Which Form of Assessment Provides the Best Information about Student Understanding of Chemistry?

Ross D. Hudson

Australian Council for Educational Research and Curtin University

David F. Treagust

Science and Mathematics Education Centre, Curtin University. Perth Australia

Background

This study developed from observations of apparent achievement differences between male and female chemistry performances in a State University Entrance examination. Male students performed more strongly than did female students especially in the higher scores. Apart from the gender of the students, two other important factors that might influence student performance were format of questions (short-answer or multiple-choice) and the type of questions (recall or application).

Purpose

The research question addressed in this study was: Is there a relationship between performance in State University Entrance examinations in chemistry and school chemistry examinations and student gender, format of questions - multiple-choice or short-answer, and conceptual level - recall or application?

Sample

Two sources of data were (1) secondary analyses of five consecutive years' data published by the examining authority of Chemistry examinations and (2) tests conducted with 192 students which provided information about all aspects of the three variables (question format, question type and gender) under consideration.

Design and methods

Both sources of data were analysed using ANOVA to compare means for the variables under consideration and the statistical significance of any differences. The tests data were also analysed using Rasch analysis to determine differences in gender performance.

Results

- Both male and female students performed better on multiple-choice-questions and recall questions than on short-answer questions and application questions respectively.

- Male students outperformed female students in both the university entrance and school tests, particularly in the higher scores.
- Rasch analysis showed that there was little difference in performance between males and females of equal ability despite significant mean differences in favour of male performance in the tests.

Conclusions

Both male and female students generally perform better on multiple-choice questions than they do on short-answer questions. However, when the questions are matched in terms of difficulty (using Rasch analysis) the differences in performance between multiple-choice and short-answer are quite small. Rasch analysis showed that there was little difference in performance between males and females of equal ability. This study shows that a simple face value score analysis of relative student performance – in this case in chemistry - can be deceptive unless the actual abilities of the students concerned, as measured by a tool such as Rasch, are taken into consideration before reaching any conclusion.

Keywords: chemistry, assessment, question type, gender

Which Form of Assessment Provides the Best Information about Student Understanding of Chemistry?

This study initially developed from the first author's observations, along with those of a number of his teaching colleagues, that there were apparent achievement differences between the student performances in the State University Entrance (SUE) examination in Chemistry. The typical SUE examination of 1.5 hours length comprises 20 multiple-choice questions and approximately eight short-answer/extended response questions with one examination per semester. The type of learning expected in the SUE examinations focused on the students' ability to recall facts, processes and related equations and/or their ability to apply their general conceptual knowledge to a particular situation, usually in the form of calculation application questions. Both forms of question are assessed in the multiple-choice section and the short-answer section of each examination. The content covered in each semester examination was different. The first semester examination included mostly stoichiometry, chemical instrumentation and equilibrium. The second semester examination covered energy, food chemistry and the history of and trends in the periodic table (VCAA, 2003-2007). The effectiveness of student responses to each form of question involving recall or application categories was noted in anecdotal observations of the student performance in the examinations and their stated preferences for the different examinations. In particular, female students achieved higher grades in the second semester examination whereas male students were more successful in the first semester examination even though the cohort of participating students in both semesters is virtually identical. Why should this be occurring? Were any observed differences actually significant?

A number of factors were relevant in considering these questions. Apart from the gender of the students, the two other important factors influencing the performance were the

format of the question,(short-answer or multiple-choice) and the conceptual level (recall or application).

The research question addressed in this study was: Is there a relationship between performance in State University Entrance examinations in chemistry and school chemistry examinations and student gender, format of questions - multiple-choice or short-answer, and conceptual level - recall or application? With a number of factors influencing student performance, this study endeavoured to determine if one factor was more significant in determining student performance.

Literature Perspectives

Question type

In reviewing the differences between multiple-choice and short-answer questions and gender issues, researchers' present mixed results. Advantages of multiple-choice questions are that they generally have less scope and complexity than short-answer questions and are likely to be less difficult. However, it is possible to have complex multiple-choice questions that do assess higher order skills whilst it is similarly possible to construct open ended questions that are essentially no more complex that simple recall (Hartman & Lin, 2011). Several authors (Bridgeman, 1992; Martinez, 1991) conclude 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 whereas multiple-choice question answers give no indication of how students arrived at their answer. Multiple-choice questions have a number of advantages, being easy and quick to score making them popular with both teachers and educational authorities (Dufresne, Leonard, & Gerace, 2002; Holder & Mills, 2001). Furthermore, multiple-choice questions do not disadvantage students with weaker

writing and spelling skills (Zeidner, 1987) as much as do short-answer questions. Equally, there are perceived disadvantages. Multiple-choice questions take more time to write correctly (Brown, Bull, & Pendlebury, 1997) and such tests tend to favour recall learning over applied learning (Martinez, 1999). Another well-known criticism of multiple-choice questions is that they allow students to guess at answers with a not impossible probability of getting the question correct (Barnett-Foster & Nagy, 1996; Bridgeman, 1992). 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).

Content

Student responses to stoichiometric questions seemed to produce different outcomes depending on whether the question was presented as multiple-choice or short-answer (Niaz & Robinson, 1995). Students were apparently influenced by the multiple-choice options and consequently provided correct answers to questions that may not have been answered correctly if presented in a short-answer form. Indeed, students commonly had difficulty in attempting to justify or explain the selection of response in any particular multiple-choice question (Barnett-Foster & Nagy, 1996). Recent studies have indicated that the content style of the questions (traditional or conceptual) does influence performance though not greatly (Holme & Murphy, 2011) and that performance on multiple-choice algorithmic questions, not unsurprisingly, decreases with the number of algorithmic steps involved in deriving the answer (Hartman & Lin, 2011).

