

Science and Mathematics Education Centre

**Multiple-Choice Questions Compared to Short-Answer Response:
Which Assesses Understanding of Chemistry More Effectively?**

Ross David Hudson

**This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University**

September, 2010

Declaration:

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material, which has been accepted for the award of any other degree or diploma in any university.

Signature: *Ross Hudson*

Date: September 2010

Abstract

The research inquires into the effectiveness of the two predominant forms of questions that are used on the Victorian Certificate of Education (VCE) Victoria Chemistry examination. These are multiple-choice questions and short-answer questions. This research examines not only the style of chemistry question but also the content type examined (recall and application questions) along with gender differences in students' responses to such questions. The research involved three phases, i) analysis of five years results from the VCE Chemistry examinations, ii) class trial testing students of both genders with structured questions that examined the same material content with each type of question (multiple-choice or short-answer) and also examined the different type of chemistry content (recall or application) and iii) interviews with students and teachers.

The first phase of the research analysed the available published VCE Chemistry results for the five years, 2003 to 2007. The findings of these data yielded statistically significant differences between the performances of students based on the type of question (multiple-choice or short-answer) and the content of the question. The second phase of analysis yielded comparative data to the VCE analysis but also provided detailed Rasch analysis of the question type and content as well as gender differences in performance.

Important findings were: i) student performance on multiple-choice chemistry questions was significantly higher than performance on short-answer questions regardless of the content and ii), the performance of males was significantly higher than that of females in upper levels of achievement but not at the lower levels of achievement. Possible factors accounting for the observed difference were noted. Implications of these findings are discussed as well as suggestions for further research.

Acknowledgements

The author wishes to express his deep and sincere gratitude for the hard work, patience and constructive criticism of Professor David Treagust. Despite the tyranny of distance, David was always willing to contribute his time and efforts beyond what was reasonable to expect. This thesis could not have been completed without his unstinting help.

The kind support of Daniel Kneebone, Gayle Petch, Maree Cody, Anne Clark, Hong Koo, Gordon Wilson, Peter Barbadonis, Louis Simimopoulos, Dr Jonathon Smith and Mark Collins for contributing their time to conducting the trial tests with their Year 11 chemistry classes and also to the students of those classes.

Mr Nick Connolly and for his invaluable assistance and advice with the Rasch analysis for the trial test data.

To my wife Karen and my children David, Victoria, Andrew and Elizabeth who quietly encouraged me and tolerated my domination of the computer.

Table of Contents	Page
Declaration	i
Abstract	ii
Acknowledgments	iii
Table of Contents	iv
List of Tables	xiv
List of Figures	xvi
Chapter 1 Introduction	1
Introduction.....	1
Research Problem and Associated Research Questions	2
Question type.....	3
Combined scale testing.....	5
Gender Differences in Performance	6
Significance.....	8
Research Methodology	8
Data Collection.....	10
Data Analysis	10
Ethical Issues.....	11
Limitations of the Study.....	12
Overview of the Thesis	13
Chapter 2 Literature Review	15
Introduction.....	15
Multiple-Choice Questions	15
Type of Examination: Multiple-Choice or Short-Answer	18
Content of Questions.....	20
A Constructivist Perspective.....	22
Other Item Methodologies	24
Two-tier testing	24
Item Response Theory methodology.....	25
Rasch and Classical test theory	27
Why Two Types of Question?	27
A Summative Comparison of Multiple-Choice Questions and Short-Answer Questions	28

Do multiple-choice questions and short-answer questions examine the same concept?.....	31
Gender Differences in Performance.....	32
Attribution Theory	37
Motivation.....	39
Summary of Chapter 2	40
Chapter 3: Research Methodology	43
Research Method.....	43
Relationship Between the Research and the Research Questions.....	43
Setting the Scene	44
Approaches to Research in Science Education	45
Triangulation.....	48
Data triangulation.....	49
Methodological triangulation	50
Interviews.....	51
Guided interview or semi-structured interview.....	52
Standardised open-ended interview.....	53
Closed quantitative interview	53
Conducting the interview	54
Are the Results Valid or Reliable?.....	56
Internal validity	56
External validity	57
Credibility and validity.....	57
Peer debriefing	58
Member checking.....	58
Transferability	59
Confirmability	59
Dependability	60
Trustworthiness	60
Summary of Practical Methodology	61
Data sources-student groups.....	61
Mixed methods	62
Interviews	63
Student interviews.....	63

Teacher interviews	64
Analysis of past papers.....	65
Determination of question categories: Recall or Application.....	65
Sample testing	68
Anonymity of participants.....	70
Data analysis	71
Data collection.....	72
Interviews.....	72
Ethical Issues.....	73
Summary of Methodology	74
Chapter 4: Results-Analysis of Past papers	75
Analysis of Past Papers	75
The 2005 VCE Chemistry Papers Examination 1 and 2	76
The 2005 VCE Chemistry Examination 1(is held in June and covers	
Unit 3 work).....	76
Classifying the questions as recall or application	82
The 2005 VCE Chemistry Paper, Unit 4 examination (held in November).	83
Detailed study of Examination 2 2005	85
Conclusions from the Analysis of the Chemistry Examinations	
(2003-2007).....	96
Analysis Using ANOVA.....	96
Response to Research Question 1: Do Students Perform More Effectively	
on Multiple-Choice Questions or Short-Answer questions?.....	99
Comparison of performance on multiple-choice and short-answer;	
both application and recall questions	99
Comparison of performance on multiple-choice and short-answer application	
type questions.....	100
Comparison of performance on multiple-choice and short-answer recall	
type questions.....	100
Response to Research Question 2: Do Students Perform More Effectively	
on Recall Type Questions or on Application Questions?.....	100
Comparison of performance on recall and application, all questions	100
Comparison of performance on recall and application	
multiple-choice questions	100

Comparison of performance on recall and application short-answer questions	101
Comparison of performance on all questions by Unit 3 and Unit 4.....	102
Comparison of performance on multiple-choice questions by Unit 3 and Unit 4.....	102
Comparison of performance on short-answer questions by Unit 3 and Unit 4.	103
Comparison of performance on application question by Unit 3 and Unit 4	103
Comparison of performance on recall questions by Unit 3 and Unit 4.....	103
Comparison of performance on all questions by Unit 3 and Unit 4.....	104
Response to Research Question 3: Does students' gender influence performance in chemistry examinations (or tests)?.....	105
Gender differences in performance.....	105
Chi-squared analysis of the VCE Chemistry examinations (2003-2007)	108
Unit 3 Examination 2003, Chi-squared analysis of male and female performance.....	108
Unit 4 Examination 2003, Chi-squared analysis of male and female performance.....	111
A ⁺ grade analysis.....	114
C ⁺ grade analysis	115
E ⁺ grade analysis	116
Examination	116
Summary of Analysis of the 2003-2007 Chemistry Examinations.....	117
Significant results.....	117
Observed results that were not statistically significant	118
Chapter 5: Results-School Trials and Interviews	119
School Trials of Chemistry Test Questions	119
Examples of paired difficulty questions.....	120
Acid-Base (recall) Question (See Appendix A).....	120
Stoichiometry (application) Question	121
Consequential multiple-choice questions.....	121
Results from the Trial Tests	123
RUMM2030 Analysis of the Trial test questions.....	123
Resolving Item and Person fit issues.....	124
Person fit	124
Item fit.....	125

Example of a weakly discriminating question compared to a strongly discriminating question.....	125
Rescoring test items	127
Final trial test construct.....	128
Test validity.....	130
Correlation between multiple-choice and short-answer items.....	130
Targeting	133
Results from the RUMM2030 Analysis.....	134
Response to Research Question 1: : Do Students Perform More Effectively on Multiple-Choice Questions or Short-Answer questions?.....	134
Comparison of all Multiple-choice with all Short-answer questions.....	136
Comparison of performance on multiple-choice and short-answer application type questions	140
Comparison of performance on multiple-choice and short-answer recall type questions.....	141
Response to Research Question 2: Do students perform more effectively on recall type questions or on application questions?.....	142
Comparison of performance on recall and application, all questions	142
Comparison of performance on recall and application multiple-choice questions.....	144
Comparison of performance on recall and application short-answer questions.....	145
Response to Research Question 3: Does students' gender influence performance in chemistry examinations (or tests)?.....	146
Comparison of Multiple-choice and Short-answer questions by gender	148
Multiple-choice and gender.....	148
Short-answer and gender.....	149
Comparison of Recall (Acid-base) questions and Application (Stoichiometry) questions by gender.....	151
Recall questions and gender.....	151
Application questions and gender	152
Response to Research Question 4: Do students have a preference for the type of question style in terms of: a) Multiple-choice or short-answer in general terms? and b) With respect to whether the question is assessing recall or application?...	154
Results from Interviews	154

The interview schedule	154
Coding responses.	155
Student responses by interview question	157
Multiple-choice advantage.	160
Multiple-choice disadvantage.	160
Short-answer advantage.	161
Short-answer disadvantage	161
Post interview member checking	162
Overall results from the student interviews	162
Response to Research question 5- Do teachers consider multiple-choice or short-answer questions to be the most effective in demonstrating their students understanding of chemistry?	163
Summary of Chapter 5	164
Chapter 6: Discussion and Conclusions	165
Introduction.....	165
Overview of Study	165
Major Findings in Response to the Research Questions.....	169
Findings in response to Research Question 1: Do multiple-choice or short-answer questions more positively emphasise student understanding?.....	169
Past VCE examination papers (RQ 1).....	169
Trial tests (RQ1).....	170
Findings in response to Research Question 2: Do students perform more effectively on recall type questions or on application questions?	171
Past VCE examination papers (RQ 2).....	171
Trial tests (R.Q.2).....	171
Findings in response to Research Question 3: Does students' gender influence performance in chemistry examinations (or tests)?.....	172
Findings in response to Research Question 4: Do students have a preference for the type of question?	174
Findings in response to Research Question 5: Do teachers consider multiple-choice or short-answer questions to be the most effective in demonstrating their students understanding of chemistry?	175

Findings in Response to the Overall Thesis Question: Multiple-Choice Questions Compared to Short-Answer Response. Which Assesses Understanding of Chemistry More Effectively?	175
Implications for Student Motivation	176
Limitations of Study.....	177
Transference	177
Student interest.....	177
Practice element	178
Limited access to VCE results in detail.....	178
Interviews	178
Suggestions for Future Research.....	179
Summary	179
References	180
Appendices	192
Appendix A: Student trial tests	192
A-1: Stoichiometry Test: Short-answer Questions	193
A-2: Acid- Base Test: Short-answer Questions	194
A-3: Stoichiometry Test: Multiple-choice Questions	195
A-4: Acid-Base Test: Multiple-choice Questions	195
Appendix B: Permission letters for trials	196
B-1: Principal permission letter request form	196
B-2: Parent permission letter request form	198
B-3: Student permission letter request form	200
Appendix C: Interview sheets, coding and transcripts	202
C-1: Student interview question list.....	202
C-2: Teacher interview question list	203
C-3 Transcripts of Student interviews	204
C-4 Distribution of student responses question type preference.....	229
Appendix D: VCE Chemistry question breakdown for Units 3 and 4	230
D-1: 2003 VCE Examination Unit 3 Question breakdown.....	230
D-2: 2003 VCE Examination Unit 4 Question breakdown.....	232
D-3: 2004 VCE Examination Unit 3 Question breakdown.....	234
D-4: 2004 VCE Examination Unit 4 Question breakdown.....	236
D-5: 2005 VCE Examination Unit 3 Question breakdown.....	238
D-6: 2005 VCE Examination Unit 4 Question breakdown.....	240

D-7: 2006 VCE Examination Unit 3 Question breakdown.....	242
D-8: 2006 VCE Examination Unit 4 Question breakdown.....	244
D-9: 2007 VCE Examination Unit 3 Question breakdown.....	246
D-10: 2007 VCE Examination Unit 4 Question breakdown.....	248
D-11: 2003-2007 VCE Examination Unit 3 Examinations combined Question breakdown	250
D-12: 2003-2007 VCE Examination Unit 4 Examinations combined Question breakdown	254
D-13: 2003-2007 VCE Examination Unit 3 &4 All Examinations combined Question breakdown.....	258
Appendix E: SPSS ANOVA results and Eta-squared results.....	270
E-1: SPSS Anova results comparing Recall v Application Multiple-choice questions	270
E-2: SPSS Anova results comparing Recall v Application Short-answer questions	271
E-3: SPSS Anova results comparing Short-answer v Multiple-choice Application questions.....	272
E-4: SPSS Anova results comparing Short-answer v Multiple-choice Recall questions	273
E-5: SPSS Anova results comparing Application v Recall All questions	274
E-6: SPSS Anova results comparing Short-answer v Multiple-choice All Questions.....	275
Appendix F: SPSS Chi-squared results comparing gender and reported grades .	276
F-1: SPSS Chi-squared results comparing Males v Females for A ⁺ , C ⁺ and E ⁺ grades 2003 Unit 3 Examination	276
F-2: SPSS Chi-squared results comparing Males v Females for A ⁺ , C ⁺ and E ⁺ grades 2003 Unit 4 Examination	279
F-3: SPSS Chi-squared results comparing Males v Females for A ⁺ , C ⁺ and E ⁺ grades 2004 Unit 3 Examination	282
F-4: SPSS Chi-squared results comparing Males v Females for A ⁺ , C ⁺ and E ⁺ grades 2004 Unit 4 Examination	285
F-5: SPSS Chi-squared results comparing Males v Females for A ⁺ , C ⁺ and E ⁺ grades 2005 Unit 3 Examination	288
F-6: SPSS Chi-squared results comparing Males v Females for	

A ⁺ , C ⁺ and E ⁺ grades 2005 Unit 4 Examination	291
F-7: SPSS Chi-squared results comparing Males v Females for A ⁺ , C ⁺ and E ⁺ grades 2006 Unit 3 Examination.....	294
F-8: SPSS Chi-squared results comparing Males v Females for A ⁺ , C ⁺ and E ⁺ grades 2006 Unit 4 Examination	297
F-9: SPSS Chi-squared results comparing Males v Females for A ⁺ , C ⁺ and E ⁺ grades 2007 Unit 3 Examination	300
F-10: SPSS Chi-squared results comparing Males v Females for A ⁺ , C ⁺ and E ⁺ grades 2007 Unit 4 Examination	303
F-11: SPSS Chi-squared results comparing Males v Females for A ⁺ , C ⁺ and E ⁺ grades All Unit 3 Examinations (2003-2007).....	306
F-12: SPSS Chi-squared results comparing Males v Females for A ⁺ , C ⁺ and E ⁺ grades All Unit 4 Examinations.....	309
F-13: SPSS Chi-squared results comparing Males v Females for A ⁺ , C ⁺ and E ⁺ grades All Unit 3 & 4 Examinations.....	312
Appendix G: Grade distributions for VCE Chemistry 2003–2007 A ⁺ , C ⁺ and E ⁺	315
Appendix H: Trial test data	316
H-1 Fit residuals and probabilities for the initial item set.....	316
H-2 Threshold map for initial item set.....	317
H-3 Person-fit characteristics for original data sample.....	318
H-4- Correlation analysis of multiple-choice compared to short-answer responses	323
H-5 Item analysis curves.....	324
H-6: Student responses to trial test, location and fit values	332
H-7:All student scores broken down by type and gender and ANOVA analysis.....	346
H-8: Grade distribution graphs for Chemistry 2008	362
Appendix I: Unit 3 examination results compared to Unit 4 examination results	363
I-1 Raw data comparing the Unit 3 and Unit 4 examinations.....	363
I-2: Comparison of Unit 3 and Unit 4 multiple-choice questions.....	369
I-3: Comparison of Unit 3 and Unit 4 short-answer questions	370
I-4: Comparison of Unit 3 and Unit 4 application questions	371

I-5: Comparison of Unit 3 and Unit 4 recall questions	372
I-6: Comparison of Unit 3 and Unit 4 all questions	373

List of Tables

Table 2.1: Advantages of multiple-choice tests over constructed response tests.....	28
Table 2.2: Disadvantages of multiple-choice tests over constructed response or short-answer question tests.....	30
Table 3.1: Relationship between the research questions and the data collection.....	44
Table 4.1: For the Unit 3 2005 part A summary results.....	78
Table 4.2: For the Unit 3 2005 part B summary results.....	81
Table 4.3: Unit 4 2005 part A summary results.....	88
Table 4.4: Mark distribution for Q 4a from Unit 4 2005 chemistry examination.....	90
Table 4.5: Mark distribution for Q 3a from Unit 4 2005 chemistry examination.....	90
Table 4.6: For the Unit 4 2005 part B summary results.....	92
Table 4.7: Summary for the 2005 Examination 1 and 2 papers.....	93
Table 4.8: Summary for the 2003, 2004, 2006 and 2007 Examination papers.....	94
Table 4.9: Overall Summary of Unit 3 and 4 performance (Years 2003-2007).....	95
Table 4.10: Combined Unit 3 and 4 performance (Years 2003-2007).....	95
Table 4.11: ANOVA and Eta-squared analysis of question performance by question type and classification VCE Chemistry 2003-2007.....	98
Table 4.12: ANOVA of Unit 3 and Unit 4 performance by question type Chemistry 2003-2007.....	102
Table 4.13: Participation numbers of male and female students in VCE Chemistry 2003-2007.....	106
Table 4.14: Frequency and Chi-squared analysis of the performance in terms of the A ⁺ , C ⁺ , E ⁺ grades 2003 Unit 3.....	109
Table 4.15: Frequency and Chi-squared analysis of the performance in terms of the A ⁺ , C ⁺ , E ⁺ grades 2003 Unit 4.....	113
Table 4.16: 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.....	115
Table 4.17: Chi-squared analysis of Male and Female performance on the Unit 3 and Unit 4 Examinations in Chemistry.....	116
Table 5.1 Individual question measured by location difficulty (from RUMM2030).....	131
Table 5.2: Summary measures of the data from the trial tests.....	135

Table 5.3: ANOVA analysis of question performance by question type and classification Chemistry trial tests.....	138
Table 5.4: Gender differences on the trial chemistry tests (means).....	147
Table 5.5: Gender differences on the trial chemistry tests (ANOVA).....	147
Table 5.6: Responses to Research Question 4 (n = 59)	155
Table 5.7: Preferences of Students for Question Type (n=100)	156
Table 5.8: Advantages and Disadvantages of Multiple-choice questions.....	159
Table 5.9: Advantages and Disadvantages of Short-answer questions.....	160

List of Figures

Figure 2.1 Causality attributions	38
Figure 3.1 Question 16 from the 2005 Unit 3 Chemistry examination.....	66
Figure 4.1: Question and answers for Q 2 from Unit 3 2005 Chemistry examination	77
Figure 4.2: Question and answers for Q 10 from Unit 3 2005 Chemistry examination	77
Figure 4.3: Question and answers for Q 5C from Unit 3 2005 chemistry examination	79
Solution	79
Figure 4.4: Question and mark distribution for Q 7b from Unit 3 2005 chemistry examination	80
Figure 4.5 Question 14 from the Unit 3 2004 Chemistry Examination	83
Figure 4.6: Question 3 from the 2005 Unit 4 Examination.....	84
Figure 4.7: Question details for Q 4 from Unit 4 2005 chemistry examination	85
Figure 4.8: Question and answers for Q 12 from Unit 4 2005 chemistry examination	86
Figure 4.9: Question and answers for Q 10 from Unit 4 2005 chemistry examination	87
Figure 4.10: Question and mark distribution for Q 3b from Unit 4 2005 chemistry examination	89
Figure 4.11: Question details for Q 3a from Unit 4 2005 chemistry examination	91
Figure 4.12: Question solution details for Q 4aii from the Unit 4 2005 chemistry examination	91
Figure 4.13: Grade distributions for the 2005 Chemistry Examination 1.....	105
Figure 4.14: Grade Distributions for the 2005 Biology Examination 1.....	107
Figure 4.15: Grade distributions for the 2005 Chemistry Examination 2.....	112
Figure 5.1: Rasch performance curves for question ST 12 using RUMM2030	126
Figure 5.2: Rasch performance curves for question AB 7 using RUMM2030	127
Figure 5.3: Rasch performance curves for question ST 12 (modified) using RUMM2030	128
Figure 5.4: Rasch performance curve for question ST03 using RUMM2030	129
Figure 5.5: Rasch distracter curve for question ST03 using RUMM2030	130

