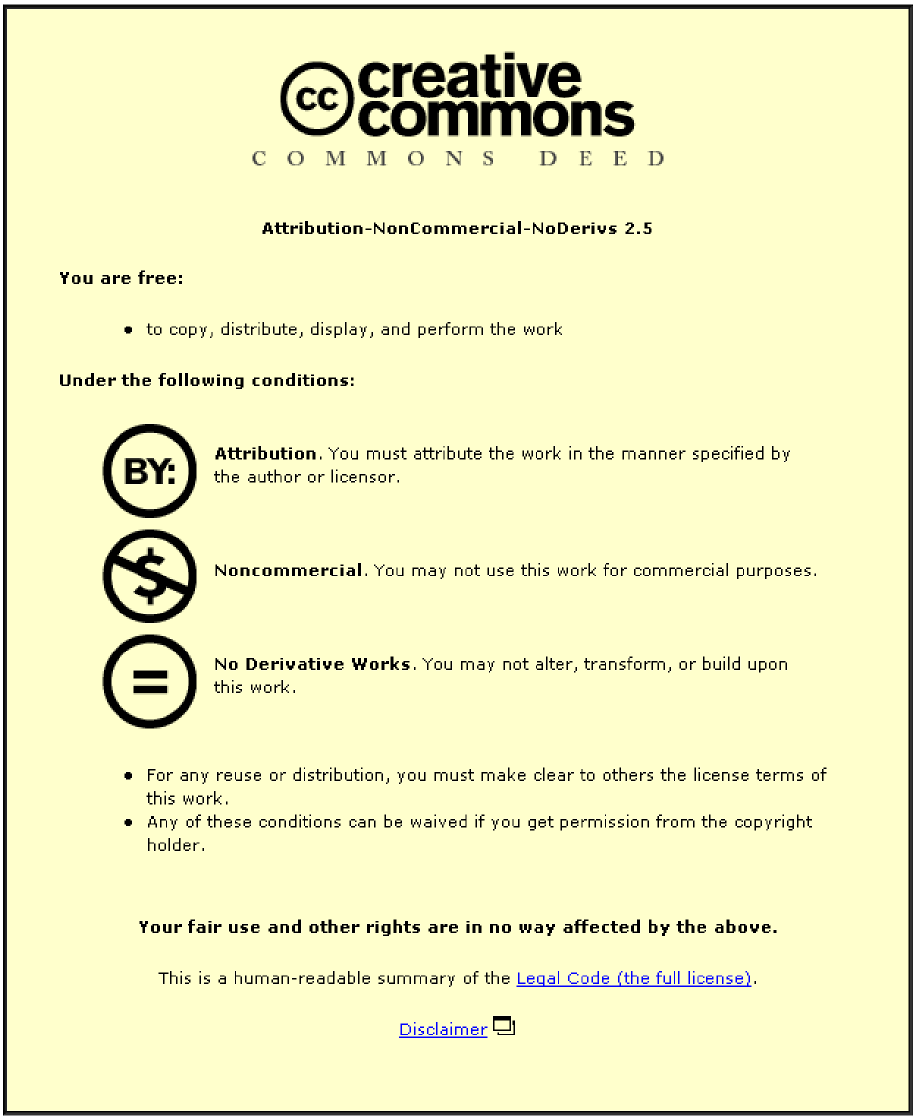


CONVERGENCE OF VALIDITY FOR THE RESULTS OF A SUMMATIVE ASSESSMENT WITH CONFIDENCE MEASUREMENT AND TRADITIONAL ASSESSMENT

Graham Farrell and Ying Leung

Convergence of Validity for the Results of a Summative Assessment with Confidence Measurement and Traditional Assessment

Graham Farrell: Usability and Innovation Group Swinburne
University of Technology
Australia

Ying Leung: Hong Kong Institute of Vocational Education
Hong Kong

Abstract:

This research examines the use in IT education of an innovative online assessment tool that incorporates confidence measurement. The tool is based on the traditional Multiple Choice Question (MCQ) format with an additional component that permits the user to register their confidence for each answer. The tool is referred to as the Multiple Choice Question with Confidence Measurement (MCQCM). A cohort of 52 Data Communication students utilized the MCQCM as their primary revision tool throughout a semester and then for a test in class. As part of the review they were then asked to give feedback on using the MCQCM as a formal summative assessment tool. The test was graded using the traditional method as well as by calculating a further grade from the student's registered confidence. The results demonstrated a good correlation and convergence of validity between the dual marks supporting the use of the system as a summative assessment option. It was also observed that the majority of high achievers scored less for the MCQCM grade than for the traditional MCQ. In addition, the students' perception of the MCQCM as a graded assessment task was quite positive. These results are very encouraging and will be further investigated as part of this ongoing research.