Gender

The role of gender in performance in chemistry and other subject areas in general has precipitated a variety of studies. Boli, Allen, and Payne (1985) explored the reasons behind the differences that were observed between the genders in undergraduate chemistry and

mathematics courses. Their exploration sought reasons behind why male students tended to outperform the female cohort, resulting in the suggestion that differences in mathematical ability were a very important consideration. 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 and also a lower anticipation of success (Buccheri, Gurber, & Bruhwiler, 2011).

A number of studies have supported the observation that male students usually outperform female students in assessments particularly in the areas of mathematics and science (Abedlaziz, Ismail, & Hussin, 2011; Korporshoek, Kuyper, van der Werf, & Bosker, 2011). The analysis of a number of large assessments has demonstrated that male students generally performed better than did female students (Beller & Gafni, 1991). 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; Halpern, 1997; Hamilton, 1998; Hedges & Howell, 1995; Linn, Baker, & Dunbar, 1991). Beller and Gafni (2000) and Hamilton (1998) noted that short-answer questions tended to favour female students whereas multiple-choice favoured male students, most probably due to females having better literacy skills than males. They also noted that in terms of higher order questions, males tended to outperform females regardless of the question type (Beller & Gafni, 2000; Bridgeman & Lewis, 1994; Hamilton, 1998; Lumsden & Scott, 1987).

Methodology

Data collection procedures and analyses

In this study, two main analyses were performed. Firstly, past SUE examinations covering the chemistry study design that operated between 2003 and 2007 were analysed (see http://www.vcaa.vic.edu.au/Pages/vce/studies/chemistry/exams.aspx). Student performance on these questions was analysed after all the questions were categorised according to whether they were multiple-choice or short-answer and as to whether they were primarily assessing recall or primarily assessing application in terms of the content being questioned. The determination of the category of each question was fundamental to this research. The first author initially determined the categories based on his many years of experience as a secondary school chemistry teacher. To assist with the classification the following qualifications were used to assign questions. Firstly, questions where the answers could not be learned beforehand and/or required calculations were assigned as application questions. Secondly, questions that could be learned as a simple fact or as an equation or structure were assigned as recall questions. Thirdly, questions that involved some degree of application and some degree of recall were allocated on the basis of the teachers' assessment as to which skill would likely play the larger part in the student's ability to answer the question. After this classification, two experienced chemistry teachers used the same classification system. A period of consultation followed whereby the first author and the two teachers reached agreement on each question. This procedure was easier with some questions than others.

A further analysis based on the SUE examination data was centred on the grade distribution data supplied for each examination. The grade distribution data shows the number of students (both as percentage and actual numbers) of male and female students awarded each grade A+ through to E. Secondly, a sample of students (197) participated in a sample test program to gain first hand data which was subjected to more extensive analysis than was possible with the SUE data.

The analysis of the past SUE examination papers and the sample testing of students in the testing program involved some interpretation but was mostly a quantitative statistical analysis to aid interpretation of the findings. The analysis of the data, particularly that of the past SUE papers was most clearly identifiable as a post hoc analysis (Myers & Well, 2003). Much of the analysis was by either ANOVA or Chi-squared analysis where the various permutations of the three different variables, question type (multiple-choice or short-answer), question content (recall or application) and genders were compared statistically for any significant differences.

The second study was of the tests conducted by the first researcher with the cooperation of the four secondary schools (two co-educational and one all male and one all female) located in an upper middle class suburb and noted for academic achievement. Virtually all students studying chemistry at the four schools participated in the study. The applicability or transferability (Anderson, 1998; Cohen, Manion, & Morrison, 2011) of the results is therefore limited to similar schools with students of similar socio-economic backgrounds. The four schools have traditionally performed well in the SUE Chemistry examinations. This was useful as it allowed the testing program to be conducted with motivated high achieving students. The importance of this lies in the fact that it is mostly at the higher grades of A^+ and A where the most notable differences in performance of males and females occur (see Figure 1). The sample students for this study provided a good representation of that high scoring group. <<*Figure 1 about here*>>

Tests conducted with 192 students provided information about student performance on two versions of essentially the same question but asked in different forms (multiple-choice or short-answer) for all aspects of the three variables (question type, question content and gender) under consideration.

Each test was designed to have identical structure and degree of difficulty when assessing in the different formats. The multiple-choice question test and short-answer questions test had the same questions (with minor changes of numbers and formulae) but were obviously structured in the different formats. To attempt to determine whether the question content or the question format were factors nearly identical questions on a particular topic point were constructed so that the multiple-choice version and the short-answer version of a pair of questions were evenly matched in terms of the content difficulty.

Example of paired difficulty - acid base (recall) question

From the Short-answer Acid Base Test, Question 3

3. Identify a conjugate pair in the following equation: (indicate which is the acid and which is the base for the conjugate pair)

(a) $HPO_4^{2-}{}_{(aq)} + H_2O_{(l)} \rightarrow OH^-_{(aq)} + H_2PO_4^-{}_{(aq)}$

From the Multiple-choice Acid Base Test Question 3

3. Consider the following equation:

$HS^-_{(aq)} + NH_{3\,(aq)} \rightarrow NH_4^+{}_{(aq)} + S^{2-}_{(aq)}$

Which of the following is an acid-base conjugate pair?

A. HS^- *and* NH_3

B. HS^- *and* S^{2-}

C. NH_3 *and* S^{2-}

D. NH_4^+ *and* HS^-

Both these questions examine essentially the same aspect of chemistry to a similar level of difficulty. Students attempting both questions need to be able to correctly identify the conjugate acid –base pairs to be able to answer the question. Ideally, the performance should be similar on both questions if the questions are measuring student ability equally. This information was important because the analysis of past SUE papers did not easily produce this sort of information. The tests offered a unique opportunity to examine this relationship as well as gender performance on the examinations. The methodology comprised quantitative

analysis (ANOVA analysis of student performance based on question type and content and Chi-squared analysis of student performance by gender) and interpretation (Anderson, 1998) throughout the analyses of the SUE examinations and the tests.