Figure 5.6 Correlation between matched item pairs of sample test items	132
Figure 5.7: Item map showing distribution of items and students	133
Figure 5.8: Multiple-choice compared to Short-answer response difference against expected score and student ability.	137
Figure 5.9: Multiple-choice (MC) compared to Short-answer (OR) responses t-test.	139
Figure 5.10: Multiple-choice (Stoichiometry) compared to Short-answer (Stoichiometry) response difference against expected score and student ability.	140
Figure 5.11: Multiple-choice (Acid-Base) compared to Short-answer (Acid-Base) response difference against expected score and student ability.	141
Figure 5.12: Recall (Acid-Base) compared to Application (Stoichiometry) response difference against expected score and student ability.	143
Figure 5.13: Acid-Base questions compared to Stoichiometry questions t-test distribution from RUMM2030.	143
Figure 5.14: Multiple-choice (Stoichiometry) compared to Multiple-choice (Acid-Base) response difference against expected score and student ability.	144
Figure 5.15: Acid-Base multiple-choice questions compared to Stoichiometry multiple-choice questions t-test distribution from RUMM2030.	145
Figure 5.16: Short-answer (Stoichiometry) compared to Short-answer (Acid-Base) response difference against expected score and student ability.	145
Figure 5.17: Acid-Base short-answer questions compared to Stoichiometry short-answer questions t-test distribution from RUMM2030.	146
Figure 5.18: Distribution of male and female scores in the trial tests.....	148
Figure 5.19: Multiple-choice questions showing gender difference against expected score and student ability.	149
Figure 5.20: Short-answer questions: showing gender difference against expected score and student ability.....	150
Figure 5.21: Recall (Acid-Base) questions: showing gender difference against expected score and student ability.	152
Figure 5.22: Application (Stoichiometry) questions: showing gender difference against expected score and student ability.....	152

Chapter 1 Introduction

Introduction

Having taught Chemistry for nearly 30 years I have become compelled to question the merits of the assessment methodology used in senior Chemistry. The examination or test assessment generally fell into two categories, multiple-choice questions and/or short-answer or extended response questions. The typical examination in Victoria set by the Victorian Curriculum and Assessment Authority (VCAA) or any of its predecessors has been of one and a half hours length with a structure of 20 multiple-choice questions and approximately eight short-answer/extended response questions. The type of learning expected in the Victorian examinations is typical of most chemistry assessment in that the students are expected to recall facts, processes and related equations and/or are expected to apply their general conceptual knowledge to a particular situation, usually in the form of calculation application questions. Both styles of question are assessed in the multiple-choice section and the short-answer section. My attention to the effectiveness of each style of question to each of the recall or application categories of question was raised by my anecdotal observations of the student performance in the examinations and stated student preferences for the different examinations together with my, again anecdotal, observations of the differences in student performances in the two examinations that were held each year. The examination or test is not the only form of assessment used in VCE chemistry. The VCAA mandates that one-third of a student's assessment is attributable to the student's performance on School Assessed Coursework. School Assessed Coursework is teacher-assessed material that is devised by the teacher to a set of criteria specified by the VCAA (VCAA, 2006d). This study is concerned only with the examination aspect of student assessment in chemistry.

What I had considered from anecdotal or observational point of view led me to consider whether these perceived differences were measurable and if the observed differences were statistically significant. The format and direction of the possible research to be addressed became clear. Essentially research into the structure and responses gained from chemistry examinations determined by the variables of question type (multiple-choice or short-answer), content (application or recall) and gender of student would provide a rich source of interest to both myself and others

interested in the field of chemistry education. These anecdotal observations are supported in the literature where a number of authors have observed differences in student performance on the two different types of questions (Aiken, 1987; Anderson, 1972; Ercikan et al., 1998; Petrie, 1986; Simkin & Kuechler, 2005).

Research Problem and Associated Research Questions

The topic explored was: “Multiple-choice Questions Compared to Short-answer. Which Assesses Understanding of Chemistry More Effectively?”

The research questions that underpin this topic were:

1. Do students perform more effectively on multiple-choice or short-answer questions?
2. Do students perform more effectively on recall type questions or on application questions?
3. Does students’ gender influence performance in chemistry examinations (or tests)?
4. Do students have a preference for the type of question style in terms of
 - a. Multiple-choice or short-answer in general terms?
 - b. With respect to whether the question is assessing recall or application?
5. Do teachers consider multiple-choice or short-answer questions to be the most effective in demonstrating their students understanding of chemistry?

The focus of the research is founded upon the researcher’s long experience of teaching chemistry in Victorian schools. The Chemistry examinations at both Year 11 and Year 12 have had essentially the same structure for all of the time the researcher taught Chemistry. The year 12 examinations are externally set by the VCAA according to the criteria outlined in the study design (VCAA, 2006d). The teachers within the school usually set the Year 11 examination. However, this decision again is dictated by the VCAA study design which outlined the content to be examined and the assessment criteria used at Year 11. The similarity between the examinations at Year 11 and at Year 12 then is mostly dependent on two factors, firstly the recommendations of the governing authority for the Victorian Certificate

of Education (V.C.E.) through the published study design for chemistry (VCAA, 2006d) and secondly, teachers wishing to prepare the students as best as they were able for the externally set Year 12 examinations. Whilst the class teacher ultimately sets the Year 11 examinations, the author's experience has been that many examinations set at Year 11 tend to mirror the structure given at Year 12. Consequently, a typical Year 11 examination consists of about 20 multiple-choice questions and about 8 to 10 short-answer questions. This is exactly the same structure as a Year 12 examination. Rather than simply accepting this wisdom, the research in this thesis seeks to explore the relationship between the two types of questions, the relationship between the type of question and the class of curriculum content (recall or application) and finally the relationship between the previous two factors and the gender of the student.

Question type

Multiple-choice questions generally have less scope and complexity than short-answer questions and therefore are 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 whereas multiple-choice question answers give no indication of how students arrived at their answer (Bridgeman, 1992). What were also of interest were the students' perceptions about which type of question measured their ability at answering analytical or at application questions. Students were evenly split over the different type of question suggesting that students recognised that short-answer or open response questions gave them a better chance of demonstrating what they actually understood about the concepts they had studied. This viewpoint is supported by a study (Pressley, Ghatala, Woloshyn, & Pirie, 1990) that found that 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 almost the same. The students felt they had (or were going to) performed better on the multiple-choice test (Pressley et al; 1990). Generally, the field of research into different test types (that is multiple-choice compared to short-

answer) has not been particularly extensive (Gay, 1980; Nungester & Duchastel, 1982).

Haynie (1994) also notes that student performance on multiple-choice questions was superior to that on short-answer questions. However, Haynie also emphasises the importance of the need for testing to support learning. 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.

Related to this study were the prior research findings by Gay (1980) and Sax and Collet (1968). These studies generally agreed with the findings of Haynie (1994), that is, students who take multiple-choice tests tended to perform somewhat better than students who took short-answer tests. Haladyna (2004) notes that the cognitive demands of multiple-choice questions are different to those of short-answer questions and as such multiple-choice questions not always perform well psychometrically depending on the way the questions and the question distracters are formulated. The element of practice was also found to influence positively the performance on multiple-choice items. One aspect of the findings was that the level of retention depended on what type of test was performed to measure the retained learning (Gay, 1980; Sax & Collet, 1968). Further studies have indicated that the type of question and the response can be different. It appears that certain types of question will be answered differently depending on the test question format (Barnett-Foster & Nagy, 1996). On the other hand, Traub (1992) found that for test items that examined inference or evaluation there appeared to be little difference on whether the question was a multiple-choice or a constructed response question.

Student responses to stoichiometric questions seemed to produce different outcomes depending on whether the question was presented as a multiple-choice as opposed to a short-answer question (Niaz & Robinson, 1995). Students were apparently influenced by the multiple-choice options and consequently provided answers to questions that were unlikely to be answered if presented in a short-answer form. Students commonly found great difficulty in attempting to justify or explain the selection of response in any particular multiple-choice question, which supports the notion that students were both willing and able to guess when presented with a multiple-choice question that the student could not 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). A possible approach to overcoming this issue of students being able to guess answers with relative freedom is the use of two-tiered tests. These tests require students to provide a multiple-choice response as usual but then the student must provide an explanation that demonstrates the method or reasoning that was used to obtain the answer (Treagust, 1995; Treagust & Chandrasegaran, 2007).

Whether or not the type of question used, be it multiple-choice questions or short-answer questions, assesses understanding better than the other will no doubt continue to be a source of debate and is a reason for conducting this research.

Combined scale testing

A number of studies have looked at evaluating whether or not the two different assessment item types can be calibrated to give an equal overall result. Essentially, the two item types (multiple-choice question and short-answer question) can be calibrated on a single scale so that student understanding is measured equally regardless of the item type involved. A number of studies have had some success in producing a combined scale (Sireci, Thissen, & Wainer, 1991; Thissen, 1993). Ercikan, Schwarz, Julian, Burket, Weber & Link (1998) found through their extensive study using an item response theory model that, in their opinion, there was sufficient similarity in the material being assessed by both item types for the creation of a common scale with a single set of scores for the responses of both item types (Ercikan et al., 1998). Generally there was little difficulty in obtaining good correlation between the multiple-choice questions and the short-answer questions except on those occasions where the short-answer questions were particularly difficult or the test item response time was too short to allow students sufficient time to fully answer the questions. This finding may be taken to mean that if there was a high correlation between the multiple-choice questions and the short-answer questions then it would be efficient to ask only multiple-choice questions. This suggests the need for the test developer to consider carefully the purpose for using a particular type of question.

Other studies have looked at the relative merits and reasons for using the two types of test items. The dominant view point would still appear from Lukhele, Thissen & Wainer (1994) to be that the multiple-choice question based assessment will continue

to be important from the aspects of time and economy but the short-answer question should be encouraged because of the diversity added to the assessment.

Do the different styles of question (multiple-choice or short-answer) examine the same thing? Several authors have indicated that certain types of learning are favoured by one type of test over the other. Typically multiple-choice question tests favour rote or recall skills whereas short-answer question tests favour critical or application skills (Aiken, 1987; Anderson, 1972; Ercikan et al., 1998; Petrie, 1986; Simkin & Kuechler, 2005). So the question, do they measure the same skills to the same measurable extent is important to this thesis report. Research on this aspect has been mixed. Some studies suggesting that the two question types (multiple-choice and short-answer) measure student understanding to more or less equal degrees whereas other findings are less certain (Aiken, 1987; Anderson, 1972; Ercikan et al., 1998). Thissen, Wainer and Wang (1994) also considered whether or not it was valid to combine the scores of a multiple-choice question test with that of a short-answer question test to give an overall combined score. Their analysis of college examinations (using a very large sample) showed that, whilst there were some differences in the way in which the short-answer questions and multiple-choice questions were answered, there appeared to be little doubt that they were in fact measuring much the same thing in terms of their reliability in demonstrating student understanding. In other words the multiple-choice questions were not more one-dimensional (measuring only one test construct) than the short-answer questions.

Gender Differences in Performance

The role of gender in performance in Chemistry and other subject areas in general has 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. Their exploration sort reasons behind why male students were tending to outperform the female cohort, resulting in the suggestion that differences in mathematical ability were a very important consideration. The most important factor, 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. These findings with regard to mathematics can be fairly evenly transferred 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 females 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. The study also showed 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. With respect to the Victorian Certificate of Education (VCE) the numbers of females attempting the VCE in general is greater than that of the males. In 2008 for example, there were 45 741 female (53%) and 40 173 male (47%) VCE students. Of interest to this study was the number of students who studied chemistry (9089), making it the ninth most popular subject (VCAA, 2008). Within Chemistry itself the number of males (4309, 48%) and females (4654, 52%) closely mirror the distribution of students overall. The difference from the total population is small but none the less it shows that females are slightly under represented in the students undertaking chemistry. These numbers do show that females are now choosing one of the major sciences in numbers that are representative of the population.

Many studies have agreed with the observation that male students usually outperform female students in assessments particularly in the areas of mathematics and science. 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). 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 & McPeck, 1993; Rennie & Parker, 1991).

Significance

This research will be of significance because of the limited research already done that addresses the full scope of the posed research questions. For teachers to encourage students to further their study of chemistry it is necessary to find the best methods of assessment that allow the students to best demonstrate understanding of chemistry. In particular, the type of testing that best demonstrates actual understanding of chemistry as opposed to gaining the highest score, and which type of testing if any is more gender neutral, is an important issue to consider in Chemistry education. Whether or not the method of testing is inadvertently disadvantaging one of the genders over the other will also be examined.

Research Methodology

The methodology used was primarily one of both quantitative and some qualitative interpretation. The data were collected from four secondary colleges. The schools chosen were both co-educational and single sex (male and female). All four schools were located in an upper middle class suburb and were noted for academic achievement. The applicability or transferability (Anderson, 1998; Cohen, Manion, & Morrison, 2000) of the results is therefore likely to be somewhat limited. The extent that any findings could be applied accurately to the wider community of coeducational or single sex schools may be questionable. Whilst the socio economic backgrounds of the students attending these schools are not typical of the general student cohort they did present a positive opportunity for this study. The schools participating in the study have traditionally performed well in the VCE Chemistry examinations. Students are also well represented in the higher study scores for Chemistry. This is 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 Figures 4.5 and 4.7). The sample students for this study provided a good representation of that high scoring group.

The phase of study that involved interviewing students had a substantial component of qualitative interpretative analysis. The interpretive nature of the research and the types of data collected made any quantitative analysis largely inappropriate although some basic summary statistical analysis was performed. The interpretive modes of data collection are clearly most suitable to this particular topic (Kemmis & McTaggart, 2000). By responding to questions, the participants had the opportunity to reflect on their own practices and evaluate their own approaches to the learning environment that they were experiencing on a day-to-day basis. As this part of the study is essentially qualitative and not quantitative the legitimacy and validity of the data are supported by the use of triangulation using several methods of data collection that focus on different aspects of the subjects' responses (Anderson, 1998; Cohen et al., 2000; Mathison, 1988). In this case the methods of interview, member checking and data analysis along with the literature review facilitated the triangulation.

The analysis of the past VCE 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 VCE papers was most clearly identifiable as a post hoc analysis (Myers & Well, 2003). Much of the analysis was either by 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 analysis of the trial tests also involved similar statistical analyses to that of the VCE examinations but also, due to the first hand nature of the data, the use of Rasch analysis (RUMM laboratory P/L, 2009) was also used to enhance the analysis of the data.

Essentially the research involved the following phases

- Statistical analysis of the past five years VCE Chemistry papers using the information provided by the VCAA focussing on question type and content. Short tests were constructed that asked essentially the same question but in both multiple-choice and short-answer form. These tests were administered randomly to students from each of the participating colleges.

- Students were interviewed about their views on Chemistry Assessment and their preferences for style of question.

Teaching staff were interviewed to ascertain their views on the Chemistry assessment.

Data Collection

Trial testing was the primary source of first hand data for this project. Tests were repeated and 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.

The interview was a significant area of data collection in this project and was also an area of concern in terms of validity and reliability in the research situation (Cohen et al., 2000). An important methodology used to improve the validity of the research through interview was to eliminate as much as possible any bias from the interview. By having a highly structured interview where the questions were constructed in a controlled environment away from the interview itself enhanced the likelihood of reliable and valid results. Different interviews will always have different environments due to the different participants. A free ranging interview may in each instance produce results that are too different to allow effective comparison and thus the reliability of the results would be compromised. Careful construction of questions to avoid problems, such as leading questions, was considered during the question preparation. The interview questions in this research were carefully constructed to elicit concise information from the participants, which allowed good comparison of results from the different interviews and simple coding of the responses. Coding allowed for some basic statistical analysis of the interview results to be performed.

Data Analysis

Where appropriate, statistical analysis (ANOVA, Chi-squared and Rasch) was performed on the data collected from the interpretation of the VCE past papers and sample student trial tests. This analysis involved some interpretive input by the researcher. Allocating the examination questions to the categories of recall or application was an important part of the analysis process. The researcher in

collaboration performed the allocation with two experienced teaching colleagues. (The collaborative process is explained in more detail in chapter 3).

With regard to the above data collection methods careful analysis was performed to ascertain with the aid of simple statistics. Means, standard deviations, Chi-squared, ANOVA and Rasch analysis were used to determine the performance characteristics of the different types of questions and the different categories. This was achieved by calculating the significance of the differences in correlations between the two variables being studied, for example, multiple-choice versus short-answer questions on topics that involve application concepts.

Examination of these measures gave an indication of the trends in question answering techniques and scores and was a reliable way of examining differences between scores on the different types of questions (multiple-choice questions and short-answer questions) and also permitted analysis of gender performance differences. Furthermore, the data provided by the VCAA with respect to the results and scores achieved on the VCE Chemistry examinations is relatively limited in its scope (only summative results and data are given). Therefore the combination of interpretive analysis by the researcher and statistical analysis was important in supporting the validity interpretive analysis. The extremely large number of students upon which the data was derived enhanced the reliability of the results.

Ethical Issues

A significant aspect of this research involved the participation of school children. Any research involving the participation of human subjects must be planned with the highest standards to ensure that no harm comes to the participants (The Coalition of Americans for Research Ethics, 1999). It is essential that the researcher is fully cognisant of all the ethical issues that may possibly pertain to the planned research. Put simply, the planned research must have the full consent and understanding of the participants and that under no circumstances should the research intentionally or unintentionally place the participants under any risk of actual *physical, psychological or environmental* harm. At the same time the researcher must ensure that his or her research *methodology, writing practice and motives* are of the highest integrity.

The most significant phase of this research, which involves ethical consideration, was the second and third phases of the research, namely the student trial tests and

student interviews. The approval of the ethical standards committee of the governing university (Curtin University of Technology) was gained before any participant research was undertaken. Participants were fully informed as to the purpose of the research and how the information gained from the trial tests and interviews would be used to further the research project. The participants' rights were respected and all participants were fully informed in writing of their rights (Appendix B). The participants were also assured that the information would remain anonymous and all information and results would be securely destroyed when no longer required. The privacy of all the participants was of paramount importance (Russell, 1998).

The researcher also has obligations in any research project. The researcher must maintain the highest ethical standards of conduct and behaviour in the conduct of the research and also in the presentation of the findings (Anderson, 1998; Halasa, 1998). Personal bias and opinion should not be allowed to influence the independence of the research findings; similarly the researcher must acknowledge and cite any assistance received. The specific issue of plagiarism is one that has been of significance in recent years and it is incumbent upon the researcher to ensure that all referenced work is correctly cited and acknowledged (Russell, 1998).

In this research project the rights of the participants was assisted by the provision of full access to all the results of the trial tests and the opportunity to review the information provided in the interviews. Every measure was taken to ensure that the trial tests were not identifiable to any particular student. At no time did the researcher have any knowledge of the participants' identity.

Limitations of the Study

This research in this instance was limited by the size of the student sample chosen. This in turn limited the depth of the generalizability of the findings of the study. This being said however, the study did focus on just one curriculum subject so, whilst the findings may not be necessarily applicable to the wider population the findings should have a reasonable level of applicability with respect to students studying Chemistry. The students chosen for the study were also a limiting factor in that they did not represent a wide cross section of the whole student population. The students came from an affluent middle class, largely Anglo-Saxon backgrounds and as such represent a fairly limited stratum of socio-economic population of Victoria. Whilst a

limitation, it was also an advantage because the students were among the high achievers in Victoria and as such tend to be over represented in the students achieving the higher grades for the subject of chemistry. As the study of the past VCE Chemistry papers showed the most significant differences occurring in the highest grades it was fortunate that the sample testing was with students that were likely to be well represented in these grades in their final VCE examinations.

