



**Queensland University of Technology**  
Brisbane Australia

This is the author's version of a work that was submitted/accepted for publication in the following source:

Mullen, Kathleen & Schultz, Madeleine (2012) Short answer versus multiple choice examination questions for first year chemistry. *International Journal of Innovation in Science and Mathematics Education*, 20(3), pp. 1-18.

This file was downloaded from: <http://eprints.qut.edu.au/54356/>

**© Copyright 2012 Institute for Innovation in Science and Mathematics Education**

**Notice:** *Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source:*

# Short Answer versus Multiple Choice Examination Questions for First Year Chemistry

**Kathleen Mullen and Madeleine Schultz\***

School of Chemistry, Physics and Mechanical Engineering  
Science and Engineering Faculty  
Queensland University of Technology, Australia  
madeleine.schultz@qut.edu.au

## Abstract

Multiple choice (MC) examinations are frequently used for the summative assessment of large classes because of their ease of marking and their perceived objectivity. However, traditional MC formats usually lead to a surface approach to learning, and do not allow students to demonstrate the depth of their knowledge or understanding. For these reasons, we have trialled the incorporation of short answer (SA) questions into the final examination of two first year chemistry units, alongside MC questions. Students' overall marks were expected to improve, because they were able to obtain partial marks for the SA questions. Although large differences in some individual students' performance in the two sections of their examinations were observed, most students received a similar percentage mark for their MC as for their SA sections and the overall mean scores were unchanged. In-depth analysis of all responses to a specific question, which was used previously as a MC question and in a subsequent semester in SA format, indicates that the SA format can have weaknesses due to marking inconsistencies that are absent for MC questions. However, inclusion of SA questions improved student scores on the MC section in one examination, indicating that their inclusion may lead to different study habits and deeper learning. We conclude that questions asked in SA format must be carefully chosen in order to optimise the use of marking resources, both financial and human, and questions asked in MC format should be very carefully checked by people trained in writing MC questions. These results, in conjunction with an analysis of the different examination formats used in first year chemistry units, have shaped a recommendation on how to reliably and cost-effectively assess first year chemistry, while encouraging higher order learning outcomes.

## Introduction

Given the growing resource constraints in the tertiary sector, many Australian universities, particularly in quantitative disciplines, use multiple choice (MC) examinations as the primary method of assessment for large (typically first year) classes of students. The advantages and limitations of MC questions in comparison to other forms of assessment, such as essay or short answer (SA) questions, have been thoroughly examined, both generally (Nicol, 2007; Struyven, Dochy, & Janssens, 2005) and within disciplines ranging from engineering (Le & Tam, 2007), management (Parmenter, 2009), medical education (Schuwirth, van der Vleuten, & Donkers, 1996), education (Scouller, 1998) to information technology (Woodford & Bancroft, 2004). There is also a significant body of literature examining the use of MC questions specifically for chemistry (Hartman & Lin, 2011; Ruder & Straumanis, 2009; Schultz, 2011). The advantages of MC examinations all derive from the automated marking process, and are:

- fast marking of large student cohorts, leading to more timely feedback for the students of their marks;
- free marking, reducing the cost of assessment;
- minimal errors in marking and data entry to student results spreadsheets;
- the ready availability of detailed statistics on student results and therefore question difficulty (Holme, 2003).

The literature cited above also lists important limitations of MC examinations. Such examinations are not considered ideal for student learning because they:

- encourage a surface approach to learning (Parmenter, 2009; Scouller, 1998; Struyven, et al., 2005);
- rarely test literacy (including chemical literacy) skills, such as writing equations and drawing chemical structures (Ruder & Straumanis, 2009);
- may be unintentionally made more difficult by examiners, by the introduction of additional algorithmic steps (Hartman & Lin, 2011);
- may introduce bias or distortion to scores if the questions and response choices are poorly written (Cheung & Bucat, 2002);
- take a long time to construct if they are to test more than simple recall or algorithmic applications of formulae (Nicol, 2007).

In addition, the use of MC examinations is problematic because they:

- do not allow students to obtain part marks by demonstrating partial understanding (Schultz, 2011);
- allow guessing or cueing to obtain a correct answer rather than using knowledge or reasoning, so assessment may not be authentic (Schuwirth, et al., 1996; Woodford & Bancroft, 2004);
- make it difficult for outstanding students to excel, because a mistake in entering the answer choice leads to zero marks (Schultz, 2011).

The difficulty in writing valid, authentic MC examinations has been dealt with in part, in the United States, by the development of standardised examinations for chemistry. For over 70 years, the American Chemical Society's Examinations Institute (ACS-EI) has, every two to four years, formed committees of academics to write examination papers in each subdiscipline. The questions are thoroughly tested and validated before release of the paper. These papers can be purchased by institutions and allow benchmarking (Holme, 2003). On a rolling 4-year basis around 2000 tertiary institutions purchase one or more university level ACS-EI examination paper. Based upon examination sales, approximately 30,000 U.S. first year and 15,000 second year university students take ACS-EI examinations each year – about 10% of all U.S. university students. Such extensive use allows detailed statistics. Use of these examinations also permits an analysis of the effectiveness of novel teaching strategies, because different class rooms can be directly compared. However, most universities are not using these examinations due to cost or the desire to choose their own content.

One alternative assessment option, which combines some of the advantages of MC (through automated marking) without some of the disadvantages (by requiring numerical or text input) involves the use of online tools (Schultz, 2011). Modern content management systems are able to recognise correct answers in text and numerical format, allowing a broader range of questions than MC. However, such tools cannot be used under examination conditions for internal students at most universities because of a lack of computer resources in the examination rooms.

Pinckard et al. have compared a MC examination with a SA examination, and with a mixed format examination that combines MC with SA questions (Pinckard, McMahan, Prihoda, Littlefield, & Jones, 2009). They found that although students performed worse in SA questions compared with their raw MC score, inclusion of SA questions in a mixed format examination improved performance on the MC section significantly. This is likely because of the effect that SA questions have on study style (Parmenter, 2009). That is, when students know that there will be some SA questions, they approach their study in a manner that leads to deeper learning, because they know that they will be required to answer questions without cueing or guessing. This approach to studying improves their performance in some MC questions, due to the resulting deeper

understanding of the material and higher order learning outcomes in the SOLO taxonomy (Biggs & Collis, 1982). Thus, a mixed format examination allows marking resources to be used selectively, while achieving the desirable student learning outcome of deep learning through better study approaches.