Results

The results from the State University Entrance examination analysis involved ANOVA analysis based on the classification of the questions and chi-squared analysis of the grade distribution data. These ANOVA results, summarised in Table 1, showed that

- Students generally performed better on multiple-choice-questions than on short-answer questions
- Students performance on recall questions was generally better than on application questions
- There was little difference between performances on the two different semester examinations except that students performed better on the short answer but better on the multiple-choice questions.

<<Table 1 about here>>

The analysis of the grade distribution data demonstrated that male students were performing very well in the higher grades in the chemistry examination. A typical grade distribution graph as shown in Figure 1 shows that male students are being awarded a greater proportion of the higher grades A+, A and B+ than are the female students. A statistical analysis (Chi squared) of all years' data focussed on three example grades across the range available namely A+, C+ and E+. The Chi squared analysis (see Table 2) showed:

- Male students outperformed female students, particularly at the higher grades.
- The difference in performance was most pronounced in the Unit 3 examination.

<<Table 2 about here>>

Rasch analysis of the test questions

Rasch analysis was conducted using software program RUMM2030. The Rasch analysis allows the comparison of different variables in an assessment test but also allows an analysis of, and differences between, item types. Rasch analysis can also measure and detect any gender differential performance in the test. Rasch analysis indicates how well a test conforms to the test construct, that is, does the test measure what it was intended to and is there any apparent bias in the test design (RUMM Laboratory, 2009b). After the tests were collected from the schools, they were marked and coded according to a coding framework. After coding had been completed, a Rasch analysis was performed on the questions to ascertain any discrepant questions that did not match the expected Rasch distribution model (Bond & Fox, 2007). In the analysis of the tests of the 192 students (99 males and 93 females) taking part only three were deleted from the Rasch analysis due to poor fit statistics.

The initial analysis using the RUMM2030 program showed that the tests, as a group, had both strengths and weaknesses. The item and person fit residuals (1.152 and 0.875) and the reliability index, the Person Separation Index (PSI), of 0.803 were all very good (Cavanagh, Romanoski, Giddings, Harris, & Dellar, 2003). The item-trait interaction measure (based on Chi-squared analysis) was, however, very low (Pr = 0.000006) suggesting that an unacceptable level of dependence existed within the item set of questions. This dependence may have been due to either poorly fitting items or poorly fitting persons or both. Closer analysis was required to ascertain the root cause of the poor interaction value so that measures could be taken to adjust the test construct (or students) to give a probability value for the person item interaction that was closer to the acceptable probability of 0.002 (Bonferroni adjusted)[1]. << *Footnote 1 on this page please*>> Adjustment of the test items took the form

of rescoring any items with reverse threshold characteristics and deleting any students whose fit residuals suggested that their performance was unreliable. Four questions were rescored and only three students deleted from the sample.

Final test construct

Following these adjustments the statistics from RUMM2030 provided good values (Cavanagh et al., 2003) for item and person fit residuals (1.136 and 0.867) and reliability index or Person Separation Index (PSI) (0.785). The item-trait interaction measure (based on Chi-squared analysis was, however, still low (Pr = 0.00005). Analysis of the item fit characteristics showed that this was most likely due to the individual probability of one item (ST05) which had a very low probability of 0.00001. As this item could not be rescored, the decision was made to remove this item from the sample set.

After removing item ST05, the sample size of items was reduced to 23 items. The summary statistics, however, vindicated the decision. Item fit residual (1.099) and the person fit residual (0.843) were improvements from the original item summary statistics. The PSI reliability index was 0.795 and most importantly the item-trait probability was now 0.124 (Cavanagh et al., 2003). This meant that the adjusted test item structure gave a valid item set demonstrating unidimensionality in the items (Cavanagh et al., 2003), a necessary condition to validate Rasch analysis (RUMM Laboratory, 2009a).

A significant factor in the use of the tests was the correlation relationship between the multiple-choice and the short-answer items. The intention of using the tests was to determine if there was any relationship between student performances on the items that were asked in both formats. That is, if the same (or in this case, nearly the same) item was asked as both a multiple-choice and as a short-answer question, would the student performance be the same or would one version be answered more successfully? To test this premise, the 22 questions

(items ST05 and ST11 were not included due to item ST05 being deleted) were arranged in ordered pairs. For example Question ST01 (multiple-choice) and ST07 (short-answer) were both questions that asked students to calculate a percentage composition. As a measure of the performance on each item, the individual location measure of relative difficulty (item location) was used. This value indicates the difficulty of the item relative to the other items in the tests (RUMM Laboratory, 2009a).

The mean values for location (see Table 3) indicate that students found the multiple-choice questions (mean location = –0.051) marginally easier than the short-answer questions (mean location = 0.048). This, however, is only a very small difference suggesting that the items were well matched overall.

<< Table 3 about here>>

When these ordered pairs (e.g. AB01:AB07) were plotted as a scatter plot, the relationship between the paired questions is evident (Figure 2). The trend in the relationship between the two variables suggests that, as the difficulty of the multiple-choice version of the question increased, the difficulty of the short-answer version also increased. The Pearson r correlation coefficient was R=0.72, indicating a reasonable correlation between the variables and accounting for nearly 52% of the variance between the variables.

