Science Education International
Vol.23, No.1, March 2012, 56-83

Is there a relationship between chemistry performance and question type, question content and gender?

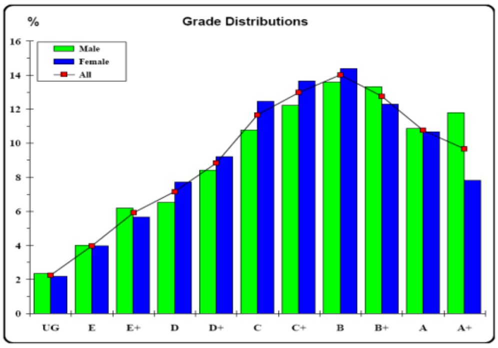
Ross D. Hudson
Australian Council for Educational Research and Curtin University, Australia

Abstract
This research inquires into the effectiveness of the two predominant forms of questions - multiple-choice questions and short-answer questions - used in the State University Entrance Examination for Chemistry including the relationship between performance and gender. It examines not only the style of question but also the content type examined (recall and application questions). The research involves class trial testing of students with structured questions that examine the same material content with each type of question (multiple-choice or short-answer) and also examines the different type of content (recall or application) and finally the influence of student gender. Rasch analysis of the class trial data, including gender difference analysis, is performed and the analyses related to performance characteristics of the State University Entrance Examination. It is found that male students achieve higher scores than female students with respect to mean scores on both tests and sub-tests. However, when student abilities, as measured by Rasch analysis were considered, male and female students of equal abilities perform equally well in each test comparison suggesting that chemistry is equally accessible to students of both genders.
Keywords: Chemistry, gender, assessment, question type.

Introduction
This study developed from the researcher's initial observations, along with those of a number of teaching colleagues, that there were apparent differences between the observed students' performances on the State University Entrance Examination for Chemistry (the Victorian Certificate of Education examinations, VCE, is a very high stakes examination as it forms the basis for selecting students for entry into tertiary courses). It was found that students were performing differently on the two semester examinations. In particular, female students achieved better grades in the second semester examination (held in November), whereas male students were much more successful in the first semester examination (held in June) than they were in the second semester examination. This difference was demonstrated in the grade distribution data released by the examining authority (Figure 1) where the distribution of the A+ grade is significantly skewed in favour of male students in both semesters, though less so in semester 2 (VCAA, 2005). Whilst there were small differences in grade distribution across all grades, the differences at the higher grades were of concern. These grades could be crucial in determining the fate of a student's entry into a particular university course and this study has focused on these grades.

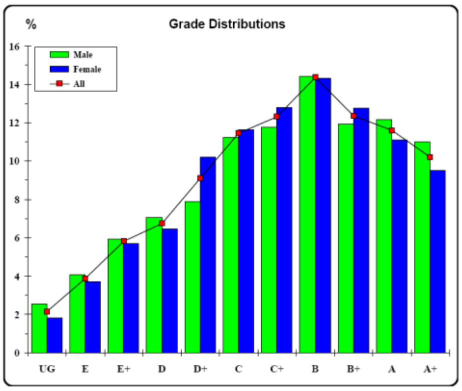
This difference raised questions. Why should this be occurring? Were any observed differences actually significant? A number of factors were relevant in considering this issue. Apart from the gender of the students, two other important factors were considered likely to influence the performance of the students. These were: type of question (short-answer or multiple-choice) and the content type of the question (recall or application).

The following research questions were addressed in the ensuing study.
1. Do multiple-choice or short-answer questions more positively emphasize student understanding?
2. Do students perform more effectively on recall type questions or on application questions?
3. Does students' gender and ability influence performance in chemistry examinations?



(VCAA, 2005)

Semester 1



(VCAA, 2005)

Semester 2

Figure 1: Grade distributions for the 2005 VCE Chemistry Examination 1 and 2*

Literature Perspectives

Question type, content and the influence of gender on performance are topics that have interested educational researchers for many years. The use of tools such as Rasch analysis to measure the influence of these factors has been a somewhat more recent, but insightful, innovation.

---

* In Figure 1 the grade UG represents the only non-pass grade in the examinations. This usually represents about 2.5 % of the participating students. The number of female students (4823) was larger than the number of male students (4207) (VCAA 2005).

Question type and content

Multiple-choice questions, generally, have less scope and complexity than short-answer questions and therefore are often likely to be less difficult. This suggestion is supported in the literature (Braswell, 1990; Bridgeman, 1992; Martinez, 1991), with these authors concluding that open ended type questions are superior in assessing student understanding of concepts because the solution methodology employed by the students in arriving at their answer can be examined. In multiple-choice questions the answers give no indication of how students arrived at their answer (Bridgeman, 1992).

Student perceptions about performance are also important. Students were generally more confident of their answers being correct when answering multiple-choice questions than short-answer questions, regardless of the fact that the actual performance on the two types of questions was sometimes almost the same. Students feel they do (or were going to) perform better on multiple-choice tests (Pressley, Ghatala, Woloshyn, & Pirie, 1990). Haynie (1994) noted that student performance on multiple-choice questions is superior to that on short-answer questions. However, Haynie also emphasised the importance of the need for testing to support learning and not simply determining summative attainment.

Of importance to this research, though, is the issue of whether the content of a particular question influences the results obtained by the students, regardless of how the question is presented. Student responses to stoichiometric questions seem to produce different outcomes, depending on whether the question is presented as multiple-choice as opposed to short-answer questions (Niaz & Robinson, 1995). Students are apparently influenced by the multiple-choice options and consequently provide answers to questions that are unlikely to be answered if presented in a short-answer form. Students commonly find great difficulty in attempting to justify, or explain the selection of response in any particular multiple-choice question. This supports the notion that students are both willing and able to guess when presented with a multiple-choice question that the student cannot answer (Barnett-Foster & Nagy, 1996). A further shortcoming of multiple-choice questions is that they do not provide insights into higher order thinking by the student (Barnett-Foster & Nagy, 1996; Frederickson, 1984; Petrie, 1986). Generally the problem with most forms of testing is that they simply focus on recalling information for a test without necessarily forming deep understanding of the material being learned. Taber (2010) argues that chemistry teaching offers many opportunities for students to develop skills in providing multilayered explanations to complex scenarios. Biggs (1996) findings point to the conclusion that, although items in objective multiple-choice tests can assess high level thinking, they rarely go beyond Bloom's comprehension level (Anderson, 1972; Marso & Pigge, 1991). With respect to this study, the first semester examination is more heavily loaded with higher order application questions (stoichiometry, rates and equilibrium) than is the semester two examination (energy) (VCAA, 2006). This difference in content loading may be a factor in explaining the observed differences in the observed grade distributions (Figure 1).

Gender Differences in Performance

The role of gender in performance in chemistry, and other subject areas in general, precipitated a variety of studies over time and will no doubt continue to do so. A study by Boli, Allen and Payne (1985) explored the reasons behind the differences that were observed between the genders in undergraduate chemistry and mathematics courses. The exploration sought reasons why male students were tending to outperform the female cohort, which resulted in the suggestion that differences in mathematical ability were a very important consideration. The most important factor, however, through an analysis of previous studies, was that the male students' natural self-confidence and belief in the importance and need for mathematics had a positive influence on male performance. Student motivation needed to impact on student participation and student engagement. Differences in motivation and attitude towards science favoured male students, although the difference was less than the difference in outcome achievement (Becker, 1989). Further to this, Becker was supported by other findings (Beller & Gafni, 1991; Hamilton, 1998; Hedges & Howell, 1995; Wolleat, Pedro, Becker & Fennema, 1980) that indicated the subject areas that most favour males over females were the more traditional pure sciences of physics and chemistry (Becker, 1989). Jones and Kirk (1990) explored the male-female differences in attitudes towards choosing sciences. Their study showed that when the issue came down to choice, females tended towards choosing the life sciences of biology and psychology rather than chemistry and physics, because females were generally more interested in studying a science they saw as a helping science, a people oriented science, or a nurturing science. Subsequent studies (Buccheri, Gurber & Bruhwiler, 2011; Francis, Hutchings, Archer & Melling, 2003; Schoon, 2001; Stobart, Elwood & Quinlan, 1992; Watson, Quatman & Edler, 2002) had all demonstrated similar patterns in enrolments and participation in higher school sciences; however, the differences were not as great or consistent as they once were. In comparing the genders with respect to motivation, the study by Lloyd, Walsh & Yailagh (2005) found that males were more confident in the study of mathematics than females. They suggested the self-efficacy of males was higher in terms of their beliefs about success in mathematics. In other words, males expected to do well in the subject. Part of this expectation was the finding that male students' expectations of a career were founded in the study of mathematics and science (Lloyd et al., 2005). Other studies have suggested, however, that there is little difference between the self-efficacy of males and females with respect to participation in the sciences (Karaarslan & Sungar, 2011) and there is some evidence that female students may be driven to succeed in subjects such as chemistry by the desire to succeed in a traditionally male dominated study (Grunert & Bodner, 2011).

These findings for mathematics were seen as fairly evenly transferable to the natural sciences (Boli, Allen & Payne, 1985). Other than mathematics, there appeared to be no directly gender-related reasons for the male students outperforming the female students, yet the evidence showed that this was the case. Boli et al. (1985) theorised that the mathematics background of the female students was less rigorous than that of the males and this was having a flow-on effect in the latter's studies of both mathematics and science. This, and other studies, have shown that females were less likely to choose mathematics and science courses at the undergraduate level, often because of lesser preparation at the prior levels of schooling (Blickenstaff, 2005; Spelke, 2005).

The analysis of a number of large assessment studies has demonstrated that male students generally performed better than did female students (Beller & Gafni, 1991; Korporshoek, Kuyper, van der Werf & Bosker, 2011; Neuschmidt, Barth & Hastedt, 2008). More detailed analysis showed that if the type of question, based on content, was considered then the differences were less pronounced; that is, male students tended to outperform female students in the areas of the physical sciences (physics and chemistry), whereas in the life sciences (biology and psychology) the differences were negligible (Beller & Gafni, 1991; Hamilton, 1998; Hedges & Howell, 1995; Linn, Baker & Dunbar, 1991). Hamilton (1998) supported the findings of other researchers showing that multiple-choice questions tended to favour males over females but, importantly to this research, the differences were less pronounced with short-answer questions. Whilst the multiple-choice questions did favour the male students, findings with respect to short-answer questions were mixed; some studies indicated that males still performed better, whereas others suggested that the females were advantaged by the short-answer format (Hamilton, 1998). In either event, if the questions asked required higher order thinking or required analysing new situations, then male students were advantaged (O'Neill & McPeek, 1993; Rennie & Parker, 1991).