The authors compared the results of individual students in SA versus MC examinations, and found that the correction for guessing (losing fractional marks for incorrect answers in MC) (Diamond & Evans, 1973) led to a better correlation between these results (Prihoda, Pinckard, McMahan, & Jones, 2006). Their conclusion was that the correction for guessing should be applied to MC examinations in order to make the results more valid. However, as the authors discuss, student exam-taking behaviour is different if they are informed that a correction for guessing will be applied. Thus, applying it retrospectively (as in this study) may not yield a true value of their expected score. Their conclusion is based on the premise that the SA results are valid; they write

*..the short-answer format examinations should provide a better measure of a student's ability to perform in clinical situations in which patients present without a set of possible choices for the diagnosis. Our use of validity refers to performance without guessing, that is, performance without "cuing."*

It can be argued that having some answers that can be chosen in MC by eliminating all other options, rather than working out the correct option, is a valid reasoning technique and also relevant in the real world. If distractors are well-chosen, this sort of question can test deep knowledge. Nonetheless, chemical literacy cannot be tested with such questions, and the students inevitably receive cues from the answer options. Although performance without cueing is important, SA questions can also be problematic, as described below.

In an effort to promote deep approaches to learning (Biggs & Tang, 2007) and achieve higher order learning outcomes (Biggs & Collis, 1982), at one Australian ATN institution, several unit coordinators replaced some MC questions with SA questions in first semester chemistry examinations in 2011. Here, we have analysed the students' responses and attempted to probe their learning outcomes.

## **Context and Approach**

The institution in this study has a fairly typical program in the first year (Mitchell Crow & Schultz, 2012). Four first year chemistry units are currently offered, as follows:

SCB111: Chemistry 1 - foundation chemistry including physical chemistry

SCB121: Chemistry 2 - bonding and organic chemistry

SCB131: Experimental Chemistry (semester 2 only)

SCB113: Chemistry for Health and Medical Science (semester 1 only)

SCB113 is a service unit taken by students enrolled in degrees including optometry, pharmacy and nutrition. These programs of study have much higher entry requirements than those for the degrees for which SCB111 and SCB121 are required, including the Bachelor of Applied Science, in which the majority of students are enrolled.

Chemistry major students who intend to progress into second and third year chemistry units are required to take SCB111, SCB121, and SCB131; SCB131 is also taken by many of the health and medical science students, and therefore involves a mixed cohort of students. These units were developed as part of a review of the Bachelor of Applied Science and were run for the first time in 2008. MC examinations have been used for the majority of summative assessment for first year chemistry at this institution since 2003, and were used for mid-semester and final examinations in all of the units above from 2008 - 2010, inclusive.

An analysis of the students' results in these units (shown in Table 1) revealed that the mean marks in SCB121 and SCB131 were consistently significantly lower than those in SCB111<sup>†</sup>. It was hypothesised that learning, and therefore student results would be improved if SA questions were included within the final examinations of SCB121 and SCB131. Such questions enable students to demonstrate partial understanding of concepts and therefore earn part marks. SA questions can also assess chemical literacy, such as the ability to draw structures of organic compounds and write balanced equations, which cannot be assessed by MC examinations. More importantly, the literature indicates that students approach the process of studying differently when they know there are SA questions (Scouller, 1998; Struyven et al., 2005). Thus, we hoped to encourage deeper learning by including this type of question. Inclusion of such questions was expected to increase the student average mark because of the potential to earn part marks.

For these reasons, in 2011, SA questions were included in the final examinations for SCB121 (both semesters, called SCB121\_1 and SCB121\_2) and SCB131 (which is offered in second semester only). Note that the first semester class size for SCB121 (SCB121\_1) is much smaller than the second semester class size (SCB121\_2), because the unit is taken in second semester in all standard programs of study (and similarly, the second semester cohort in SCB111 is much smaller); the exact numbers are included in Table 1. Both of the second semester units in the trial, SCB121\_2 and SCB131, have over 200 students.

**Table 1.** Mean examination results in first year chemistry units 2009 – 2011\* (SD: standard deviation)

| <b>Year, Semester</b> | <b>SCB111</b>                                    | <b>SCB113</b>                                    | <b>SCB 131</b>   | <b>SCB121</b>  |
|-----------------------|--|--|--|--|
| 2009, Sem 1           | 69 % (SD 16)<br>range: 23 – 98<br><i>n</i> = 428 | 67 % (SD 16)<br>range: 28 – 98<br><i>n</i> = 237 | -  | 50 % (SD 15)<br>range: 23 – 96<br><i>n</i> = 84  |
| 2009, Sem 2           | 66 % (SD 17)<br>range: 30 – 95<br><i>n</i> = 72  | -  | 60 % (SD 17)<br>range: 23 – 97<br><i>n</i> = 250                               | 53 % (SD 16)<br>range: 22 – 94<br><i>n</i> = 331   |
| 2010, Sem 1           | 68 % (SD 15)<br>range: 23 – 98<br><i>n</i> = 440 | 63 % (SD 16)<br>range: 21 – 96<br><i>n</i> = 237 | -  | 55 % (SD 17)<br>range: 25 – 91<br><i>n</i> = 92  |
| 2010, Sem 2           | 68 % (SD 16)<br>range: 30 – 92<br><i>n</i> = 43  | -  | 55 % (SD 16)<br>range: 20 – 97<br><i>n</i> = 248                               | 51 % (SD 16)<br>range: 19 – 100<br><i>n</i> = 324  |
| 2011, Sem 1           | 67 % (SD 16)<br>range: 20 – 97<br><i>n</i> = 336 | 62 % (SD 17)<br>range: 20 – 98<br><i>n</i> = 390 | -  | 51 % (SD 16)<br>range: 16 - 83<br><b>MC</b><br>51 % (SD 12)<br>range: 28 - 80<br><b>SA</b><br>51 % (SD 26)<br>range: 5 - 97<br><i>n</i> = 50 |
| 2011, Sem 2           | 64 % (SD 16)<br>range: 25 – 90<br><i>n</i> = 74  | -  | 60 % (SD 16)<br>range: 21 - 98<br><b>MC</b><br>64 % (SD 17)<br>range: 22 - 100 | 49 % (SD 19)<br>range: 12 - 97<br><b>MC</b><br>50 % (SD 17)<br>range: 10 - 97  |

<sup>†</sup> Unpaired two tailed t tests were used to test for significance throughout this work.

|  |  |  |   |  |
|--|--|--|---|--|
|  |  |  | <b>SA</b><br>59 % (SD 18)<br>range: 14 - 97<br><i>n</i> = 269 | <b>SA</b><br>48 % (SD 23)<br>range: 3 - 98<br><i>n</i> = 200 |
|--|--|--|---|--|

\*note: only students who attempted the final examination were included in these mean values.