Introduction and Literature Review:

Instructors are required to assess their students in a number of ways to ascertain their level of knowledge. An important aspect of assessment is to supply timely feedback to both the students and the instructor enabling revaluation of the students learning path. The term "Assessment" often refers to the group of activities that are undertaken by both teachers and students, providing both grades and feedback to modify teaching (Black and William 1998). Instructors will choose assessment tasks to best suit the needs of the student and themselves, often utilizing various tools for the purpose. (See Assessment tools for assessment, evaluation and curriculum redesign for examples 2003) It is accepted wisdom that assessment should be an integral part of the learning activities rather than an interruption. (Principles and Standards for School Mathematics 2003)

There are many assessment options available to educators and there is often a dilemma in choosing the appropriate mix for feedback and evaluation purposes. Schuwirth and Van Der Vleuten (1996) state “a well designed assessment program will use different types of questions appropriate for the content being assessed”. There is a need to allow a student to demonstrate their knowledge in various ways, as individuals show preference to particular assessment types. Some of the options presently available to the instructors include multiple choice questions (MCQ), short answer questions (SA), long essay questions, case study reports, presentations and other equally effective and proven choices. In this particular research the author’s have limited the testing option to the MCQ style.

Gardner-Medwin and Gahan (2003) state that “to measure knowledge we must measure a persons degree of belief”. They postulate that a student, when registering their degree of belief for a true statement demonstrates either

- Knowledge
- Uncertainty
- Ignorance
- Misconception
- Delusion

Their introduction of confidence based assessment was to assist their students to identify where they lie on the above scale. They, like Davidoff (1995), consider misconception and miss-calibration of knowledge to be a serious impediment in the building of higher levels of knowledge. Uncertain correct answers, or lucky guesses, do not equate to knowledge. Further more; registration of high levels of confidence in wrong answers definitely deserves special attention (Gardner-Medwin and Gahan, 2003).

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 1996). 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 have enjoyed resurgence of late due to the advancement of technology, offering online tests with immediate scoring and feedback capability at the convenience of the student. MCQs are also traditionally used for summative assessment, grading the students. They are extensively used in exams offering a quick assessment of the students' knowledge of a broad area of the content (Wilson and Case 2003) while being timesaving for the examiner when grading. 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).

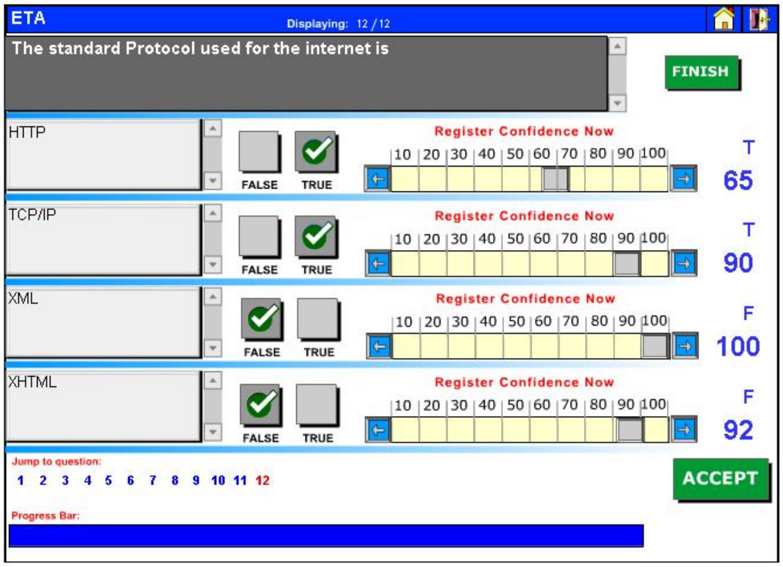
The grading of MCQs is "Objective" avoiding the obvious lack of reliability of essay tests (Ashburn 1938). There are those who criticize the extensive use of MCQs stating that they often rely heavily on recognition (Schuwirth and Van Der Vleuten 1996) and knowledge of isolated facts (Wilson and Case 2003). There is also the great concern that students will learn in the mode of testing, consequently learn at the lower level of recall when faced with an MCQ test. (Schuwirth and Van Der Vleuten 1996). Well crafted MCQs have the ability to access the higher levels of Blooms Taxonomy (1956) but the greater the level the more difficult to construct. Bloom and his colleagues defined six levels of learning outcomes being: Knowledge, Comprehension, Application, Analysis, Synthesis and Evaluation (Bloom,1956).