The other major limitation in terms of the generalizability of the study is the relatively limited size of the field study of Chemistry students (192 students). Whilst some advantage to the structure of the group is identified in the previous paragraph, these students, in performing the series of small tests provided data that had the intent of more carefully examining how students performed on the different types of assessment mentioned earlier. This unique study compared performance of individual students on questions of equal difficulty but different presentation. A much larger study group would have enhanced the applicability of the results to the general student population but in this instance the research should, perhaps, be considered as more of a pilot study that will provide a foundation for later more extensive studies if the findings warrant this. Further discussion of the limitations will occur later in this thesis.

Overview of the Thesis

Chapter 2 provides a review of the literature regarding the nature of and findings of prior research, some of which has been briefly mentioned in this chapter, into the field of chemistry research as well as the latest views on improving the assessment models in chemistry. Included is an assessment of the limitations of this established research and the specific applicability to the research problem and questions associated with this research program. Chapter 3 considers the research methodology and theory relevant to this research and an overview of the practical research that was undertaken in pursuit of the problem and its associated questions.

Chapters 4 and 5 present the results of the study divided into several sections. The statistical analysis and review of the 2003 to 2007 VCE chemistry papers used to assess Chemistry at Year 12 in the state of Victoria are presented in this chapter. The findings of the field study to examine student performance on a set of carefully constructed chemistry tests are described in chapter 5. The tests aim to consider

student ability and performances taking into account the variables of question type, content and gender of student. The results of the student interviews in regard to various aspects of the chemistry assessment that they have experienced and the views of their teachers are presented in Chapter 5.

Chapter 6 provides a review of the findings of chapters 4 and 5 and discusses their importance. Any significant findings are explained using the literature sources for commentary where necessary and where appropriate. This chapter also presents the overall findings and conclusions of the research and presents that information in response to the initial research problem and addresses each of the research questions. A consideration to any specific further research is discussed.

Chapter 2 Literature Review

Introduction

This chapter considers some of the literature that has examined aspects of assessment that are relevant to this research. First is an examination of the literature with respect to multiple-choice questions and general aspects of their use as an assessment tool. Next, the development of assessment instruments using multiple-choice and/or short-answer questions is considered along with some of the reasons for the use of each type of question. The content of the questions (application or recall) is considered along with several different assessment methodologies including item response theory and two-tier testing. A comparison of the literature in terms of why there are two types of testing (multiple-choice and short-answer) used follows. The issue of student gender and student performance is examined together with the factors of attribution and motivation to conclude the chapter.

Multiple-Choice Questions

Research into multiple-choice questions and short-answer questions has focussed on a number of different aspects. Bridgeman's (1992) study showed some interesting anomalies when examining responses to standard mathematical questions in the Graduate Record Examination (GRE). The performance of students regardless of gender or race was, from an overall score point of view, virtually the same. There were, however, quite marked differences between similar content questions that were answered as multiple-choice or short-answer. In other words, it seemed that students were able to demonstrate different skills and hence were scoring differently on the different type of question even though the content of the questions were much the same. While the study conducted did not look at the differences with respect to the type of question, that is recall as opposed to application questions, it did find some significant aspects with regard to student performance on the different types of questions.

There are many types of multiple-choice question. Haladyna (2004) describes the advantages and disadvantages of 8 question types. Of these 8 types only two are commonly used in VEC chemistry. Conventional multiple choice: usually consisting of a question stem and 4 choices (correct response and three distracters). Typically

about 18 or 19 of the 20 multiple-choice questions in the VCE examinations are of this type. The other type of multiple-choice question used is the complex multiple-choice. This question type has a stem which includes three or four statements and the question options give various combinations from which the student must choose the correct combination (Haladyna, 2004).

Several issues were identified with the use of multiple-choice questions. Students were able to guess, that is, if there are four alternative answers then logically 25% of students could get a question correct simply by guessing. Bridgeman (1992) considered that this aspect of question response could not be easily eliminated. A further issue arose from the fact the correct response was included within the question and was able to guide the student response. Put simply the students were armed with the knowledge that the correct answer was in front of them; they just had to identify it. This was evident in the study where tests were constructed in such a way that the students had to present the working for the question in a typical short-answer mode then select the multiple-choice answer corresponding to their working. One question appeared to mislead students to an incorrect answer in the short-answer version of the question. The incorrect response was not one of the multiple-choice responses, so students who may have initially worked towards this incorrect response were forced to reconsider their working and rechoose an answer (Bridgeman, 1992). If the multiple-choice responses had not been included in the test structure it is highly probable that the students who had worked out the incorrect answer would have left their response at that point and not reconsidered their response. The presence of the multiple-choice answer mode had prompted them to look again because they had to appreciate that their initial working must have been in error. In a sense they were given a second opportunity to produce the correct answer. This approach has some similarities with two-tier testing (Treagust, 1988; Treagust & Chandrasegaran, 2007). This approach to assessment is considered in more detail later in this chapter.

On the other hand, the short-answer questions did give more insight to the students' methods of solving the problem and largely negated the possibility of guessing the correct answer. Bridgman felt that this was an aspect of short-answer questions that was highly desirable in terms of its analysis of student learning.

Other authors have reached similar conclusions in measuring the effectiveness of multiple-choice questions (Braswell, 1990; Martinez, 1991). These authors have concluded that open-ended type questions are superior in assessing students' understanding of concepts because the students' methodology in arriving at their answers can be examined. Multiple-choice questions give no indication of how students' arrived at their answers (Bridgeman, 1992). When considering the students' attitudes towards the different styles of testing, the students were strongly in favour of sitting multiple-choice test items (80%) compared to other styles of test (Bridgeman, 1992). This perhaps, is not overly surprising and supports this researcher's anecdotal view that students feel less threatened by the prospect of a multiple-choice question because they at least know the answer is there (somewhere!). Of interest too were the students' responses to which type of question measured their ability at analytical or application questions. The students were evenly divided over the different type of question suggesting that they recognised that short-answer or open response questions gave them a better chance of demonstrating what they actually understood about the concepts they had studied. This viewpoint is supported by a study that found that students were generally more confident of their answers being correct when answering multiple-choice questions than when answering short-answer questions regardless of the fact that the actual performance on the two types of questions was almost the same. The students felt they had (or were going to) perform better on the multiple-choice test (Pressley et al., 1990). The authors believed that this was due to the inclusion of plausible distracters in the multiple-choice questions, which gave the students a feeling of confidence even when they had selected the wrong answer.

A number of studies have indicated that the gender of the students' influences question type preferences; females have a preference and ability to perform well on short-answer or open ended response questions when compared to males and those males perform better on multiple-choice questions than females (Burton, 1996; Hildebrand, 1996; Walding, Fogliani, Over, & Bain, 1994). This outcome is significant in light of the Victorian Chemistry Examination that has two thirds of the available marks awarded for short-answer questions and one-third for multiple-choice questions (VCAA, 2006d). This observation would suggest that as most of the marks are awarded for a type of question that favours females then females could

potentially perform better on the examinations. This generally appears to be the case because females have performed better on the Year 12 Chemistry examination for a number of years. However, a general trend in the Victorian VCE examinations, not just Chemistry (VCAA, 2006d) is for females to perform better in general terms. Nevertheless, males are typically the best performers at the higher end of the grade structure (VCAA, 2002, 2003c, 2004c, 2005e, 2006c, 2007c).

Type of Examination: Multiple-Choice or Short-Answer

This researcher's general observation is that students tended to perform better (in terms of marks) on the short-answer questions in spite of their general preference for multiple-choice questions. This was one of the significant areas to be explored in this research.

This observation led to the second major focus of my interest, namely, the type of question with respect to the content. For example, do students prefer to answer recall content questions as multiple-choice or do they prefer short-answer questions as a vehicle to demonstrate their understanding. A similar question can be framed regarding student preferences in responding to application/calculation questions. Application questions are those (for the purposes of this investigation), which specifically involve students applying concepts or processes to new situations. Calculation questions (or chemical stoichiometry) form a significant proportion of these types of questions, with a greater proportion of the semester one chemistry examination than the semester two chemistry examination involving chemical calculations. This difference in proportion could lead to differences in levels of performance in the two examinations due to the differences in weighting of these types of questions in the two examinations. Recall questions are primarily based around student's ability to recall and reproduce known facts, processes and structures in answering relevant questions. This aspect is covered in more detail later in this thesis.

The researcher's hypotheses, as already alluded to, suggests that (from anecdotal observations) students prefer recall multiple-choice because the answer is in the question options and that the options may provide the necessary prompt to recognise the answer where as in a short-answer question the prompt is not present.

In considering testing in general, regardless of the type of question, the overall aim of the testing needs to be remembered. Do students learn effectively; are they retaining what they learn or are they learning what they need to know to pass the test and then forget the material? Haynie (1994) tested students after the initial phase of teaching, to see if the type of test made any difference to their students noting that *No difference was found between short-answer and multiple-choice tests as learning aids on the subtest of information, which had not been tested on the initial tests. (Haynie, 1994, p.32)*

Elaborating on this observation Haynie found that the multiple-choice tests did have some advantages in terms of the quality of student learning and he did offer some suggestions as to why this might be.

multiple-choice tests appear to be more effective in promoting retention learning than are short-answer tests as shown by the finding of significantly higher scores This may be because the correct answer to each item is provided along with the distracters in the multiple-choice items, but students had no cues to help them remember the answers, or even reconsider the issues, in the short-answer test items. (Haynie, 1994, p. 40)

The overall results of Haynie's research are supported by the findings of other researchers. Although the area of test type (that is multiple-choice or short-answer) does not appear to be well researched (Gay, 1980; Haynie, 1994; Nungester & Duchastel, 1982), Haynie's research emphasised the importance of the need for testing to support learning. However, in terms of the research reported in this thesis it does not address whether the content of the question was related to the results obtained by the students or whether the student performance was different for recall as opposed to application questions. Similarly, this research does not address the interaction effect between the question type and question content. The main intent of this research is to compare actual student performance on each type of question and on each type of content.

Related to this study were similar findings of Gay (1980) and Sax and Collet (1968) showing that students who take multiple-choice tests tended to perform somewhat better than students who took short-answer tests. The level of retention depended on what type of test was performed to measure the retained learning (Gay, 1980).

Students who tested initially with multiple-choice tests performed better in subsequent tests if the test was again multiple-choice; similarly students who tested with short-answer tests performed better when again retested with short-answer tests. Students who initially responded to short-answer questions performed better on multiple-choice questions than did the students who initially responded to multiple-choice tests were able to perform on subsequent short-answer questions. This finding suggested that the quality of learning from short-answer questions might be better in the longer term even though performance tends to be generally better on multiple-choice tests (Gay, 1980; Sax & Collet, 1968).

In this study, a review of the final assessment in Year 12 chemistry was conducted. The purpose of the Year 12 examinations is to provide a relative ranking of students based on their performance on the examinations. No follow up testing of student performance occurs, so any assessment of the examinations as a measure of longer-term retention or learning cannot be made. The chemistry study design emphasises the importance of developing student understanding of chemistry concepts but does not indicate clearly how this can be determined (VCAA, 2006d).

Content of Questions

Further studies have indicated that certain types of question will be answered differently depending on the test question format. Traub (1992) found that for test items that examined inference or evaluation there appeared to be little difference on whether the question was a multiple-choice or a constructed response question. This, however, he qualified by finding that if the questions did not measure the same mental characteristics then comparing relative performance on multiple-choice questions and short-answer questions was of doubtful value. However, questions that required response strategies to test explicit knowledge produced variable responses from students depending on the test format. Differences in strategy were evident when looking at the different test formats (Niaz & Robinson, 1995; Traub, 1992). Students tended to take more risks with their attempts to answer multiple-choice questions. As mentioned previously this risk taking is prompted by the knowledge that the correct answer is in front of them and students are well drilled to never leave a multiple-choice question unanswered (Barnett-Foster & Nagy, 1996), these findings were based on a study of 300 college chemistry students.

One particular area of chemistry appears to produce distinct differences in the way students respond to questions. Stoichiometric questions seemed to be answered differently depending on whether the question was formed as a multiple-choice as opposed to a constructed response question (Niaz & Robinson, 1995). The study by Niaz & Robinson (1995) measured student performance on stoichiometric calculations involving the gas laws.

Students appeared to be guided by the multiple-choice options and were thus able to provide answers for questions that could not be answered in the short-answer form. That students were often unable to fully explain or justify their choice of answer for the multiple-choice questions suggested that they were prepared to take more risks with the multiple-choice options (Barnett-Foster & Nagy, 1996).

Generally speaking the multiple-choice question is, and has been, a contentious issue in the field of education. Clearly its ease of use makes it appealing, however, in terms of the broader issues of whether it is an effective learning tool, is not so clear-cut. Studies are contradictory in what they reveal about the multiple-choice questions and a number of studies suggest that multiple-choice questions are a more reliable source of assessment than short-answer questions (Barnett-Foster & Nagy, 1996; Ebel & Frisbie, 1986; Traub, 1992). A number of other studies suggest the opposite with the most common cause of complaint about multiple-choice questions being that they encourage guessing and do not test higher order thinking; students are only being tested on their recall (Aiken, 1987; Barnett-Foster & Nagy, 1996; Frederickson, 1984; Petrie, 1986). Critics claim that apart from guessing, students are able to work out answers by eliminating incorrect options or working backwards from the given choices. These options suggest that students are using some higher order thinking skills to work out the answer even though they have not been able to recall or work out the correct answer in the first instance. The general finding is that the short-answer question requires a constructed response whereas the multiple-choice question requires only discrimination between the presented alternatives: *the test format effects should really be assessed within a particular discipline and separate analyses made for different question types* (Barnett-Foster & Nagy, 1996, p 178).

Whether or not the type of question used, be it multiple-choice questions or short-answer questions, assesses understanding better than the other will no doubt continue

to be a source of debate and it is an important reason for conducting this research. A study by Barnett-Foster and Nagy (1996) reached the conclusion, amongst other things, that problem solving type questions appear to be largely unaffected by the style of question used (although some small differences were noted). They also found that the students' strategies for answering questions remained remarkably similar regardless of the test format used (Barnett-Foster & Nagy, 1996). An obvious difference between the two types of test questions, multiple-choice and short-answer is that short-answer questions do allow examination of the answering process more clearly than does the multiple-choice question and also allows the student to gain partial marks for a question.

A Constructivist Perspective

In considering the validity of testing within a constructivist framework Biggs (1996) suggested that most forms of student assessment are flawed if viewed from a constructivist framework (Blais, 1988; Driver, Asoko, Leach, Mortimer, & Scott, 1994; Gunstone, 1995). Generally the problem with most forms of testing is that they focus on students simply recalling information for a test without necessarily forming deep understanding of the material being learned. Biggs (1996) found 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).

A taxonomy of cognitive learning was introduced by B.S Bloom in 1956 as a method of distinguishing and classifying the types of cognitive learning processes (Bloom, 1956). Bloom proposed that there were six hierarchical levels of cognitive learning.

- Knowledge: recalling data or information
- Comprehension: understanding the meaning of instructions or problems
- Application: using a concept in a new situation
- Analysis: separating a concept or material into its component parts so that the links between the parts may be understood
- Synthesis: building a structure from diverse components to form as whole with a new meaning.
- Evaluation: making judgements about the value of an idea or concept,

These were the six levels of the cognitive domain. Bloom also defined two other learning domains; the affective and the psychomotor but discussion of these latter two domains is not within the aims of this research. The relationship between Bloom's taxonomy is deeper than that suggested by Biggs (1996). Using the taxonomy in conjunction with constructivist principles can lead to a learning program that allows a student to develop or construct understanding. This can be achieved at the instructional planning stage and developing a program that allows knowledge construction as a meaningful student outcome (Lee, 1999).

Later reviews and evaluations of Bloom's taxonomy have suggested that to suggest that the later three cognitive levels were hierarchical was too prescriptive and that Analysis, Synthesis and Evaluation were on an equal plateau sitting above the three lower levels of Knowledge, Comprehension and Application (Anderson et al., 2001). The suggestion by Anderson (1972) and Marso and Pigge (1991) that multiple-choice questions rarely went beyond the comprehension level is somewhat of an oversimplification. It is quite possible to have questions that are multiple-choice that can extend students into the higher levels of thinking. Application multiple-choice questions are certainly possible. The three top levels of thinking are more difficult to assess with multiple-choice questions because the abstract thinking required is somewhat obscured by the prompting that the distracters and key in the response options provide for the students.

A more critical view would suggest that if multiple-choice questions are assessing knowledge, it is in terms of the least demanding process, recognition of the correct answer, not even its recall as would be required to successfully answer a short-answer question. Both multiple-choice and short-answer tests further exemplify an insurmountable problem with quantitative approaches to assessment: the content of knowledge is treated as having been learned in binary Units (correct/incorrect), which are then summed, each Unit being seen as equivalent to any other Unit. Not only does this reflect a bizarre epistemology, it nudges the student to focus on details (Biggs, 1996).

Other Item Methodologies

Attempts to provide a more focussed analysis of the types of questions used and their development have led to a variety of methodologies that aim to better understanding the structure and intent of questions. In particular the development of these methodologies is intended to provide a more thoughtful and balanced construct that better and more fairly and accurately measures students understanding. Two of these methods will be considered at this point to reflect the diversity of the research in the field of assessment.

Two-Tier Testing

Two-tiered testing was developed over twenty years ago and is intended to explore in depth the students understanding of science concepts whilst utilizing the convenience of multiple-choice question techniques. The research by Bridgeman (1992) examined some aspects of two-tiered assessment in which students had to explain or justify the answers they had given as their multiple-choice response. This approach has a long history with the idea of explaining the selection of the response going back to the work of Tamir in 1971. A reasonable criticism of this approach could be that the selection of the multiple-choice option serves little particular purpose if the students must then explain the answer. The students may as well simply approach the question as a constructed response or short-answer style question. The main difference in a two-tier multiple-choice test and a short-answer test is that one of the alternative answers informs the solution process and prompts students to reevaluate the method if they do not arrive at one of the offered solutions (Tamir, 1971).

More recently Treagust (1988,1995) has promoted the more refined method of two-tiered test development that uses some of the approach used by Bridgeman but has taken the idea to a more advanced level. The method has gained the support of many academics and has been promoted as an excellent diagnostic tool for students' understanding (Treagust & Chandrasegaran, 2007). A straightforward multiple-choice question gives no indication as to whether the student selected the correct answer for the right or wrong reasons. Conversely students who select the wrong answer may have understood the appropriate method but have made a simple error that has led the student to choosing the incorrect response. The two-tiered test requires students to select the correct answer to the question but they must also select the correct reason for that response: the second tier. Students are only considered to

have understood the concept if both tiers are correctly answered (Treagust, 1988, 1995; Treagust & Chandrasegaran, 2007).

The method of producing a two-tiered test of this nature is more involved than the earlier two-tiered test methods. To offer a credible set of alternative reasons in the second tier requires a degree of trialling and preparation. Typically the conceptual content of the questions must be established. Secondly the typical student explanations and misconceptions must be obtained. This can be achieved using a pilot group of students and through research of the literature. From this the second tier reasons can be formulated and trialled.