In order to evaluate any differences due to the use of SA questions, the examination performance in all four first year chemistry offerings at this university was examined from 2009-2011. This indicated whether any differences in examination performance were due to differences in the cohort of students, or differences due to altering the assessment style. For the two units SCB111 and SCB113, which did not trial the use of SA questions, the same examination paper was used from 2008 - 2011. Thus, these units provide a baseline from which student performance in the remaining units can be compared. Table 2 contains the formats of the final examinations for the four units in the period 2008 - 2011.

**Table 2:** Format of final examinations in first year chemistry units from 2008 - 2011

| unit | year, semester | number of MC questions | % of examination for MC | number of SA questions | % of examination for SA | % of overall mark that examination is worth |
|------|----------------|------------------------|-------------------------|------------------------|-------------------------|---|
| 111  | ongoing        | 60                     | 100                     | N/A                    | N/A                     | 55  |
| 113  | ongoing        | 80                     | 100                     | N/A                    | N/A                     | 50  |
| 121  | up to 2010     | 90                     | 100                     | N/A                    | N/A                     | 55  |
| 121  | 2011, S1       | 50                     | 62.5                    | 10                     | 37.5                    | 55  |
| 121  | 2011, S2*      | 30                     | 60                      | 10                     | 40                      | 45  |
| 131  | up to 2010     | 30                     | 100                     | N/A                    | N/A                     | 30  |
| 131  | 2011*          | 15                     | 25                      | 15                     | 75                      | 30  |

\*in these semesters, mid-semester examination papers also included SA format questions.

The remainder of the assessment in each unit consists of practical/laboratory reports, problem based assignments and progress examinations.

It should be noted that the unit coordinators who included SA questions did so independently, and therefore the point value of a SA question relative to a MC question varies between the units. In SCB121\_1 and in SCB131, a SA question was worth three times the value of a MC question. In SCB121\_2, a SA question was worth double the value of a MC question. In addition, the SCB121\_1 examination had a three hour time limit, whereas both SCB131 and SCB121\_2 adopted a two hour format, because in these two cases, the examination accounted for less than 50% of the final percentage of the overall assessment. There is no current institutional policy relating the length of examinations to their percentage worth of total assessment.

## Results and Discussion

### Final Examination Performance

A detailed analysis of the examination performance of chemistry students enrolled in each of the four chemistry subjects on offer from 2009-2011 was performed and is tabulated in Table 1. Only students who sat the final examination were taken into consideration when analysing the examination results. For SCB111, the final examination performance was relatively constant over this period, with mean scores between 64 - 69 % over the six semesters. The mean for SCB113 in 2009 fell in the same range, while the mean scores in SCB113 in 2010 and 2011 were lower than those for SCB111 (2010 semester 1:  $t(675) = 4.04, p < 0.0001$ ; 2011 semester 1:  $t(724) = 4.06, p < 0.0001$ ).

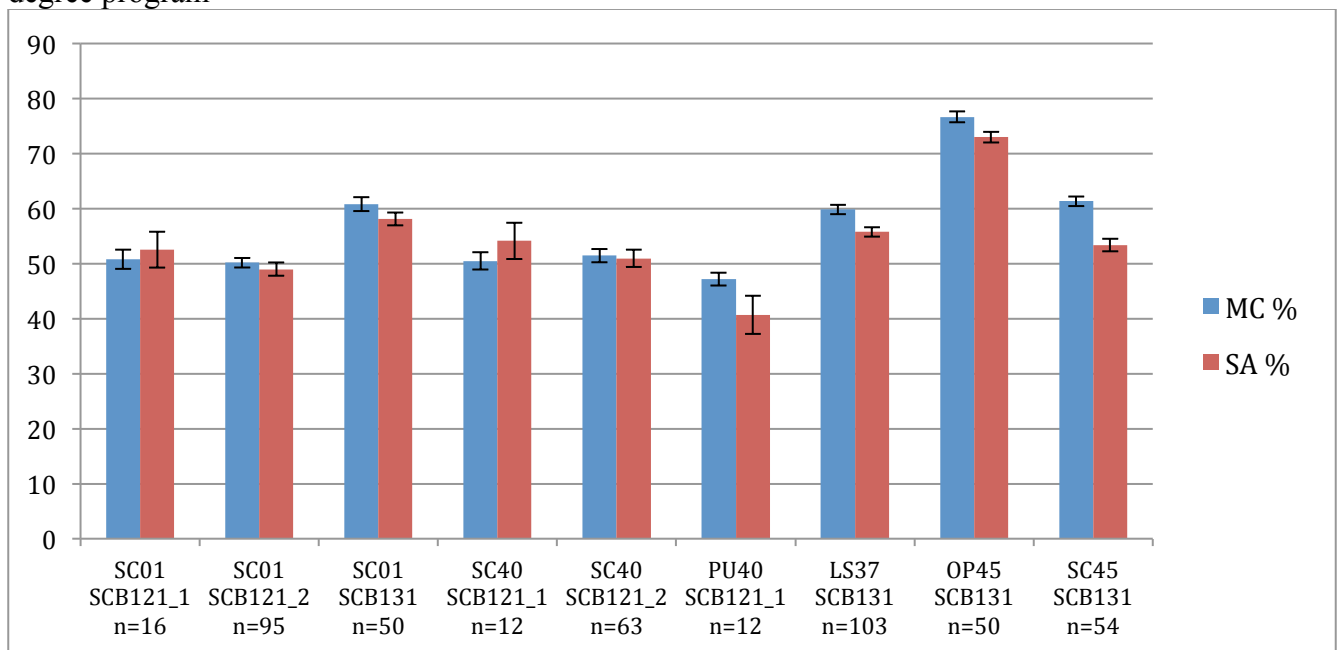
Looking at final examination performance in SCB131 and SCB121, the mean scores were significantly lower than those in SCB111 and SCB113 in every semester, despite these units including overlapping cohorts of students. This observation partially prompted this study, in order to see whether including SA questions benefitted weaker students, who in SA could obtain part marks. However, given that the examination instruments are neither standardised nor validated at this institution, there is no reason to expect the same performance across the different units, especially because the content is also different in each unit.