Item Response Theory and Classical Test Theory

Item response theory is a measure of the performance of each student on each item in a test. Classical test theory, however, places emphasis on a large number of items with high correlations between items to enhance the interpretation of results (Andrich, 2005, Pallant, 2010) and analysis is less dependent on gaining a normal distribution of examination scores.

The typical test used in the VCE examinations consists of a combination of multiple-choice questions and short-answer questions, which is considered a suitable outcome as it increases the reliability of the assessment tool and also allows a wider range of material to be covered (Ercikan et al., 1998). A conflict exists between choosing between the two types of assessment modes, in that multiple-choice questions allow a wider range of content, which typically assess recall content, whilst the short-answer questions allow better assessment of problem solving skills (Barnett-Foster & Nagy, 1996; Ercikan et al., 1998).

The position of item response theory has been enhanced through the application of models such as Rasch analysis (Andrich, 1988). Rasch analysis examines and measures the probability that the participants will answer items correctly. An important aspect of Rasch analysis is the establishment of the uni-dimensional nature of each item, (each item must be shown to have little, if any, dependence on any other item). Compared to classical test theory, item response theory focuses more on measuring the quality of the items and the capability of each item to distinguish or discriminate between participants for the particular traits under analysis. That is, for any given item, a relationship exists between the ability of a student and the probability of the student getting the item correct (Pallant, 2010). Rasch measurements take into account two measures, test item difficulty and person ability. The measures are assumed to be interdependent, but separation between the measures is also assumed (Andrich, 2005). Important aspects of Rasch analysis are the establishment of the uni-dimensional nature of each item, the placing of all items on a scale of relative difficulty with respect to each item, and placing all participating students on a similar, relative ability scale (Andrich, 2005).

The uni-dimensional nature of the trial tests in this research fits well with item response theory, as a relatively small sample of students was used. The items were tested using Rasch analysis to demonstrate the validity of the test construct in terms of uni-dimensionality. With a small number of items involved in the test, the supportive correlation of items, which is a feature of classical test theory, was not possible in the trial tests (Pallant, 2010). The external validity of the test construct is limited in terms of the relatively small sample size, but will have application to students with a similar background to that of the sample group (Cohen, Manion & Morrison, 2000).

Methodology

Data Sources

Data was collected from 192 participating Year 11 students from four secondary colleges. The participation rate by the students was very high, with few cases detected in the subsequent Rasch analysis suggesting instances of students not attempting the tests in a purposeful manner, three students were removed from the data set for this reason. Also, a further five students' data was not able to be used in the subsequent Rasch analysis as they did not complete sufficient questions due to absence and other reasons. Overall 184 students' data was included in the Rasch analysis. The test results were analysed using both ANOVA and Rasch analysis. The Rasch analysis determined the validity and uni-dimensionality of the trial tests while the ANOVA analysis provided summative data about the tests.

Sample Testing Procedure

The researcher constructed short tests that asked essentially similar questions but in both multiple-choice and short-answer form. That is, pairs of questions were constructed so that the content loading of each was similar, but one was presented as a multiple-choice question and the other as a short-answer question. Validity was established by the students' teachers participating in the process of checking that the question pairs were as much as possible of equal content loading. The equal loading of each question was later demonstrated by the excellent correlation found during the analysis of the trial tests. Whilst some researchers (Anderson, 1972; Marso & Pigge, 1991; Simkin & Kuechler, 2005) have made conclusions about the advantages and disadvantages of each type of question, there appears to have been few studies directed at examining the effectiveness of each type of question as to how well they assess student understanding in chemistry. Only a limited number have explored student performance where the questions are very similar in content but framed in the two question types (Chan & Kennedy, 2002).

To test the understanding of students, the following testing structure was adopted.

- Each class was divided into two groups. Each group did essentially identical tests (in terms of the curriculum content), except that one group's test required multiple-choice responses and the other group short-answer responses.
- About one week later, a second test on similar material was administered except that the type of test the two groups received was reversed.
- Consequently, each student completed a multiple-choice and short-answer test on the subject matter being tested. The purpose of dividing the groups into two halves was to reduce the effect of learning and enhancement (or possibly reduced retention) that may have occurred between the two tests. Splitting the groups allowed each type of test to be examined under similar circumstances.

Example of paired difficulty questions

The following represent examples of paired difficulty stoichiometry (application) questions.

From the Stoichiometry multiple-choice Test: Question 1

1. *The percentage by mass of oxygen in $Mg(NO_3)_2$ is closest to:*
   A. 11%
   B. 48%
   C. 65%
   D. 78%

From the Stoichiometry Short-answer Test: Question 1

1. *What is the percentage by mass of Zn in $Zn_3(PO_4)_2$?*

The examples shown in the above, test student understanding of the concept of percentage composition by mass. One is framed as a multiple-choice question and the other as a short-answer question. This allows an investigation of performance, where question format is the discriminating variable. Such information is important because an analysis of VCE examination papers does not easily produce this sort of comparative information; it is not the practice of examiners to ask the same question twice in the same paper.

Analysis of the examination papers may show that students answered certain types of questions more successfully than others, but it is difficult to ascertain whether this was a result of the general difficulty of the question content, or due to the question type itself. The trial tests offer a unique opportunity to examine this relationship. The trial papers also elicit valuable information with respect to gender performance on the examinations.

Results/Analysis

The results from the trial test analysis were analysed using ANOVA and Rasch analysis, based on the classification of the questions. To ensure the uni-dimensional nature of the final trial test, adjustments were undertaken to remove students from the subject group whose person fit residuals were above 2.5 (Pallant, 2010). Of the 192 students taking part, only three students were deleted from the Rasch analysis due to poor fit statistics, a further five were removed because they provided insufficient data leaving 184 students in the data set. The initial analysis also resulted in one multiple-choice question being eliminated due to poor fit statistics. The final Rasch summary statistics were; Item fit residual standard deviation (1.099), Person fit residual standard deviation (0.843), PSI reliability index (0.795) and, most importantly, the Item-trait interaction probability was 0.124. The Item fit residual standard deviation and Person fit residual standard deviation measure how well each item and each person fit the Rasch model. Standard deviation of greater than 1.5 suggests that some items or persons do not fit the Rasch model (Pallant, 2010). The PSI (Person Separation Index) measures the internal consistency of the Rasch scale and how well the analysis can discriminate between the persons taking part in the analysis. The PSI can be interpreted in a similar way as Cronbach alpha (Pallant, 2010). The Item-trait interaction probability is based on the Chi-squared probability value and measures the interaction between item difficulty and person ability level. A non significant value indicates that the rank order of the items is consistent across all levels of the underlying trait (Cavanagh, Romanoski, Giddings, Harris & Dellar, 2003; Pallant, 2010). The adjusted test item structure gave a valid item set demonstrating strong uni-dimensionality in the items (Cavanagh et al., 2003; Pallant, 2010), a necessary condition to validate Rasch analysis.

The initial correlation analysis importantly supported the premise of the trial tests. That is, the tests would be testing the same skills, but in different formats (see Figure 2).

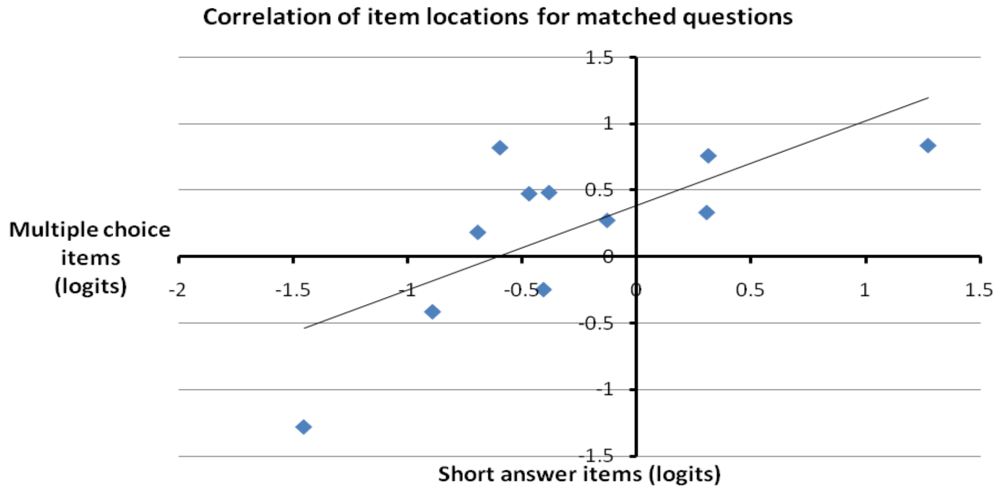


Figure 2. Correlation between the difficulties of the multiple-choice items and short-answer items

This graph essentially showed that the easiest multiple-choice question was also the easiest short-answer question and so on. The Pearson’s correlation value of 0.72 supports this interpretation. The score distribution (Figure 3) was very similar to that of the grade distribution of the VCE examinations (Figure 1) suggesting that conclusions drawn from the trial tests were likely to have application to the larger university entrance cohort.