Finally, the final test instrument can be developed, which typically starts with a multiple-choice question examining the desired concept followed by the second tier of multiple-choice option reasons or justifications (Treagust & Chandrasegaran, 2007). This method of construction has the advantage of being easy and quick to assess, one of the accepted advantages of multiple-choice assessment but it also has the advantage of the two-tiered test in that it can examine the answer and the reason for the choice of answer. In particular, this approach helps overcome one of the perceived disadvantages of the multiple-choice question, that is, students guessing the answer. Whilst not eliminating the problem it reduces the likelihood of a student guessing too many questions. The probability for any individual item goes from $\Pr(\text{correct}) = 0.25$ to $\Pr(\text{correct}) = 0.0625$ assuming 4 options at both tier one and tier two.

Item Response Theory methodology

Item response theory measures the performance of each student on each item in a test. Student performance is calculated using sophisticated mathematical models such as Rasch analysis. Rasch analysis is considered by some researchers to be superior to classical test theory that relies upon a large sample base for the purposes of calculating student ability (Andrich, 2005). As mentioned earlier the typical test used in Victorian VCE examinations consists of a combination of multiple-choice questions and short-answer questions, which is considered a suitable outcome for a variety of reasons, namely 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 next problem to consider is whether breadth of cover is more important than depth. A number of studies have looked at evaluating whether the two different assessment item types can be calibrated to give an equal overall result. Essentially, can they be calibrated on the basis that the two item types (multiple-choice question and short-answer question) are essentially assessing the same material? Studies by Sireci and Thissen have had some success in producing a combined scale (Sireci et al., 1991).

Through their extensive study using an item response theory model, Ercikan et al. (1998) concluded that there was sufficient similarity in the material being assessed by both item types question (multiple-choice and short-answer) to allow the creation of a common scale that would provide a single set of scores for the responses of both item types (Ercikan et al., 1998). Generally there was a good correlation between the multiple-choice questions and the short-answer questions except on those occasions where the short-answer questions were particularly difficult or when the allowed response time for the test item was too short to allow students sufficient time to fully answer the questions. Furthermore, the two types of questions, whilst being amenable to a common scale, did provide different information about certain types of students. For example, the short-answer question was able to provide more exact information about the understanding of both the very low achieving and very high achieving students that was not evident from the multiple-choice question. The multiple-choice questions were also a more effective measure when comparing student outcomes across years or regions because the marking of them is not subjective in the way that short-answer question marking may be (Ercikan et al., 1998). This is a particularly useful aspect of the method. Whilst multiple-choice questions are machine marked, or at least objectively marked, constructed response questions are normally evaluated by human markers and therefore there will naturally be variation between years and areas with regard to the consistency of the evaluation. Having calibrated the multiple-choice questions with the constructed response items, these multiple-choice questions can then be used to anchor the marking of the next set of items in subsequent years without the need for the more time consuming equating of constructed response items (Ercikan et al., 1998).

Rasch and Classical test theory

The position of item response theory has been enhanced through the application of models such as Rasch analysis (Andrich, 1988). Compared to classical test theory, item response theory focuses more on the quality of the items and the individual items ability to distinguish between participants for particular traits under analysis. 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, 1996). At the same time Rasch analysis places all the items on a relative common scale. Models such as Rasch examine and measure the probability that the participants will answer items correctly. Classical test theory places emphasis on a large number of items with high correlations between items to enhance the interpretation of results (Pallant, 2010). The unidimensional nature of the trial tests in this research fits well with item response theory. The items were tested using Rasch analysis to demonstrate the validity of the test construct in terms of unidimensionality, and with a small number of items involved in the test, the supportive correlation of the items which is a feature of classical test theory was not possible in the trial tests (Pallant, 2010).

Why Two Types of Question?

Other studies have looked at the relative merits and reasons for using the two types of test items. This is significant because many large-scale testing programs rely heavily on multiple-choice questions as the basis of the test regime. The previously mentioned reasons such as scope of questions and quickness and ease of marking are usually considered as being significant in the choice of this method. The difficulty of marking short-answer questions arises due to the time consuming and expensive use of trained markers or assessors. These questions also take longer, generally, for the examinee to answer (Lukhele, Thissen, & Wainer, 1994). In spite of this issue, short-answer questions are still used because of the perception that they are better at testing in-depth knowledge (Anderson, 1972; Lukhele et al., 1994; Marso & Pigge, 1991). Another feature of the short-answer question is that it can better represent the tasks that the students are likely to face in future academic and work settings. So aside from the direct purpose of the testing, students gain from the indirect benefit of responding in writing much as they may have to do in a typical work environment. Furthermore, the necessary literacy skills involved in constructing a written response

are enhanced in a way that would not occur in a purely multiple-choice testing program. Thus the benefits arising from this form of testing goes beyond the pure assessment tool for which it may have originally been designed (Chan & Kennedy, 2002; Lukhele et al., 1994). The dominant view point would still appear to be that the multiple-choice question based assessment will continue to be important from the aspects of time, economy and breadth of cover but the short-answer questions should continue to be encouraged because of the diversity added to the assessment and the belief that by using appropriate item response theory some of the correlation problems between the two item modes can be overcome.

A Summative Comparison of Multiple-Choice Questions and Short-Answer Questions

Any evaluation of the two types of assessment involves a comparison of the two methods of assessment. In attempting to summarise this information, Tables 2.1 and 2.2 outline some of the differences found by various researchers who examined various aspects of the comparison of multiple-choice questions versus short-answer questions (constructed response) or essay questions as well as gender related differences.

Table 2.1: Advantages of multiple-choice tests over constructed response tests.

Issue	Researchers
Machine gradable, thereby increasing scoring accuracy	Holder & Mills, 2001; Kniveton, 1996; Walstad, 1998; Walstad & Becker, 1994
Efficient way to collect and grade examinations from large numbers of test takers, Dufresne	Dufresne, Leonard, & Gerace, 2002
Helps certification examiners agree on questions to ask a large number of test takers	Bridgeman, 1991; Bridgeman & Rock, 1993; Holder & Mills, 2001; Snyder, 2003
Facilitates referencing the correct answer in textbook or other source,	Bridgeman & Lewis, 1994

Table 2.1 continued...

Perceived objectivity in grading process,	Becker & Johnson, 1999; Wainer & Thissen, 1993; Zeidner, 1987
Facilitates timely feedback for test takers in classes, and immediate feedback in web-based systems,	Delgado & Prieto, 2003; Epstein & Brosvic, 2002; Epstein, Epstein, & Brosvic, 2001; Kreig & Uyar, 2001
Enables instructors to ask a large number of questions on a wider range of subject materials	Becker & Johnson, 1999; Lukhele et al., 1994; Walstad & Becker, 1994; Walstad & Robson, 1997
Helps students avoid losing points for poor spelling, grammar, or poor writing ability,	Zeidner, 1987
Easier preparation by test takers	Carey, 1997; Fredericksen & Collins, 1989; Ramsden, 1988; Scouller, 1998
Does not require deep understanding of tested material (student advantage),	Beard & Senior, 1980; Biggs, 1973; Entwistle & Entwistle, 1992
Reduces student anxiety,	Snow, 1993
Multiple versions of the same multiple-choice examination helps thwart cheating	Kreig & Uyar, 2001; Wesolowsky, 2000
Helps avoid inconsistent grading of essays,	Kniveton, 1996
Availability of computerized multiple-choice test banks, answer keys, and test generators. Test takers can increase the probability of guessing the right answer to a question by eliminating unlikely choices (student advantage),	Bush, 2001; Hobson & Ghoshal, 1996
Electronic test items can easily be edited, pre-tested, stored, and reused	Haladyna & Downing, 1989

Table 2.2: Disadvantages of multiple-choice tests over constructed response or short-answer question tests.

Issue	Researchers
Difficulty in preparing questions without access to a suitable item bank,	Brown, Bull, & Pendlebury, 1997
Poorly written questions actually hiding true student ability,	Dufresne et al., 2000; Becker & Johnson, 1999
The difficulty of writing suitable distracters that do not adversely affect the students ability to engage with the question	Haladyna, 2004
Disadvantage to students where English is not their first language	Paxton, 2000
Examinations that are too easy and provide an inaccurate indicator of student understanding. This gives a misleading indication of students' grasp of course concepts or mastered course materials	Chan & Kennedy, 2002
Students become use to doing multiple-choice test and develop strategies that enable them to do well without actually having a good understanding of the course material. E.g. eliminating the obvious incorrect options to improve the chances when guessing	Fenna, 2004; Martinez, 1999; Rogers & Harley, 1999; Zimmerman & Williams, 2003
Gender bias in favour of males in multiple-choice question test,	Bell & Hay, 1987; Bolger & Kellaghan, 1990; Bridgeman & Lewis, 1994; Lumsden & Scott, 1987
Multiple-choice question test favour recall learning over application or critical thinking	Martinez, 1999

The reasons identified in Table 2.1 are those advantages of multiple-choice tests perceived by researchers and examiners. Students tend to have different perspectives on testing which include:

- (a) the perception that multiple-choice question tests are objective and, therefore, avoid instructor bias;
- (b) the ability to guess correct answers, usually without penalty;
- (c) the ability to get credit even if he or she is a slow test taker, presumably from the ability to at least guess answers when they are running out of time; and
- (d) the perceived ability to do better on multiple-choice question tests than on essay or other forms of short-answer or constructed response tests.

Student preferences for a specific test format appear to have functional validity as well, because students have been demonstrated to perform on these multiple-choice tests at a level beyond their perceived ability (Simkin & Kuechler, 2005).

Do multiple-choice questions and short-answer questions examine the same concept?

Several authors have indicated that certain types of learning are favoured by one type of test over the other. Typically multiple-choice question tests favour rote or recall skills whereas short-answer question tests favour critical or application skills (Aiken, 1987; Anderson, 1972; Chan & Kennedy, 2002; Ercikan et al., 1998; Petrie, 1986; Simkin & Kuechler, 2005). Thus the question, do they measure the same skills to the same level is critical to this thesis report. A study by Simkin and Kuechler demonstrated that whilst they concluded it was not possible to construct a multiple-choice question test that was able to reach the highest levels of what may be obtained by a sophisticated short-answer question test, they noted that with careful preparation the multiple-choice question test can be constructed so that it correlates well with the short-answer question or constructed response item test (Simkin & Kuechler, 2005).

One aspect of the consideration of short-answer questions and multiple-choice questions is whether or not they are measuring the same thing with the same degree of validity. Research on this aspect has shown some degree of uniformity in the responses to those studies (Chan & Kennedy, 2002; Simkin & Kuechler, 2005; Thissen, Wainer, & Wang, 1994). Other authors have been less certain of this mutual support of purpose (Aiken, 1987; Anderson, 1972; Ercikan et al., 1998). Thissen et al. (1994) also considered whether or not it was valid to combine the scores of a multiple-choice question test with those of a short-answer question test to give an overall combined score. Their analysis of college examinations using a very large

sample showed that, whilst there were some differences in the way in which the short-answer questions and multiple-choice questions were answered, there appeared to be little doubt that they were measuring much the same thing in terms of their reliability in demonstrating student understanding.

The authors also determined that it was not detrimental to combine the scores because it did not adversely influence the overall outcome of the testing process. However, if the purpose or intention of the short-answer questions was to show or demonstrate different outcomes to the multiple-choice questions then the scores should not be combined. Rather a different factor model of scoring such as that suggested by Bennet et al. (1991) would be more appropriate, their research showed that a single factor model provide the most efficient fit for data gained from an advanced placement test. Bennet et al. (1991) initially started with a two-factor analysis treating short answer and multiple choice questions as separate factors, however, the findings showed that a single factor analysis allowed adequate comparison of the two item types. If a test is designed where it is intended to show some overall level of understanding in a particular subject and that test was formulated with both short-answer questions and multiple-choice questions then it would not be unreasonable to combine the two parts of the test as they were both intended to be demonstrating a similar outcome. On the other hand, if the test was designed in such a way as to demonstrate different skills through the different styles of questions, for example, multiple-choice questions may have been used only to show recall knowledge but the short-answer questions designed to demonstrate interpretation or application of theory, then it would not be appropriate to combine the results because they are measuring different learning processes (Bennet, Rock, & Wang, 1991).

Gender Differences in Performance

Another aspect of the VCE examinations that provided interest arose from the observation that students tended to perform better on the second examination (that is the semester two examination-which only examines semester two work) than they did on the semester one examination. There was an important difference with respect to the work covered in each semester. Semester one heavily focussed on application questions strongly based around calculations with only about a one-third of the curriculum being described as descriptive chemistry which relied on recall of detail

to answer questions on the examination. Semester two was the reverse in balance, with approximately 75% of the examination based on descriptive chemistry. If the better performance on the semester two examinations could have been explained in terms of the increased maturity of the students, then the increase in performance could have been reasonably understood. Discussions with teachers of at other schools suggested that this was not necessarily the case at their own schools. A significant factor here seemed to be the gender of the students involved. The researcher taught at an all female school whereas his colleagues, where the differences were noted, taught at all male (or predominantly all male) schools.

Whilst there have been numerous studies examining various aspects of the gender participation and performance by males and females in mathematics and the sciences, these reports have focused either on science or mathematics in general terms. For example a study that examined Australia's participation in the Trends in International Mathematics and Science Study (TIMSS) in 2007 showed that Australian female students outperformed their male counterparts in Science at Years 4 and 8 decisively (Thomson, Wernert, Underwood, & Nicholas, 2007). More pertinent to this research were the findings of a study that examined the final year participation in science subjects in the Victorian Certificate of Education (VCE). The study found that females formed the larger proportion of all students studying all the sciences except physics, and, in terms of study score rankings, the female students outperformed male students in all subjects except Chemistry (Cox, Leder, & Forgasz, 2004). A study of the GCSE examination in England over a three-year period demonstrated significant gender differences in both participation and performance. Generally females performed better in most subjects except some of the sciences (chemistry and physics) and mathematics. Using English and Mathematics as benchmarks the researchers found that females significantly outperformed boys in obtaining A to C grades (54.6% of females obtaining these grades whereas only 41.5% of boys achieved A to C grades). In mathematics the situation was reversed, 38.9% of males obtaining A to C grades and only 34.6 % of females (Stobart, Elwood, & Quinlan, 1992). Similar patterns were found in each individual science subject though the differences tended to be a little less dramatic due to the more selective entry to these (Stobart et al., 1992).

The importance of gender in performance in science and mathematics subjects has been the starting point of numerous research studies. It is an emotive issue along with socio-economic and racial backgrounds and as such the number of studies focussing on gender and performance is large. A common theme, which has prompted much of the research, was the observation that male students tended to outperform female students in the science subjects. Research by Boli et al. (1985) attributed this difference to performance in mathematics. Not only were male students more likely to participate in the higher mathematics studies, these students were also more likely to perform well in such studies. Female students in the final years of secondary schooling support this finding in British studies, which examined participation and performance. The studies found that females were less likely to choose the sciences compared to males (Francis, Hutchings, Archer, & Melling, 2003). A contributing reason for this lower participation rate and performance rate was partly attributable to a general lack of interest by females in pursuing careers in the science field (Schoon, 2001). Similar findings were also made in a study conducted in the USA where the researchers found that interest in science subjects waned along with their career aspirations as female students progressed through secondary school (Watson, Quatman, & Edler, 2002). The situation in Australia is somewhat different with respect to the Victorian VCE examinations. In recent years (beyond 2000) the participation rate by females in the science and mathematics subjects has been consistently high. The number of females studying the more rigorous traditional science of Chemistry has usually outnumbered the males. This may in part be a reflection of the importance of Chemistry as a required prerequisite subject for entry into many University course. There is also a larger number of females participating in the Victorian VCE overall. However, in all the sciences and mathematics subjects the males consistently outnumber the females in the achievement of the highest grades in the subjects (VCAA, 2002, 2003c, 2004c, 2005e, 2006c, 2007c).

More detailed analysis of the results of previous studies did show that there were both positive and negative differences between the performances of the males and females. Boli et al. (1985) observed that males tended to perform better in tasks that involved higher level cognition and that the females out performed males in computational tasks. Overall there appeared to be no directly gender related reasons for the males outperforming the females yet the evidence showed that this was the

case. Boli et al. and later researchers theorised that the mathematics background of the females was less rigorous than that of the males and this was having a flow on effect in the later studies of both mathematics and science. 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 (Boli et al., 1985).

A repeated study by Beller and Gafni (2000) examined the performance of males and females in mathematics on the same examination over a three-year interval. The findings from this research demonstrated that the performances of females and males were not consistent. In the first study, correlations between performance and gender demonstrated that male students outperformed female students most significantly in multiple-choice questions but there was less difference on open-ended questions. This supported earlier findings that suggested that the open-ended questions tended to be more effectively done than multiple-choice by female students. The follow up study in 1991 demonstrated that in this instance the open-ended questions most significantly favoured male students' performances, essentially contradicting the first set of findings. What was consistent between the two studies were the findings that the harder the questions were (regardless of format) the male students were more likely to do well (Beller & Gafni, 2000).

If the female students received equivalent levels of preparation to the male students then there was little significant difference in performance (Boli et al., 1985). What this study did not address, and which this research attempts to, is whether or not the style of question in both presentation (short-answer questions or multiple-choice questions) and the content (analytical or recall) will exhibit gender related differences in performance. Little work appears in the literature on this aspect of multiple-choice questions versus other question formats in relation to gender. One study with economics students found that in some instances students answered multiple-choice questions more successfully but in other instances there seemed little difference. This study also found that there was little to separate the performance on multiple-choice questions of male students and female students and also between lower achieving and higher achieving students (Chan & Kennedy, 2002).

Whilst most studies continue to support the idea that males outperform females as mentioned above the picture is not quite that clear-cut. A large study of a variety of large-scale assessments showed that males generally performed better than females;

however, closer examination showed that the content of the question did influence the performance. Males tended to outperform females in the areas of the physical sciences whereas in the life sciences the differences were negligible (Beller & Gafni, 1991; Hamilton, 1998; Hedges & Howell, 1995). Such differences have been supported by numerous other studies (Francis et al., 2003; Hamilton, 1998; Watson et al., 2002). The underlying cause is open to debate but one belief is that the males' generally better ability on spatial measures is considered to be significant, however, it is unclear whether and how increased spatial ability should affect performance in the physical sciences. What is noted is that there is some correlation between spatial ability and science performance (Halpern, 1997; Hamilton, 1998; Linn et al., 1991). Hamilton (1998) also reinforced the findings of other researchers that multiple-choice questions tended to favour males over females whereas the differences were less marked in short-answer questions. With respect to short-answer questions the findings were mixed; some studies indicated that males performed better whereas others suggested that the females were advantaged by the short-answer format (Hamilton, 1998). Further examination of this apparent trend has suggested that the advantage female students tended to have with the constructed or short-answer responses has been usually associated with questions that had significant language content. Overall those questions that require higher order thinking or required attempting questions that were dissimilar to those that they had already attempted were seen to favour the male students (O'Neill & McPeck, 1993; Rennie & Parker, 1991).

Research has also indicated some of the underlying causes that may result in differing performances on multiple-choice questions by male students and female students. The differences appear to stem from attitudinal approaches to the questions. Male students tend to have a more literal approach to questions, answering as they see them and are more willing to guess rather than leave a question unanswered. Females tend to take fewer risks in answering questions and hence are less likely to guess answers (Ben-Shakhar & Sinai, 1991).

The study by Hamilton (1998) showed several important conclusions supporting some of the assertions already mentioned. Questions that involved higher levels of spatial ability appeared to significantly favour males and with respect to this it appeared to matter little whether the questions were presented as multiple-choice or

short-answer. Females were least disadvantaged by questions that were heavily dependent on school-based learning as opposed to experiential learning that may have taken place outside the classroom environment. Thus activities that males undertook outside the classroom enhanced their spatial abilities that in turn favoured their performance on related test items (Hamilton, 1998). Hamilton also believed that the differences in performance between the males and females could not be fully accounted for by these observations and the author concluded that other cognitive processes were important in accounting or helping to explain the differences. This experiential advantage of males is supported by research of others and was consistent with the findings of Hamilton's research (Entwistle, Alexander, & Olsen, 1994; Linn et al., 1991). Hamilton (1998) concluded that in constructing large scale tests the format and source of questions should be varied so as not to disadvantage or advantage either males or females or any particular socially experienced group.