As can be seen in the bottom right quadrant of Table 1, no significant difference was observed in average student performance in SA questions compared to MC formats for either of the SCB121 examinations in which they were trialled, and the overall mean was also unchanged. Notably, the ranges and standard deviations were much larger in the SA sections of both semesters of SCB121 than in any of the MC examinations, reflecting the range of abilities in the diverse student cohorts. However, for SCB131 the difference in performance in the two sections was significant,  $t(536) = 3.06, p = 0.0023$ , with the MC question having a higher mean. This result is analysed in more detail below.

### Relationship between Examination Performance and Enrolled Degree

It was of interest to see if the relative performance in SA and MC questions was dependent on the enrolled degree of the student. Specifically, we were interested in whether students in degrees with lower entry cut-offs performed significantly better in SA questions, from which they could obtain marks for partial understanding of a concept, or whether they benefitted from the cueing and guessing aspects of MC questions. The results are presented in Figure 1.

**Figure 1.** Comparison of mean scores in MC and SA sections of mixed format examinations, by degree program\* §



\*The number of students in each course for each examination is included below the columns. Only degrees with 12 or more students have been included.

§The error bars show the standard error of the mean.

The courses, along with their associated 2011 Australian Tertiary Admissions Ranks (ATAR) are as follows:

- SC01: Bachelor of Applied Science (ATAR 75)
- SC40: Bachelor of Biomedical Science (ATAR 80)
- PU40: Bachelor of Health Science (Nutrition) (ATAR 80)
- LS37: Bachelor of Medical Laboratory Science (ATAR 75)
- OP45: Bachelor of Vision Science (Optometry) (ATAR 97.5)
- SC45: Pharmacy (ATAR 92)

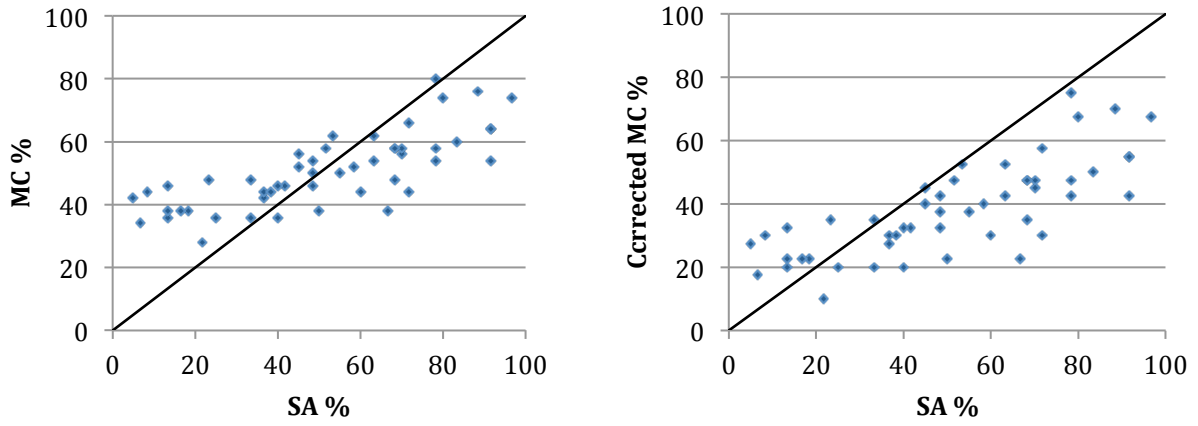
The percentage scores in the two sections of the examinations did not differ significantly within each degree program. The only exception was the pharmacy students; these students performed significantly better,  $t(106) = 2.7834$ ,  $p = 0.0064$ , on the MC section than the SA section for the SCB131 examination. This suggests that there are cohort-related metacognitive factors at work. These most likely relate to the perceived relevance of the content to their degree, which will affect how students prioritise their study. Of interest in this context is current work by Gwen Lawrie and her group, who have found that pharmacy students have a particularly routine-based approach to their science studies (Lawrie, 2012), which makes them quite different from science students, who are more likely to see avenues for inquiry in their learning. Importantly, the Figure also highlights that many of the cohorts were very weak, with means scoring around 50% in both sections of the mixed format examinations; in contrast, the optometry students, with a competitive entry requirement, performed much better.

### **Analysis of Relative Performance in MC vs SA for Individual Students**

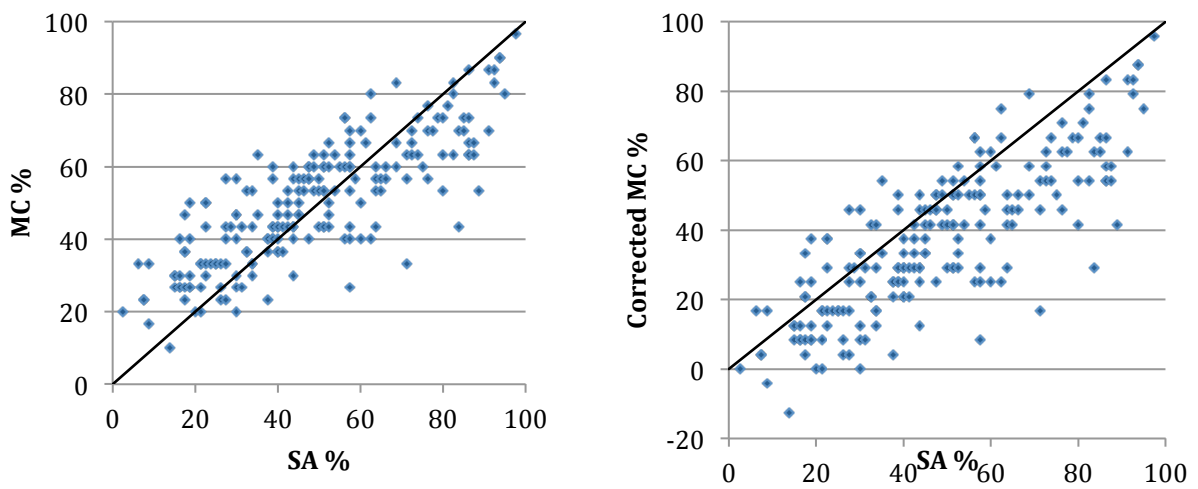
Confounding the expectation that inclusion of SA questions would improve student results, not only were the average results not improved in comparison to previous cohorts for SCB121, but the analysis also showed that the mean score was equal to or worse for SA than for MC sections in the mixed format examinations. This observation, along with an awareness of the literature in this area (Diamond & Evans, 1973; Prihoda, et al., 2006), lead us to investigate the influence of the correction for guessing. This was appropriate because an uncorrected MC score is biased towards a higher score for low-performing students, given that they can expect a score of 20% by pure random guessing (Diamond & Evans, 1973). We have retrospectively applied the correction for guessing to the scores of the students in order to better understand the relationship between SA and MC scores.

