A COMPARISON OF AN INNOVATIVE WEB-BASED ASSESSMENT TOOL UTILIZING CONFIDENCE MEASUREMENT TO THE TRADITIONAL MULTIPLE CHOICE, SHORT ANSWER AND PROBLEM SOLVING QUESTIONS

Graham Farrell

A Comparison of an Innovative Web-based Assessment Tool Utilizing Confidence Measurement to the Traditional Multiple Choice, Short Answer and Problem Solving Questions

Graham Farrell
Usability and Innovation Group
Swinburne University of Technology
Australia
gfarrell@ict.swin.edu.au

Key Words

Innovative Web-based Assessment Tool
Confidence Measurement
Multiple Choice, Short Answer and Problem Solving Assessments
Comparative Analysis
Convergence Validity
Reliability

Abstract

Computerized assessment is playing a major role in IT education, with extensive utilization of the multiple choice question (MCQ) format. This is mainly due to the ease of adaptation of MCQs into the internet environment, offering extensive advantages to both the student and the instructors. This study analyzes the results of students' grades using an alternative web-based assessment tool and the more traditional modes of assessment, being Multiple Choice Questions, Short answers and Problem Solving (Scenario) questions. The Multiple Choice Questions with Confidence Measurement (MCQCM) is a web based assessment tool that permits the student to register their level of confidence in their answer, and was included as a revision tool for the duration of the semester and as a component of the final exam. Additionally the exam also contained questions using more traditional methods for assessment. A total 43 students sat the final exam producing some interesting results. The statistical analysis indicated that the correlation between the MCQCM and the other alternatives ranges from strong to medium. In addition it appears that the MCQCM demonstrated equal to slightly stronger convergence of validity compared to the traditional MCQ method and the other alternative assessment methods.

Introduction and Literature Review

Educational institutions utilize a variety of assessment options to grade their students and assess the effectiveness and validity of subject content. A critical component of sound educational programs is to assess the learning outcomes throughout the duration of the course, as both a means of giving timely feedback and as a mechanism to grade the students. Black and William (1998) use the term “Assessment” as referring to the group of activities that are undertaken by both teachers and students in self assessment, providing both grades and feedback to modify teaching. Educators appreciate that each kind of assessment has its purpose (Assessment Tools, 2003). It is accepted wisdom that assessment should be an integral part of the learning activities rather than an interruption. (See Principles and Standards for School Mathematics (2000) for example.)

An issue facing educators is what methods of assessment should they be using and what would be the appropriate mix to maximize the feedback and evaluation process? Schuwirth and Van Der Vleuten (2003) state “a well designed assessment program will use different types of questions appropriate for the content being assessed”. The options presently available to the instructors include multiple choice questions (MCQ), short answer questions (SA), longer problem solving questions (PS), case study reports, presentations and other equally effective and proven choices. In the majority of cases the final grade is calculated by combining each separate mark from assessment tasks completed during the subject. The utilization of multiple assessment methods recognizes the need to permit students to demonstrate their knowledge in various methods throughout their learning experience.

Multiple choice questions (MCQs) are highly regarded by instructors (Bacon 2003) and consequently utilized extensively, with world wide experience in their construction (Schuwirth and Van Der Vleuten 2003). In addition, the ease of adaptation to the computer assessment environment has been swift and effective. There are two roles that MCQs play in the balanced educational program. Firstly, MCQs are used extensively as a means of formative assessment (self assessment), where the feedback influences the direction of the students as they journey along their learning path. MCQs are a popular self-assessment option being readily available to the students due to the advancement of technology that now supports its functions. Web based MCQ self-assessment packages permit the student to self assess their knowledge at any time convenient to them, providing instant feedback and in many cases recommended change in directions to their learning path. Secondly, MCQs are also traditionally used for summative assessment for the grading of students, being strategically placed in the exams with various mark allocations directly contributing to the students’ final grade. Their popularity can be attributed to their ability to “yield equivalent reliability and validity in a shorter amount of time” as they have an “economy of scale not found in constructed-response” (Bacon 2003). In addition they are considered to have the ability to test many topic areas in relatively shorter time (Wilson and Case 1993). Bacon (2003) also identifies one advantage of using MCQs is the “Objective” marking as a method of avoiding the “obvious lack of reliability of essay tests”,