Attribution Theory

The difference between the genders in performance has been most studied in mathematics, where males tend to outperform and also outnumber females in the higher levels of education. This trend is also repeated to a lesser extent in the sciences, particularly in chemistry and physics. The number of males studying physics and specialist mathematics are significantly larger than females. Numbers of males and females studying chemistry and biology have declined slightly over time as a percentage of the Year 12 cohort but have generally remained even in terms of males and females (VCAA, 2008). To some extent this is probably due to growth in the numbers of students undertaking psychology and biology where the number of females is much greater than male numbers (Cox et al., 2004; Dekkers & De Laeter, 1997, 2001; Dobson & Calderon, 1999; Wolleat, Pedro, Becker, & Fennema, 1980). That is, females are tending to avoid these perceived more difficult subjects.

Whilst performance level can be influenced by the students' own attitudes and expectations about a particular subject, the attitudes of the teachers themselves can have an impact. Teachers have been found to attribute increased or decreased performance by students to factors other than improvement in performance by the student. For example, teachers may attribute an improvement by students to be a result of an easier test or better teaching by himself or herself rather than genuine effort or talent from the student. This effect is even more pronounced when students

are categorised. For example, teachers are more likely to attribute increased performance by students categorised as less achieving or disadvantaged to factors other than the students efforts than they would be if the student were categorised as higher achieving. They have also been shown to view improvement by males differently to that of females (Rolison & Medway, 1985).

Difference in gender attributions to success or failure were demonstrated in a study that found male students, for example, were less likely to attribute success or failure to the level of help from their teacher; female students on the other hand were more likely to attribute failure to a lack of teacher support and a general belief that they were not likely to do well in the first place (Lloyd, Walsh, & Yailagh, 2005). This attribution of success to gender was also demonstrated in a study of sociology students that found considerable unintended gender bias amongst the students. Students attributed success in their course more often to the instructor if that instructor was a male as compared to the instructor being female. Students were noted also to have a belief that the male instructors were more likely to be more highly qualified than the female instructors. These attitudes were even exhibited in the way the students referred to the instructors. Male instructors were referred to as the professors whereas the female instructors as teachers (Miller & Chamberlin, 2000).

Attribution theory suggests that causality or attributions fall into one of four categories: ability, effort, task difficulty and luck. The four categories are derived from two factors, each of which has two levels: causation and stability. The attributions can be expressed in a two by two matrix shown in Figure 2.1.

		Causation	
		Internal	External
Stability	Stable	Ability	Task
	Unstable	Effort	Luck

(Wollett et al., 1980)

Figure 2.1 Causality attributions

The theory predicts that performance that is consistent with the student expectations will attribute to a stable situation whereas results that are contrary to expectations will lead to an unstable situation. The actual outcome, success or failure is not

important. That is, someone who expects to fail and does will result in a stable situation because the result was as expected. The importance of this is that if a successful outcome is attributed to a stable factor, then the same outcome can be expected in the future. The effect of such an outcome on student motivation is an entirely separate issue and is discussed in the following section. If, however, the outcome is attributed to an unstable situation, even if successful, the same outcome cannot be reasonably expected in the future (Wolleat et al., 1980). This leads to a situation of learned helplessness, which describes the situation where failure is viewed as inevitable and leads to the student developing a lower motivation to persist. Apparently, females are more likely to display the condition of learned helplessness, which can reduce the motivation to persist with the harder mathematics and sciences, and as a result they are likely to either underperform or simply opt out of these subjects (Lloyd et al., 2005; Wolleat et al., 1980).

Motivation

Motivation is a key factor in determining success in education. The level of students' commitment to learning is seen as a key factor in determining students' levels of motivation and success in any course of study. The factors that affect motivation, both positively and negatively, are therefore important (Buehl & Alexander, 2005). Success is important to motivation. Students who believe that their success is in their own hands are described as *internals* whereas students who believe their success in school is out of their control are described as *externals*. Success and motivation are more strongly linked to *internal* students than *externals* (Hattie, 2009).

A study by Ryan and Patrick (2001) found that motivation and engagement by students was strongly determined by the motivation and engagement that had been developed in prior years. Factors such as gender, race and prior achievement were less important in determining the level of motivation in subsequent years. The most important factors in determining motivation were found to be teacher support of the students and teacher encouragement of mutual respect and interaction. Teacher promotion of performance goals was seen to have a negative impact on motivation and engagement (Ryan & Patrick, 2001). A study that focussed on mathematics found that the motivation of students was often determined early in the students' education and was strongly linked to successful achievement by the students as being an identifiable long-term purpose to the study. In other words if success was

achieved early in their study of a subject then it was important that success and purpose be fostered in later years for motivation levels to be maintained in subsequent years (Middleton & Spanias, 2005).

The relationship between motivation educational successes was examined in a review of studies conducted by Becker (1989). In this analysis or re-examination of nearly 30 previous studies Becker concluded a clear advantage in achievement levels of male students over female students in the broad areas of science. Differences in motivation and attitude towards science again favoured male students but the difference was less than the difference in outcome achievement (Becker, 1989). Further to this Becker was also able to support other findings (Beller & Gafni, 1991; Hamilton, 1998; Hedges & Howell, 1995; Wolleat, Pedro, Becker, & Fennema, 1980) that have indicated that the subject areas that most favour males over females are the more traditional pure sciences of physics and chemistry (Becker, 1989). Jones and Kirk (1990) explored the differences by males and females 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 studying a science they saw as a *helping* science, a *people oriented* science or a *nurturing* science. Subsequent studies by Francis et al. (2003), Schoon (2001) Stobart et al. (1992), and Watson et al. (2002) have all demonstrated similar patterns in enrolments and participation in higher school sciences; however, the differences are not as great as they once were. In recent years in Victoria female enrolments have achieved levels in Chemistry for example that have parity with male enrolments. The figures for Physics and Specialist Mathematics are still significantly disproportionate in favour of male students (VCAA, 2007d).

In comparing the genders with respect to motivation the study by Lloyd et al. (2005) found that males were more confident in the study of mathematics than females. 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 the male students' expectations of a career were founded in the study of mathematics and science (Lloyd et al., 2005).

Summary of Chapter 2

The studies reviewed can be simplified to two general foci, which are consistent with the aims of this study.

The first focus related to the type of questions and the styles of questions.

- A number of the studies examined the effect of the advantages and disadvantages of asking multiple-choice questions as opposed to short-answer questions.
- These studies were generally consistent in the conclusion that females were more often favoured by the latter style of question in comparison to males and that males tended to be advantaged by the former.
- The studies also generally found that students more often liked multiple-choice questions over short-answer questions and a variety of possible reasons were proposed; the answer could be prompted from the options given seemed to be a significant factor.
- Multiple-choice questions are often favoured in large scale testing programs by administrators due to the efficiency of their application and scoring.

The second aspect of the reviews focussed on the differences in performance by female students as compared to male students. The studies were somewhat more unanimous in their findings but less so with respect to the reasons.

- Generally male students tended to outperform female students in tests that focussed on the harder sciences (Physics, Chemistry and Mathematics) and this was particularly evident in the more challenging levels of questions.
- Female students usually outperformed males or at least scored on the same levels, as did the male students in the softer sciences of Psychology and Biology. The reasons offered tended to suggest that it was a combination of teaching and attitude both personal (with respect to the females having lower self confidence levels) and societal (females not being expected to do sciences).
- This trend is, however, not wholly demonstrated in Victorian schools where females generally outperform males in the VCE. However, the very top performances in subjects like Chemistry are usually from male students

whereas larger numbers of female students perform well in the higher levels overall (Victorian Curriculum and Assessment Authority, 2002, 2003, 2004, 2005, 2006, 2007).

The merging of these two foci gives an overall picture of the purpose of this study. Put broadly:

- Firstly, is the general performance of males and females in chemistry in the VCE affected by the nature and structure of the examination in terms of the proportion of multiple-choice and short-answer questions and also the distribution of questions by the content type (recall or application)?
- Secondly why is it that despite outperforming male students generally in VCE Chemistry, female students are not well represented in the distribution of the higher grades?

Chapter 3: Research Methodology

Research Method

This chapter includes a summary of the general principles of the research methodology with specific reference to that employed in this thesis and also includes a discussion of the data collected, the methods employed and the analysis involved. A discussion of the trustworthiness of the process and related ethical issues are also included. The chapter considers the relevance of theories aligned with both qualitative and quantitative analyses and examines the use of triangulation as a method of data validation. Interviews form one aspect of the data collection and the chapter discusses the various approaches to the interview process and the conduction of the interview. A discussion of the important considerations to the research of validity and credibility and some related issues also takes place. A description of the practicalities of the research that was undertaken then follows. This involves a description of the three data sources in this research - interviews, analysis of past papers and the sample-testing program. These three data sources provided the information necessary to answer the research questions posed below. Finally a discussion of the ethical considerations that impact on this research is addressed.

Relationship Between the Research and the Research Questions

The topic being explored in this thesis is: Multiple-choice Questions Compared To Short-answer. Which Assesses Understanding Of Chemistry More Effectively?

The research questions that underpin this are:

1. Do students perform more effectively on multiple-choice or short answer questions?
2. Do students perform more effectively on recall type questions or on application questions?
3. Does students' gender influence performance in chemistry examinations (or tests)?
4. Do students have a preference for the type of question style in terms of
 - a. Multiple-choice or short-answer in general terms?

- b. With respect to whether the question is assessing recall or application?
5. Do teachers consider multiple-choice or short-answer questions to be the most effective in demonstrating their students understanding of chemistry?

Table 3.1: Relationship between the research questions and the data collection

Summary Research Questions	Information enabling a response to the research questions provided by		
	Interviews	Analysis of past papers	Sample tests
1. Effectiveness of multiple-choice questions	Yes	Yes	Yes
2. Effectiveness of recall and short-answer questions	Yes	Yes	Yes
3. Gender influences on student performance	Yes	Yes	Yes
4. Student preference of question type	Yes		
5. Teacher opinions	Yes		

Setting the Scene

How you see the world is largely a function of where you view it from, what you look at, what lens you use to help you see, what tools you use to clarify your image, what you reflect on and how you report your world to others (Anderson, 1998, p.3)

Anderson's observation clarifies one of the major difficulties of educational research. How can researchers divorce themselves from their own history of learning and social experience to be able to reflect on an issue in an unbiased manner? This of course is very difficult, as one cannot remove oneself from the past. What the researcher must then do is devise methods of research that can be as independent from their life experience as possible even if the problem being explored is derived from that experience.

Anderson (1998) recognises that the type of research will in some way influence that nature of the research and the interpretation of the results. The empirical researcher would appear to have the most likely chance of producing independent research, that is, research conducted in such a way as to be as little influenced in its outcomes by the researcher's own beliefs and views. However, researchers will still be subject to their personal life experiences once they reach the point of interpreting that research.

Approaches to Research in Science Education

The difference between the methodology of the research and the methods employed in the research can often be confused (Guba & Lincoln, 1989). It is important to understand that the methodology attempts to understand, explain and justify the reasons and approaches of the research rather than the methods actually used. The methodology should inform the approaches taken and these should be in a sense self-justifying if the methodology employed is sound. As Guba and Lincoln explain *far from being merely a matter of making selections from among methods, methodology involves the researcher utterly-from unconscious worldview to enactment of that worldview via the enquiry process* (Guba & Lincoln, 1989, p183).

Quantitative research has become a powerful and accessible research tool through the introduction of sophisticated statistical packages such as SPSS, Conquest and RUMM2030 which have allowed the researcher to statistically validate the results of enquiry methods such as the use of tests, interviews and questionnaire responses (Andrich, 1988). The use of mathematical models has had a wide appeal to researchers and many have employed their use (Burns, 1994; Denzin & Lincoln, 2005; Mason, 1993; Ryan & Bernard, 2000). The ability to use a purely mathematical approach to analysing an educational environment is not without some risk. The classroom is not a scientific laboratory operating in an always-predictable manner. That there are differences between the approaches necessary for pure scientific research and for educational research is now recognised (Bell, 1993). Consequently the use of qualitative research methods in the field of educational research has wide application.

Quantitative and qualitative methods of enquiry each have their place and particular strengths in research design and methodology. Quantitative methods on the one hand require the application of clearly defined and rigorous approaches so that the responses of the subjects will deliver results and observations in a defined and

expected range. A method that produces results that have a scattered response range are unlikely to reproduce observations that will allow valid or reliable conclusions to be drawn. Thus quantitative research is likely to draw on controlled data that has depth and can be subjected to statistical analysis tools such as RUMM2030 (RUMM Laboratory P/L, 2009) and other similar statistical packages. Qualitative research is more open-ended and is usually able to examine issues in much greater depth and detail. However, being more open the boundaries of qualitative research are less constrained and to some extent will be guided by the ongoing research itself. Typically qualitative research produces considerable data about a small number of participants over a range of particular focus points. Quantitative research produces a large quantity of data usually from a larger number of participants but in a clearly defined range of responses (Patton, 1990; Peshkin, 1993).

In this research both methods of research were employed in their particular areas of relevance. Data were collected quantitatively through the analysis of past test papers from the VCAA (VCAA, 2003-2007) and through the application of trial papers with volunteer subjects in the school environment. The details of the methodology of these are discussed in detail later in this chapter. Qualitatively the analysis and application in this study come from the results of the interviews with students about their attitudes and feelings about the different types of testing methods that they have typically experienced. Comments from their teachers were also considered in reviewing the opinions of the students in response to the questions asked of them. The details of the application and design of this section of the study are also discussed in later in the chapter.

Sound research designs, whether they are qualitative or quantitative are necessary to ensure that scientifically valid conclusions can be reached and are able to pass the scrutiny of validity and relevance. Traditional research draws evidence from previous studies and combines it with current research to reach conclusions, which allow the combination of previous wisdom with current research. This process is enhanced by the use of a mixed methods approach, which allows both considered evaluation by the researcher combined with statistical calculations that can emphasise trends in the data. In other words, the best research design is a mixed method design that integrates qualitative and quantitative research. As described by Condelli and Wrigley (2004)

This type of design begins with a strong research methodology with quantitative methods that are enhanced with qualitative measures of key processes and outcomes. Qualitative methods, such as interviews and case studies, improve the design by providing data that can give insights into how findings work and how findings can be translated to practice. By itself, a quantitative method can identify what works, but has limited explanatory power.....By combining the two methods; we can obtain a much richer understanding. In other words, using a rigorous design the quantitative methods can tell us what works, while the qualitative methods can tell us how it works. (Condelli & Wrigley, 2004, p. 2)

The conflict between uses of the two methodologies stems from the differences in the paradigms governing each method. The quantitative approach is supported by the positivists, whereas the qualitative approach is supported by the constructivists (Tashakkori & Teddlie, 1998). This dichotomy of paradigm approaches has resulted in researchers tending to support the methodology that most closely aligns with their own viewpoints and beliefs. Consequently support for the combined approach was initially somewhat limited. Denzin and Lincoln have only one chapter in their book, *The Sage Handbook of Qualitative Research*, devoted to mixed methodology (Denzin & Lincoln, 2005). Other authors support the viewpoint that because the philosophical and theoretical differences in the paradigms supporting each method they are incompatible and should not be combined (Burrell & Morgan, 1979; Smith, 1983).

The number of authors, however, that are now supporting the mixed methodology approach is growing (Crabtree & Miller, 1999; Creswell, 2008; Mason, 1993; Mertens, 2005). The argument being that the two apparently incompatible methods can be integrated to give both depth and diversity to a study, which may otherwise be limited by the use of only one method. The method or methodologies used should be informed by the research questions asked and not by a fundamental philosophical viewpoint held by the researcher.

Methods such as interviews observation and questionnaires have fully taken their places as valid methods of research alongside the more quantitative and numerically measurable approaches such as achievement testing. Even then, the interview and

survey can be statistically validated, making it a quantitative method based on naturalistic inputs.

As mentioned previously there are advantages to using a combination of the two types of methods. Research may often have multiple purposes and therefore the research methodology should support these approaches or intended outcomes. Consequently it is likely that a purely qualitative or quantitative approach may not be satisfactory (Fraser, 1994; Fraser & Tobin, 1991, 1998). Understanding data collected is more meaningful if there is an understanding of the inherent mathematics of the data collected, that is, numbers without understanding are just numbers. Having a qualitative understanding of the quantitative data enhances the understanding elicited by the researcher (Carver, 1993; Fraser & Tobin, 1991). In this study the need for the two approaches is evident in that there is a quantitative measurement aspect to the research as well as a qualitative aspect. In other words this study incorporates the assessment and understanding of the data as well as the interpretation of the meaning behind that data. The use of combined methods is also useful in that it enhances the credibility and reliability of the results obtained. Having many sources of data improves the credibility of the conclusions through the demonstrable triangulation of the data sources (Anderson, 1998; Cohen et al., 2000; Mathison, 1988).

Triangulation

Good research practice obligates the researcher to triangulate, that is to use multiple methods, data sources and researchers to enhance the validity of research findings (Mathison, 1988, p. 13)

To have various methods of validating the data and interpretations of the presented data is particularly important where only a single researcher is involved with the collection of the data. The use of triangulated data reduces the likelihood of bias permeating the conclusions, which may happen regardless of the best intentions of the researcher. In this case data will come from a number of sources each of a different nature and each will have an element of interpretation by the researcher involved.

The data sources are

- Evaluation of past examinations of VCE chemistry.
- Chemistry tests of current students.

- Interviews of current students of chemistry.

Of these three sources of data, the source with the greatest potential for unreliability is the interview process. Many factors, often out of the students' or researcher's control, may influence the responses on a particular day. To improve the reliability of these results, particularly in view of the relatively small sample size involved, a number of procedures were employed to assist in gaining a satisfactory and dependable outcome. For example, the adequacy of the results from the initial interview process was referred to the original students interviewed as check for consistency, reliability and dependency. The data were collected over an extended period of time to allow member checking to help ensure the accuracy of the data with respect to the views expressed in the interview process. In this way responses could be corrected or clarified so that they more truly represent the opinions of the interviewees. Finally peer debriefing also offered an opportunity for another researcher to offer an independent assessment of the process and suggest possible refinement or adjustment to the procedures being followed. There are several recognised methods of data triangulation that may be employed to assist in the validation of the results. (Investigator triangulation involves the use of multiple investigators; however, as previously indicated this process was not used in this study).

Data triangulation

Data triangulation includes the use of different data sources for the study. These typically are of three types: time (this may involve cross sectional studies or longitudinal studies), space (this involves the situation of the participants, for example, the use of different cultural groups) and person (this may be defined further by the three levels associated with it, aggregate, interactive and collective (Cohen et al., 2000; Denzin & Lincoln, 2005). In this study a cross sectional longitudinal study was implemented. All students attempted the same test and were evaluated over time with repeated testing (albeit a relatively short time). The test subjects were drawn from the three systems of senior education present in Victoria, that is, Government, Independent and Catholic.

Methodological triangulation

Methodological triangulation has two distinct variants, within-method triangulation and between-method triangulation. Within method triangulation involves in its simplest sense the use of method repetition as way of enhancing reliability. This method is strongly employed in the pure science community. Between methods triangulation relies upon different independent sources of data that all support the pursuit of the given objective. In this study between-method triangulation was employed as the primary source of triangulating the data sets (Cohen et al., 2000; Denzin & Lincoln, 2005; Mathison, 1988). Triangulation can enhance the outcomes of research by demonstrating that several different mutually supportive sources of data are providing evidence for a particular conclusion and this gives much greater applicability and relevance to the findings (Cohen et al., 2000). In spite of the apparent good sense of triangulation it is not without its critics. Patton (1990) and Fielding and Fielding (1986) have suggested that having multiple data sources does not necessarily increase the likelihood of reliability or replication of the results and may even compromise the reaching of a conclusion as the researcher attempts to artificially find a conclusion that is supported by the different sources of data, thus compromising their objectivity (Cohen et al., 2000; Denzin, 1997; Fielding & Fielding, 1986; Mathison, 1988; Patton, 1990). In spite of these critics it is difficult to dispute the inherent value of triangulation.