Figures 2 - 4 show scatter plots of individual students' scores in MC vs. SA format; the lines are the predicted 1:1 relationship. The analysis used the starting point described by Pinckard, et al. (2009), that SA results were valid, because they are marked by people able to judge the level of student understanding. Thus, the SA result was considered the independent variable, and the correlation with the dependent variable, the score in the MC part of the same examination by the same student, was being tested. Figures 2A, 3A and 4A show the relationship between raw MC scores and SA scores. Figures 2B, 3B and 4B include the standard correction for guessing to the MC scores; that is, 1/4 of a mark has been subtracted for each wrong answer. It is worth noting that retrospective application of the correction for guessing is useful for data analysis purposes only, because if this were to be applied to the awarded scores, the students would have been informed and their exam-taking behaviour would be modified (Prihoda, et al., 2006). It should also be noted that a student with a low mark can actually achieve a negative score when the correction for guessing is applied in this way; two students in SCB121\_2 obtained this result due to their poor performance in the MC section, with scores of 3/30 (10%) and 5/30 (17%), leading to corrected scores of -12.5% and -4%, respectively. Such scores are lower than would be expected if a student guessed every answer randomly from the five choices (expected score: 6/30), and indicate that the distractor answers have been well-chosen. The same two students received scores of 5.5/40 (14%) and 3.5/40 (9%), respectively, for the SA section, which confirms their poor understanding of the content.

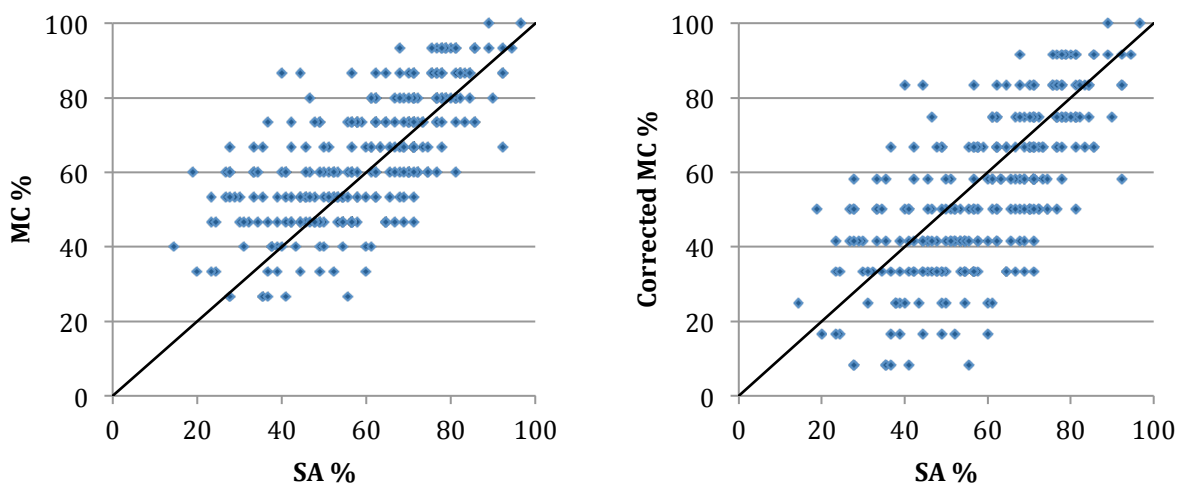




**Figure 2.** SCB121\_1 MC vs SA percentage scores.  
 Left hand side: Uncorrected MC (2A) Right: Corrected MC (2B)



**Figure 3.** SCB121\_2 MC vs SA percentage scores.  
 Left hand side: Uncorrected MC (3A) Right: Corrected MC (3B)



**Figure 4.** SCB131 MC vs SA percentage scores.  
 Left hand side: Uncorrected MC (4A) Right: Corrected MC (4B)

The first point that is clear upon examination of these plots is that for SCB121\_1 and SCB121\_2, the scatter is reasonably evenly distributed above and below the line before the correction for

guessing, and the majority of points lie below the 1:1 line after the correction is applied. This indicates that the MC sections were relatively harder than the SA sections in these examinations; this is discussed in more detail below. In contrast, for SCB131, the majority of the data points lie above the line before correction, showing the increased score in MC over SA for most students, presumably due to guessing and cueing. Application of the correction for guessing leads to a more even distribution of data points above and below the expected line.

Considering the SA result as a good measure of the students' understanding, it was possible to use these plots to determine the validity of the MC sections. If the MC section was a valid measure of understanding, a slope of 1.0 and an  $R^2$  value close to 1 would be expected, with an intercept close to zero. Table 3 summarises these values for the calculated lines of best fit for the six scatter plots in Figures 2 – 4.

**Table 3.** Slope, y intercept and correlation coefficients of lines of best fit for MC vs SA plots

|                    | slope  | y intercept | $R^2$  |
|--------------------|--------|-------------|--------|
| SCB121_1           | 0.3534 | 32.488      | 0.5859 |
| SCB121_1 corrected | 0.4417 | 15.61       | 0.5859 |
| SCB121_2           | 0.6118 | 20.595      | 0.6687 |
| SCB121_2 corrected | 0.7648 | 0.7436      | 0.6687 |
| SCB131             | 0.6744 | 23.796      | 0.4763 |
| SCB131 corrected   | 0.8430 | 4.763       | 0.4763 |

This table shows that the MC sections in these three examinations were not very good measures of understanding according to this criterion. In particular, the slope was much too flat for SCB121\_1, because these MC questions did not differentiate the students very well; none received very low or very high scores, and this is discussed below. The slope of SCB131 was reasonable, but the  $R^2$  value was less than 0.5, showing that there was a large amount of scatter in the scores. SCB121\_2 had the best MC section by this measure, with a slope over 0.5 and an  $R^2$  value above 0.65. Application of the correction for guessing made each slope steeper, as expected, and reduced the value of the y intercept; however, only for SCB121\_2 did the value of the y intercept become close to zero.