as he sites previous work Ashburn's (1938) where subjective marking of short essay answers yielded significant difference in grades when remarked. Schuwirth and Van Der Vleuten (1996) emphasize the growing dissatisfaction with the MCQ format as they rely on recognition of the correct answers, while some see MCQs as only demonstrating knowledge of isolated facts (Wilson and Case 1993). Wilson and Case (1993) also state that they fear this "undue emphasis on recall" will "stimulate students to learn in a like mode". Schuwirth and Van Der Vleuten (2003) go on to recommend variation in the question formats due to the likelihood that students will prepare depending on the types of questions used. Bacon (2003) discusses at length the concerns of some that the MCQ format is too simple and does not assess the complex levels of knowledge, in particular the higher levels of Bloom's (1956) taxonomy of educational objectives (Knowledge, Comprehension, Application, Analysis, Synthesis, Evaluation). Bacon (2003) does recognize the examples of MCQs in Blooms (1956) work that demonstrate the application of MCQ testing designed to assess outcomes at every level. It is also recognized that this level of MCQ is difficult to construct. However, some educators argue strongly that research has demonstrated that the question format is of limited importance and that the construction of the question is critical (Schuwirth and Van Der Vleuten 2003).

The Short Answer (SA) assessment format has equal popularity as the MCQ alternative. Short answer assessment strategies can offer more flexibility, with greater ability to test creativity and higher levels of Bloom's (1956) taxonomy of educational objectives, as outlined previously. However, SAs are resource intensive when grading and are subject to poor reliability due to subjective marking.

The longer Problem Solving (PS) questions are often included in the final exam as it permits the instructor to assess the highest of Blooms levels. The format of these questions usually present the student with a scenario situation which requires the student to call upon many aspects of the subject material to analyze, synthesize and evaluate, offering alternatives in some situations. These are clearly more difficult to grade consistently as there is often not a prescribed correct solution but a number of equally valid alternatives.

In this study we introduce a fourth assessment option. The students are required to complete a formal assessment task utilizing the MCQCM, contributing to their final grade. The MCQCM is a web-based assessment that has been developed over a period of years designed to permit the student to register their confidence in each of their choices and consequently be rewarded or penalized proportionally. (Farrell, Leung, 2004) The MCQCM format is similar to the MCQ display where each question has a stem followed by four options (Klohe 1995, Frary 1993). Once the student commits to an answer ("level") they are required to register their confidence in that choice ("strength"). (Bandara 1983, Betz & Hacket 2002)
Each option of the question must be committed to either correct or incorrect.
The confidence is registered as a %, with 100% stating complete certainty in the choice and a low % representing extreme doubt. Fig 1 demonstrates the tool in action.