The particular aspect of triangulation that is relevant to this study is the mixed methodology approach of using both qualitative and quantitative paradigms that increase the scope for the data meeting triangulation requirements. In spite of the apparent advantages of the mixed methodology approach the case for it in the literature (as mentioned earlier) has, at times, been limited. Several reasons are offered for the preference of purely qualitative or quantitative research. By involving both qualitative and quantitative approaches it is believed that the length of the investigation will be extended beyond what a purely qualitative or quantitative enquiry may require and that by following two methods of research the clarity of purpose of the research may be compromised. The costs associated with mixed methods enquiries may also be unnecessarily prohibitive (Maor & Fraser, 1996; Patton, 1990; Reichardt & Cook, 1979).

Overall though the choice of method should be dictated by the needs of the research and utilise the most appropriate tools to elicit the information needed to reach a satisfactory conclusion to that research (Patton, 1990).

Interviews

Since the time of Piaget interviewing children to find their views and beliefs has been an integral part of developing understanding of how children learn and understand their learning environment (Anderson, 1998). Of particular concern in this research is to gain an understanding of students' views about the methods used to assess understanding in chemistry classrooms. Specifically the questions are designed to determine

- the purpose of the assessment that students typically experience in the science classroom.
- do students regard multiple-choice and short-answer questions as assessing the same thing (from their point of view)?
- which of the two question types do students prefer?
- which of the two question types do students find easier or friendlier?
- which of the two question types do students think more accurately assess their understanding?

The specifics of the actual interview and the questions asked will be covered in more detail later in this chapter. Interviewing is a complex procedure and because of the individual interaction involved in is possible that this interaction could influence the outcomes of the interview (Bell, 1993; Patton, 1990). Cross checking or validating the interview results is necessary to ensure the validity of the interview results, thereby contributing to the triangulation of the results obtained (Cohen et al., 2000; Mathison, 1988).

As a process to validate the interviewees' responses, it is important in an effort to attain acceptable validity, to cross check the responses of the participants (Patton, 1990). To this end a sample of the interviewees were reinterviewed after a period of time to check on their responses. Ideally, of course, the interviewees should not change their responses (or at least change very little). The cross checking process involved reinterviewing a small sample of the original interviewees (about 10%). In

this process almost no changes between the original views and the interviewees' subsequent views were detected. This was probably due, to some extent, to the fairly structured and limited scope of the questions asked. The consistency of the responses did, however, give a degree of confidence in the remainder of the responses obtained during the whole initial interview process.

Interviews may be conducted in a number of different ways depending on the circumstances required. Largely the process chosen depends significantly on how predetermined the interview will be (Patton, 1990). The conversational interview is informal in nature and the nature of the questions flows from the answers and responses of the both the interviewer and interviewee. There is limited predetermined nature in the interview. This approach is particularly common in ethnographic inquiries where it is part of the overall process of observation and fieldwork. This type of interview has the advantage of being able to match the interview to the particular respondent and the respondent's circumstances. However, because of the informal nature it is theoretically possible that no two interviews will proceed in the same way and therefore may produce a series of apparently unconnected and therefore difficult to summarise results (Cohen et al., 2000; Patton, 1990; Silverman, 1993).

Guided interview or semi-structured interview

The guided interview or semi-structured interview has a more formalised structure. Topics for discussion are prepared in an outline form first, the interviewer then determines the nature and progress of the interview and which order of questioning will take place. There is still a substantial informal nature to the interview which makes the interview both responsive and adaptive but, at the same time, allows a structure to exist making sure that the data collection will have common themes and points between interviews. That the order of the questions and themes can be variable between the interviews may be a disadvantage because the responses from one sequence of questions may influence the responses to subsequent questions meaning that different responses may have been elicited had the questions been asked in a different order. This situation adds some uncertainty to the results obtained and may potentially affect any conclusions drawn (Patton, 1990; Silverman, 1993).

Standardised open-ended interview

In the standardised open-ended interview, the exact nature and wording of the questions is determined in advance. Also the questions are asked in a predetermined order. Little flexibility can be allowed in this process and the interview has a number of advantages, in that the results can be more easily correlated and analysed. The interview will by its structured nature be more predictable in its timing making it more efficient as a process. Essentially the process is similar to being a verbal questionnaire and is useful for large scale interviews enabling more than one interviewer to be used with the confidence that all interviewees will be subjected to a similar experience and therefore similar standards of responses can be obtained (Fraser, 1991; Kvale, 1996; Patton, 1990). This technique, however, has the disadvantage of not allowing the interviewer or interviewee to follow thread that may arise in the formal questions because the structure will not permit further exploration of these ideas. Thus the naturalness of the interview process is constrained (Cohen et al., 2000).

Closed quantitative interview

The most structured type of interview is the closed quantitative interview. This technique, based on the standardised open-ended form except with the nature or the responses controlled, requires interviewees to choose responses to questions from a given list of possible responses. This technique has wide acceptance in the market research field where interviews can be made over the phone by relatively unskilled interviewers. It is essentially a multiple-choice or LIKERT scale questionnaire. This technique has the advantage of having very structured responses making statistical analysis a relatively straightforward process. However, the approach is very restrictive to both the interviewer and, the even more so, the interviewee because the options may not necessarily quite match the true feelings or views of the interviewee. Thus the responses elicited may well have a significant element of compromise attached to them (Cohen et al., 2000).

Of these techniques the open-ended structured interview is the most suitable for this particular research project. The targeted response areas are clearly defined thus suiting a formalised sequence of questions. The particular area of research is confined so the responses of students with regard to the types of questions they experience in chemistry assessment is, in itself, limited. This technique also

minimises the interview time required, an important consideration when viewed in the light of minimising disruption to senior students educational programs and also when only one interviewer is available to conduct the interviews. That being said, however, should particular points of interest be raised that are of relevance to the research topics and which are unlikely to be covered by the remaining questions then the interviewer reserves the right to occasionally vary the interview questions on those occasions. A response from an interviewee may possibly cause an alteration and review of some of the questions being asked for all subsequent interviews. In other words the researcher must make some allowance to be responsive to the interview process, but still maintain a formalised structure to allow consistency and ease of interpretation.

Conducting the interview

With the open-response quantitative interview being restricted in terms of the questions to be asked, it might be argued that converting the interview to a questionnaire may yield the same outcomes and be easier and quicker to administer. This is far from the case in spite of the structured nature of the closed interview. Whilst the structure is very formal and rigid for reasons mentioned above it is also worth appreciating that the interview process allows participants to more freely express their viewpoints and allows the interviewer and interviewee to seek clarification on issues as they arise. These possibilities will not occur in the questionnaire situation. The obvious result of this is that a misunderstanding of a question by a number of interviewees may result in data that are misleading or contradictory, thus making the drawing of valid conclusions more problematic.

All of these issues depend significantly on the interviewer's ability to draw suitable data from the interviewees. Much has been written about the issues concerned with a good interview process (Cohen et al., 2000).

Problems arise when

- the attitudes, opinions, and expectations of the interviewer impose on the process.
- the tendency of the interviewer to see the interviewee in their own image.
- the tendency of the interviewer to ask questions and seek answers that support a preconceived notion or idea.

- misinterpretations on the part of both the interviewer and/or the interviewee (Cohen et al., 2000).

It is therefore important that the interviewer does not allow or at least minimises the influence of the factors mentioned above. Most importantly the establishment of trust between the interviewer and interviewee is essential. The interviewer should attempt to remain passive in the interview, not ask leading questions, and not be judgemental of the interviewee's responses (Fontana & Frey, 1994; Kvale, 1996).

The list of features of the good interview has been well explored and includes the following;

- Maintaining informality and establishing trust.
- Being knowledgeable of the subject matter and having a clear purpose or structure in the interview.
- Making allowances for a lack of articulation in the interviewees particularly where children are involved.
- Ensuring that respondents have the opportunity to respond fully and not only as far as the interviewer wishes to hear. That is, do not just listen for the responses that suit the interviewer's point of view.
- Being empathetic to the interviewee and being responsive to the manner of their responses.
- Being in control of the interview ensuring that the process remains on track.
- Ensuring that time is not wasted and being able to interpret quickly the interviewee's responses and seek clarification immediately.
- Finally being self-critical. Checking the reliability, consistency and validity of the responses (Cohen et al., 2000; Kvale, 1996; McCormick & James, 1988; Simons, 1982).

With respect to the last point of reliability and validity, there is some contention as to whether both are equally achievable. The reliability of the interview is increased by greater control of the interview through a less open-ended response format (thus increasing the consistency of the responses) but it does so at the expense of validity. The less the interviewee is able to articulate his or her own particular views (and this

could be the case in the closed response interview) then the validity of the responses must necessarily decrease. If responses do not truly represent the interviewees' real beliefs and ideas, then they cannot be considered truly valid (Cohen et al., 2000). The role of the interviewer is indeed a challenging one, however, being forewarned of the potential problems heightens the opportunities to plan to avoid, or at least reduce the problems.

In this research project the researcher was aware of the potential problems; hence the questions to be asked were intended to be very specific and very precise. Ambiguity is of course to be avoided. The questions are short allowing the interviewees to be clear about the aim of the question. The interviewer allowed the interviewees to seek clarification of questions at every instance and avoided making judgements about the responses given, that is, the interviewer was a passive and encouraging listener.

Are the Results Valid or Reliable?

Ensuring validity and reliability is essential to ensure the transferability and applicability of the results (Anderson, 1998). Using triangulation, research data drawn from several sources and, sources within sources in the case of the interviews (that post interview checking of responses), were able to support each other. The quality of the research can be judged by examining the validity of that research. Validity has several guises and each form seeks to establish its own credibility and purpose in the research. However, two forms, internal and external, are of particular significance.

Internal validity

Questions to be answered include: How well do the results and outcomes match reality? How well does the data collected actually conform to the reality of the problem being examined? To a large extent internal validity reflects the accuracy of the research being undertaken. Will there be confidence in the findings? If the answer is no then the research is failing internal validity tests (Guba & Lincoln, 1989). The cause of internal validity is enhanced where the researcher takes steps to ensure that the data collection process results in data that is plausible and credible, takes account of the type and amount of evidence required to support the findings and that there is clarity in the findings that is clearly linked to the data (Hammersley & Atkinson, 1995). In the case of interviews, Lincoln and Guba (1985) propose that reflecting and critical questioning of the findings will enhance their internal

credibility. For example, the use of peer debriefing or member checking where the responses can be queried or clarified will help remove any bias (intentional or otherwise) or misinterpretation by the interviewer from the findings.

External validity

External validity refers to the extent to which the results or findings of the research can be generalised or applied to the wider population. Generalizations of findings, which can be a difficult, are interpreted as the degree to which they are comparable and transferable. If the measurements and tests applied in the research are repeated do they produce essentially the same results (Cohen et al., 2000)? As with internal validity, external validity has its own threats and impediments. Circumstances related to selection effects (where the selection of participants may mean that the results are applicable to that group only), setting effects (the setting of the research precludes the likelihood of applicability outside that environment), history effects (where the particular time and place of the research mean that it is unlikely to have wider applications), and construct effects (where the nature of the research and measurements being conducted may be applicable to the particular group being studied) are amongst those identified by Lincoln and Guba (1985). Other writers have identified a large range of possible threats to external validity some of these are: failure to describe the independent variables explicitly, lack of representativeness of the chosen sample reaction to experimental conditions and unreliability of the instruments used (Cohen et al., 2000).

The objectivity of the researcher (and possibly the participants) can also be a factor in the determination of both internal and external validity. If either or both have particular biases or points of view then this will influence the outcomes of the research and then, necessarily, its validity and applicability. For results to be useful they need to be confidently applicable to the outside world beyond the confines of the research group. They must be both internally and externally valid, this can only occur if the results are internally valid, an area that the researcher has the most control over, for without internal validity the results of the research can never be externally valid (Pilliner, 1973).

Credibility and validity

The validity of the results of a research project or inquiry might be best judged if the results can be objectively assessed as being credible, dependable, confirmable and

transferable. If these objectives are met then the results can be confidently assessed as being valid (Guba & Lincoln, 1989; Lincoln & Guba, 1985).

The credibility of the research can be enhanced with a number of techniques related to the constructs of the research. Prolonged engagement by the researcher and the participants is significant. The research in this instance was conducted over a period of time with breaks between each part of the research. This mitigated the effects of particular circumstances that may have affected participants at a particular time. This essentially reduces the impact of random errors. Member checking and peer debriefing were designed to enhance the credibility of the results obtained (Guba & Lincoln, 1989).

Peer debriefing

Peer debriefing enhances the credibility, or truth, of a qualitative study, by providing an external check on the inquiry process (Guba & Lincoln, 1989). Peer debriefing is particularly advisable because of the vital importance of the researcher in the research process. The individual researcher is the primary means for data collection and analysis and as such may inadvertently influence the outcomes of the research. Each investigator brings a different combination of subjective knowledge, skills, and values to the research process. An independent reflective, independent viewpoint can be most instructive for the researcher to have as a tool to reflect on his or her own work. In this inquiry the emphasis on achieving internal validity was of paramount importance and by using peer debriefing in the interview process and post interview checks, was by design a method of ensuring the validity of the outcomes from this section of the research (Peshkin, 1993).

Member checking

Member checking is another important tool to assist the researcher in ensuring the credibility of his or her results. Member checking, a particularly important validation tool in the interview process, can increase the credibility and validity of any interview-based qualitative study. The researcher should intend and affect the best possible practices in the interview process to ensure that the requirements of a fair and valid interview are met as discussed earlier. This will reduce the likelihood of any serious concerns being raised in the member check process. Interview participants are given the opportunity to reflect on the responses they have previously given and are allowed to review or change those viewpoints if they feel

that they did not properly reflect their feelings. If the interviewees can affirm the accuracy and completeness of the initial interviews, then the credibility of the study is greatly enhanced. Member checking is not foolproof but it certainly increases the likelihood of credibility in the results obtained (Cohen et al., 2000; Guba & Lincoln, 1989).

Transferability

Transferability reflects how well the data obtained in the research can be applied to a new but related situation (Merriman, 1988). A major impact on the reliability of the results in terms of their transferability is the sample chosen for the research, which must reflect and represent the target group for transferability to be meaningful. There must be some practical limitations because the researcher cannot be in all places and have a sample that absolutely represents the entire population. The research then must make allowances and acknowledge openly the limitations that may arise from the chosen sample in terms of its likely applicability to the wider population. The researcher must therefore describe the circumstances of the inquiry or research and the practical methods chosen. An independent observer or reader is then able to make his or her own assessment as to whether the results obtained will have applicability or transferability to the population. This process can be facilitated by the use of thick description whereby substantial sections of the interview outcomes are reported so that judgements can be made independently by others as to their transferability (Geertz, 1977; Guba & Lincoln, 1989; Merriman, 1988). The use of thick description, a process that was employed in this research, will aid the reader in determining the applicability of the findings. This procedure is particularly important in this research due to the limited size of the research group.

Confirmability

The nature of qualitative research tends to mean that the researcher brings his or her unique perspective to an investigation. This is unavoidable. What needs to be minimised is the influence of this uniqueness on the outcomes of the findings. The results are confirmable (and the influence of the researcher minimised) if the same results are obtained by another researcher using the same methodology (Guba & Lincoln, 1989). Enhancing confirmability can be achieved by a number of methods. Amongst these is the full documentation of the data collection procedures used throughout the study. The researcher can also look to finding contradictory results in

the findings and conduct a full data audit trail to enhance the confirmability of the findings.

Dependability

The quantitative view of reliability is based on the assumption of repeatability. Essentially it is concerned with whether or not the same results are obtained when observing the same event on another occasion. In a naturalistic inquiry it is not possible to measure the same event twice. This would be akin to seeing if one school was better than another by sending one student to both schools to do the same work. The influence of one school would naturally affect the performance at the other, making comparison impossible. The three types of reliability usually followed in quantitative research are; the degree to which a measurement exhibits repeatability; the stability and repeatability of the measurement over time; and the similarity of measurements within a given time period. The issue of reliability in qualitative research receives less attention from such researchers, as the focus tends to be on the validity of the work (Kirk & Miller 1986).

The idea of dependability, on the other hand, emphasizes the need for the researcher to account for the ever-changing context within which research occurs. One method of achieving this end is to maintain a thorough inquiry audit that will allow the data to be tracked and be trackable. This process provides transparency to the research process (Lincoln & Guba, 1985). Whilst it cannot be seen in the same light as repeatability of quantitative research it does provide a degree of clarity to the process followed.

Trustworthiness

There is some belief that traditional measures of reliability are not applicable at all in qualitative research because of the nature of the methods and because the epistemological assumptions of the research are unique to a particular study. However, regardless of any assumptions behind the research or the methods of validation employed, the issue of trustworthiness cannot be avoided whatever the epistemological approach of the research (Gibbs, 2002).

Setting the criteria for the evaluation of the research and its ensuing trustworthiness must be a priority in any research if the work is to have any validity in a wider audience.

Previously, the trustworthiness of the research was inextricably linked to the validity in its various guises. Validity is dependent on the factual accuracy of the account of the research, the degree that the participants' viewpoints, thoughts, intentions, and experiences are accurately reported and independently reported by the researcher so that personal bias is not evident and the way in which the explanation fits the data. Johnson (1997) distinguishes between qualitative and quantitative research and presumes that the validity of quantitative research is the starting point and that it should aim to avoid any subjectivity in the results, but qualitative research by its very nature will be subjective in nature and therefore will need more careful strategies to ensure validity and trustworthiness (Johnson, 1997). Johnson suggests some 13 strategies to promote validity and trustworthiness in qualitative research. Virtually all of the strategies mentioned by Johnson (1997) were employed in this research program. The only strategy that was not used was that of extended fieldwork. This was a consequence of the research being conducted over a relatively short period of time (about 10 weeks) so in that sense the term extended was not appropriate.

Summary of Practical Methodology

Data sources-student groups

The methods of research were both quantitative (statistical analysis of the VCE examinations from 2003 to 2007, the sample student tests) and qualitative (interpretation of the interview responses). The data were gathered at four local secondary colleges covering the spectrum of secondary schools in Victoria. The college at which the researcher had recently taught was a Catholic girls' secondary college in an affluent middle class suburb. The applicability or transferability (Anderson, 1998; Cohen et al., 2000) of the results was therefore limited in the extent to which any findings could be applied accurately to a coeducational or males school environment. To ameliorate this possibility the study included participation from a nearby boys' grammar school, a government girls' secondary college and a coeducational secondary college. As a result, 194 Year 11 students were able to participate in the trial testing process. The 194 students represented almost the entire cohort of students studying Year 11 chemistry at the participating schools. A smaller number (59) of these same students also agreed to participate in the interview process even though they had taken part in the testing process.

These colleges are all in an affluent area of Melbourne and draw students from affluent middle and upper class suburbs. Consequently these schools' students do not entirely reflect the range of abilities in the state secondary education system. The students attending the schools are generally motivated students who traditionally have performed well in the VCE examinations. Drawing the sample from these schools should result in students attempting more fully the questions and tasks asked of them. This was important in terms of the research being undertaken. To effectively compare how well the two question types tested student understanding it was important to be confident that motivated students participated in the study. The selection of these students increased the likelihood of this occurring and also gave a higher likelihood that these students would match the performance characteristics of Year 12 students. Most of the participating students were highly likely to take Year 12 chemistry as part of their final year's study. This will add to the reliability and transferability of the outcomes obtained. The main aim is to gauge student performance and student preferences on a largely individual basis. The overall results and findings will most directly reflect on the students involved in the study. The applicability and transferability (Cohen et al., 2000) of these results will be limited by the size of the sample; however, they will provide information that can be a rich basis for further study.