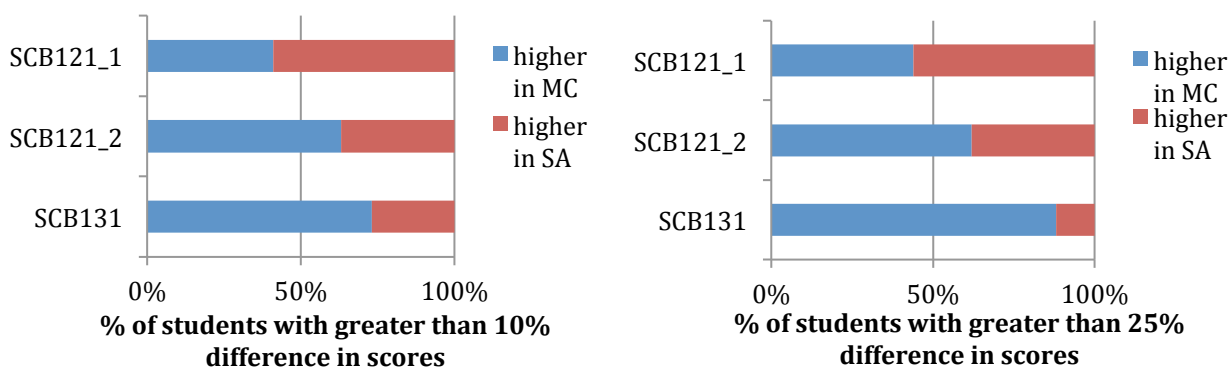
### **Direction of swing for individual students with a large discrepancy between MC and SA scores**

Inspection of the scatter plots in Figures 2 - 4 shows that while some students achieved similar results in their MC and SA sections, many have a wide discrepancy and are far away from the 1:1 line. This was investigated for SCB121\_1; the difference in the percentage marks between the MC and SA sections was greater than 10% for 29 students (58% of the class), reflecting the low slope found for the line of best fit above. Of these students 17 (or 59%) of them had a positive swing, indicating that they performed better in the SA than MC section. For the 9 students (18% of the class) with a greater than 25% difference in scores, 5 (56%) had a positive swing. That is, students with a large difference in their scores on the two sections in this examination, were reasonably evenly spread between achieving better scores in the SA or better scores in the MC section.

A similar analysis for SCB131 paints a different picture. In that unit, the difference in the percentage marks between MC and SA was greater than 10% for 114 students (42% of the class). Of these students, 83 (73%) had a negative swing, indicating that they performed worse in the SA than MC section. More extreme swings were associated with very weak performance in SA; 25 students (9% of the class) had a greater than 25% difference in their scores in the two sections, of which 22 (88%) had a negative swing. The cueing effect of MC is likely responsible for weak students performing better in that format in this examination (Schuwirth, et al., 1996).

The results in SCB121\_2, the large class for this unit, lie in between those found for SCB131 and SCB121\_1. In particular, of the 87 students (44% of the class) with a difference of 10% or greater in their scores on the two sections, 55 (63%) of them had a negative swing, indicating that they performed worse in the SA than MC. The 13 students (7%) with a greater than 25% difference in scores, were nearly evenly split into positive and negative swings, with 8 (62%) having a negative swing.

These results are presented graphically in Figure 5. Inspection of the Figure shows the dramatic difference between the relative performance in the two sections, across the three examinations, for students whose performance was very different in the two sections.



**Figure 5.** Direction of swing for students with large differences in their performance on the two sections of examinations.

Thus, although the overall average scores for SA and MC sections did not differ significantly in the examinations containing both formats for SCB121, it is clear that individual performance varied significantly between the two formats for some students. The SCB121 examinations allowed weak students to achieve better scores on the SA sections due to part marks being awarded, whereas the SA section of the SCB131 examination was more difficult for weak students.

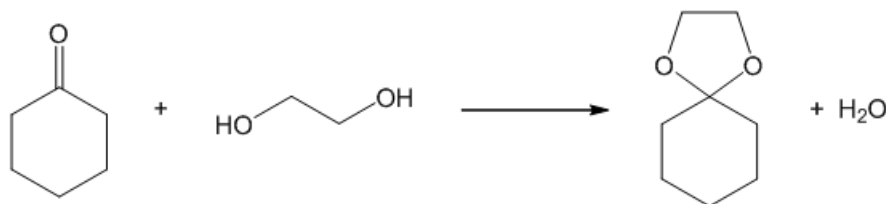
Inspection of Table 1 shows that the SCB131 students performed significantly better ( $t(515) = 6.1856, p < 0.0001$ ) in the MC section of their examination in 2011, compared to their performance in 2010 on the MC-only examination. The difference is also significant compared to the means in other years that this unit has been run (2009:  $t(517) = 2.6784, p = 0.0076$ ; 2008:  $t(436) = 3.5957, p = 0.0004$ ). This important observation supports the inclusion of SA questions as a way to improve study practices (Parmenter, 2009). The 15 questions used in the MC format section of the mixed format examination in 2011 included nine previously asked in the MC examination papers, while the remaining six were similar to previous MC questions. Five of the 15 SA questions were also taken from previous MC examination papers, while the others were specifically designed to require drawing of chemical structures, equations, predicted data or apparatus and were therefore different to any previous MC questions.

### Detailed Analysis of one SA question

It was of interest to examine how students performed on one of the SA questions that had been taken directly from the MC examination paper used from 2008 – 2010, and how this performance compared with their performance on that question when it had been asked in MC format. The question itself is shown in Figure 6.

### QUESTION 25

A student performs the following reaction using toluenesulfonic acid as a catalyst in refluxing toluene.



If the student starts with 1.45 mL of cyclohexanone (density 0.9478 g/mL) and 1.00 mL of ethylene glycol (density 1.1132 g/mL), what is the theoretical maximum yield?

**Figure 6.** Question used until 2010 as MC, and in 2011 as SA, in SCB131 examination paper

The marking scheme used for the SA section in this examination gave each SA question a mark from 0 - 3 (including half-marks). Marking was performed by a team of 15 postgraduate students over three hours, each of whom was assigned one question to mark on all scripts in order to minimise differences in marking. Due to time differences in marking, some of the questions were marked by two or three people, and during the present analysis a number of inconsistencies was observed, including marking differences where two identical answers were given different marks. These differences amounted to typically 0.5 out of 3 marks for the question.

Table 4 presents the results of the analysis for this question, which was performed as follows: First, the examination scripts were divided by score on this question (ranging from 0 - 3). Scores of 0 were further divided into blank, or no marks in spite of some writing. For scores from 0.5 - 2.5, the analysis process was to look at the answer given, and determine whether the student would likely have got the correct answer in MC format. This was usually because they had the correct or close to correct value but with the incorrect units or significant figures, and had lost partial marks. Students with an answer that was very different from the correct answer would likely have given an incorrect response in MC. Note that the analysis is not perfect, because a student who calculated an answer that was different from any of the MC responses may have been prompted to re-calculate their response and thus could have obtained the correct answer. This shows the importance of cueing.