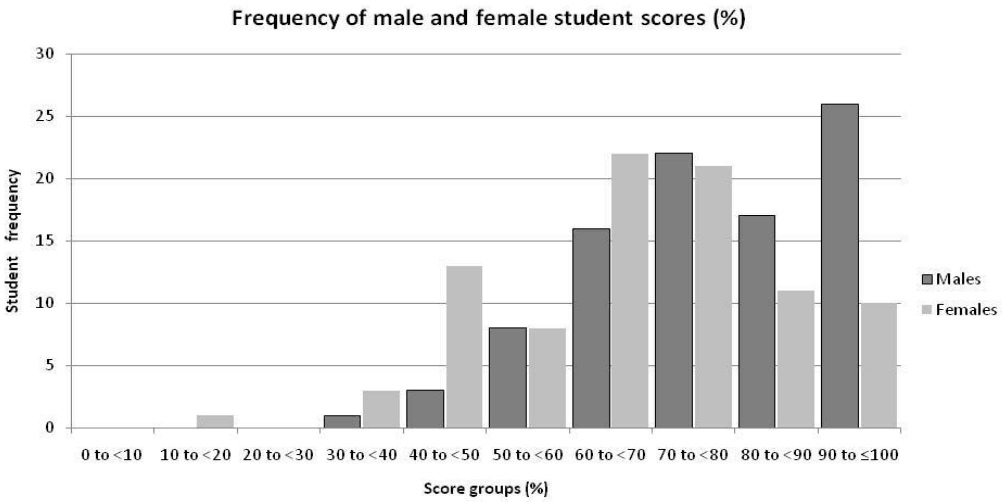


Figure 3. Distribution of male and female scores in the Chemistry trial tests (n=184) Targeting – item against student ability

A measure of the effectiveness of the trial tests is the target matching of items against student ability. The item map (Figure 4) shows how the item difficulties matched the student abilities on a common scale. While on a test that covers the entire population (the VCE examinations), the spread of item difficulties should approximately match the ability spread of the persons taking the test, the students participating in the trial tests were from schools that usually performed well in chemistry in the VCE examinations. As a result, it would be expected that there would be a mismatch of the item difficulty with student ability; that is, the student abilities were likely to be generally higher than the item difficulty. This result was substantiated, but the item distributions and person distributions both conformed to a normal distribution.

Targeting – item against student ability

A measure of the effectiveness of the trial tests is the target matching of items against student ability. The item map (Figure 4) shows how the item difficulties matched the student abilities on a common scale. While on a test that covers the entire population (the VCE examinations), the spread of item difficulties should approximately match the ability spread of the persons taking the test, the students participating in the trial tests were from schools that usually performed well in chemistry in the VCE examinations. As a result, it would be expected that there would be a mismatch of the item difficulty with student ability; that is, the student abilities were likely to be generally higher than the item difficulty. This result was substantiated, but the item distributions and person distributions both conformed to a normal distribution.

As shown in the graph, the test items showed some weakness in properly assessing students of high ability (students above 2 logits). It is evident from the item map that there were no items that were sufficiently difficult for the most able students. However, as mentioned earlier, this was expected. The tests did match student ability well at the lower end of the ability range.

Overall, the trial tests had (with some modification) produced a valid measure of student ability. The use of the Rasch analysis gave credibility to the internal validity of the test items (Andrich, 2005; Cavanagh et al., 2003). An initial analysis of the trial papers was conducted using ANOVA. The result of this analysis is shown in Table 1.

```
Location                 Students     Items [uncentralised thresholds]
(logits)
--------------------------------------------------------------------
  4.0                             |
                              ooo |
                                  |
                                  |
                                o |
  3.0                        oooo |
                                  |
                                o |
                             oooo |
                                  |
  2.0                   ooooooooo |
                             oooo |
                           oooooo |
                            ooooo |  ST12.1
                      ooooooooooo |  AB06.1
  1.0                    oooooooo |  ST07.2   AB02.1
                           oooooo |  AB10.1   AB09.2   AB07.2   ST11.1
                      ooooooooooo |  AB10.2   AB12.2   AB12.1
                           oooooo |
                          ooooooo |  ST10.1   AB03.1   AB04.1   AB11.2
  0.0                       ooooo |  ST11.2   ST12.2   AB11.1
                              ooo |  ST07.1   AB05.1   AB08.2
                              ooo |  ST01.1   ST04.1   ST09.1   AB09.1
                               oo |  ST06.1   AB08.1   ST05.1
                                o |  ST10.2   AB01.1   AB07.1
 -1.0                           o |  ST03.1
                                  |
                                  |  ST08.1
                                o |  ST02.1
                                  |
 -2.0                             |
         o = 2 or fewer persons
-----------------------------------------------------------------------------
```

Figure 4. Item map showing distribution of items and students†

Summarising, differences in performance on different question types did occur even when the questions were matched for difficulty. While the main thrust of this paper is towards the analysis of gender effects on student performance, measurement of question type and content is important to provide validity to the outcomes through the correlation of the data with established literature findings. A brief summary of the question type and question content analysis follows.

Question Type

The difference of performance means between multiple-choice and short-answer questions overall is small (mean-multiple-choice=71.0, standard deviation = 20.8 and mean-short-answer = 69.1, standard deviation = 22.3). The difference (Table 1) is not statistically significant F (1,366 = 0.72; $p > 0.05$. It thus appears that the multiple-choice questions do not offer any particular advantage to students over short-answer questions, except for the subset of recall questions where the difference is significant.

The Rasch analysis show that students in the lower ability ranges find short-answer questions slightly more difficult than multiple-choice questions (see Figure 5). However, at the higher end of student abilities, the difference between performances is negligible. This result could be explained by the likelihood of a good student making an inadvertent error in selecting the multiple-choice response, whereas in the short-answer version of a question this would be far less likely. As these students were more likely to gain full credit for his or her efforts in the short-answer items, the narrowing of observed differences, at the top of the ability range, is understandable.

† **Key**: ST refers to a stoichiometry question, AB refers to an acid-base question. The two numerals immediately after AB or ST are the question numbers on the trial tests. The decimal numbers indicate whether the item was code 1 or code 2. E.g. ST11.1 and ST11.2 indicate that stoichiometry Q11 was code 2.

Table 1. ANOVA analysis of question performance by question type and classification Chemistry trial tests

(Refer to Appendix 1 for data)

				ANOVA results (N = 368, df = 1)		
Comparison variables	**Mean (%)**	**Std.deviation**	**Discriminating variable**	**Sum of Squares**	**F**	**p (sig)**
Multiple-choice Short-answer	81.5 69.9	24.5 30.7	Application Questions	12312	15.95	0.000
Multiple-choice Short-answer	62.3 68.6	28.8 25.9	Recall Questions	3660.3	4.86	0.028
Multiple-choice **Short-answer**	**71.0** **69.1**	**20.8** **22.3**	**Application and Recall Questions**	**333.84**	**0.72**	**0.39**
Application Recall	81.5 62.3	24.5 28.8	Multiple-choice Questions	33925	47.4	0.000
Application Recall	69.9 68.6	30.7 25.9	Short-answer questions	162.0	0.20	0.65
Application **Recall**	**74.8** **66.4**	**25.3** **22.8**	**Multiple-choice and Short-answer questions**	**6450.3**	**11.10**	**0.001**

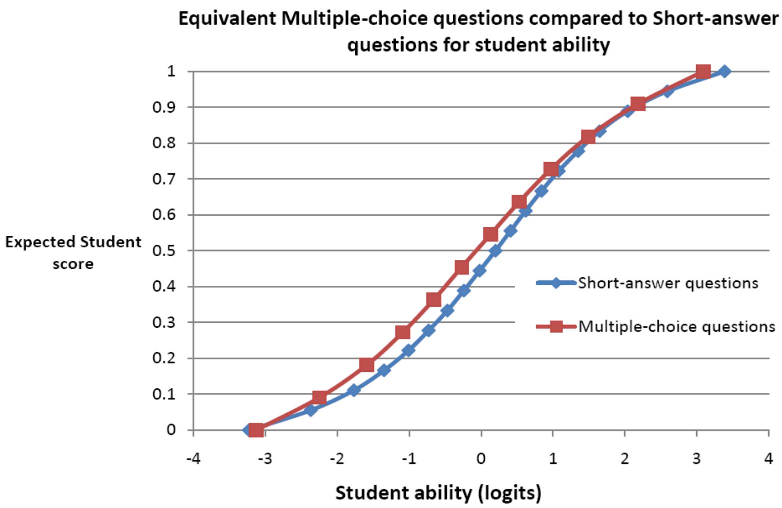


Figure 5. Multiple-choice compared to Short-answer response difference against expected score and student ability.

To compare the 11 multiple-choice questions (each code 1; total: 11 marks) with the 12 short-answer questions (some code 1 and some code 2; total: 18 marks) the score range for each set of questions was reduced proportionally to a range of 0 to 1 (see Figure 5). The expected score effectively shows the score a student of a particular ability is likely to obtain. For example a student with a student ability of logit = 0 is likely to score 0.5 on the short-answer test and 0.55 on the multiple-choice test. (A similar rescaling was performed on the acid-base to stoichiometry comparison (Figure 6).

Question Content

The content of the questions, however, does raise issues. The content of the question does affect the performance outcome of the student response (particularly if the question is presented in multiple-choice format). As shown in Table 1, students' performance on recall questions (mean = 66.4 and standard deviation = 22.8) is weaker than on application questions (74.8 /25.3). The ANOVA results show a statistically significant difference $F\ (1,366) = 11.1$; $p < 0.001$ in performance on recall questions compared to application questions.

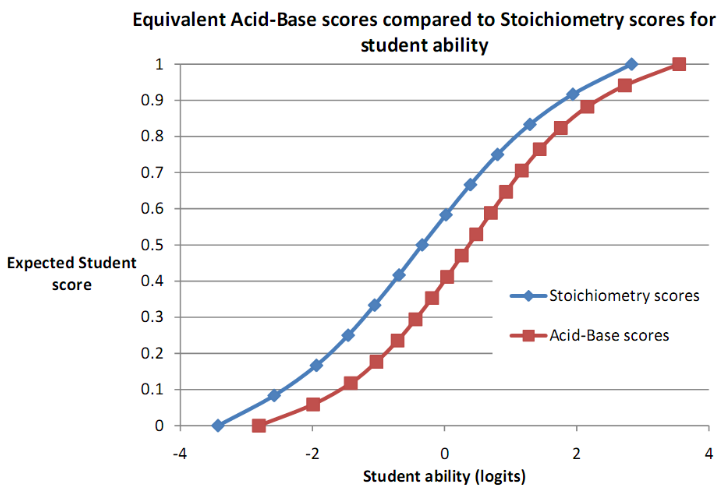


Figure 6. Recall (Acid-Base) compared to Application (Stoichiometry) response difference against expected score and student ability.

The graph (Figure 6) shows that students of all abilities find application type questions easier than recall questions. This outcome was largely affected by the result from the multiple choice application questions which was substantially higher than all other means. Overall however, with respect to question type and content, the trial tests generally reflect the findings of previous research and show that the trial tests have credibility as a valid test construct (Barnett-Foster & Nagy, 1996; Niaz & Robinson, 1995; Pressley et al., 1990).