<< Figure 2 about here>>

The graph in Figure 2 essentially showed that the easiest multiple-choice question was also the easiest short answer question and so on. The score distribution (Figure 3) was very similar to that of the grade distribution of the SUE examinations (Figure 1) suggesting that conclusions drawn from the tests were likely to have application to the larger State University Entrance cohort. *<< Figure 3 about here>>*

The mean performance by males and females is shown in Table 4, showing that, as with the SUE examinations, males were performing at a higher level than were the female students (mean (males) = 78.3 and mean (females) = 68.6).

<< Table 4 about here>>

The analysis of the tests was conducted using ANOVA as were the SUE examination papers. The results of this analysis are shown in Table 5. The differences apparent between the SUE examination analysis and that of the test analysis are also included in Table 5 which shows the comparison with the results from the SUE analysis. It is important to note that many of the findings demonstrated in the SUE analysis were generally replicated in the test analysis. *<< Table 5 about here>>*

Question type and content Rasch analysis

The Rasch analysis allowed comparison of the test students' responses to these questions on a number of levels. The initial analysis was of student performance on all the multiple-choice questions compared to that on all short-answer questions. This analysis was achieved using the final test structure and using the equating test option of RUMM2030 (RUMM Laboratory, 2009a). Two examples of this analysis are shown below.

Comparison of all multiple-choice with all short-answer questions. The ANOVA test results show that there was only a small difference in performance when comparing the multiple-choice answers to the short-answer responses and the difference was not statistically significant ($F(1,366) = 0.72$; $p > 0.05$). The performance on the multiple-choice questions, based on the means, was greater than on the short-answer questions: mean (multiple-choice) = 71.0, standard deviation = 20.8 compared to mean/standard deviation (short-answers) = 69.1/22.3. This difference between the means is small. The RUMM2030 graphical analysis

(Figure 4) shows several interesting aspects of students' abilities with respect to answering multiple-choice and short-answer questions. << *Figure 4 about here*>>

Students in the lower ability ranges found short-answer questions slightly more difficult than multiple-choice questions. However, at the higher end of student abilities the difference between performances was negligible. This result may 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 and the student would be more able to obtain full credit for his or her efforts, so the narrowing observed differences at the top of the ability range is understandable.

Comparison of performance on recall and application, all questions. Students' performance on recall questions (mean = 66.4 and standard deviation = 22.8) was weaker than on application questions (74.8 /25.3). The ANOVA results show a statistically significant difference $(1,366) = 11.01$; $p < 0.001$ in performance on recall questions compared to application questions. This result is at odds with the results of the analysis of the SUE examinations that showed the opposite to be the case.

The difference is likely due to the impact of the difference in performance on the multiple-choice questions (see Table 5) where the response to multiple-choice stoichiometry questions was unexpectedly high. Figure 5 shows the RUMM2030 analysis for the two question categories with adjustment for student abilities. The difference is quite marked in that students of equal ability were more likely to score better on the stoichiometry questions than on the acid-base questions. << *Figure 5 about here*>>

Gender difference Rasch Analysis

Rasch analysis of the tests allowed a comparison using Differential Item Functioning (DIF) analysis of relative male and female performances on different subsets of questions (RUMM Laboratory P/L, 2009).

The subset of short-answer questions. Male performance was significantly different ($F(1,182) = 11.85$; $p < 0.001$) to female performance on the short-answer questions as shown by the means for males (mean = 74.5, standard deviation = 20.2) compared to females (63.5/23.2). However, when this was analysed using Rasch, the graph was enlightening (see Figure 6) showing that for males and females of equivalent ability, performance on short answer questions was essentially the same despite the means showing a substantial dominance by the male students. *<< Figure 6 about here>>*

Each of the DIF analyses (involving multiple-choice, short-answer, application or recall) showed similar outcomes, that is male performance measured by means was significantly greater than female performance yet when ability was considered there was essentially no difference. This finding provides some important information with respect to the relationship between student chemistry performance and gender. A likely explanation for this is that a greater proportion of higher achieving males choose to study chemistry than do higher achieving females. Presumably, the higher achieving females choose to study other subjects. If this conclusion is applied to the State University Entrance cohort it goes someway to explaining the higher achievement of the male students at the top three grades of chemistry.

Conclusions

The purpose of this research was to address the question: Is there a relationship between performance in State University Entrance examinations in chemistry and school

chemistry examinations and student gender, format of questions - multiple-choice or short-answer, and conceptual level - recall or application? The results suggest that some definite conclusions can be stated; however, there is some qualification once the results of the Rasch analysis are considered. Both male and female students generally do perform better on multiple-choice questions than they do on short-answer questions; however, when the questions are matched in terms of difficulty the differences in performance are quite small. This is an important finding because it provides a different viewpoint to the strong belief that males are apparently more able in the more technically demanding sciences such as Chemistry and Physics than are females (Beller & Gafni, 2000; Bridgeman & Lewis, 1994; Cox et al., 2004; Hamilton, 1998; Lumsden & Scott, 1987). However, it is clear that the multiple-choice questions in the SUE Chemistry examinations are of a lesser difficulty than the short-answer questions (Table 4). A wider ranging and more extensive study of the SUE examinations in terms of the relative difficulties of the two question types would be needed to more definitively answer this question.

Implications for Student Motivation

A concern of this research was how the performance in chemistry would impact on student motivation. It has been well recorded that male students both prefer sciences like chemistry and physics and have generally been shown to outperform female students (Beller & Gafni, 1991; Cox et al., 2004; Hamilton, 1998; VCAA, 2009a). These findings, which are unlikely to encourage female participation in these subjects, generally show that male students have a clear domination in the awarding of the higher grades (Buccheri et al., 2011). Were this information to be taken at face value then the motivation of female students to enrol in chemistry may be diminished and could damage efforts to promote participation in the sciences by female students. However, it appears that there may be an explanation offered from the analysis of the gender performance. Initial analysis showed that the males

statistically outperformed the females as may have been expected in terms of previous research; however, the Rasch gender differential analysis (RUMM laboratory P/L, 2009) showed that when latent student abilities are included there was little difference between the performance of the male and female students, that is, male and female students of equal ability performed very similarly in chemistry. This suggests that the male students choosing chemistry include a greater proportion of high achieving males compared to the proportion of high achieving females choosing chemistry and would account for the males performing better in terms the number of A^+ and A grades awarded.