**Table 4.** Data for Q25 in SCB131 examination papers

|    | % of students who got full marks | % of students with zero | % of students with part marks who would get full marks in MC | % of students with part marks who would get zero marks in MC | % of students who left blank |
|----|----------------------------------|-------------------------|--|--|------------------------------|
| MC | 80                               | 20                      | N/A  | N/A  | 0                            |
| SA | 55                               | 13                      | 10   | 15   | 7                            |

Several conclusions can be drawn from this analysis. First, the marking inconsistencies are of some concern. It is difficult to ensure consistency, even with the same marker, over a very large number of examination scripts, and this is one reason for the use of MC examinations. Second, the percentage of students receiving the full three marks was much less in the SA format (55%) than those who received the mark in MC format (80%). This may appear to be a weakness of SA format, but in fact shows the more sophisticated analysis that is done when marking this format, and in particular the extra depth required in the response. The 10% of students who received partial credit for this question who would have received the mark in MC format, had errors including incorrect use of significant figures, missing units, or small calculation errors. These students have not demonstrated a mastery of this content and so it is appropriate that partial marks are lost. No student

could receive marks for guessing in SA format, although partial marks were awarded even for commencing the solution, so a weak student was able to nevertheless demonstrate some understanding, whereas they would have received no marks in MC (unless they guessed correctly).

A final result of the analysis was the surprising finding that many students were unable to answer this particular question because they were not able to convert the line structure into a correct chemical formula. This question was not designed to test this understanding, which forms part of the fundamentals of organic chemistry (taught in SCB121, not SCB131). Thus, the analysis elucidated an unexpected misconception, which we hope to combat in future semesters.

### SA questions in SCB121

In the SA sections of the SCB121 examinations, the questions were chosen to test chemical literacy, including drawing the structures of organic compounds and writing equations. No questions were asked that are directly comparable to the MC questions, so the above analysis cannot be performed. In particular, most questions in the SA sections tested several concepts (drawing structures, nomenclature, reaction products), and so three or more MC questions would be required to test the same set of understanding. However, the structures asked are simpler in the SA format than in the MC format sections, because the visual cues are lacking in SA and therefore the MC compounds can be more complex.

Two examples of SA questions asked in SCB121 semester 1 and 2 examinations are shown in Figures 7 and 8.

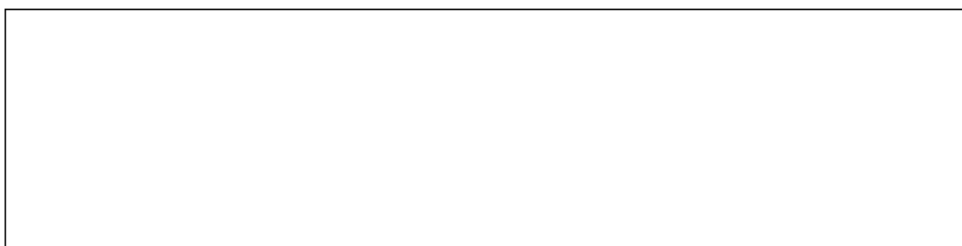
#### QUESTION 56

Suggest a synthetic route (draw an equation showing the starting material, product and all necessary reagents) for preparation of the following:

- (a) 2-Butanol from 2-butene



- (b) Bromobenzene from benzene



**Figure 7.** Typical question in the SA section of the final examination paper for SCB121\_1.

Each part of this question was worth 1.5 times a MC question in the same examination paper.

### QUESTION 3

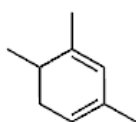
- (a) Draw the structure of 5-methylhexan-2-ol.



- (b) Draw the two possible products from the heating of this compound with concentrated sulfuric acid, and indicate which one is the major product.



- (c) Draw the structure of the product of the incomplete equation below.



**Figure 8.** Typical question in the SA section of the final examination paper for SCB121\_2.

Part (a) of this question was worth 0.25 of an MC question, part (b) was worth 0.75 of an MC question and part (c) was worth the same as an MC question in the same examination paper.

It is clear from these typical questions that it is impossible to test this sort of functional understanding in a MC examination; no question could exist without cues, and the ability to draw something is quite different from being able to select the correct image from a list. In marking the SA sections of these examinations, it was noted that the last few questions were left blank by many students, which indicates that the time available was insufficient.

### MC questions in SCB121

During the analysis of the results presented here, it was noted that the MC sections of the SCB121 examinations appeared difficult, relative to their SA sections; in addition, the MC section of the SCB121\_1 examination had a very narrow spread of scores, compared to all other MC examinations (see Table 1). For these reasons, we examined the questions in this MC section in detail, as well as looking at the MC section of the SCB121\_2 examination, the SCB111 and the

SCB113 examinations. We found that a very large number of questions in the SCB121 examinations, and the section of SCB113 that covers organic chemistry, did not follow basic guidelines for setting MC questions (Cheung & Bucat, 2002). In particular, a large number had a complex format, many were negatively phrased, and unnecessarily complex vocabulary was used. As Cheung and Bucat state, "The purpose of chemistry MC tests is to assess students' knowledge, understanding and problem solving, not reading proficiency." It is clear that improvements can be made to the MC assessment of organic chemistry; this could include training for academic staff. However, it should also be noted that some chemistry topics are more easily tested by high quality MC questions, such as concepts requiring calculation. Due to the nature of organic chemistry, it is poorly suited to the MC format, and the poor questions are partly a consequence of the nature of the material under examination.

### **Conclusions, Lessons Learnt and Suggested Modifications**

One of the difficulties of this study is that there was no consistency in the examination format or weighting across the four first year chemistry units. The numbers of MC questions and the examination weighting vary between the units. Even the approach taken in introducing SA questions (number and weight), has been at the discretion of unit coordinators, and a consistent approach was not used. Thus, a suggested modification, from an institutional perspective, is that a more consistent approach to first year assessment strategies is needed. This is important if the impact of any educational improvement is to be accurately assessed.