Gender Differences

Previous analysis showed that males performed significantly better in the examinations than did females, particularly at the A+ (see Figure1) end of the score range (Hudson, 2011). An initial analysis of the student performance in the trial tests showed that males (mean = 78.3) again outperformed females (mean 68.6) on the trial tests (Table 2). The ANOVA results (Table 3) of the raw scores showed that the difference was significant ($F(1,182) = 15.9$; $p < 0.01$). The distribution of scores in the trial tests (Figure 3) closely matched that of the VCE examination distributions (Figure 1).

Table 2. Gender differences on the trial chemistry tests (means)

Groups	*Count*	*Sum*	*Average (%)*	*s.d.*
Male	94	7364.2	78.3	15.8
Female	90	6174.4	68.6	17.3

Table 3. Gender differences on the trial chemistry tests (ANOVA)

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	4360.3	1	4360.3	15.9	0.0001
Within Groups	49813.4	182	273.7		

Whilst the distribution (Figure 3) is less uniform than the VCE distribution (the smaller sample size in the trial tests partly account for this), the significant aspects are the negative skewness and the peak in the performance of the males compared to females in the 80 to 100% score range (equivalent to the A and A+ grades in the VCE grade distribution). This distribution demonstrates that the trial tests are useful in that the tests appear to have reasonably mirrored the distribution of students' abilities, even though the sample size is relatively small compared to that of the VCE examinations (Hudson, 2011).

The trial tests, however, allow a finer examination of student performance on the various category types of question asked. That is not possible on the VCE examination. The student performance was able to shed some specificity on where males outperformed female students. The following analysis attempts to identify where, within the test structure, male students performed differently from females. Four sub-tests were created within RUMM2030 to examine the relationship between gender and question type and between gender and question content.

The four subtests were:

1. [ST01]; sub-test 1: comparing gender and all multiple-choice question performance.
2. [ST02]; sub-test 2: comparing gender and all short-answer question performance.
3. [AB]; sub-test Acid-Base: comparing gender and all recall (acid-base) question performance.
4. [ST]; sub-test Stoichiometry: comparing gender and all application (stoichiometry) question performance.

Multiple-choice and gender

The ANOVA test results (Appendix 2) showed that there is a significant statistical significance ($F(1,182) = 13.65$; $p < 0.001$) in performance on multiple-choice questions between males and females. Male student performance was better than female performance on the multiple-choice questions as supported by the means (males) = 76.4, standard deviation = 19.8 compared to mean (females) = 65.5/20.4. The Rasch graphical analysis showed, however, that the difference was less marked when the scores were adjusted for latent student ability as measured by RUMM2030 software.

The graph (Figure 7) showed that male students, within their ability ranges, find the multiple-choice questions slightly more difficult than did female students. In other words, in spite of the higher raw scores obtained by the males, male students find multiple-choice questions somewhat more challenging than expected. Put simply, females found multiple-choice questions easier than did males for students of equal ability.

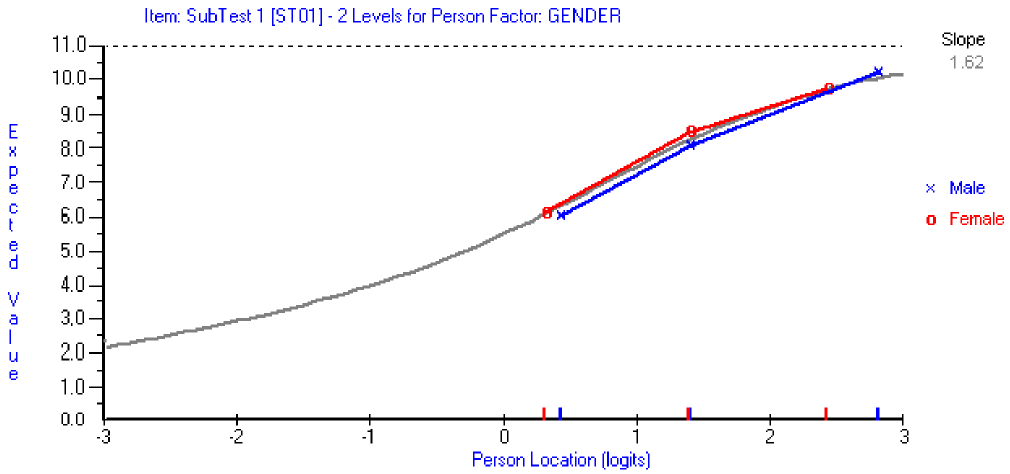


Figure 7. Multiple-choice questions showing gender difference against expected score and student ability. (Note: The expected value is the score a student of a particular ability would be expected to achieve out the 11 marks available on the multiple-choice questions on the trial tests).

Short-answer and gender

An examination of the subtest covering the short-answer questions and gender shows a similar result to that of multiple-choice in terms of the raw score analysis.

The ANOVA test results (Appendix 3) show that there is a statistically significant difference ($F(1,182) = 11.85$; $p < 0.001$) in performance between the male and female students on short-answer questions. Male performance is better than the females on the short-answer questions as shown by the means (for males mean = 74.5, standard deviation = 20.2 compared to that of females 63.5/23.1). The Rasch graphical analysis shows that the difference is less marked when the scores are adjusted for latent student ability, as measured using RUMM2030 software.

The Rasch graphical analysis shows a different result than with the multiple-choice subtest. The graph (Figure 8) shows that male students, within their ability ranges, find the short-answer questions slightly easier than do female students. This result is perhaps not surprising when referred back to the previous analysis. If male students found multiple-choice questions relatively more difficult, then it is reasonable to assume that the short-answer questions would appear to be relatively easier for them. The reverse appears to be the case for the female students.

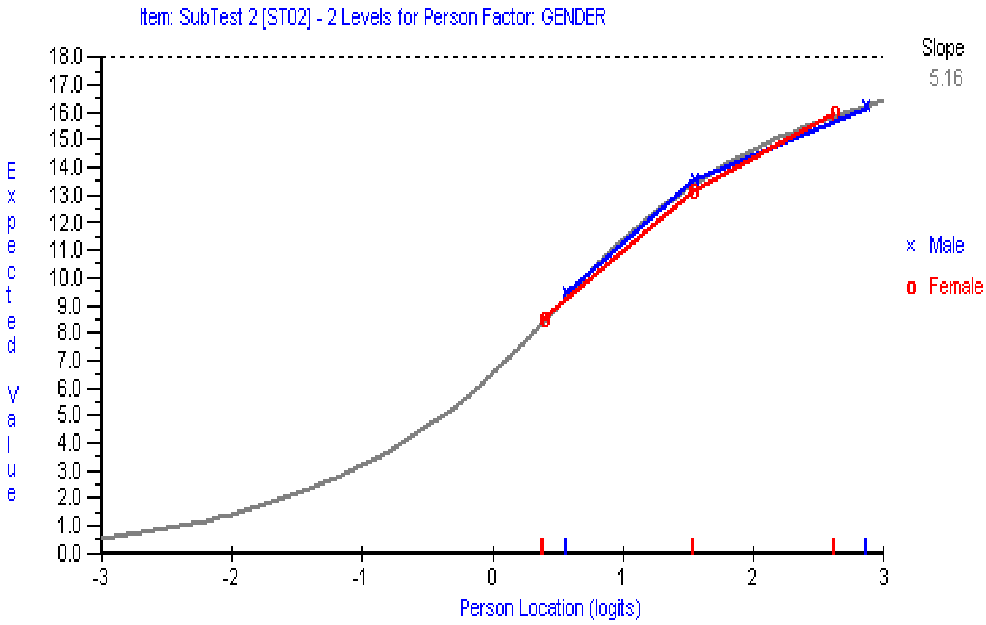


Figure 8. Short-answer questions: showing gender difference against expected score and student ability (Note: The expected value is the score a student of a particular ability would be expected to achieve out the 18 marks available on the short-answer questions on the trial tests).

However, in both instances, the males outscore the females on both tests. The difference between the male and female student scores is statistically significant. An interesting observation in Figure 8 was that female students find the short-answer questions easier than did the male students at the very top end of the ability scale. This outcome is explained as being consistent in terms of the reportedly greater ability of female students in language expression and recall questions (Beller & Gafni, 2000). The very best female students might be more capable of expressing answers coherently, than do the male students, of equally high ability.

Comparison of Recall (Acid-base) questions and Application (Stoichiometry) questions by gender

The initial anecdotal observations from the researcher's teaching career suggests that the female students tend to find recall type questions easier than application questions. By creating an analysis of subtests within RUMM2030, to separate the recall (Acid-Base) questions into one subtest and the application (Stoichiometry) questions into a second subtest, allows this notion to be tested.

All Recall questions and gender

The ANOVA test results (Appendix 4) show that there is a statistically significance (F (1,182) = 6.27; $p < 0.05$) in performance on recall type questions between the male students and female students. Male performance is better on the recall questions than the females, as supported by the means (for males, mean = 70.5, standard deviation = 23.3 compared to females 62.2/21.6). This difference is relatively smaller and less significant than in the two previous analyses. This tends to show that the females are more able (or the males less so) with these types of questions. Considering that the difference in the overall ability is 10 percentage points (mean difference on all items), the difference of 8 percentage points between the means here suggests that males do not have the same advantage when questions are recall based (see Table 2).

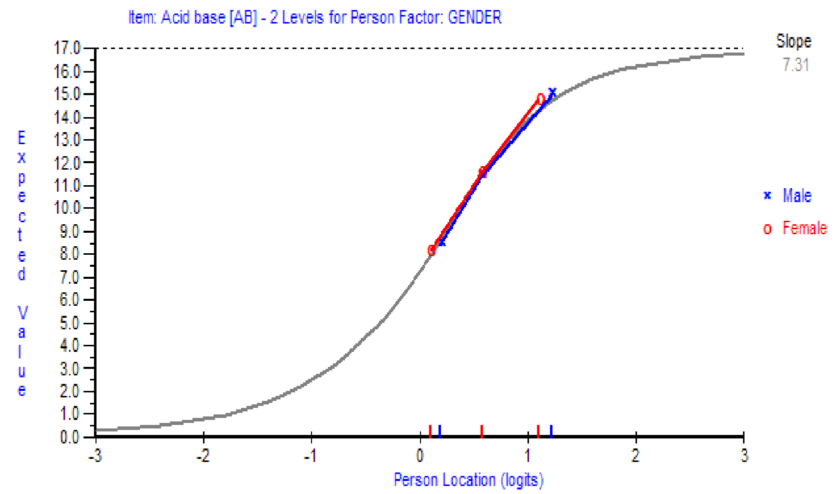


Figure 9. Recall (Acid-Base) all questions: showing gender difference against expected score and student ability (Note: The expected value is the score a student of a particular ability would be expected to achieve out the 17 marks available on the acid-base questions on the trial tests).