Mixed methods

The research will be both qualitative and quantitative in nature, depending on the particular focus of the study being researched. The qualitative research, involving the interviews will be substantially interpretative in nature. Some basic quantitative analysis (mean) took place with the interview responses however this was to be only indicative of any trends due to the small size of the interview group. Extensive quantitative analysis under these circumstances would have been inappropriate. The quantitative methods of data analysis were most suited to this topic particularly with respect to the analysis of past papers (which included statistical analysis), and the sample tests where the validity of the questions was able to be tested as well as analysis of student performance using Rasch analysis (RUMM laboratory P/L, 2009).

Interviews

Student interviews

All willing student participants (59) were interviewed using an open ended standardised interview (Fraser, 1991; Kvale, 1996; Patton, 1990). Interviewees were each asked the same series of questions (refer to Appendix C for interview question details), which were recorded by the interviewer. The interviewees were asked to give their views about the following aspects of chemistry assessment

- Types of assessment they typically have experienced
- Preferences regarding that assessment
- Specific questions about multiple-choice and short-answer questions
- Preferences about these types of questions
- Perceived advantages and disadvantages about these question types

The participants, by responding to the questions, had the opportunity to reflect on their own practices and evaluate their own approaches to the learning environment that they were experiencing on a day-to-day basis. The interviewees were given the opportunity to reflect and clarify their answers and the interviewer sought clarification when this was warranted. To help facilitate the validity of the process, a peer review took place at a later date.

A few weeks after the initial interviews were conducted, a random sample of the original participants were re-interviewed and the interviewees given the opportunity to reflect upon and change/clarify their initial responses. If this process had revealed substantial differences between the initial and review interviews then further clarification was sought from the interviewees. Fortunately this was not a significant issue. This re-interview process enhanced the validity of the initial interview process in that it was possible to show that the responses initially made by the interviewees were of a considered nature, therefore enabling both credibility and trustworthiness in the interview process.

The interview process facilitated responses and evidence for the interpretation of all the interview questions (see Table 3.1)

In particular the student interviewees were crucial in providing responses that were used to answer research questions 4 and 5.

4. Do students have a preference for the type of question style in terms of
 - a. Multiple-choice or short-answer in general terms?
 - b. With respect to whether the question is assessing recall or application?
5. Do teachers consider multiple-choice or short-answer questions to be the most effective in demonstrating their students understanding of chemistry?

Teacher interviews

To gauge the teachers' view and opinions about the two types of questions (multiple-choice and short-answer), a similar set of interview questions were asked of the teachers whose classes were participating in the testing and interview process (Appendix C). This was important in that it enabled clarification of any patterns that emerged from one class to the next. The teacher's attitudes and opinions provided a useful reflection when considering the student responses. The students have had little experience of assessment in chemistry other than that which they have experienced in the previous years of schooling where chemistry was only one part of their science program. Comparing the two sets of responses allowed some reflection as to whether the student responses matched their experience as set by the teachers or whether they were at odds with the teachers' opinions.

However, the students' experiences of testing in previous years are likely to have influenced the responses. Their life experience in education will have formed a large part of their understandings and beliefs in education and this would have exhibited itself in the interviews. The influence of their current chemistry teacher should not therefore be a major consideration or influence. However, the comparison of the teachers' views and that of the students' was likely to prove enlightening if any patterns emerged within class responses to the interviews. As this part of the study was essentially qualitative and not quantitative the legitimacy and validity of the data was supported by the use of triangulation in the data collection. This is a valid method of cross supporting the data collection by using several methods of data collection that focus on different aspects of the subjects' responses (Anderson, 1998; Cohen et al., 2000; Mathison, 1988). In this case the methods of interview, member checking, peer review along within literature review facilitated the triangulation.

Essentially the research involving testing took two parts.

Analysis of past papers

An analysis of the past five years VCE Chemistry papers (VCAA, 2003a, 2003b, 2004a, 2004b, 2005c, 2005d, 2006a, 2006b, 2007a, 2007b) was performed using information provided by the VCAA. They provide a statistical breakdown of each examination, question by question; showing the average number of marks awarded for each short-answer question and the percentage correct responses for the multiple-choice questions.

Determination of question categories: Recall or Application.

The determination of the category of each question was fundamental to this research. As demonstrated below some questions are necessarily ambiguous in that they are neither simply recall and simply application.

The researcher initially performed the determination of the categories. Many years of experience as a Chemistry teacher assisted greatly in assigning questions or question parts to one category or the other. After this classification the researcher sort the opinions of two experienced Chemistry teachers to check the classifications assigned.

A period of consultation then followed where issues of differing classifications were resolved to determine the final classification. The researcher met with the two consulting teachers and the classifications were discussed until general agreement on each question was reached. This was easier with some questions than others.

To assist with the classification the following qualifications were used to assign questions.

1. Questions where the answers could not be learned beforehand and/or required calculations were assigned as application questions.
2. Questions that could be learned as a simple fact or as an equation or structure were assigned as recall questions.
3. Questions that involved some degree of application and some degree of recall (as shown in the example Q16 from Unit 3 2005 Examination-Figure 3.1) were allocated on the basis of the teacher's assessment as to which skill would play the larger part in the student's ability to answer the question. This included questions where application of a principle may have been involved

but the student could also have learned the process and been able to recall the answer.

Some questions were difficult to assign and as a result the decision to place a particular question in to either the recall or application group became a balance of opinion between the researcher and his colleagues. Undoubtedly other teachers may have differed with some of the classifications; however, with such a large data set of questions small errors were unlikely to have any substantial bearing on the overall findings of this research, as with only two categories the probability of the errors nullifying each other is high.

Clearly there is a matter of professional judgement involved in making such assessments but as an experienced chemistry teacher this researcher feels justified in making such assessments and where doubt existed the opinion of trusted peers was employed. An example of such a question is Question 16 from part a of the 2005 Unit 3 (Semester 1) Chemistry examination (VCAA, 2005c).

Question 16

A representation of a section of a polymer chain that has been produced from two different monomers is given below.



The two monomers are

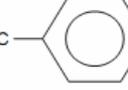
- A.  and $\text{HOOCCH}_2\text{CH}_2\text{COOH}$
- B.  and $\text{HOCH}_2\text{CH}_2\text{COOH}$
- C.  and $\text{HOCH}_2\text{CH}_2\text{OH}$
- D.  and HOCH_2OH

Figure 3.1 Question 16 from the 2005 Unit 3 Chemistry examination

This question was likely to have been intended as an application of the condensation reaction that occurs between the carboxyl and alcohol functional groups. It can be

solved this way by using an understanding of ester formation and applying that knowledge to this particular example. However, this question is identical to the example of polyester formation that appears on the pages of the most commonly used text in Victoria; Chemistry two, Figure 15.13, page 156 (Commons et al., 1999).

Many students would have been instructed to remember this example, as it is one of the simplest cases of polyester formation. Consequently a large number of students would consider this question a recall rather than application question. The question had a facility of 68% making it one of the easier questions on the paper suggesting that many students did simply recall the correct answer. The figure of 68% is almost identical to the average facility for recall questions on this particular examination compared with the average facility of 51% for questions that were identified as being application (Appendix D).

The data itself, once the questions had been classified, was analysed using simple statistical comparisons to determine whether or not there were significant differences in the performance of the students when comparing the type of question (multiple-choice or short-answer), the scope of the question (descriptive/recall or analytical) and the gender of the student (male or female). Analysis took form of mean, standard deviation, ANOVA and Chi-squared as appropriate.

The analysis provided a long-term assessment of examination performance by Year 12 chemistry students over the five years being considered. The results were largely quantitative; however a considerable impact was made by some initial qualitative analysis of the papers. This is all that can be reasonably gleaned from the publicly released data, as individual question data is not released by the VCAA. That being said some 75,000 students' data contributed to the analysis so minor misunderstandings and individual poor questions or responses should have little effect on the overall analysis of the examinations. The analysis provided statistical summaries and measures of a range of aspects of the data.

- A breakdown of the performance of students on the multiple-choice questions compared to the short-answer questions.
- A breakdown of student performance of recall questions in the multiple-choice and short-answer sections of the examinations.

- A breakdown of student performance of analysis questions in the multiple-choice and short-answer sections of the examinations.
- A comparative analysis of student performance on the Unit 3 and Unit 4 examinations. This analysis was significant because of the curriculum difference between Unit 3 and Unit 4. Unit 3 comprises of a much more substantial analysis and application component in the examinations whereas Unit 4 is substantially descriptive chemistry requiring more recall of material.
- A comparative analysis of student performance by gender on the Unit3 and Unit 4 examinations. This again was important due to the perceived differences in performance of males as compared to females in questions involving recall as compared to application questions (Beller & Gafni, 1991; Boli et al., 1985; Fennema, 1979; Hamilton, 1998; Hedges & Howell, 1995).

Overall the analysis provided a descriptive breakdown on student examination performance that focus on question type, content type and student gender. This was used to provide answers and analysis with respect of the research questions 1, 2 and 3.

1. Do students perform more effectively on multiple-choice or short answer questions?
2. Do students perform more effectively on recall type questions or on application questions?
3. Does students' gender influence performance in chemistry examinations (or tests)?

Sample testing

The researcher constructed short tests that asked essentially the same question 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 presented as a multiple-choice question and the other as a short answer question. The multiple-choice question was usually written first and then the short answer variation of the multiple-choice question. The teachers that were involved in the classification of the examination question questions were also involved 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 in the analysis of the trial tests. The testing is a crucial part of this study as it seeks to examine an area of research that has not been extensively studied. Whilst some researchers (Anderson, 1972; Lukhele et al., 1994; Marso & Pigge, 1991; Simkin & Kuechler, 2005) have made some assessments and conclusions about the advantages and disadvantages of each type of question there has not been any study directed at examining the effectiveness of each type of question in 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). This testing may provide some further insights into this matter. To test the understanding of students the following testing structure was adopted.

- Each student participated in a series of short (approximately 10 - 15 minutes) tests.
- On each test day the class was divided into two groups.
- Each group did essentially identical tests (in terms of the curriculum content) excepting that one group's tests 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.
- Two sets of such tests were prepared for each student. The first set of tests examined either descriptive/recall content or application/calculation content depending on what was determined to be most appropriate by the class teacher. The second pair of tests examined the content area not examined in the first tests.

- In this way the effects of the questions being asked as either multiple-choice or short-answer and the content domains of descriptive/recall chemistry and application chemistry could both be studied.
- After all students had completed the tests they were marked and the tests for each student were combined as if it were one larger test to allow psychometric analysis of the results using Rasch modelling.
- The Rasch analysis allowed some determination of the effectiveness of each question and their discrimination of student ability. Any differences that were related to gender may also have become apparent with some questions. The assistance of two psychometricians from the researcher's employer was used to provide assistance and advice during the Rasch analysis.
- ANOVA diagnosis allowed an examination of the trends within the data with respect to significant differences in performance attributable to various factors being considered, namely the type of question (multiple-choice or short-answer), the scope of the question (descriptive or analytical) and the gender of the student (male or female). ANOVA was preferred to t-tests as it allows comparison of multiple groups to see if they differ on one or more variables (Levin, 1991).
- At the conclusion of the tests the interview process took place.

The tests were administered randomly to 192 Year 11 students from secondary schools (mentioned above). The intent was to see if the format of the question affected the level of student response. The problem of consequential errors needed to be carefully allowed for in the multiple-choice questions. This process (testing) was repeated several times to reduce random errors in the choice of students. The four sample tests are shown in Appendix A.

Anonymity of participants

To ensure that the tests were anonymous from the students' point of view, each student was assigned a code that was known only to the students' teachers. The researcher did not have any knowledge of any particular student. The code provided the researcher only with information about which school the students attended and their gender. The correlation list was known only to the students' teachers and was

destroyed at the conclusion of the testing. This analysis was essential in providing real data in responding to research questions 1, 2 and 3.

1. Do students perform more effectively on multiple-choice or short answer questions?
2. Do students perform more effectively on recall type questions or on application questions?
3. Does students' gender influence performance in chemistry examinations (or tests)?

Data analysis

Data were analysed using a combination of methods, mostly quantitative combined with some qualitative and subjective analysis (Cohen et al., 2000; Ryan & Bernard, 2000). Careful inductive analysis was performed to ascertain, with the aid of simple statistics means, standard deviation, Chi-squared and ANOVA, the performance characteristics of the different types of questions and the different categories. Deeper understanding of the results of the tests was enhanced through the use of Rasch modelling and analysis of the trial tests. This form of analysis is significant with the tests in that it enhanced the credibility and validity of the results that were produced from a relatively small sample that was tested. This type of deeper analysis was not possible with the past paper analysis because of the simplified group data provided by the VCAA.

However, the very large sample size of the past paper analysis did provide a degree of confidence in the validity of the trends determined from the analysis of the results.

Analysis of the examination through simple statistical measures gave an indication of the trends in question answering techniques and the scores and measures obtained can be reasonably seen as a reliable way of examining differences between scores on the different types of questions (multiple-choice questions and short-answer questions). Differences in gender performance were also examined in the analysis process. Furthermore, the data analysis provided by the VCAA with respect to the results of the scores achieved on the VCE Chemistry papers are relatively simple and therefore attempting to perform higher statistical analysis would probably not be justified.

Data collection

Testing was the primary source of data for this project. Tests were repeated and each test was designed to have identical structure and degree of difficulty when assessing in the different formats. Each test consisted of six questions. Two tests were on acid-base chemistry and two tests on stoichiometry. Each pair of tests had matching questions. For example Question one on Stoichiometry Tests A and B (Appendix A) asked students to calculate the percentage composition of a simple compound. The difference between the actual questions was minimal, only the chemical compound and required element were different. The main difference was that one required a multiple-choice response and the other a short-answer response. The multiple-choice question test and short-answer questions test had the same questions but obviously structured in the different formats. Where questions had a consequential mark structure, typical in stoichiometric questions of the short-answer format, then the equivalent multiple-choice questions were structured so that a consequential marking scheme was possible. After administration of the tests they were marked and copies returned to the students, teachers for their own diagnostic purposes.

Interviews

The interview was a significant area of data collection in this research and it was also an area of concern in terms of its validity and reliability in the research situation (Cohen et al., 2000). The most important method that can be used to improve the validity of the research through interview is to eliminate as much as possible any bias from the interview. By having a highly structured interview where the questions were asked in a calm and controlled environment was a way of enhancing the likelihood of reliable and valid results. Careful construction of questions was used to avoid problems such as leading questions appearing in the interview. Students were interviewed soon after the testing but before the return of the results. This was done to avoid students' opinions being unduly influenced by test performance. After the transcription of the data the responses were coded to allow overall trends in the student responses to be ascertained. Students who generally favoured multiple-choice questions were coded 1 and students who generally favoured short-answer questions were coded 2.

Ethical Issues

With regard to the ethics of this issue I firmly believe that the overriding parameter is that of do no harm. That this has not always been the case has been a source of often bitter debate (The Coalition of Americans for Research Ethics, 1999). In approaching any form of research the researcher must have a full understanding of the ethical issues involved. The ethical considerations are broadly divided into two areas of concern involving the interviewees of the research and the researcher themselves.

The researcher must consider with respect to the interviewees that no physical, psychological or environmental harm shall befall the subjects. At the same time the researcher must ensure that their research methodology, writing practice and motives are beyond reproach.

In conducting research interviews in the second part of the research, the participants needed to be fully informed as to the purpose of the research and how the information gained from the questionnaire or interview would be used. The interviewee's time, right to withdraw and anonymity of the research would all need to be respected. The interviewees would also need to be assured that their information would be adequately disposed of when no longer required for further research. This being said, the anonymous nature of a sample does not completely subsume the participants right to privacy (Russell, 1998).

In conducting the interview/questionnaire the researcher needs also to be cognisant of any cultural or societal implications of the question being asked. A simple example may be that the choice of school by some religious or ethnic groups may be significantly skewed by the influences of religious belief or the gender of the school/child. The personal views of the researcher must not be allowed to impact on the process and the rights and views of the research subjects must be respected for what they are (Anderson, 1998; Halasa, 1998). As mentioned earlier the very real and determined attempt to keep the results completely anonymous will also enhance the ethical nature of the research being performed.

In this type of research, the informed consent of the participants is the most important issue, as it will (hopefully) allay any fears that the participant may possibly have formed. At the end of the research the participants should be supplied with the results of the research to ensure that they are satisfied with the way in which the data

was used with respect to any privacy issues that may be apprehensive about. Ethical considerations that relate to the authorship of the report must be considered. The specific issue of plagiarism is one which has been of significance in recent years and it is the responsibility of the researcher to ensure that all referenced work is correctly cited and acknowledged in the final written report (Russell, 1998).

Summary of Methodology

The research methodology defines the validity and reliability of any research. This study examines the trends and patterns in chemistry assessment in Victoria. The study involves a quantitative analysis of past papers from the Victorian VCE Chemistry examinations as well as a quantitative analysis of a set of trial tests involving Year 11 students. Further data was obtained from interviewing a number of the Year 11 students. This involved an essentially qualitative analysis. The merits of using a mixed methods approach to the study were discussed in this chapter with extensive discussion of the merits of the two approaches as well as the use of triangulation as a key tool in the validation of the data sources.

The interviews of the year 11 students were an important part of this study. Discussion and consideration of the different approaches to interview techniques concluded with the reasons for selecting the method of standardised open-ended interview used in this study. Further discussion then takes place into the methods of ensuring the validity, credibility, reliability and trustworthiness of the findings.

A summary of the details of the practical methods used in this study then follow. The study involved three separate data collection stages mentioned above.

Stepwise procedures are outlined showing the practical steps followed in this study to obtain the data for, the past-paper analysis, trial-test analysis and the student interviews.

The analysis steps that were performed on each of the data were also outlined. The chapter concludes with a discussion of the ethical considerations of this research.

Chapter 4: Results-Analysis of Past papers

The research supporting this thesis is based on three main sources of information. This chapter deals with the first of these, which is the analysis of the Year 12 Chemistry examinations from 2003 to 2007. The analysis measures the papers for any significant differences with respect to the content of the questions that is recall or application, the style of the questions short-answer or multiple-choice and finally whether any patterns exist in the grade allocations awarded to male and female students.

Analysis of Past Papers

The Victorian Curriculum and Assessment Authority (VCAA) sets and administers the Victorian Certificate of Education (VCE) examinations in all Year 12 (Unit 3 and 4) subjects.

Each examination is set by an independent panel and is based on the study design for each particular subject. In Chemistry the examination is conducted in two parts. Unit 3 is examined in June and assesses only Unit 3 work. The Examination is 90 minutes long and usually consists of 20 multiple-choice questions (Part A) and approximately 8 short-answer questions (Part B). Each multiple-choice question is worth 1 mark and the whole of part A is worth between 25-30% of the total examination. Part B questions usually total about 55-60 marks and are worth the remaining 70-75% of the marks of the paper. The questions range in difficulty from simple recall to more difficult analysis and application questions. The Unit 4 examination is held in November and examines only Unit 4 work; however, it can encompass concepts from Unit 3 if they are appropriate to the material being examined in Unit 4. For example any question on Faraday's laws may well involve some stoichiometric calculations, which is primarily examined in Unit 3 but would be considered appropriate in the context of a Faraday's laws question in Unit 4.

The current structure of the examinations has been in place for over 15 years and is generally well accepted by the teachers in the state.