Suggestions for future research

The results of this research provide the framework for further analysis, particularly into the performance of males and females in different state university examinations. The most informative outcome was the possibility that the differences in performance of the male students compared to female students may be due more to the abilities of the students actually choosing the subject than it has to do with either the nature of the assessment (test structure issues) or any latent ability advantage that male students have over female students. That notwithstanding, further deeper analysis of additional state university examination data may provide some useful insights into this issue.

The limited size of the sample and the low stakes nature of the tests placed some limitations on the transferability of the results. However, sufficient information was gained to suggest the need for a wider scale test program, which analysed performance on matched multiple-choice and short-answer items to clearly determine whether the item type influences success or demonstration of chemistry understanding. Whilst there were substantial similarities between the state university examination analysis and the test analyses, the variations mentioned suggest that a more detailed analysis and larger numbers of students

taking the tests may give more certainty to the observation that students perform better on recall questions. Such a test program would also help provide insights into differing performances of the male and female students. This line of examination may lead to findings that will ultimately suggest an examination structure that more evenly assesses the performances of male and female students. These findings have applicability to other states within Australia and in other countries that depend on multiple choice and short answer questions to assess students for university entrance.

Acknowledgements

The authors wish to acknowledge the assistance of Dr Julia Pallant (University of Melbourne) and Mr Nick Connolly (Australian Council for Educational Research) for providing guidance and assistance in the use and interpretation of the RUMM2030 software and outputs.

References

Abedlaziz, N., Ismail, W., & Hussin, Z. (2011). Detecting a gender-related DIF using logistic regression and transformed item difficulty. *US-China Education Review B 5,* 734-744.

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

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

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.

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.

Bond, T., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. London, UK: Routledge.

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.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th ed.). London: Routledge Falmer.

Dufresne, R. J., Leonard, W. J., & Gerace, W. J. (2002). Making sense of students' answers to multiple-choice questions. *The Physics Teacher, 40*(174-180).

Frederickson, N. (1984). Implications for cognitive theory for instruction in problem solving. *Review of Educational Research, 54*, 363-407.

Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52*, 1091-1102.

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.

Hartman, J. R., & Lin, S. (2011). Analysis of student performance on multiple-choice questions in general chemistry. *Journal of Chemical Education, 88,* 1223-1230.

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

Holder, W. W., & Mills, C. N. (2001). Pencils down, computer up: The new CPA exam. *Journal of Accountancy, 191*(3), 57-60.

Holme, T., & Murphy, K. (2011). Assessing conceptual and algorithmic knowledge in general chemistry with ACS exams. *Journal of Chemical Education, 88,* 1217-1222.

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.

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

Martinez, M. (1991). A comparison of multiple choice and constructed figural response items. *Journal of Educational Measurement, 28,* 131-145.

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

Myers, J. L., & Well, D. W. (2003). *Research design and statistical analysis.* (2nd. ed.). Mahwah, NJ: Erlbaum.

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.

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

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

RUMM Laboratory. (2009a). Getting started with RUMM2030, Duncraig, WA: RUMM Laboratory P/L.

RUMM Laboratory. (2009b). Interpreting RUMM2030, Duncraig, WA: RUMM LaboratoryP/L.

Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student's perspective. *Journal of Educational Research, 80*(6), 352-358.

Footnotes

[1]The Bonferroni adjustment is calculated against a base probability value of 0.05 as the minimum acceptable value for dependence within a set of test questions that are intended to be unidimensional. The calculation for this set of items is Pr (Bonferroni = 0.05 / 24 (number of items) = 0.00208. (Pallant, 2007)

Table 1.ANOVA and Eta-squared analysis of question performance by question type and classification SUE Chemistry 2003-2007

					ANOVA results						
Comparison variables	**Mean**	**Standard deviation.**	**Discriminating variable**	**N**	**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig (p)**	**Eta**	**Eta^2**
Multiple-choice Short-answer	62.2 58.9	15.1 16.0	Application and Recall Questions	487	1280.94	1	1280.94	5.24	0.023	0.103	0.011
Multiple-choice Short-answer	56.2 57.1	13.6 16.2	Application Questions	257	55.36	1	55.36	0.24	0.624	0.031	0.001
Multiple-choice Short-answer	69.3 60.9	13.8 15.5	Recall Questions	230	3931.96	1	3931.96	17.80	0.000	0.269	0.072
Recall Application	64.2 56.7	15.4 15.2	Multiple-choice and Short-answer questions	487	6845.76	1	6845.76	29.37	0.000	0.239	0.057
Recall Application	69.3 56.2	13.8 13.6	Multiple-choice Questions	200	8550.55	1	8550.55	45.71	0.000	0.433	0.188
Recall Application	60.9 57.1	15.5 16.2	Short-answer questions	287	1001.59	1	1001.594	3.96	0.047	0.117	0.014

Table 2. A^+, C^+ and E^+ grade results of Chi-squared analysis (males compared to females) of the Unit 3 and Unit 4 examinations for 2003-2007

