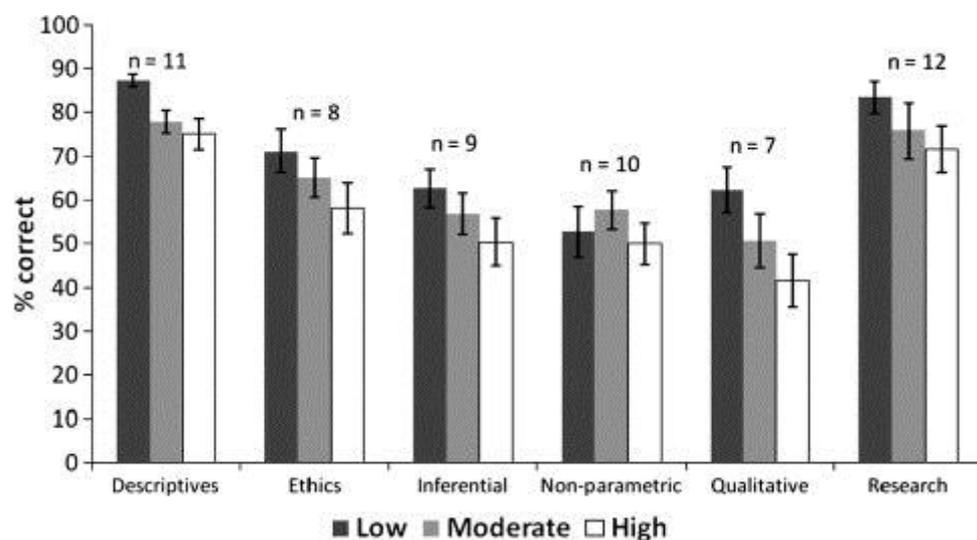


Fig. 2. Mean percentage correct for each of the three levels of difficulty for each quiz. Error bars represent ± 1 standard error. The number of quizzes contributing to each category is indicated above the columns.

4. Discussion

The results of this pilot study generally suggest that postgraduate students may benefit from the use of online formative assessment. Of those who agreed to participate in the study, 92%

of the postgraduates completed at least one quiz, compared to 64% of academics and 36% of undergraduates, suggesting that postgraduate students believe that they will derive some benefit from participation. That postgraduate students may be in need of such an online tool is supported by the observation that these students performed worse than both academic staff members and undergraduate students, despite presumably having formally studied similar material more recently than the former, while having more practical experience with the topics covered by the material than the latter. The discrepancy of approximately 13 points between postgraduates and the other two groups was striking. The relatively poor performance of postgraduate students confirms the need for targeted training for this group. It is worth noting that the size of the present sample was limited; as such it may be that the use of this small sample resulted in inadequate power. That these results achieved (or approached) statistical significance in spite of the small number of participants recruited makes these results all the more striking.

Low participation rates are of a larger concern with regard to large-scale roll-out of this programme within the Masters programme. A number of earlier studies have reported that use of online formative assessment can be low, especially amongst important subgroups. For example, Henly (2003) reported that low ranked students, who may have been most likely to benefit from use of eLearning programmes, were less likely to make full use of resources supplied online. Similarly, Buchanan (2000) reported that of 223 undergraduate students who had access to online formative assessment, only 27 made use of the tool during semester. As such, it appears that motivating students to access formative assessment tools may be an important intermediate step when attempting to improve performance via such tools. Methods by which to motivate students to complete formative assessment tasks is an important direction for future research in this area. In light of this, the very high participation rate

amongst our postgraduate participants suggests that this group are an appropriate target for an online formative assessment tool.

It was predicted that that in each of the six quizzes, the total score should decrease as participants progressed from the lowest to the highest level of difficulty. This trend was observed for two of the six content domains assessed the in eLearning module (descriptive statistics, qualitative research), while a further two were marginal (ethics, research methods). It appears that the allocation of questions to three levels of difficulty was generally appropriate. However, for the remaining two domains (namely, inferential statistics and non-parametric tests, correlation and regression) it appears that some reorganisation of these questions is required in order for this approach to be used successfully. Since these two categories also produced the lowest performance scores, it may be the case that the topics are inherently more difficult than the others.

Such categorisation of questions into levels of difficulty was employed by Peat and Franklin (2002) in their online tool developed for undergraduate Biology students. In their online tool, students were able to choose the level of difficulty they wished to attempt, irrespective of their performance. In contrast, in the version of the present online tool that will be available to Masters students, progression to the next level of difficulty will be contingent upon a certain level of achievement in the lower level of difficulty, according to a pre-defined criterion. Although not tested in this pilot study, this system of progression should ensure that students have a clear idea about their level of knowledge relative to that which is expected of them, and should also encourage these students to make use of the suggested resources. Given the constraints of the tool and its evaluation in the pilot study it was not possible to determine the extent to which individuals used suggested resources and readings to extend

their knowledge before and after assessment. This is an issue that researchers may wish to investigate in future research.