This paper initially considers the validity of the innovative MCQCM as a summative testing method. It utilizes a comparison with another traditional MCQ testing method where validity is "whether the question actually tests what it is purported to test".(Schuwirth and Van Der Vleuten 1996) Validation is assessed by comparing the correlations between two methods of testing that are supposed to measure the same construct (Bacon 2003). 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 1996).

The Design of the MCQCM Tool:

The MCQCM tool was designed by the author and used extensively for a number of units for revision. MCQCM is a Web based assessment tool available to the students via a URL for the duration of the course, 24/7. The MCQCM 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 and Leung 2002). The MCQCM format is similar to the MCQ display where each question has a stem followed by four options (Frary, 1995). Once the student commits to an answer ("level") they are required to register their confidence in that choice ("strength").(Bandura, 1983) . In previous studies (Farrell and Leung 2002) it was demonstrated that the MCQCM is 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 of concern, 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 a summative assessment tool to the standard accepted MCQ format.

When using the MCQCM the student is presented with each option of the question requiring a commitment to either correct or incorrect. They then are required to register their confidence as a %, where 100% defines complete certainty in the choice and a low % representing extreme doubt. Fig 1a demonstrates the tool in action followed by Fig 1b demonstrating the consequential score and feedback from the particular exercise.



ETA

Displaying: 12 / 12

The standard Protocol used for the internet is

HTTP — □ FALSE ☑ TRUE — Register Confidence Now — T 65

TCP/IP — □ FALSE ☑ TRUE — Register Confidence Now — T 90

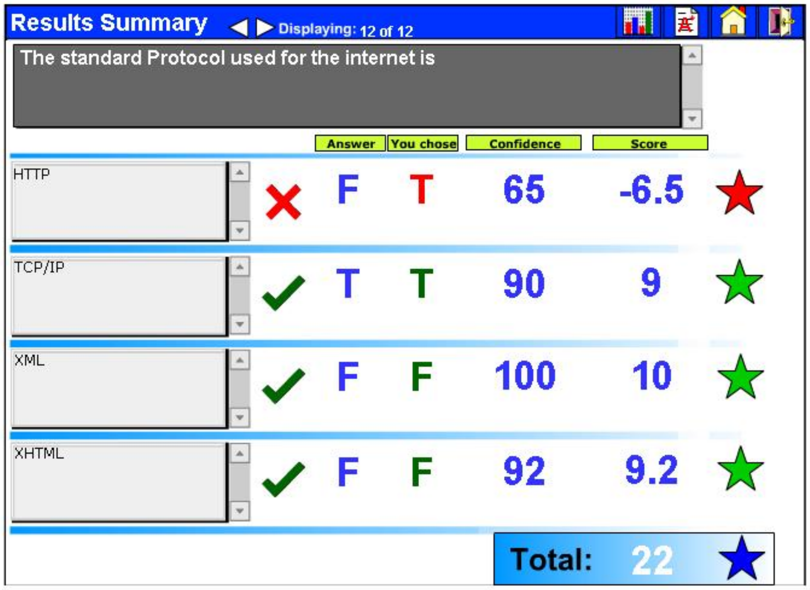
XML — ☑ FALSE □ TRUE — Register Confidence Now — F 100

XHTML — ☑ FALSE □ TRUE — Register Confidence Now — F 92

Jump to question: 1 2 3 4 5 6 7 8 9 10 11 12

Progress Bar:

Fig 1a: Screen shot demonstrating the tool in use. In this case the student demonstrates that even though they are quite confident that the answer is option 3, TCP/IP, they think that it might also be the first option HTTP. In this choice they are not as confident.



The standard Protocol used for the internet is

		Answer	You chose	Confidence	Score	
HTTP	✗	F	T	65	-6.5	★
TCP/IP	✓	T	T	90	9	★
XML	✓	F	F	100	10	★
XHTML	✓	F	F	92	9.2	★

Total: 22 ★

Fig 1 b is the consequential score from the above example. The level of confidence for each option is included in the calculated score as explained below. Notice that they are all converted from the percentage recorded to a score of out of 10 for each option, giving a range from -40 to 40 as a possible final score for each question.