The Rasch graphical analysis (Figure 9) shows that there is little difference between the students in this sample, with respect to answering recall type questions. Within in the relative student abilities, both males and females find item difficulty to be about the same.

All application questions and gender

The ANOVA test results (Appendix 5) show that there is a significant statistical significance ($F(1,182) = 17.03$; $p < 0.001$) in performance on application (Stoichiometry) type questions between the male students and female students. Male student performance is substantially better on the stoichiometry questions than the female students, as supported by the means (for males, mean = 82.0, standard deviation = 21.4 compared to females 67.2/26.9). This is the largest difference between the means of the different question classifications and it is notable that the standard deviation of the female student scores is also very large in comparison to the previous analyses. This difference supports the notion that the females find the stoichiometry somewhat harder than do the males and consequently goes someway to explaining the difference in performance between the female students in the Semester 1 examination and the Semester 2 examination, where the proportion of recall questions is greater.

The Rasch graphical analysis (Figure 10) shows that there is a difference between the male and female lines on the graph. The graph also has a very substantial slope, indicating that these questions are highly discriminating (Pallant, 2010). Essentially these questions divide students into two groups, those who can do stoichiometry and those who cannot. The graph also shows that these questions are harder for females than for males, even when allowing for ability.

This result, as demonstrated in Figure 10, is particularly important in terms of the relationship to the VCE examination results and may help to explain the difference in performance of the male students and female students when comparing the Semester 1 examinations to the Semester 2 examinations. Whilst the male students outperform the female students in both examinations, the differences are less pronounced than in Semester 2. As mentioned previously, the content structure of Semester 1 is heavily loaded with stoichiometric calculations. Semester 2 has a much higher proportion of descriptive chemistry. It is likely, given the results shown in Figure 10, that this is a factor in explaining the differing semester achievements shown in Figure 1.

Given the observations demonstrated in Figure 10, it is reasonable to conclude that males will perform more strongly compared to the females in an examination that has a greater loading of stoichiometry or application questions. This effect seems to have been demonstrated in the comparison of the semester 1 and semester 2 examinations (see Figure 1).

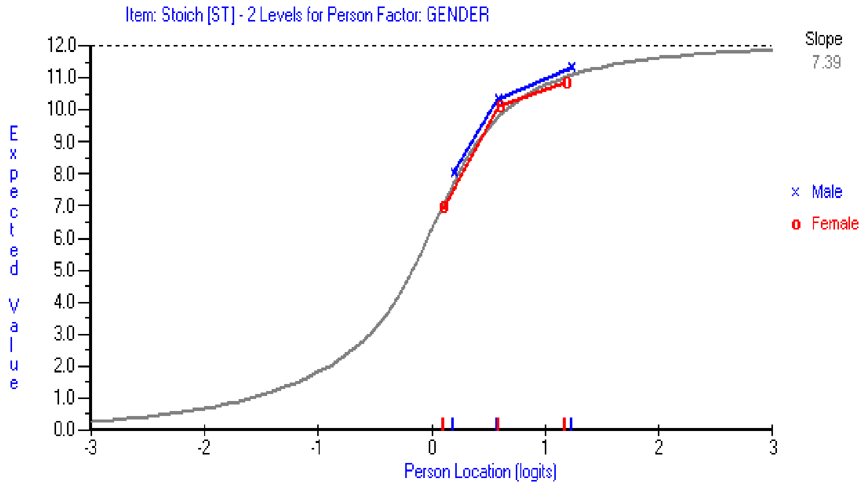


Figure 10. Application (Stoichiometry) all questions: showing gender difference against expected score and student ability (Note: The expected value is the score a student of a particular ability would be expected to achieve out the 12 marks available on the stoichiometry questions on the trial tests).

Conclusions

The trial test analysis show similar patterns to the VCE examinations in terms of grade distribution (Hudson, 2011). However, when gender difference analysis is undertaken using RUMM2030 software, the difference in performance is much smaller. The Rasch analysis compares student performance according to gender, but allows for student ability. When this result is taken into account, the performance by the students differ little by gender, even though the mean scores are significantly in favour of males. In some instances, female performance (allowing for ability) is better than that of the male students, even though the mean scores suggest otherwise. For example, in comparing multiple-choice question performance, the mean for males = 76.4% (s.d. = 19.8) is higher than that of the females, mean = 65.5% (s.d. = 20.4) (see Appendix 2). However, when analysed using Rasch, the gender difference analysis shows that females generally performed slightly better on multiple-choice than did males, once student ability is taken into account (Figure 7). Another significant observation shows that males outperformed females on the stoichiometry questions (Figure 10). This observation, when taken with the grade distributions for the VCE Chemistry examinations and trial tests (see Figures 1 and 3), may explain the fact that male students outperform female students much more so in the semester 1 examination than in the semester 2 examination (2003-2007). The implication of this finding is that the performance of females in chemistry could be enhanced, if there were less stoichiometric application questions in the examinations. Such an observation has been supported by numerous findings from the literature (Beller & Gafni, 2000; Cox, Leder & Forgasz; 2004; Hawkes, 2004).

Another important conclusion is implied by the trial test results. The initial findings of the trial test analysis support the observations of other researchers (Beller & Gafni, 2000; Bridgeman & Lewis, 1994; Cox et al., 2004; Hamilton, 1998; Lumsden & Scott, 1987) in that male students achieve higher scores than do female students (Table 2). When the performance allows for student ability, however, the differences are quite small (Figures 7 to 10). This suggests that perhaps the ability of the male students is (at the top end) greater than that of the female students taking chemistry. It may well be that a greater proportion of high performing males are choosing chemistry than are high performing females. The reasonable assumption is that significant numbers of high achieving females are choosing to do other subjects and not chemistry. This would account for the skewed appearance of the results. This proposition certainly warrants further investigation.

Overall, however, the clear outcome from this the analysis is that male students achieve higher grades in chemistry than do female students, particularly at the top end of the grade scale. Two significant factors appear to be the use of stoichiometric questions (favouring male students) and the possibility that more high ability males study chemistry.

The findings of this research show that there is not any underlying advantage of multiple-choice to short-answer questions. Performance on both is similar, when the question content is similar. Performance on multiple-choice questions is only slightly better than on short-answer even when allowances are made for the difficulty of the items and abilities of the students. Further research with a larger, more extensive cohort is suggested to clarify these findings.

References

Anderson, G. (1998). *Fundamentals of educational research* (2nd. ed.). Bristol: PA: Falmer Press.

Anderson, R. C. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research, 42*(2), 145-170.

Andrich, D. (1988). *Rasch models for measurement*. London: SAGE.

Andrich, D. (2005). The Rasch model explained. In S. Alagumalai, D. D. Curtis & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars.* Dordrecht: Springer.

Barnett-Foster, D. & Nagy, P. (1996). Undergraduate student response strategies to test questions of varying format. *Higher Education, 32*(2), 177-198.

Becker, W. E. & Johnson, C. (1999). The relationship between multiple-choice and essay response questions in assessing economics understanding. *Economic Record, 75*(231), 348-357.

Beller, M. & Gafni, N. (1991). The 1991 International Assessment of Educational progress in mathematics and science. The gender differences in perspective. *Journal of Educational Psychology, 88*, 365-377.

Beller, M. & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles, 42*(1-2), 1-21.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347-364.

Blickenstaff, J. C. (2005). Women and science careers: leaky pipeline or gender filter? *Gender and Education*, 17(4), 369-386.

Boli, J., Allen, M. L. & Payne, A. (1985). High-ability women and men in undergraduate mathematics and chemistry courses. *American Educational Research Journal, 22*(4), 605-626.

Braswell, J. (1990). *A comparison of item characteristics of multiple-choice and grid-in type questions*: Paper presented at the Annual Meeting of the American Educational Research Association. Boston, MA.

Bridgeman, B. (1992). A Comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*(3), 253-271.

Bridgeman, B. & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31*(1), 37-50.

Buccheri, G., Gurber, N. & Bruhwiler, C. (2011). The impact of gender on interest in science topics and the choice of scientific and technical vocations. *International Journal of Science Education,* 33(1), 159-178.

Cavanagh, R., Romanoski, J., Giddings, G., Harris, M. & Dellar, G. (2003). *Application of Rasch model and traditional statistics to develop a measure of primary school classroom learning culture.* Paper presented at the International Education Research Conference AARE - NZARE, Auckland, New Zealand.

Chan, N. & Kennedy, P. E. (2002). Are multiple-choice exams easier for economics students? A comparison of multiple-choice and equivalent constructed response exam questions. *Southern Economic Journal, 68*(4), 957-971.

Cohen, L., Manion, L. & Morrison, K. (2000). *Research methods in education* (5th ed.). London: Routledge Falmer.

Cox, J., Leder, J. & Forgasz, H. (2004). Victorian Certificate of Education: Mathematics, science and gender. *Australian Journal of Education, 48*(1), 27-46.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M. & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35*(2), 137-154.

Francis, B., Hutchings, M., Archer, L. & Melling, L. (2003). Subject choice and occupational aspirations among pupils at girls' schools. *Pedagogy, Culture & Society, 11*(3), 425-442.

Grunert, M. L. & Bodner, G. M. (2011). Underneath it all: gender role identification and women chemists' career choices. *Science Education International, 22*(4), 292-301.

Hamilton, L. S. (1998). Gender differences on high school science achievement tests: Do format and content matter? *Educational Evaluation and Policy Analysis, 20*(3), 179-195.

Hawkes, S.J. (2004), Reaction to why do we teach equilibrium calculations? *Journal of Chemical Education, 81*(9), 1265.

Haynie, W. (1994). Effects of multiple-choice and short-answer tests on delayed retention learning. *Journal of Technology Education, 6*(1), 32-44.

Hedges, L. V. & Howell, A. (1995). Sex differences in mental scores, variability, and numbers of high scoring individuals. *Science, 269*, 41-45.