Masters students differ in a number of important ways from their undergraduate peers and as such understanding the way in which they interact with online formative assessment, and developing and implementing tools for use in this student population in an important research task. The evaluation of this online tool prior to its implementation proved useful in that the aim of obtaining detailed feedback upon which to base refinements to the tool was achieved. The responses to the Likert-scale items show that the tool was generally regarded in a positive light, while at the same time the qualitative feedback contained many useful suggestions as to ways the tool could be improved. These findings support the further development and implementation of the eLearning formative assessment tool for Masters students, particularly with regard to subjects that Masters students found particularly challenging.

5. Conclusion

In Australia, as in other developed countries around the world, tertiary institutions are increasingly making use of computers and the internet to deliver learning materials and assessment. This growing trend reflects changes in the characteristics and expectations of students and increases in computing accessibility and usability. This paper highlights several challenges faced by educators when attempting to develop and evaluate such modules, particularly when material covered by online formative assessment is not summatively assessed as part of course requirements and when tools are developed for use by postgraduate students. Overall, this research does suggest that online formative assessment is a feasible method of self-assessment in this context. However, the overall low participation rates and challenges faced when designing suitable assessment items do emphasise the importance of

extensive development and pilot testing in order to ensure that tools are appropriate and engaging for the intended student group.

Acknowledgements

The development of this eLearning tool was funded in part by a Teaching and Improvement Grant from the Teaching Improvement and Equipment Scheme at the University of Sydney.

The authors would also like to thank the staff and students who took part in this study, and also Tom Buchanan, who provided advice and resources that assisted in the development of the online tool.

References

- Anderson, D., Johnson, R., & Saha, L. (2002). Changes in academic work: Implication for Universities of the changing age distribution and work roles of academic staff. <http://www.dest.gov.au/sectors/higher_education/publications_resources/summaries_brochures/changes_in_academic_work.htm> Retrieved 30.06.09.
- Buchanan, T. (1998). Using the World Wide Web for formative assessment. *Journal of Educational Technology Systems*, 27(1), 71–79.
- Buchanan, T. (2000). The efficacy of a World Wide Web mediated formative assessment. *Journal of Computer Assisted Learning*, 16, 193–200.
- Clariana, R. B. (1993). A review of multiple-try feedback in traditional and computer-based instruction. *Journal of Computer-Based Instruction*, 20(3), 67–74.
- Crossouard, B., Pryor, J., & Torrance, H. (2004). Creating an alternative assessment regime with online formative assessment: Developing a researcher identity. In Paper presented at the European conference on education research, Crete.
- Department of Education Science and Training (2004). Selected Higher Education Statistics. <<http://www.dest.gov.au/highered/statpubs.htm>> Retrieved 30.06.09.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Gipps, C. V. (2005). What is the role for ICT-based assessment in universities? *Studies in Higher Education*, 30(2), 171–180.
- Gregory, V. L., & Wohlmuth, S. R. (2002). Planning for the internationalization of a postgraduate professional degree programme in library and information science. *Higher Education in Europe*, 27(3), 261–268.

- Harding, R. (2001). What have examinations got to do with computers in education? *Journal of Computer Assisted Learning*, 17, 322–328.
- Henly, D. C. (2003). Use of Web-based formative assessment to support student learning in a metabolism/nutrition unit. *European Journal of Dental Education*, 7, 116–122.
- Klecker, B. M. (2007). The impact of formative feedback on student learning in an online classroom. *Journal of Instructional Psychology*, 34, 161–165.
- Krause, K.-L., Hartley, R., James, R., & McInnis, C. (2005). The first year experience in Australian Universities: Findings from over a decade of national studies. University of Melbourne.
- Lyall, R., & McNamara, S. (2000). Learning tool or potplant stand? Students' opinions of learning from a CAL program in a distance education context. *Australian Journal of Educational Technology*, 16(2), 126–146.
- O'Reilly, M., & Patterson, K. (1998). Assessing learners through the WWW. In Paper presented at the 7th international World Wide Web conference, Brisbane.
- Organisation for Economic Co-operation and Development (2009). OECD stats extracts – students enrolled by type of institution.
<<http://stats.oecd.org/Index.aspx?DatasetCode=RENRL>> Retrieved 30.06.09.
- Peat, M., & Franklin, S. (2002). Supporting student learning: The use of computer-based formative assessment modules. *British Journal of Educational Technology*, 33(5), 515–523.
- Ring, G. (1993). Characteristics of state-of-the-art computer-administered testing systems. *Australian Educational Computing*, 8(2), 25–28.

Velan, G. M., Kumar, R. K., Dziegielewska, M., & Wakefield, D. (2002). Web-based self-assessments in pathology with questionnaire perception. *Pathology*, 34, 282–284.

Zakrzewski, S., & Bull, J. (1998). The mass implementation and evaluation of computer-based assessments. *Assessment & Evaluation in Higher Education*, 23(2), 141–152.