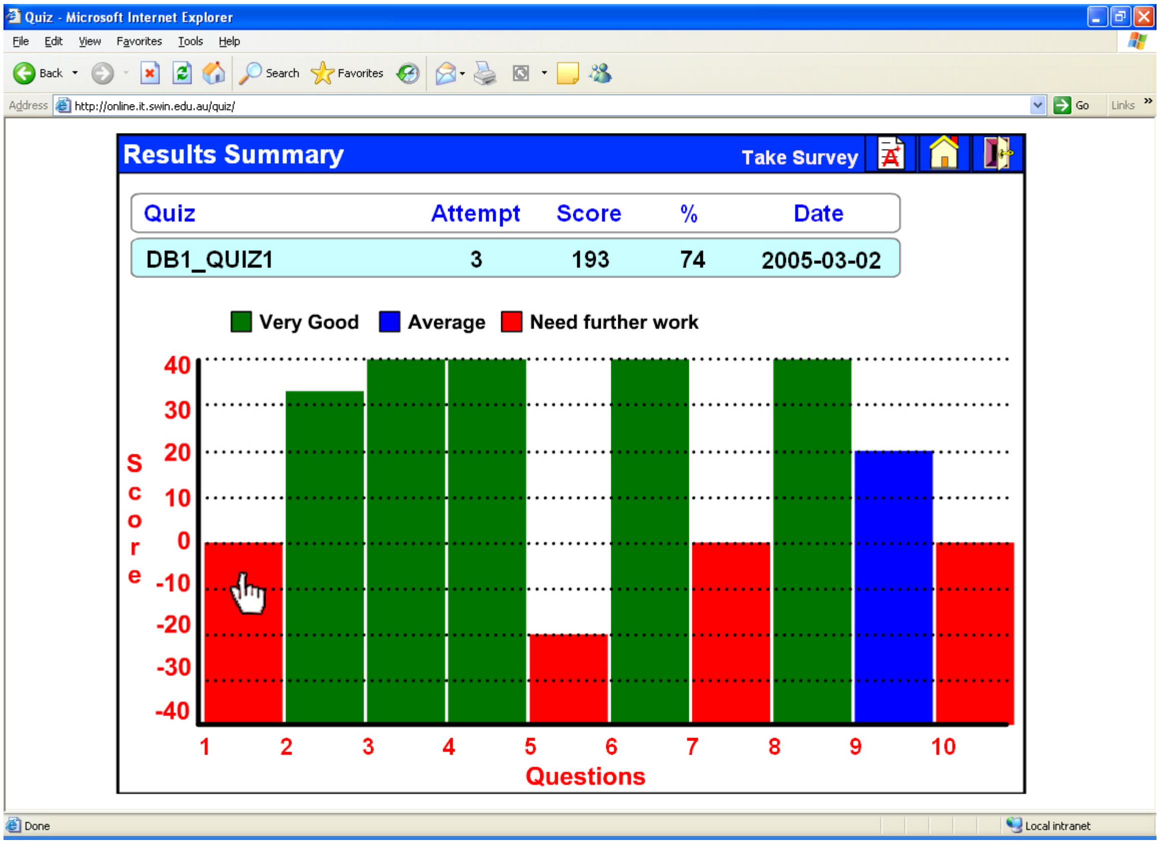
Scoring:

The scoring mechanism for MCQCM is a simple linear one. If the student registers a high level of confidence for a correct answer this produces a high positive score. (Eg. 100% gives 10 marks) The grade decreases 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). Hence, demonstrating high confidence in an incorrect choice is heavily penalized, as recognized by Davidoff (1995) and Gardner-Medwin and Gahan (2003), which is the justification of giving such a high weighting to the measurement of confidence in the score.

The total score for each question is then calculated by the tally of all of the option answers giving a grade from -40 to 40 for each question.

On completion of the self test the students would be presented with a graphical display which used effective colour codes to highlight the areas where they have demonstrated good, reasonable or poor knowledge. See Fig 2.