Hudson, R. (2011). *Multiple-choice questions compared to short-answer response: Which assesses understanding of chemistry more effectively?* Perth: Curtin University of Technology.

Jones, A. T. & Kirk, C. M. (1990). Gender differences in students' interests in applications of school physics. *Physics Education*, 25, 308-313.

Karaarslan, G. & Sungar, S. (2011). Elementary students' self-efficacy beliefs in science: Role of grade level, gender, and socio-economic status. *Science Education International, 22*(1), 72-79.

Korporshoek, H., Kuyper, H., van der Werf, G. & Bosker, R. (2011). Who succeeds in advanced mathematics and science courses*? British Educational Research Journal*, 37(3), 357-380.

Linn, M. C., Baker, E. L. & Dunbar, S. B. (1991). Complex, performance based assessment. Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Lloyd, J. E. V., Walsh, J. & Yailagh, M. S. (2005). Sex differences in performance attributions, self-efficacy, and achievement in mathematics: If I'm so smart, why don't I know it? *Canadian Journal of Education / Revue canadienne de l'education, 28*(3), 384-408.

Lumsden, K. G. & Scott, A. (1987). The economics student reexamined: Male-female difference in comprehension. *Journal of Economic Education, 18*(4), 365-375.

Marso, R. N. & Pigge, F. L. (1991). An analysis of teacher made tests: Item-types, cognitive demands and item construction errors. *Contemporary Educational Psychology, 16*, 279-286.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207-218.

Neuschmidt, O., Barth, J. & Hastedt, D. (2008). Trends in gender differences in mathematics and science (TIMSS 1995-2003). *Studies in Educational Evaluation*, 34, 56-72.

Niaz, M. & Robinson, W. R. (1995). From algorithmic mode to conceptual gestalt in understanding the behaviour of gases: An epistemological approach. *Research in Science and Technological Education, 10*, 53-64.

O'Neill, K. A. & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* (pp. 255-276). Hillsdale: Erlbaum.

Pallant, J. (2007). *SPSS Survival Manual* (3rd. ed.). Maidenhead: Open University Press, McGraw-Hill Education.

Pallant, J. (2010). Rasch analysis advice. In R. Hudson (Ed.). (pp. J Pallant provided analysis advice to the author R. Hudson). Canberra.

Petrie, H. (1986). Testing for critical thinking. In D. Nyberg (Ed.), *Philosophy of Education* (pp. 3-19). Normal, IL.: Philosophy of Education Society

Pressley, M., Ghatala, E. S., Woloshyn, V. & Pirie, J. (1990). Sometimes adults miss ideas and do not realize it: Confidence in responses to short-answer and multiple-choice comprehension questions. *Reading Research Quarterly, 25*(3), 232-249.

Rennie, L. J. & Parker, L. H. (1991). Assessment of learning in science: The need to look closely at item characteristics. *The Australian Science Teachers Journal, 37*(4), 56-59.

Schoon, I. (2001). Teenage job aspirations and career attainment in adulthood: A 17 year follow up study of teenagers who aspired to become scientists, health professionals, or engineers. *International Journal of Behavioural Development, 25*(2), 124-132.

Simkin, M. & Kuechler, W. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education, 3*(1), 73-98.

Spelke, E. S. (2005). Sex Differences in Intrinsic Aptitude for Mathematics and Science?: A Critical Review. American Psychologist, 60(9), 950-958.

Stobart, G., Elwood, J. & Quinlan, M. (1992). Gender bias in examinations: How equal are the opportunities? *British Educational Research Journal, 18*(3), 261-276.

Taber, K. S. (2010). Challenging gifted learners: general principles for science educators; and exemplification in the context of teaching chemistry. *Science Education International, 21*(1), 5-30.

VCAA. (2005). *Grade distribution report for 2005*. Retrieved 9th June 2008, from http://www.vcaa.vic.edu.au/vce/statistics/2005/section3/vce_chemistry_ga05.pdf

VCAA. (2006) *VCE Chemistry Units 1-4:* Victorian curriculum and Assessment Authority

Watson, C., Quatman, T. & Edler, E. (2002). Career aspirations of adolescent girls: Effects of achievement level, grade and the single sex environment. *Sex Roles: A Journal of Research, 46*(9), 857-871.

Wolleat, P. L., Pedro, J. D., Becker, A. D. & Fennema, E. (1980). Sex differences in high school students' causal attributions of performance in mathematics. *Journal for Research in Mathematics Education, 11*(5), 356-366.

Appendices

Appendix 1:
All student scores broken down by type and gender and ANOVA analysis and student location. (N=184)

Note in all tables MC= multiple-choice and SA = short-answer

Gender	Acid-base score	Acid-base score %	Stoichiometry score	Stoichiometry scores %	Multiple-choice (MC) score	MC scores %	Short-answer (SA) score	SA scores %	Acid-base MC score	Acid-base scores %	Acid-base SA score	Acid-base SA scores %	Stoichiometry MC score	Stoichiometry MC scores %	Stoichiometry SA score	Stoichiometry SA score %	Student Location (logits)
Possible score	17		12		11		18		6		11		5		7		
male	3	17.65	11	91.67	8	72.73	6	33.33	3	50.00	0	0.00	5	100.0	6	85.71	1.282
male	11	64.71	3	25.00	4	36.36	10	55.56	1	16.67	10	90.91	3	60.00	0	0.00	0.704
male	7	41.18	5	41.67	5	45.45	7	38.89	2	33.33	5	45.45	3	60.00	2	28.57	-0.258
male	9	52.94	10	83.33	6	54.55	13	72.22	2	33.33	7	63.64	4	80.00	6	85.71	0.757
male	15	88.24	12	100.0	11	100.0	16	88.89	6	100.0	9	81.82	5	100.0	7	100	2.577
male	12	70.59	10	83.33	10	90.91	12	66.67	5	83.33	7	63.64	5	100	5	71.43	1.253
male	13	76.47	10	83.33	8	72.73	15	83.33	4	66.67	9	81.82	4	80.00	6	85.71	1.445
male	11	64.71	11	91.67	8	72.73	14	77.78	4	66.67	7	63.64	4	80.00	7	100	1.253
male	5	29.41	11	91.67	5	45.45	11	61.11	0	0.00	5	45.45	5	100	6	85.71	0.317
male	13	76.47	10	83.33	10	90.91	13	72.22	5	83.33	8	72.73	5	100	5	71.43	1.445
male	3	17.65	10	83.33	7	63.64	6	33.33	3	50.00	0	0.00	4	80.00	6	85.71	0.971
male	3	17.65	8	66.67	7	63.64	4	22.22	3	50.00	0	0.00	4	80.00	4	57.14	0.426

male	8	47.06	7	58.33	7	63.64	8	44.44	3	50.00	5	45.45	4	80.00	3	42.86	0.174
male	15	88.24	12	100	11	100	16	88.89	6	100	9	81.82	5	100	7	100	2.577
male	13	76.47	12	100	10	90.91	15	83.33	5	83.33	8	72.73	5	100	7	100	1.903
male	0	0.00	9	75.00	4	36.36	5	27.78	0	0.00	0	0.00	4	80.00	5	71.43	0.801
male	10	58.82	8	66.67	6	54.55	12	66.67	2	33.33	8	72.73	4	80.00	4	57.14	0.606
male	15	88.24	10	83.33	9	81.82	16	88.89	5	83.33	10	90.91	4	80.00	6	85.71	1.903
male	10	58.82	12	100	10	90.91	12	66.67	5	83.33	5	45.45	5	100	7	100	1.253
male	12	70.59	9	75.00	8	72.73	13	72.22	3	50.00	9	81.82	5	100	4	57.14	1.077
male	11	64.71	10	83.33	4	36.36	17	94.44	1	16.67	10	90.91	3	60.00	7	100	1.077
male	12	70.59	10	83.33	8	72.73	14	77.78	3	50.00	9	81.82	5	100	5	71.43	1.253
male	6	35.29	9	75.00	5	45.45	10	55.56	0	0.00	6	54.55	5	100	4	57.14	0.64
male	13	76.47	10	83.33	10	90.91	13	72.22	5	83.33	8	72.73	5	100	5	71.43	1.445
male	14	82.35	9	75.00	8	72.73	15	83.33	4	66.67	10	90.91	4	80.00	5	71.43	1.445
male	11	64.71	12	100	7	63.64	16	88.89	2	33.33	9	81.82	5	100	7	100	1.445
male	13	76.47	11	91.67	9	81.82	15	83.33	4	66.67	9	81.82	5	100	6	85.71	1.658
male	4	23.53	10	83.33	6	54.55	8	44.44	1	16.67	3	27.27	5	100	5	71.43	0.032
male	11	64.71	9	75.00	7	63.64	13	72.22	2	33.33	9	81.82	5	100	4	57.14	0.913
male	12	70.59	11	91.67	8	72.73	15	83.33	3	50.00	9	81.82	5	100	6	85.71	1.445
male	7	41.18	8	66.67	8	72.73	7	38.89	3	50.00	4	36.36	5	100	3	42.86	0.174
male	10	58.82	6	50.00	6	54.55	10	55.56	3	50.00	7	63.64	3	60.00	3	42.86	0.317
male	6	35.29	6	50.00	5	45.45	7	38.89	2	33.33	4	36.36	3	60.00	3	42.86	-0.258
male	12	70.59	11	91.67	7	63.64	16	88.89	3	50.00	9	81.82	4	80.00	7	100	1.445
male	15	88.24	12	100	10	90.91	17	94.44	5	83.33	10	90.91	5	100	7	100	2.577
male	14	82.35	12	100	11	100	15	83.33	6	100	8	72.73	5	100	7	100	2.198
male	7	41.18	12	100	7	63.64	12	66.67	2	33.33	5	45.45	5	100	7	100	0.757
male	10	58.82	9	75.00	8	72.73	11	61.11	4	66.67	6	54.55	4	80.00	5	71.43	0.757
male	9	52.94	12	100	9	81.82	12	66.67	4	66.67	5	45.45	5	100	7	100	1.077
male	7	41.18	9	75.00	4	36.36	12	66.67	0	0.00	7	63.64	4	80.00	5	71.43	0.388
male	11	64.71	9	75.00	7	63.64	13	72.22	4	66.67	7	63.64	3	60.00	6	85.71	0.913
male	13	76.47	12	100	10	90.91	15	83.33	5	83.33	8	72.73	5	100	7	100	1.903