	A^+ grade	C^+ grade	E^+ grade
Examination	Chi-squared	Chi-squared	Chi-squared
2003 Unit 3	27.464*	1.207	1.361
2003 Unit 4	7.230	7.545	1.600
2004 Unit 3	23.300*	5.851	0.306
2004 Unit 4	7.150	1.130	1.398
2005 Unit 3	36.46*	4.654	1.084
2005 Unit 4	4.820	1.942	0.194
2006 Unit 3	52.142*	0.322	8.251#
2006 Unit 4	33.072*	0.304	0.007
2007 Unit 3	54.765*	5.697	5.350
2007 Unit 4	18.767*	3.922	1.530

*: $p < 0.001$ # $p < 0.05$

Table 3. Individual question measured by location difficulty (from RUMM2030)

Multiple-choice questions		Short-answer questions	
Q no.	ability location	Q no.	ability location
AB01	-0.636	AB07	0.27
AB02	1.265	AB08	-0.212
AB03	0.412	AB09	0.418
AB04	0.385	AB10	0.032
AB05	-0.061	AB11	0.335
AB06	1.366	AB12	0.937
ST01	-0.3	ST07	0.585
ST02	-1.356	ST08	-1.189
ST03	-0.814	ST09	-0.349
ST04	-0.31	ST10	-1.14
ST06	0.512	ST12	1.175
Mean	**-0.051**		**0.048**

Table 4. Gender differences on the chemistry tests (means)

Groups	*Count*	*Sum*	*Mean*	*s.d.*
Male %	92	7364.2	78.3	15.8
Female %	90	6174.4	68.6	17.3

Table 5. ANOVA analysis of question performance by question type and classification Chemistry tests (N = 368)

Comparison variables	Mean	Standard deviation	Discriminating variable	Sum of Squares	F-value	p (sig)	Similarity to State University Entrance analysis
Multiple-choice Short-answer	81.5 70.0	24.5 30.7	Application Questions	12312	15.95	0.000	Generally similar pattern
Multiple-choice Short-answer	62.3 68.6	28.8 26.0	Recall Questions	3660.3	4.86	0.028	Average score on multiple-choice higher in State University Entrance examinations
Multiple-choice Short-answer	71.0 69.1	20.8 22.3	Application and Recall Questions	333.84	0.72	0.39	Generally similar pattern with some small differences
Recall Application	62.3 81.5	28.8 24.5	Multiple-choice Questions	33925	47.4	0.000	Result was the reverse in the State University Entrance examination
Recall Application	68.6 69.9	26.0 30.7	Short-answer questions	162.0	0.20	0.65	Very similar result to State University Entrance examination
Recall Application	66.4 74.8	22.8 25.3	Multiple-choice and Short-answer questions	6450.3	11.10	0.001	Result was the reverse in the State University Entrance examination