Quiz - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://online.it.swin.edu.au/quiz/
Go Links
Results Summary
Take Survey
🟩 Very Good 🟦 Average 🟥 Need further work
40
30
20
10
0
-10
-20
-30
-40
Score
1
2
3
4
5
6
7
8
9
10
Questions
Done
Local intranet

Quiz	Attempt	Score	%	Date
DB1_QUIZ1	3	193	74	2005-03-02

Fig 2: The Feedback screen showing the students score for each question, colour coded to represent the level of concern. The student can link directly to the question, correct answers and their response for assistance (See Fib 1b).

Importantly the students utilize the system as a formative assessment option during the semester and are familiar with the functionality and scoring mechanism.

Method

A cohort consisting of 52 Data Communication students, were required to sit a test during the semester contributing to their final grade. The test consisted of 10 Multiple Choice Questions (MCQ) testing the students on the fundamentals of Network design. The author of the test was mindful of Bloom's taxonomy (1956) of educational objectives when constructing the questions to facilitate the assessment of various levels, in particular testing at the application level.

The students sat the test under supervision during the tutorials.

They were instructed that the test would be graded in two ways. Firstly using the MCQCM grading where the registered confidence for each response would be included in the calculation of the mark and secondly they would be graded using the traditional method. They were informed that the final allocated grade would be the greater of the two. This was done to alleviate any possible additional stress experienced by the students using a new grading system.

In addition the students utilized the system consisting of questions similar to the ones being asked on the actual test for the duration of the semester as a means of revision,

On completion of the test the students were given the opportunity during the follow up tutorial to give feedback as part of the unit review process, facilitated via questions and written format.

Results:

The results section has been divided into two areas.

The first being the actual grades analysis. Here we compare the test results for both methods and investigate the convergence of validity.

The analysis is as follows:

Analysis of Grade Averages:

	Average Grade	Standard Deviation
MCQCM	67.60%	24.80%
MCQ	60.58%	22.73%
MCQCM - MCQ	7.02%	20.23%

Table 1: The Average and Standard Deviation for both the marking schemes applied as well as the average difference between them.

On analysis of the data in Table 1 it is noted that the average grades for the two marking schemes are reasonably close, as too are most of the standard deviations. It is observed that the MCQCM has the greater Average Grade and Standard Deviation. Instructors would be quite pleased with these outcomes at this stage.

Analysis of Individual Grades:

As the data violates the Kolmogorov-Smirnov test for normality the Wilcoxon Signed Rank Test for non parametric data with repeated measures is used within subjects to compare the two samples. The results are shown in Table 2 below.

	MCQ - MCQCM
Z	-.646(a)
Asymp. Sig. (2-tailed)	.518

Table 2: Results of Wilcoxon Signed Rank Test

As the results of Table 2 demonstrates the associated significance level is more than .05 (.518) indicating the two sets of results are not statistically different.

However, when considering the different individual results for the test it was observed that 81% of the MCQCM scores are different than that of the corresponding MCQ scores.

In particular it is observed that there is close to an even distribution of those who benefited from the MCQCM marking scheme (42%) and those who benefited from the MCQ marking scheme (39%), while the remaining 19% achieved the same.

Further investigation identifies that of those who scored well in the traditional MCQ marking scheme (>65) only 32% scored higher for the MCQCM marking scheme with 52% scoring worse and 16% scoring the same.

In addition, of the 10 students who achieved 90% or more for the MCQ grading scheme 6 of them scored less, 0 scored higher and 4 scored the same for the MCQCM. This is an important observation as the higher achieving students do not necessarily score better using MCQCM, suggesting that the MCQCM might be a more discerning indicator of knowledge for those higher achieving students. However, it could also represent the level of confidence they are prepared to register.

Convergence of Validity:

To ascertain the convergence validity we investigated the correlation between the MCQCM and the MCQ scores. The observed result (.629 (p>.01) confirmed that there is a correlation between the grade for the MCQCM and the grade for the MCQ. This correlation gains strength when considering the .722 Cronbach's Alpha Reliability Coefficient for this assessment, demonstrating the internal consistency, slightly above the minimum of 7.0. Whilst this correlation is not strong, it is satisfactory and similar to results experienced by the authors in previous studies when comparing the scores of traditional assessment items.

Analysis of Feedback from Students:

The second component of the results section of this paper analyzes the students' feedback on using the MCQCM as an assessment tool.