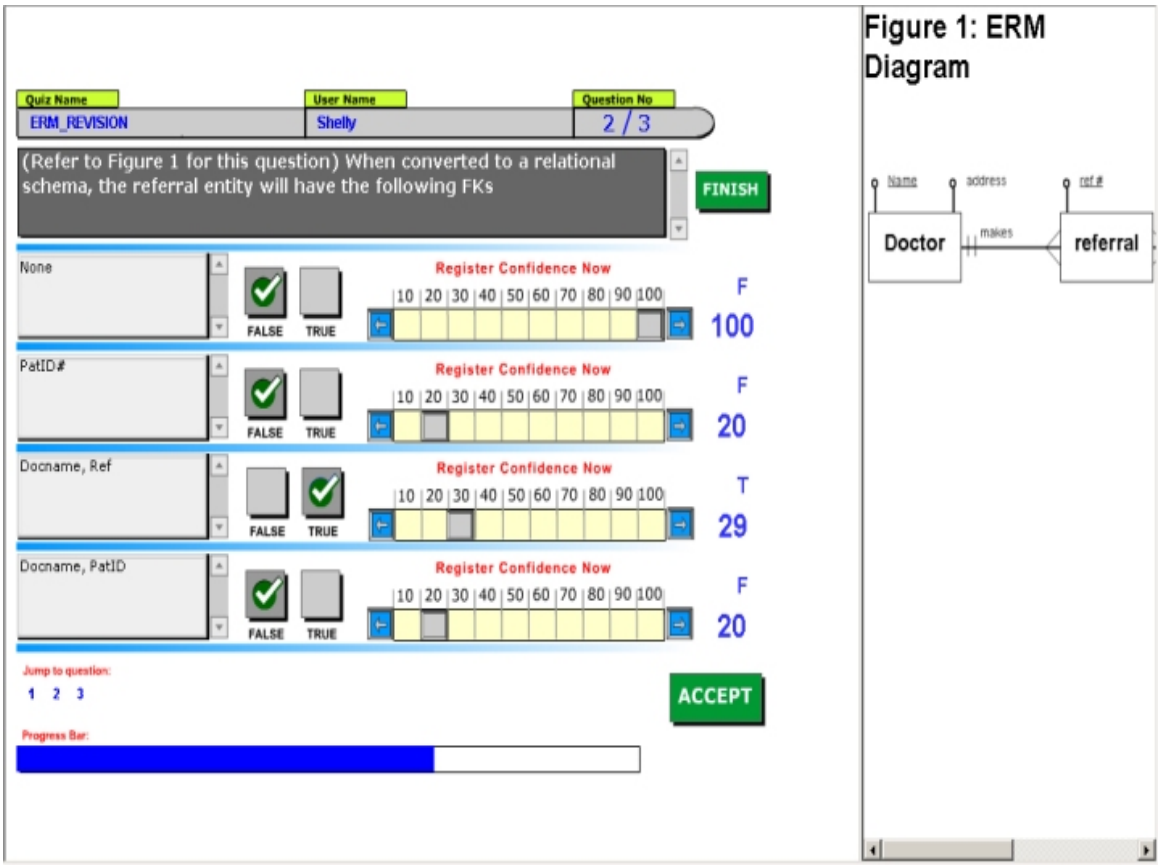

Fig 1: Screen shot demonstrating the tool in use. In this case the ERM is given on the side and the student is required to identify the Foreign Key. This example demonstrates very little confidence by the student in the subject material.

Scoring

Registering a high level of confidence for a correct answer results in a high positive score. (Eg. 100% gives 10 marks), decreasing in increments of 1 for less confidence (90% gives 9, 80% gives 8 etc).
In comparison registering a high % for an incorrect answer gives a large negative result with the same increment (Eg. 100% gives -10, 90% gives -9 etc).
Importantly the students utilize the system as a formative assessment option during the semester and are familiar with the functionality and scoring mechanism.

The Validity of any testing method is mainly assessed using comparison with other test methods (Schuwirth and Van Der Vleuten 1996), yet is often a point of debate (Bacon 2003). Schuwirth and Van Der Vleuten (2003) define the validity as “whether the question actually tests what it is purported to test”. A recognized method of assessing validity is by comparing the correlations between methods of testing that are supposed to measure the same construct (Bacon 2003).

In addition, the Reliability of any testing method is defined as the accuracy of which a score on a test is determined, or more precisely, a score that a student obtains should indicate the score that this student would obtain in any other given (equally difficult) test in the same field ("parallel test") (Schuwirth and Van Der Vleuten 2003).

In previous study (Farrell & Leung 2005) it was demonstrated that the MCQCM provided a rich formative assessment tool, guiding both student and instructor to areas of concern in the student's learning path. The student using MCQCM is not only able to alert the instructor to any areas where knowledge is lacking or incorrect (as in MCQ's), but can also demonstrate areas where they have partial knowledge and/or lack confidence in their knowledge. While the MCQCM proved to be beneficial in its feedback objective it remained to show that it was at least equivalent in its convergence of validity as an assessment tool to the standard accepted MCQ format.
This paper will firstly present an examination which includes four separate methods of assessment. It will then statistically compare the results for each student across each method. A discussion and conclusion will follow to determine the validity of MCQCM as an assessment tool.