Figures

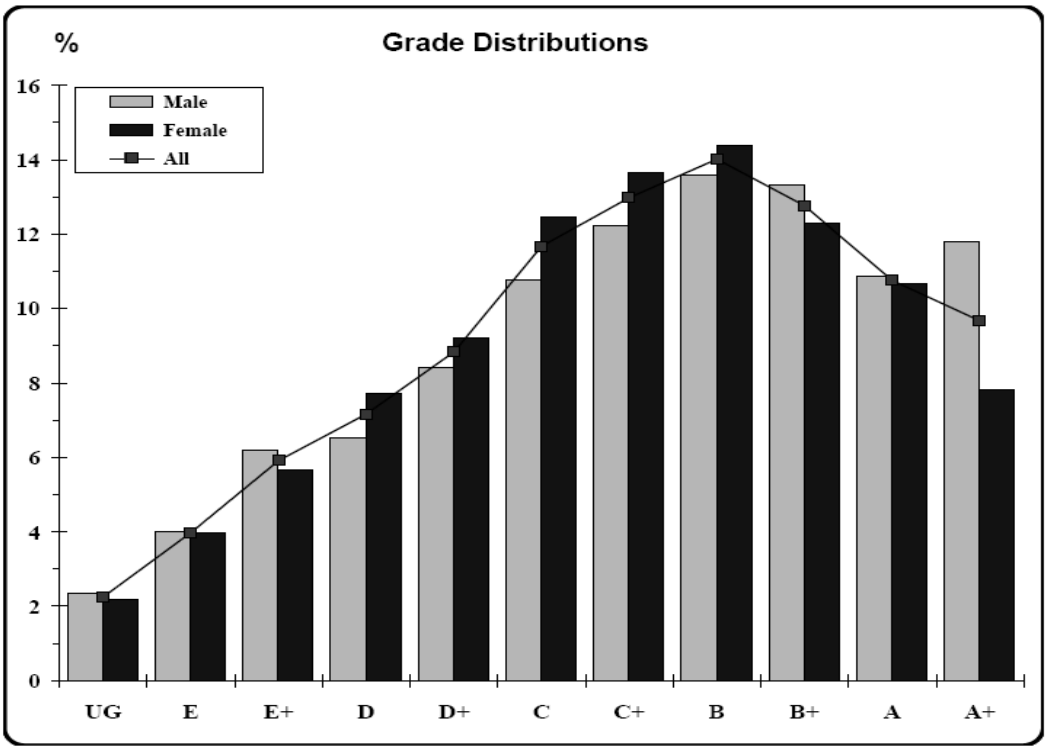

Figure 2. Grade distributions for the 2005 Chemistry Examination 1

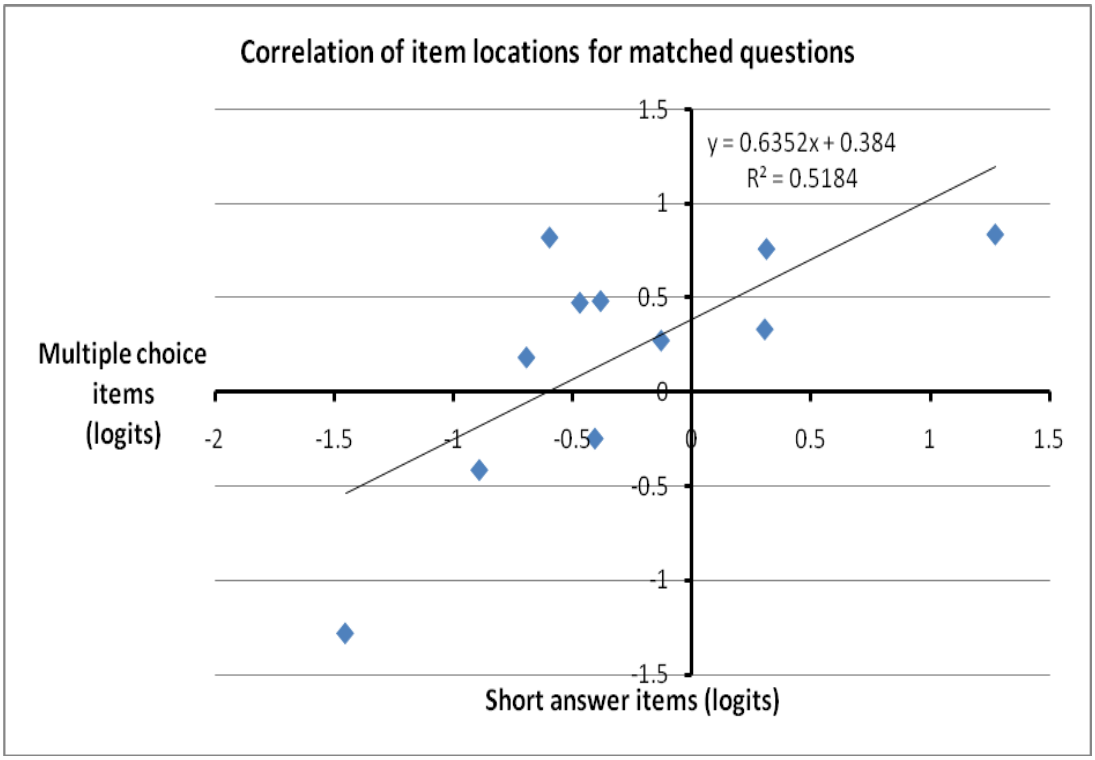

Figure 2. Correlation of multiple-choice and short-answer questions in the tests

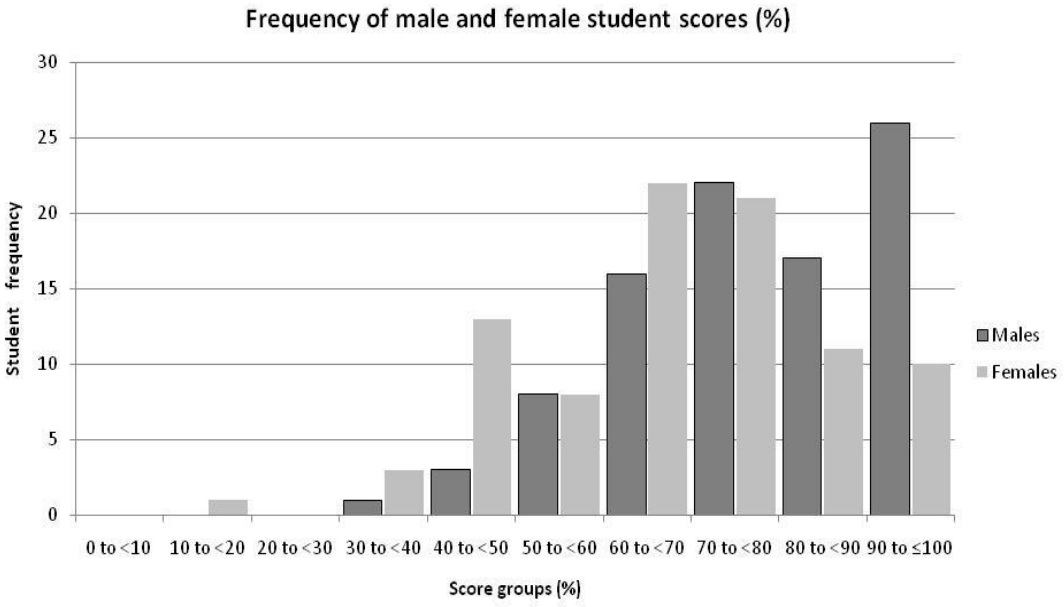

Figure 3. Distribution of male and female scores in the tests

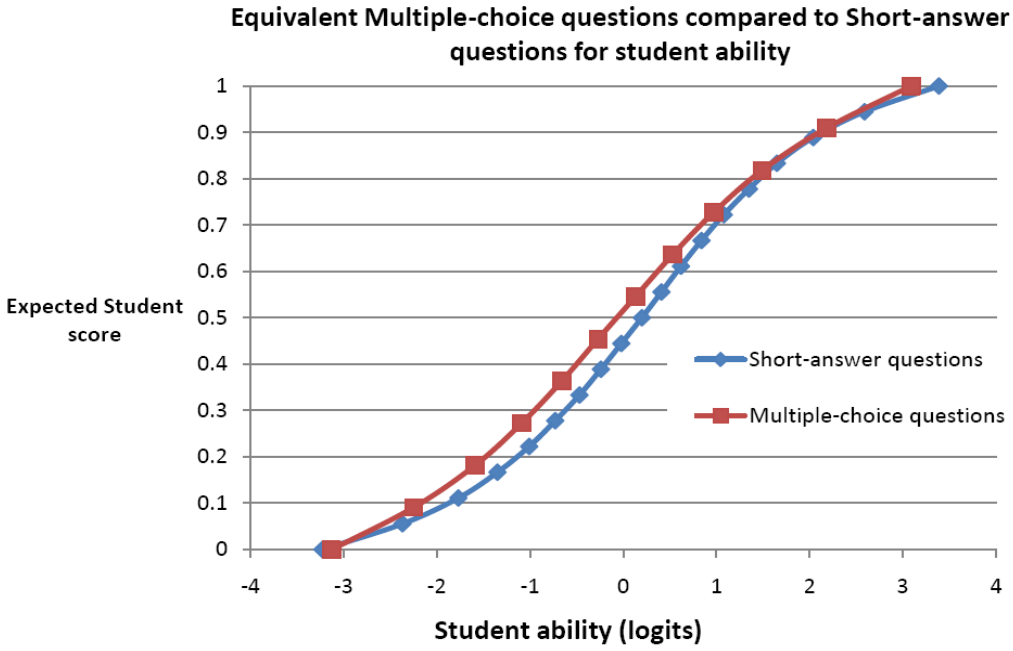