male	13	76.47	12	100	10	90.91	15	83.33	5	83.33	8	72.73	5	100	7	100	1.903
male	8	47.06	9	75.00	5	45.45	12	66.67	2	33.33	6	54.55	3	60.00	6	85.71	0.46
male	12	70.59	10	83.33	8	72.73	14	77.78	3	50.00	9	81.82	5	100	5	71.43	1.253
male	13	76.47	12	100	10	90.91	15	83.33	5	83.33	8	72.73	5	100	7	100	1.903
male	17	100	12	100	11	100	18	100	6	100	11	100	5	100	7	100	3.937
male	15	88.24	11	91.67	10	90.91	16	88.89	5	83.33	10	90.91	5	100	6	85.71	2.198
male	15	88.24	12	100	10	90.91	17	94.44	5	83.33	10	90.91	5	100	7	100	2.577
male	6	35.29	7	58.33	6	54.55	7	38.89	6	100	0	0.00	0	0.00	7	100	3.414
male	12	70.59	9	75.00	7	63.64	14	77.78	3	50.00	9	81.82	4	80.00	5	71.43	1.077
male	7	41.18	9	75.00	4	36.36	12	66.67	0	0.00	7	63.64	4	80.00	5	71.43	0.83
male	14	82.35	12	100	10	90.91	16	88.89	5	83.33	9	81.82	5	100	7	100	2.198
male	9	52.94	6	50.00	8	72.73	7	38.89	4	66.67	5	45.45	4	80.00	2	28.57	0.174
male	7	41.18	2	16.67	3	27.27	6	33.33	2	33.33	5	45.45	1	20.00	1	14.29	-0.725
male	17	100	12	100	11	100	18	100	6	100	11	100	5	100	7	100	3.937
male	15	88.24	11	91.67	10	90.91	16	88.89	5	83.33	10	90.91	5	100	6	85.71	2.198
male	15	88.24	12	100	10	90.91	17	94.44	5	83.33	10	90.91	5	100	7	100	2.577
male	7	41.18	12	100	7	63.64	12	66.67	2	33.33	5	45.45	5	100	7	100	0.757
male	10	58.82	9	75.00	8	72.73	11	61.11	4	66.67	6	54.55	4	80.00	5	71.43	0.757
male	9	52.94	12	100	9	81.82	12	66.67	4	66.67	5	45.45	5	100	7	100	1.077
male	17	100	12	100	11	100	18	100	6	100	11	100	5	100	7	100	3.937
male	17	100	12	100	11	100	18	100	6	100	11	100	5	100	7	100	3.937
male	16	94.12	11	91.67	11	100	16	88.89	6	100	10	90.91	5	100	6	85.71	2.577
male	17	100	10	83.33	11	100	16	88.89	6	100	11	100	5	100	5	71.43	2.577
male	16	94.12	11	91.67	10	90.91	17	94.44	5	83.33	11	100	5	100	6	85.71	2.577
male	13	76.47	12	100	10	90.91	15	83.33	5	83.33	8	72.73	5	100	7	100	1.903
male	16	94.12	11	91.67	11	100	16	88.89	6	100	10	90.91	5	100	6	85.71	2.577
male	16	94.12	3	25.00	8	72.73	11	61.11	5	83.33	11	100	3	60.00	0	0.00	0.757
male	17	100	5	41.67	11	100	11	61.11	6	100	11	100	5	100	0	0.00	3.658
male	17	100	10	83.33	11	100	16	88.89	6	100	11	100	5	100	5	71.43	2.577
male	15	88.24	10	83.33	8	72.73	17	94.44	4	66.67	11	100	4	80.00	6	85.71	1.903

male	16	94.12	11	91.67	11	100	16	88.89	6	100	10	90.91	5	100	6	85.71	2.577
male	15	88.24	11	91.67	10	90.91	16	88.89	5	83.33	10	90.91	5	100	6	85.71	2.198
male	17	100	12	100	11	100	18	100	6	100	11	100	5	100	7	100	3.937
male	15	88.24	12	100	10	90.91	17	94.44	5	83.33	10	90.91	5	100	7	100	2.577
male	15	88.24	11	91.67	9	81.82	17	94.44	4	66.67	11	100	5	100	6	85.71	2.198
male	16	94.12	12	100	10	90.91	18	100	5	83.33	11	100	5	100	7	100	3.13
male	13	76.47	2	16.67	7	63.64	8	44.44	5	83.33	8	72.73	2	40.00	0	0.00	0.899
male	17	100	11	91.67	10	90.91	18	100	6	100	11	100	4	80.00	7	100	3.13
male	11	64.71	0	0.00	4	36.36	7	38.89	4	66.67	7	63.64	0	0.00	0	0.00	0.927
male	14	82.35	11	91.67	9	81.82	16	88.89	4	66.67	10	90.91	5	100	6	85.71	1.903
male	12	70.59	9	75.00	8	72.73	13	72.22	4	66.67	8	72.73	4	80.00	5	71.43	1.077
male	15	88.24	10	83.33	10	90.91	15	83.33	6	100	9	81.82	4	80.00	6	85.71	1.903
male	17	100	11	91.67	11	100	17	94.44	6	100	11	100	5	100	6	85.71	3.13
male	10	58.82	11	91.67	8	72.73	13	72.22	4	66.67	6	54.55	4	80.00	7	100	1.077
male	10	58.82	9	75.00	7	63.64	12	66.67	3	50.00	7	63.64	4	80.00	5	71.43	0.757
male	17	100	12	100	11	100	18	100	6	100	11	100	5	100	7	100	3.937
male	10	58.82	11	91.67	8	72.73	13	72.22	3	50.00	7	63.64	5	100	6	85.71	1.077
male	16	94.12	12	100	11	100	17	94.44	6	100	10	90.91	5	100	7	100	3.13
male	16	94.12	12	100	11	100	17	94.44	6	100	10	90.91	5	100	7	100	3.13
male	15	88.24	12	100	9	81.82	18	100	4	66.67	11	100	5	100	7	100	2.577
male	15	88.24	10	83.33	10	90.91	15	83.33	5	83.33	10	90.91	5	100	5	71.43	1.903
male	10	58.82	6	50.00	7	63.64	9	50.00	6	100	4	36.36	1	20.00	5	71.43	0.317
female	9	52.94	11	91.67	9	81.82	11	61.11	4	66.67	5	45.45	5	100	6	85.71	0.913
female	6	35.29	7	58.33	7	63.64	6	33.33	3	50.00	3	27.27	4	80.00	3	42.86	-0.112
female	16	94.12	10	83.33	10	90.91	16	88.89	6	100	10	90.91	4	80.00	6	85.71	2.198
female	11	64.71	11	91.67	9	81.82	13	72.22	4	66.67	7	63.64	5	100	6	85.71	1.253
female	13	76.47	12	100	9	81.82	16	88.89	4	66.67	9	81.82	5	100	7	100	1.903
female	9	52.94	8	66.67	7	63.64	10	55.56	3	50.00	6	54.55	4	80.00	4	57.14	0.46
female	11	64.71	8	66.67	9	81.82	10	55.56	5	83.33	6	54.55	4	80.00	4	57.14	0.757
female	16	94.12	12	100	10	90.91	18	100	5	83.33	11	100	5	100	7	100	3.13

female	10	58.82	11	91.67	8	72.73	13	72.22	3	50.00	7	63.64	5	100	6	85.71	1.077
female	11	64.71	11	91.67	8	72.73	14	77.78	3	50.00	8	72.73	5	100	6	85.71	1.253
female	7	41.18	6	50.00	4	36.36	9	50.00	1	16.67	6	54.55	3	60.00	3	42.86	-0.112
female	12	70.59	11	91.67	9	81.82	14	77.78	4	66.67	8	72.73	5	100	6	85.71	1.445
female	15	88.24	9	75.00	10	90.91	14	77.78	5	83.33	10	90.91	5	100	4	57.14	1.658
female	10	58.82	6	50.00	2	18.18	14	77.78	0	0.00	10	90.91	2	40.00	4	57.14	0.83
female	15	88.24	4	33.33	8	72.73	11	61.11	5	83.33	10	90.91	3	60.00	1	14.29	0.882
female	11	64.71	11	91.67	10	90.91	12	66.67	5	83.33	6	54.55	5	100	6	85.71	1.253
female	13	76.47	8	66.67	9	81.82	12	66.67	5	83.33	8	72.73	4	80.00	4	57.14	1.077
female	16	94.12	11	91.67	9	81.82	18	100	5	83.33	11	100	4	80.00	7	100	2.577
female	12	70.59	7	58.33	9	81.82	10	55.56	5	83.33	7	63.64	4	80.00	3	42.86	0.757
female	10	58.82	6	50.00	2	18.18	14	77.78	0	0.00	10	90.91	2	40.00	4	57.14	0.83
female	10	58.82	8	66.67	7	63.64	11	61.11	3	50.00	7	63.64	4	80.00	4	57.14	0.606
female	13	76.47	0	0.00	6	54.55	7	38.89	6	100	7	63.64	0	0.00	0	0.00	1.443
female	13	76.47	12	100	9	81.82	16	88.89	4	66.67	9	81.82	5	100	7	100	1.903
female	13	76.47	6	50.00	5	45.45	14	77.78	3	50.00	10	90.91	2	40.00	4	57.14	0.757
female	9	52.94	0	0.00	3	27.27	6	33.33	3	50.00	6	54.55	0	0.00	0	0.00	0.478
female	11	64.71	0	0.00	5	45.45	6	33.33	5	83.33	6	54.55	0	0.00	0	0.00	0.927
female	6	35.29	4	33.33	3	27.27	7	38.89	0	0.00	6	54.55	3	60.00	1	14.29	-0.246
female	9	52.94	11	91.67	5	45.45	15	83.33	0	0.00	9	81.82	5	100	6	85.71	1.8
female	10	58.82	11	91.67	5	45.45	16	88.89	0	0.00	10	90.91	5	100	6	85.71	2.176
female	15	88.24	11	91.67	9	81.82	17	94.44	4	66.67	11	100	5	100	6	85.71	2.198
female	9	52.94	11	91.67	6	54.55	14	77.78	2	33.33	7	63.64	4	80.00	7	100	0.913
female	10	58.82	4	33.33	7	63.64	7	38.89	4	66.67	6	54.55	3	60.00	1	14.29	0.032
female	13	76.47	11	91.67	9	81.82	15	83.33	4	66.67	9	81.82	5	100	6	85.71	1.658
female	0	0.00	5	41.67	2	18.18	3	16.67	0	0.00	0	0.00	2	40.00	3	42.86	-0.483
female	12	70.59	11	91.67	7	63.64	16	88.89	3	50.00	9	81.82	4	80.00	7	100	1.445
female	6	35.29	2	16.67	4	36.36	4	22.22	2	33.33	4	36.36	2	40.00	0	0.00	-0.401
female	10	58.82	10	83.33	8	72.73	12	66.67	4	66.67	6	54.55	4	80.00	6	85.71	0.913
female	16	94.12	11	91.67	10	90.91	17	94.44	6	100	10	90.91	4	80.00	7	100	2.577