The result of the innovation of introducing SA questions to examinations was negative from the perspective of increasing the overall scores and pass rates for the units, because the SA results were not significantly higher than the MC results in any unit where the mixed format was trialled (and in fact were lower in SCB131). One lesson learnt is that the weighting of SA questions relative to MC questions is extremely important to the final scores of the students if they perform very differently in the two sections (which many students did). Although it is unlikely that each question requires the same amount of time to answer correctly (because some require a conceptual response, whereas others require calculations), we suggest that a more accurate estimate of time required to correctly answer both the MC and the SA sections is required when the weighting is chosen. The point value of each question should be proportional to the time required, the difficulty and also the importance of the concept being tested. Further to this, we propose that the examination instruments should be kept the same each year, except for incremental changes to remove errors and possibly change the length. Retaining and modifying an existing examination instrument allows the time allowed to better match the time required, based on an analysis of student responses (such as not attempting the final few questions).

The data do not directly indicate whether the study habits were different for the units and semesters with mixed format examinations compared with previous semesters and other units. However, the significant improvement in the MC scores for the SCB131 students in the mixed format examination compared with the previous MC examinations is some evidence that the students did study more effectively. The overall marks were unchanged in the three units with mixed format examination papers.

The detailed analysis of one SA question showed that the difficulty that many students had was with an aspect of the content that we did not realise was examined by this question. Such an analysis would be too time consuming to perform on every question type, but the result was eye-opening in this case. Our analysis of the MC questions covering organic and coordination chemistry content showed that these questions were poorly worded and likely contributed to the low MC scores in SCB121. Thus, it is vital to have a thorough checking process for any examination instrument, with adequate time and a mixture of experienced and junior academics involved.

Timely and constructive feedback, which helps clarify what good performance is and gives students the opportunity to close the gap between current and desired performance, has been shown to greatly enhance student learning (Nicol, 2007; Sadler, 1989). In the context of our study, an overriding question is whether it is better to use limited resources in marking SA examination questions, from which the students do not receive feedback other than their mark, or whether it makes more sense for these resources to be used in progress examinations and assignments, where detailed feedback can be given. The answer to this question depends on many factors including the specific content being taught, the resources available and institutional policies around assessment. We propose that, given the demonstrated increase in learning when some SA questions are included (shown by improved MC scores relative to previous years in SCB131), no examination should be MC only. A small number of well-designed SA questions will lead to improved learning approaches without a great marking cost. These are particularly important for organic chemistry, because they allow chemical literacy to be assessed.

Our final suggested modification is the development, in Australia, of a suite of agreed, shared MC examination questions for first year chemistry, which are fully tested and validated. These questions would not have the issues of unintentional difficulty (Hartman & Lin, 2011), bias (Cheung & Bucat, 2002), and difficulty of construction (Nicol, 2007) described in the introduction. Use of such questions in a mixed format examination would allow inexperienced academics to avoid some of the pitfalls of writing MC questions, while also permitting benchmarking between cohorts and institutions. Although there are arguments against such a proposal, largely because institutions have different specialities and focii, the majority of the content taught at the first year level is common and similar text books are used (Mitchell Crow & Schultz, 2012). Optional adoption of some shared questions would improve the quality of chemical education (Holme, 2003).

### **Acknowledgements**

The authors acknowledge the generous assistance of our colleagues in the first year teaching team who have shared their examination instruments and results. In particular, we have had fruitful discussions about the philosophy of assessment with Dennis Arnold, John McMurtrie and Eric Waclawik.

### **References**

- Biggs, J., & Collis, K. (1982). *Evaluating the Quality of Learning: the SOLO Taxonomy*. New York: Academic Press.
- Biggs, J., & Tang, C. (2007). *Teaching for Quality Learning at University: What the Student Does* (3rd ed.). Maidenhead: SRHE and Open University Press.
- Cheung, D., & Bucat, R. (2002). *How can we construct good multiple-choice items?* Paper presented at the Science and Technology Education Conference, Hong Kong.
- Diamond, J., & Evans, W. (1973). The Correction for Guessing. *Review of Educational Research*, 43, 181-191.
- Hartman, J. R., & Lin, S. (2011). Analysis of Student Performance on Multiple-Choice Questions in General Chemistry. *Journal of Chemical Education*, 88, 1223-1230.
- Holme, T. (2003). Assessment and Quality Control in Chemistry Education. *Journal of Chemical Education*, 80, 594-596.
- Lawrie, G. (2012). Personal communication, 2012.
- Le, K. N., & Tam, V. W. Y. (2007). A survey on effective assessment methods to enhance student learning. *Australasian Journal of Engineering Education*, 13(2), 13-19.
- Mitchell Crow, J., & Schultz, M. (2012). Report on a Mapping Exercise of Chemistry at Australian Universities. Melbourne, Australia. Retrieved June 10, 2012 from <http://chemnet.edu.au>



- Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and Higher Education*, 31, 53-64.
- Parmenter, D. A. (2009). Essay versus multiple choice: Student preferences and the underlying rationale with implications for test construction. *Academy of Educational Leadership Journal*, 13(2), 57-71.
- Pinckard, R. N., McMahan, C. A., Prihoda, T. J., Littlefield, J. H., & Jones, A. C. (2009). Short-Answer Examinations Improve Student Performance in an Oral and Maxillofacial Pathology Course. *Journal of Dental Education*, 73, 950-961.
- Prihoda, T. J., Pinckard, R. N., McMahan, C. A., & Jones, A. C. (2006). Correcting for Guessing Increases Validity in Multiple-Choice Examinations in an Oral and Maxillofacial Pathology Course. *Journal of Dental Education*, 70, 378-386.
- Ruder, S. M., & Straumanis, A. R. (2009). A Method for Writing Open-Ended Curved Arrow Notation Questions for Multiple-Choice Exams and Electronic-Response Systems. *Journal of Chemical Education*, 86, 1392-1396.
- Sadler, D. R. (1989). Formative Assessment and the Design of Instructional Systems. *Instructional Science*, 18, 119-144.
- Schultz, M. (2011). Sustainable assessment for large science classes: Non-multiple choice, randomised assignments through a Learning Management System. *Journal of Learning Design*, 4, 50-62.
- Schuwirth, L. W. T., van der Vleuten, C. P. M., & Donkers, H. H. L. M. (1996). A closer look at cueing effects in multiple-choice questions. *Medical Education*, 30, 44-49.
- Scouller, K. (1998). The Influence of Assessment Method on Students' Learning Approaches: Multiple Choice Question Examination versus Assignment Essay. *Higher Education*, 35(4), 453-472.
- Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: a review. *Assessment and Evaluation in Higher Education*, 30(4), 325-341.
- Woodford, K., & Bancroft, P. (2004, 5-8 December). *Using multiple choice questions effectively in Information Technology education*. Paper presented at the 21st ASCILITE Conference, Perth.