Figure 4. Multiple-choice compared to short-answer response difference against expected score and student ability.

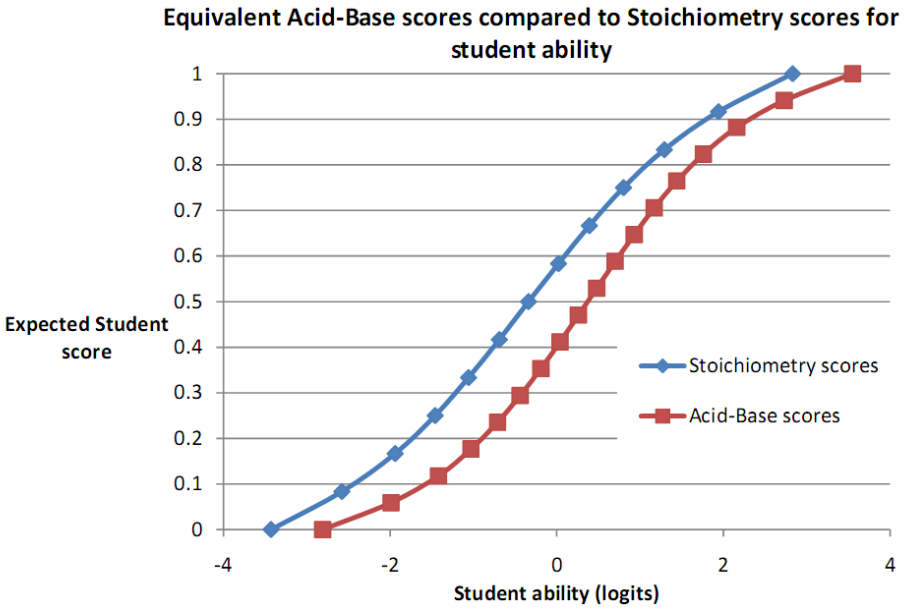

Figure 5. Recall (Acid-Base) compared to Application (Stoichiometry) response difference against expected score and student ability.

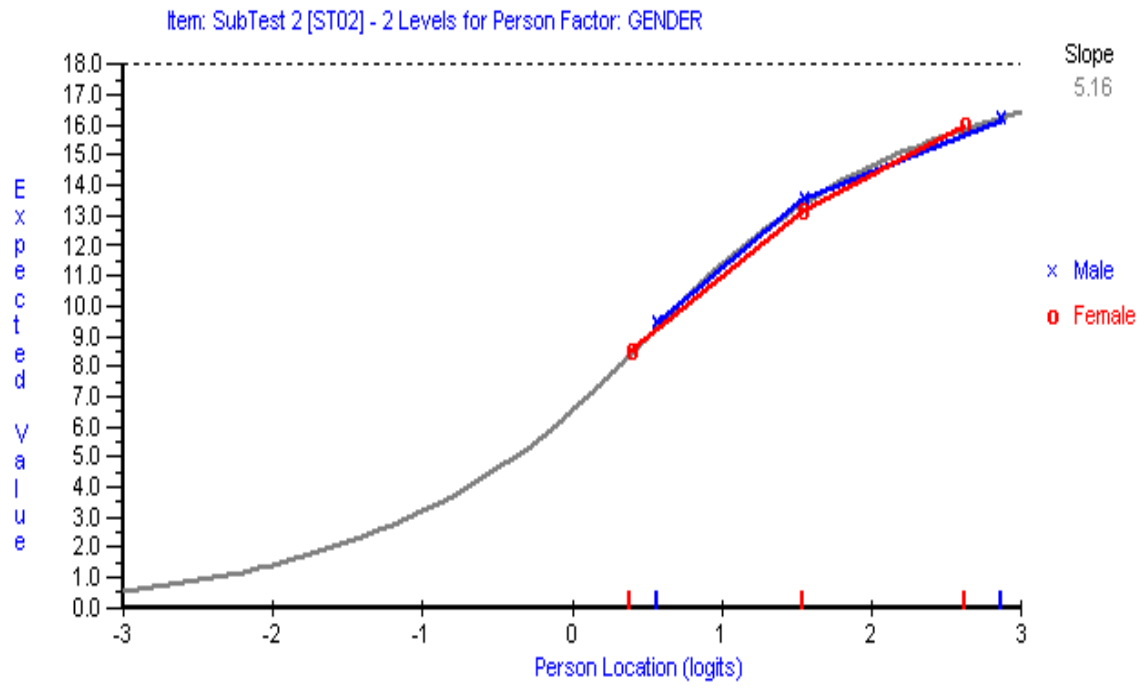

Figure 6: Short-answer questions: showing gender difference against expected score and student ability.