female	10	58.82	4	33.33	9	81.82	5	27.78	5	83.33	5	45.45	4	80.00	0	0.00	0.704
female	13	76.47	8	66.67	9	81.82	12	66.67	4	66.67	9	81.82	5	100	3	42.86	1.077
female	11	64.71	9	75.00	4	36.36	16	88.89	0	0.00	11	100	4	80.00	5	71.43	1.8
female	12	70.59	11	91.67	8	72.73	15	83.33	3	50.00	9	81.82	5	100	6	85.71	1.445
female	17	100	12	100	11	100	18	100	6	100	11	100	5	100	7	100	3.937
female	13	76.47	2	16.67	5	45.45	10	55.56	3	50.00	10	90.91	2	40.00	0	0.00	0.899
female	9	52.94	8	66.67	4	36.36	13	72.22	0	0.00	9	81.82	4	80.00	4	57.14	1.032
female	10	58.82	12	100	7	63.64	15	83.33	2	33.33	8	72.73	5	100	7	100	1.253
female	7	41.18	9	75.00	7	63.64	9	50.00	3	50.00	4	36.36	4	80.00	5	71.43	0.317
female	10	58.82	8	66.67	7	63.64	11	61.11	3	50.00	7	63.64	4	80.00	4	57.14	0.606
female	10	58.82	7	58.33	6	54.55	11	61.11	3	50.00	7	63.64	3	60.00	4	57.14	0.46
female	8	47.06	9	75.00	6	54.55	11	61.11	3	50.00	5	45.45	3	60.00	6	85.71	0.46
female	6	35.29	6	50.00	7	63.64	5	27.78	2	33.33	4	36.36	5	100	1	14.29	-0.258
female	4	23.53	8	66.67	4	36.36	8	44.44	2	33.33	2	18.18	2	40.00	6	85.71	-0.258
female	4	23.53	10	83.33	7	63.64	7	38.89	2	33.33	2	18.18	5	100	5	71.43	0.032
female	13	76.47	7	58.33	8	72.73	12	66.67	3	50.00	10	90.91	5	100	2	28.57	0.913
female	11	64.71	11	91.67	8	72.73	14	77.78	4	66.67	7	63.64	4	80.00	7	100	1.253
female	4	23.53	8	66.67	5	45.45	7	38.89	2	33.33	2	18.18	3	60.00	5	71.43	-0.258
female	7	41.18	4	33.33	5	45.45	6	33.33	4	66.67	3	27.27	1	20.00	3	42.86	-0.407
female	8	47.06	11	91.67	7	63.64	12	66.67	2	33.33	6	54.55	5	100	6	85.71	0.757
female	10	58.82	11	91.67	9	81.82	12	66.67	4	66.67	6	54.55	5	100	6	85.71	1.077
female	11	64.71	11	91.67	10	90.91	12	66.67	5	83.33	6	54.55	5	100	6	85.71	1.253
female	7	41.18	11	91.67	9	81.82	9	50.00	4	66.67	3	27.27	5	100	6	85.71	0.606
female	10	58.82	3	25.00	5	45.45	8	44.44	3	50.00	7	63.64	2	40.00	1	14.29	-0.112
female	9	52.94	3	25.00	8	72.73	4	22.22	5	83.33	4	36.36	3	60.00	0	0.00	0.335
female	17	100	8	66.67	10	90.91	15	83.33	6	100	11	100	4	80.00	4	57.14	1.903
female	10	58.82	7	58.33	7	63.64	10	55.56	4	66.67	6	54.55	3	60.00	4	57.14	0.46
female	4	23.53	1	8.33	3	27.27	2	11.11	2	33.33	2	18.18	1	20.00	0	0.00	-1.511
female	14	82.35	4	33.33	6	54.55	12	66.67	4	66.67	10	90.91	2	40.00	2	28.57	0.606
female	4	23.53	5	41.67	7	63.64	2	11.11	4	66.67	0	0.00	3	60.00	2	28.57	-0.725

female	12	70.59	8	66.67	8	72.73	12	66.67	4	66.67	8	72.73	4	80.00	4	57.14	0.913
female	11	64.71	8	66.67	6	54.55	13	72.22	3	50.00	8	72.73	3	60.00	5	71.43	0.757
female	13	76.47	11	91.67	9	81.82	15	83.33	4	66.67	9	81.82	5	100	6	85.71	1.658
female	12	70.59	4	33.33	7	63.64	9	50.00	3	50.00	9	81.82	4	80.00	0	0.00	1.107
female	10	58.82	5	41.67	7	63.64	8	44.44	4	66.67	6	54.55	3	60.00	2	28.57	0.174
female	14	82.35	9	75.00	8	72.73	15	83.33	4	66.67	10	90.91	4	80.00	5	71.43	1.445
female	16	94.12	11	91.67	10	90.91	17	94.44	5	83.33	11	100	5	100	6	85.71	2.577
female	11	64.71	9	75.00	9	81.82	11	61.11	5	83.33	6	54.55	4	80.00	5	71.43	0.913
female	14	82.35	7	58.33	5	45.45	16	88.89	3	50.00	11	100	2	40.00	5	71.43	1.077
female	9	52.94	4	33.33	7	63.64	6	33.33	5	83.33	4	36.36	2	40.00	2	28.57	-0.112
female	13	76.47	8	66.67	7	63.64	14	77.78	4	66.67	9	81.82	3	60.00	5	71.43	1.161
female	8	47.06	5	41.67	5	45.45	8	44.44	1	16.67	7	63.64	4	80.00	1	14.29	0.124
female	13	76.47	12	100	11	100	14	77.78	6	100	7	63.64	5	100	7	100	1.903
female	8	47.06	11	91.67	8	72.73	11	61.11	3	50.00	5	45.45	5	100	6	85.71	0.757
female	13	76.47	10	83.33	8	72.73	15	83.33	4	66.67	9	81.82	4	80.00	6	85.71	1.445
female	14	82.35	9	75.00	9	81.82	14	77.78	5	83.33	9	81.82	4	80.00	5	71.43	1.445
female	4	23.53	10	83.33	5	45.45	9	50.00	1	16.67	3	27.27	4	80.00	6	85.71	0.032
female	7	41.18	11	91.67	6	54.55	12	66.67	1	16.67	6	54.55	5	100	6	85.71	0.733
female	17	100	12	100	11	100	18	100	6	100	11	100	5	100	7	100	3.937
female	14	82.35	11	91.67	11	100	14	77.78	6	100	8	72.73	5	100	6	85.71	2.474
female	0	0.00	5	41.67	5	45.45	0	0.00	0	0.00	0	0.00	5	100	0	0.00	1.376
female	17	100	10	83.33	10	90.91	17	94.44	6	100	11	100	4	80.00	6	85.71	2.577
Possible	17		12		11		18		6		11		5		7		
Means	11.3	66.4	9	74.8	7.8	71.0	12.5	69.1	3.7	62.3	7.5	68.6	4.1	81.5	4.9	69.9	
S.D.	3.9	22.8	3	25.3	2.3	20.8	4.02	22.3	1.7	28.8	2.9	25.9	1.2	24.5	2.2	30.7	

Appendix 2

Sub-test [ST01]: gender and all multiple choice question performance

Anova: Single Factor: Comparing male and female all multiple-choice responses.

SUMMARY

Groups	***Count***	***Sum***	***Average (%)***	***Variance***	***s.d.***
Male MC scores	94.00	7181.82	76.40	391.57	19.8
Female MC scores	90.00	5890.91	65.45	416.38	20.4

ANOVA

Source of Variation	***SS***	***df***	***MS***	***F***	***P-value***	***F crit***
Between Groups	5510.67	1.00	5510.67	13.65	0.00029	3.89
Within Groups	73474.24	182.00	403.70			
Total	78984.91	183.00				

Appendix 3:

Subtest [ST02]: gender and all short-answer question performance

Anova: Single Factor: Comparing male and female all short-answer responses.

SUMMARY

Groups	***Count***	***Sum***	***Average (%)***	***Variance***	***s.d.***
Male SA scores	94.00	7005.56	74.53	407.81	20.2
Female SA scores	90.00	5716.67	63.52	535.48	23.1

ANOVA

Source of Variation	***SS***	***df***	***MS***	***F***	***P-value***	***F crit***
Between Groups	5572.14	1.00	5572.14	11.85	0.00072	3.89
Within Groups	85583.92	182.00	470.24			
Total	91156.07	183.00				

Appendix 4:

Sub-test [AB]: gender and all recall (acid-base) question performance

Anova: Single Factor: Comparing male-female all acid-base scores

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*	*s.d.*
Male Acid-base scores %	94.00	6623.53	70.46	542.46	23.3
Female Acid-base scores %	90.00	5594.12	62.16	467.36	21.6

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	3172.19	1.00	3172.19	6.27	0.0131	3.89
Within Groups	92043.51	182.00	505.73			
Total	95215.70	183.00				

Appendix 5:

Sub-test [ST]: gender and all application (stoichiometry) question performance

Anova: Single Factor: Comparing male and female all stoichiometry scores

SUMMARY

Groups	*Count*	*Sum*	*Average (%)*	*Variance*	*s.d.*
Male Stoichiometry scores	94.00	7708.33	82.00	460.43	21.4
Female Stoichiometry scores	90.00	6050.00	67.22	725.34	26.9

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	10045.68	1.00	10045.68	17.03	0.00006	3.89
Within Groups	107375.44	182.00	589.97			
Total	117421.12	183.00				