Method and Objectives

A total of 43 students sat the final exam as part of the formal grading process of an IT subject.
The exam consisted of an 8 Multiple Choice Question (MCQ) section followed by 8 MCQCMs, 8 Short Answer Question (SA) section and a 2 part Longer Problem Solving questions (PS). The students sat the final 3 Hr exam at the same time on campus. The MCQ and MCQCM sections carried 20% each of the final exam grade, the SA section carried 33% while the longer PS section the remaining 27%. The author of the exam was mindful of Bloom's (1956) taxonomy of educational objectives when constructing the questions to facilitate the assessment of various levels.

The results were collected on the completion of the exam and each question's mark was carefully recorded for analysis.

Results and Discussion

To facilitate this study we investigated the exam results of a cohort of 43 Information Technology students enrolled in the optional subject.

Section	Average Grade	Standard Deviation
MCQ	73%	17.7%
MCQCM	67%	21.0%
SA	85%	9.8%
Problem Solving	75%	14.5%

Table 1:Means and Standard Deviations for each of the sections of the exam

On analysis of the data in Table 1 it is noted that the average grades for all sections of the paper are close, as too are most of the standard deviations. It is observed that the SA section has the greater average grade with a smaller Standard Deviation. Instructors would be quite pleased with these outcomes at this stage.

On further examination and analysis of the data it was found that in most cases there appears to be a good relationship between each of the grades allocated for each of the sections for the individuals. (In a few instances this is not the case) Again this is very pleasing for the instructor as there appears to be a good convergence for each of the assessment areas under consideration. As educators we rely on a reasonable convergence of the grades for each of the sections. Failure to achieve this might indicate poor question construction in a particular section. In this case there does not appear to be any one area of concern.

At this stage, a statistical analysis is appropriate to identify the true relationship between these results.

The correlation for the scores for each of the sections was used to test the convergent validity, using Spearman's Rank Order correlation test.

Due to the number of pairs for comparison the results are displayed in Table 2:

Correlations

			MCQ	PS	MCQCM
Spearman's rho	PS	Correlation Coefficient	.235		
		Sig. (2-tailed)	.129		
		N	43		
	MCQCM	Correlation Coefficient	.436(**)	.302(*)	
		Sig. (2-tailed)	.003	.049	
		N	43	43	
	SA	Correlation Coefficient	.447(**)	.442(**)	.544(**)
		Sig. (2-tailed)	.003	.003	.000
		N	43	43	43

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Table 2 Correlation table for the sections of the exam

The following observations can now be discussed. All of the levels of correlation are as defined by Pallant (2005) reference to (Cohen 1998))

Firstly, let us consider the correlation between the MCQCM and the other sections of the exam paper.

There is a reasonably strong correlation between the MCQCM and the SA section ($r=.544$, $n=43$, $p<.01$).

MCQCM also has a medium correlation with MCQ and PS ($r=.436$, $n=43$, $p<.01$ and $r=.302$, $n=43$, $p<.05$) respectively).

These statistics confirm that there is a convergence of validity for the MCQCM and all of the other sections of the exam. Additionally, these correlations gain strength when considering the Cronbach's Alpha reliability coefficient for the results, demonstrating the internal consistency of .692, (slightly below the recommended minimum of 7.0).

Further, it is interesting to see that the grades for the MCQ section demonstrate a medium correlation to SA ($r=.447$, $n=43$, $p<.01$) and a small correlation to PS ($r=.235$, $n=43$, $p<05$).

SA and PS has a large correlation ($r=.442$, $n=43$, $p<.01$).