70% of the students felt they were given the opportunity to improve their grade by using the slide bar to register their confidence. In addition they appreciated the chance to demonstrate partial knowledge whilst having a “Safe Zone” if unsure. The other 30% did not consider the gain worth effort required, as the increase in the cognitive load might have a detrimental effect on the outcome

Despite the figures above, 60% used the slide bar to register their confidence, while the other 40% preferred to set it at the 100% level, effectively using it to state outright whether their answer was true or false. The drop in numbers of usage is explained by the response by some, that even though they used the slide bar to their advantage during the practice sessions they did not want to use it under test conditions.

During further discussion it was noted that 72% of the cohort have no hesitation in registering 100% for an answer if they were convinced it was either right or wrong, while the remaining 28% would be hesitant to outright declare such confidence. This is an interesting observation as Gardner-Medwin and Gahan (2003) discuss this as being psychometrics deserving research but of less significance in this context of education.

A large number of the cohort, 75%, stated that the system forces you to think more carefully about your options as the element of guessing becomes more critical. The remaining 26% felt that either “you know it or you don’t” and registering a confidence is not worth considering.

Discussions and Conclusions:

In conclusion, this study has identified a convergence of validity between the two types of grading schemes being investigated, Multiple Choice Questions with Confidence Measurement (MCQCM) and the traditional Multiple Choice Question (MCQ) format, for this subject of Information Technology. Hence the MCQCM appears to be an acceptable summative assessment option to be included in the suite of assessment tools available to the instructor. Previous work by the author has demonstrated that the MCQCM delivers a rich feedback and guidance to the students when used as a formative assessment tool. In addition they have well documented the perceived advantages of utilizing it in preparation for their exams from the students’ point of view.

It is pleasing to observe that even though the grades do correlate there appears to be encouraging results for the MCQCM to possibly offer a more critically discerning grading system. Although the evidence is not overwhelming, it appears that the higher achievers in the group do not score as well for the MCQCM. This could either be that MCQCM forces the students to “show their hand” giving a true indication of their knowledge or that it is really acting as a statement of their own personal confidence in their choices. In light of this the author’s would recommend that ongoing application of the system to increase the data gathered, investigating to see if this observed results occur again and if so interview the participants to ascertain the reason/s.

The survey revealed further interesting observations which are worth noting. There was an overall support from the majority of the students, which was pleasing to the authors and instructor. The majority of the students acknowledged the benefits of the system stating that they appreciated the opportunity to demonstrate partial knowledge and optimize their grade by lessening the impact of an incorrect choice and increasing the grade for a correct one. The observation that a number of the cohort (28% in this case) do not have the confidence to register 100% for any answer should be always considered in the analysis of future observations, as that particular group will never be able to maximize their grade utilizing this system. An important observation is the decrease in the use of the slide bar during the summative assessment than when used for the formative assessment, as some of the students' consider it to be a distraction and perceived to be not beneficial enough.

The results from this study are encouraging and the author intends on utilizing MCQCM as part of the assessment in the future. The overall positive response from the students towards MCQCM as both a formative and summative assessment tool increases the enthusiasm of the designers and instructors who are keen to pursue its usage in the classroom. A singular advantage of using the MCQCM as a summative assessment tool is that it requires the students to utilize it during the semester for revision. This introduces a strategy which motivates the students' to actively revise the course material in preparation for the test, which in most cases greatly increases the likelihood of their success in the subject.

References

[1].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*

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

[3].Assessment tools for Assessment, Evaluation and Curriculum Redesign workshop: *month 7 http://www.thirteen.org/edonline/concept2class/month7/index_sub2.html (Last accessed Feb 2008)*

[4].Lambert W.T. Schuwirth and C.P.M. van der Vlueten (1996): Quality Control: Assessment and Examinations:

[5].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*

[6].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.*

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

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

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

[10].Frary R (1995): More multiple-choice item writing do's and don'ts. Practical Assessment, assessment and evaluation ERIC Clearinghouse on Assessment and Evaluation. Vol 4 (11)

[11]. Bandura, A (1983): Self-Evaluation and Self-Efficacy Mechanisms Governing the Motivational Effects of Goal Systems. Journal of Personality and Social Psychology Vol 45,No 5, P 1017-1028 (1983)

[12].Davidoff F (Feb 1995): Confidence Testing- How to Answer a Meta-Question. American College of Physicians Observer

[13]. Gardner-Medwin and Gahan (2003): Formative and Summative Confidence-Based Assessment *Conference Proceedings 7th Computer Aided Assessment (CAA) (2003)*