Discussions and Conclusions

In conclusion, this study has identified a convergence of validity between MCQCM and all of the other sections of the exam paper, with the strongest correlation being between MCQCM and SA. This observation is very encouraging as the MCQCM was primarily designed as a formative assessment tool to support the learner along the learning path (Farrell& Leung 2002).

Interestingly, the traditional MCQ section of the paper has medium correlation with the SA but only has a small correlation to the PS section. Hence, whilst there is convergence of validity between MCQ and SA there is no significant convergence of validity between the MCQ section and the PS section. This means that a good performance in either section would not predict a good performance in the other.

As a result of these initial observations MCQCM appears to be a valid assessment option, producing grades that have equal reliability as the more traditional methods of assessment. However, MCQCM does not appear to offer any great advantage over the rest of the methods of summative assessment. The question then must be asked, why bother?

Previous investigative work in using MCQCM as a formative assessment tool (Farrell, Leung 2005) has proved that utilizing MCQCM can be highly beneficial to both the student and the instructor as its feedback is often reflective of their confidence in their knowledge of a particular subject material. This often influences the learning path of the individual to address the areas of concern, encouraging management of the learning by the student. (Farrell, Leung, 2005)

This study encourages the utilization of the MCQCM as a summative testing option in the future. It is proposed that the tool continue to be utilized as a formative assessment method for the duration of the semester and be included as part of the final exam, producing more data for analysis. In addition the authors intend on gauging the students' acceptance or rejection of MCQCM as a standard method for summative assessment.

References

Ashburn, Robert (1938). An experiment in essay-type question. *Journal of Experimental Education 7 (1): 1-3*

Assessment tools for Assessment, Evaluation and Curriculum Redesign workshop: *month 7 http://www.thirteen.org/edonline/concept2class/month7/index_sub2.html (Last accessed Aug 2003)*

Bacon, Donald R (2003): Assessing Learning Outcomes: A Comparison of Multiple-Choice and Short-Answer Questions in a Marketing Context: Journal of Marketing Education. Vol 24. No 22. Sage Publications

Black. P and William D (1998): Inside the Black Box: Raising Standards Through Classroom Assessment. Phi Delta Kappan October 1998· Volume 80· Number 2 P 139-149 http://www.pdkintl.org/kappan/kbla9810.htm

Black. P and William D (March 1998): Assessment and Classroom. Learning *Assessment in Education, Vol 5 March* P. 7-74

Bloom, Benjamin S (1956): Taxonomy of educational objectives, hand book 1: Cognitive domain. *New York: Longman Green.*

Bridgeman, Brent, and Charles Lewis (1994): The relationship of essay and multiple-choice scores with grades in collage courses. *Journal of educational measurement 31 (1): 37-50*

Cohen, J. Statistical Power Analysis for the Behavioural Sciences. Hillsdale, NJ: Erlbaum. (1988)

Farrell, G and Lung, Y:Designing an Online Self-Assessment Tool Utilizing Confidence Measurement. Conference Proceedings *IFIP 8.4 WG (2002) P.525-537*

Farrell, G and Lung, Y:Improving the Design an Online Self-Assessment Tool Utilizing Confidence Measurement. Conference Proceedings *Web-based Learning (2002) P.149-159*

Lambert W.T. Schuwirth and C.P.M. van der Vlueten (1996): Quality Control: Assessment and Examinations: *http://www.oeghd.or.at/zeitschrift/1996h1-2/06_art.html (Last accessed Aug 2003)*

Pallant. J. SPSS Survival Manual, 2nd Edition, Allen & Unwin (2005)

Principle and Standards for School Mathematics (2000): *National Council of Teachers of Mathematics - Standards 2000 Project Chpt 2*

Wilson, R. B. and Case, S. M.: Extended Matching Questions: An Alternative to Multiple-choice or Free-response Questions: *Journal of Veterinary Medical Education. Volume 20:3. http://www.utpjournals.com/jour.ihtml?lp=jvme/jvme203/ExtendedMatchingQuestions.html (Last accessed Aug 2003)*