In this section each of the Unit 3&4 Examinations from 2002 to 2007 have been analysed for the content in terms of whether the questions are essentially recall or

essentially analysis/application. This sorting of the questions has been done for both Part A (multiple-choice questions) and part B (short-answer questions). The VCAA provides an assessment report for each examination that includes correct answers and an analysis of the marks awarded on a state-wide basis. The breakdown was used to analyse performance on a number of criteria, multiple-choice and short-answer questions, recall and application questions and gender (Appendix D).

For example, for each multiple-choice question the correct answer and the percentage of students getting that response are given. The percentage choosing each of the distracters is also given but this was not used in the analysis of the questions. Thus the marks awarded for each question can be simply given as a percentage, which were then totalled, and an average mark and standard deviation for the multiple-choice questions calculated for each examination. For part B questions the analysis provided is somewhat different. For each question the percentage of students getting each of the possible marks is provided; if a question is worth two marks the percentage of students getting 0, 1 or 2 marks are provided. In the VCE examinations half marks are not awarded.

As an example of the analysis completed in this study a thorough description of the 2005 examination papers now follows.

The 2005 VCE Chemistry Papers Examination 1 and 2

The 2005 VCE Chemistry Examination 1 (is held in June and covers Unit 3 work).

Each question in part A was assigned as either a recall question or an analysis/application question as described in Chapter 3. Most of the questions in each examination, either section, tended to be application types but a significant number are recall or recall with some interpretation involved. An example of a relatively simple recall multiple-choice question is shown in Figure 4.1; correct answer A. The Assessment report for the 2005 VCE Chemistry Examination 1 provides the following information about this question (Figure 4.1); the mark awarded for this question as a state average would be 86% or as a mark out of 1: 0.86.

Part A Question 2

A mixture extracted from honey contains two different sugars. The most appropriate way of separating these sugars would be with the use of

- A. high-performance liquid chromatography
- B. atomic absorption spectroscopy
- C. UV-visible spectrophotometry
- D. flame tests.

Question	%A	%B	%C	%D	%No Answer	Comments
2.	86	7	5	2	0	Of the techniques listed, only chromatography (option A) can be used to separate components of the mixture. Since sugars are heat sensitive and likely to decompose in a GC, HPLC is the most appropriate technique.

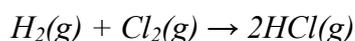
Figure 4.1: Question and answers for Q 2 from Unit 3 2005 Chemistry examination

Note: The table indicates the percentage of students who chose each option. The correct answer is shaded

A typical simple application question provided by the VCAA is shown in Figure 4.2.

Part A Question 10

Hydrogen and chlorine react according to the equation



3 mole of H_2 and 2 mole of Cl_2 are placed in a vessel and sealed.

When reaction is complete the vessel will contain

- A. 5 mole of HCl
- B. 6 mole of HCl and 1 mole of Cl_2
- C. 4 mole of HCl and 1 mole of Cl_2
- D. 4 mole of HCl and 1 mole of H_2

Question	%A	%B	%C	%D	%No Answer	Comments
10.	12	6	11	70	0	$H_2(g) + Cl_2(g) \rightarrow 2HCl(g)$ Initially (mol) 3 2 Reacting (mol) 2 2 \rightarrow 4 Finally (mol) 1 - 4 When complete, 1 mol of H_2 and 4 mol of HCl are present.

(VCAA, 2005a, 2005c)

Figure 4.2: Question and answers for Q 10 from Unit 3 2005 Chemistry examination

Note: The correct answer is D and on this question some 70% of students were able to select the correct answer.

In summary the results for part A of the 2005 Examination 1 paper are shown in Table 4.1

Table 4.1: For the Unit 3 2005 part A summary results

Part A

Recall Questions		Application Questions	
Question Number	% Correct	Question Number	% Correct
1	84	4	58
2	86	5	52
3	86	6	34
7	42	8	47
9	55	10	70
16	69	11	59
18	79	12	39
19	55	13	36
20	61	14	61
		15	33
		17	69
Mean score	68.6	Mean score	50.7
Standard deviation	16.1	Standard deviation	13.8

(VCAA, 2005a)

From this one paper students scored higher on the recall based questions with a mean score of 68.6% as opposed to the application questions with a mean score of 50.7%.

This outcome is not entirely unexpected due to the generally simpler nature of the recall question as opposed to the application question (Niaz & Robinson, 1995). What is also important is a comparison between the performances on the two different halves of the paper.

A similar analysis was performed on part B of the 2005 examination.

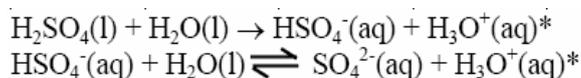
Here the analysis provided is somewhat different, but easily reconciled with that provided for part A.

Part B Question 5 c.

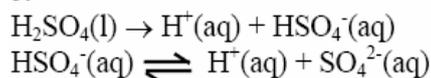
Sulfuric acid is a diprotic acid. The first ionisation is that of a weak acid. Give the chemical equations for both the first and second ionisation reactions of sulphuric acid. (2 marks)

Marks	0	1	2	Average
%	33	28	39	1.1

Solution:



or



Comment: The range of mistakes in this part of the question was extensive, particularly the use of incorrect chemical formulae, including incorrect charges. Equations involving $\text{H}_2\text{S}_2\text{O}_7$ were not uncommon, which suggests that some students simply defaulted to learned material. Students needed to include \rightleftharpoons in the second equation to get both marks

(VCAA, 2005a, 2005c)

Figure 4.3: Question and answers for Q 5C from Unit 3 2005 chemistry examination

This is a relatively uncomplicated recall question with 67% of students gaining at least one mark for the question. The course specifically states that students must know the acidic properties of sulfuric acid and the relevant equations. Students would be reasonably expected to be able to write these equations.

The solution provided in Figure 4.3 shows the acceptable equations and acceptable variations and provides a mark results table: 33% of students received zero marks, 28% received 1 out of 2 and 39% received full marks. The average mark was 1.1 out of 2. This mark is used for comparing each question.

An application question from this paper was **Question 7b** is shown in Figure 4.4.

Part B Question 7

Ethanoic acid (CH_3COOH) is a weak acid in water

7b. A 0.100 M solution of ethanoic acid in water at 25°C has a pH of 2.88.

- i. Calculate the hydrogen ion concentration in a 0.100 M solution of ethanoic acid.
- ii. Calculate the acidity constant of ethanoic acid at 25°C .

1 + 2 = 3 marks

Marks	0	1	2	3	Average
%	38	11	13	38	1.6

Solution

7bi.
 $[\text{H}^+] = 10^{-2.88} = 1.32 \times 10^{-3} \text{ M}^*$

7bii.

$$\begin{aligned} K_a &= \frac{[\text{CH}_3\text{COO}^-][\text{H}^+]}{[\text{CH}_3\text{COOH}]} \\ &= \frac{[\text{H}^+]^2}{[\text{CH}_3\text{COOH}]} \\ &= \frac{(1.32 \times 10^{-3})^2}{0.100} * \quad \text{or} \quad \frac{(1.32 \times 10^{-2})^2}{0.100 - 0.00132} \\ &= 1.74 \times 10^{-5} * \quad \text{or} \quad 1.77 \times 10^{-5} \end{aligned}$$

Comment: A significant number of students used the concentration of the CH_3COOH , rather than the pH, to determine the $[\text{H}^+]$. When asked to calculate the $[\text{H}^+]$, $1.32 \times 10^{-3} \text{ M}$ is a more appropriate answer than $10^{-2.88}$. If students incorrectly calculated $[\text{H}^+]$ in part i. But used this value correctly in part ii., they received both marks for ii.

(VCAA, 2005a, 2005c)

Figure 4.4: Question and mark distribution for Q 7b from Unit 3 2005 chemistry examination

From this analysis 38% of students were not awarded any marks for the question, but 38 % earned the full marks available (3 marks).

Each part of each question is usually given a mark breakdown as exemplified with Question 7b above. The remaining questions for part B of the 2005 Unit 3 examination are shown in Table 4.2.

The overall result here is different to that of part A. There is a reduced performance in the recall section of the paper and an improved performance in the application questions.

Analysis of this one paper appears to support the initial hypotheses that students perform more effectively on multiple-choice questions that are testing recall and perform better on application questions when they are asked in a short-answer or extended format. An initial explanation for this finding could be that the multiple-choice question format does help students in recall questions because the answer is in front of them and may provide a recognition cue assisting students in selecting the answer (Haynie, 1994).

Table 4.2: For the Unit 3 2005 part B summary results

Recall Questions	Average correct	Total marks	%	Application Questions	Average correct	Total marks	%
Question Number				Question Number			
1a	0.6	1	60.0	1e	0.9	2	45.0
1b	0.7	2	35.0	2b	2.6	4	65.0
1c	0.5	1	50.0	3a	3.2	4	80.0
1d	0.5	1	50.0	3b	1.2	2	60.0
1e	0.9	2	45.0	3c	1.3	2	65.0
2a	0.6	1	60.0	4a	0.6	1	60.0
2c	3.3	4	82.5	4b	0.5	1	50.0
3d	1.8	3	60.0	6a	0.9	1	90.0
4c	1.1	2	55.0	6b	1.8	3	60.0
4d	1.4	2	70.0	6c	1.0	2	50.0
5a	0.6	1	60.0	7b	1.6	3	53.3
5b	2.6	4	65.0	8c	0.4	1	40.0
5c	1.1	2	55.0				
5d	0.5	1	50.0				
7a	0.7	1	70.0				
7c	0.8	3	26.7				
8a	0.8	1	80.0				
8b	0.8	1	80.0				
Mean score			58.6	Mean score			59.9
Standard deviation			15.0	Standard deviation			14.2

(VCAA, 2005a)

In a short-answer question such cues do not exist and students may not be able to recall the answer in time to answer it even if they do know the answer. Similarly the short-answer mode is likely to assist students in application questions because they have the opportunity to gain marks for correct working even when they make a small error. In a multiple-choice question this same small error could result in the likelihood of a zero mark.

Unfortunately this one example alone is insufficient to reach any meaningful conclusion and similar analyses were applied to other examinations (2003, 2004, 2006 and 2007 VCE Chemistry examinations) to see if the same findings were repeated. A further complication is that the comparison does not and is unable to allow for the differing degrees of difficulty in the questions. For example, if the recall multiple-choice questions are easier than the application multiple-choice questions, then it is highly likely that the average mark on the multiple-choice questions will be higher in this section. Ideally the most appropriate way to test this hypothesis is to have identical questions framed in both formats. This is an unlikely occurrence in an official examination such as the Victorian VCE. This aspect will be explored in the student trial test phase of this study.

Classifying the questions as recall or application

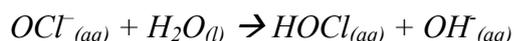
In this section the analysis of the examination papers according to whether the questions are recall or application is presented. The classification process is described in Chapter 3 page 64.

Unfortunately the results released by the VCAA on each examination do not allow full psychometric analysis because they are summary results and as such cannot be analysed with regard to individual student performance as compared to the whole examination. If individual student examination data were available, Rasch analysis could be used which allows each individual question to be examined for student performance on the examination. Rasch analysis was used to determine the effectiveness of individual questions in the student sample test phase of the research (see Chapter 5). Thus the effectiveness of each question in discriminating between students on the basis of the individual performance can be determined. Such analysis would enable each individual question to be assessed on how effectively it assessed the content in relationship to the actual performance of the student on an overall

basis. A question such as Question 15 in part A (2005 unit 3 examination) has a correct response percentage of only 33%. It is not unusual to have questions that have on a 25% (or less) correct percentage.

For example the multiple-choice question 14 from part A from the 2004 Unit 3 paper is shown in Figure 4.5.

NaOCl is completely dissociated in water to form $\text{Na}^+_{(aq)}$ and $\text{OCl}^-_{(aq)}$. In solution, OCl^- hydrolyses according to the equation



100 mL of pure water at constant temperature is added to a 100 mL solution of 0.10 M NaOCl.

Question 14

When the solution reaches equilibrium again, the

- A. $[\text{H}^+]$ has decreased.*
- B. pH of the solution has decreased.*
- C. concentration of HOCl has increased.*
- D. value of the equilibrium constant has halved.*

(VCAA, 2004a)

Figure 4.5 Question 14 from the Unit 3 2004 Chemistry Examination

The correct answer to this question was **B: pH of the solution has decreased**. However, only 24% of students chose this option, some 47% chose *C: concentration of HOCl has increased*. This would almost suggest that a student would have as much chance of getting the question correct by purely guessing. This doesn't mean, however, that Question 14 was a poor question. If the top 24% of students, based on the remainder of the paper, were the ones who were generally correct on this item then the question has served its purpose by discriminating between the students on the basis of ability. (Whether the question did in fact discriminate effectively is unknown, as the data are not released by the VCAA).

The 2005 VCE Chemistry Paper, Unit 4 examination (held in November).

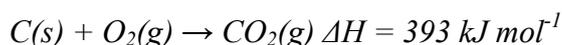
It is important to note that whilst Examination 2 only assesses material studied in Unit 4 (that is the second semester in the relevant year), any material taught in semester 1 (Unit 3) becomes assumed knowledge. This means that any material that

was covered and is directly applicable to second semester can be indirectly examined in the Unit 4 examination. Two simple illustrative examples that demonstrate this indirect assessment are shown below.

i. Stoichiometry. Stoichiometry and its direct applications are directly examined in semester I and form a significant block of work. Refer to Q10 Part A multiple-choice Examination 1 2005 (Figure 4.2) above or Q4 part B short-answer (Figure 4.7) above. Whilst these are reasonable questions within the context of Examination 1 they would not be acceptable in Examination 2. That being said it is quite reasonable to expect that a question that involves some stoichiometric calculations in relation to a question on Heats of reaction would be acceptable in the context of the Unit 4 examination. For example, Q3 from the 2005 Examination 2

Question 3 (2005 Unit 4 examination)

3a. *Coke, which is essentially pure carbon, is widely used as a fuel. Its complete combustion can be represented by the following equation.*



However, under certain conditions, the combustion is incomplete and the following reaction also occurs.



Calculate the energy, in kJ, released when 2.00 tonne (1 tonne = 10^6 gram) of coke is reacted with oxygen if 80% of the coke is oxidised to carbon dioxide and the remaining 20% is oxidised to carbon monoxide.

(4 marks) (VCAA, 2005d)

Figure 4.6: Question 3 from the 2005 Unit 4 Examination

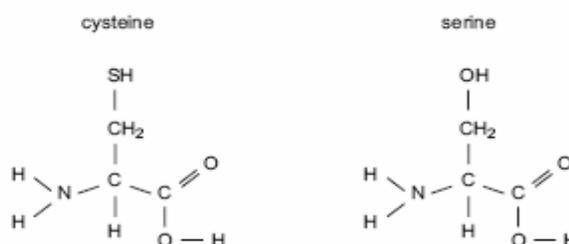
The question in Figure 4.6 is focussed on an understanding of the use of ΔH values to determine the amount of energy released in a particular reaction involving the combustion of coal. To calculate this, the student must be able to use the stoichiometric skills that were specifically examined in Unit 3. Whilst it may be considered re-examining material it must be understood that the use of simple stoichiometry is only a tool in examining the relevant area of heats of reactions.

ii. Organic reaction mechanisms. In Unit 3 the students cover the range of simple addition, substitution and condensation reactions and are examined accordingly

on that knowledge. It is reasonably assumed that in the section in Unit 4 on food chemistry which involves a number of reaction sequences involving amino-acids, saccharides and fats that the types of reactions examined in Unit 3, most importantly the condensation reactions should be understood in answering food chemistry questions. An example of this is Question 4 part B Unit 4 2005 Examination 2.

Question 4

a. Two common α amino acids (2-amino acids) are cysteine and serine. Their structural formulas are given below.



- i. What chemical feature must an amino acid have in order to be classified as an α amino acid?*
-
- iii. Cysteine and serine can combine together to form **two** different dipeptides. Draw the structural formulas of these two dipeptides.*

(1 + 2 = 3 marks)(VCAA, 2005d)

Figure 4.7: Question details for Q 4 from Unit 4 2005 chemistry examination

The reactions shown in Figure 4.7 can be learned but understanding the mechanism of a condensation reaction is a far more useful approach to answering the question. As with the example shown in Figure 4.6, the question is not directly assessing the understanding of the condensation reaction it is using that understanding as a tool to answering a question about peptide formation.

Detailed study of Examination 2 2005

As mentioned in the description of Examination 1 each question in part A was assigned as either a recall question or an analysis/application question. Most of the questions in each examination, either section, tend to be application types but a

significant number are recall or recall with some interpretation involved. The following examples illustrate the analysis of the second semester examination (Tables 4.7 and 4.8).

The correct answer for this question is **A**, and this question was a recall question.

Part A: Question 12

An electrolytic cell is used commercially to extract aluminium from its ore. The anode and cathode of this electrolytic cell are composed of

	<i>anode</i>	<i>cathode</i>
A.	<i>carbon</i>	<i>carbon</i>
B.	<i>carbon</i>	<i>iron</i>
C.	<i>iron</i>	<i>carbon</i>
D.	<i>iron</i>	<i>iron</i>

<i>Question</i>	<i>%A</i>	<i>%B</i>	<i>%C</i>	<i>%D</i>	<i>%No Answer</i>	<i>Comments</i>
12.	54	26	17	2	0	<i>This is a direct reference to the standard electrolytic method of producing aluminium commercially. In this process, the anode is carbon and the cathode is a carbon lining over a steel base. Response B incorrectly indicated a steel cathode</i>

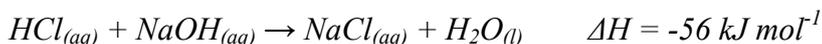
(VCAA, 2005b, 2005d)

Figure 4.8: Question and answers for Q 12 from Unit 4 2005 chemistry examination

The mark awarded for this question, as a state average, was 54% or as a mark out of 1:0.54. This particular question was a direct recall, as the study of the Hall-Heroult cell is a required part of the course. That a significant number of students incorrectly picked B suggest some confusion with the similar Downs cell, which has carbon and iron electrodes.

In Examination 2, typically, there tend to be more recall type questions, which reflects the change in balance between the two examinations in terms of the type of chemistry being studied.

The reaction between solutions of hydrochloric acid and sodium hydroxide can be represented by the following equation.



60.0 mL of 2.0 M HCl, at 21°C, is mixed with 40.0 mL of 2.0 M NaOH, also at 21°C, in a well-insulated calorimeter. The calibration factor for the calorimeter and contents is 420 J K⁻¹.

Question 10. *The final temperature, in °C, of the resultant solution in the calorimeter would be closest to*

- A. 11
- B. 32
- C. 37
- D. 52

Question	%A	%B	%C	%D	%No	Comments
						<i>Answer</i>
10.	12	61	20	6	0	<p><i>There is an excess of 20 mL of the 2.0 M HCl.</i></p> <p><i>Only 40 mL of the HCl is needed to react with the 40 mL of NaOH.</i></p> <p><i>The amount of heat evolved will thus be</i> $40/1000 \times 2 \times 56\ 000 = 4480 \text{ J}.$</p> <p><i>The temperature rise must then be</i> $4480/420 = 10.7^\circ\text{C}.$</p> <p><i>This is added to the initial temperature to give</i> $21 + 10.7 = 31.7^\circ\text{C},$ <i>which is the closest to the correct response, B.</i></p>

(Note: The correct answer is B and on this question some 61% of students were able to select the correct answer (VCAA, 2005b, 2005d)

Figure 4.9: Question and answers for Q 10 from Unit 4 2005 chemistry examination

This is a good response rate for such a question and reflects the relatively straightforward nature of the question.

In summary the results for part A of the 2005 paper examination 2 are shown in Table 4.3. As was demonstrated in the Unit 3 Examination 1 paper it would seem that the students scored more highly on the recall based questions with an average mark of 65.5% as opposed to the application questions, which had an average score of 55.8%.

Table 4.3: Unit 4 2005 part A summary results

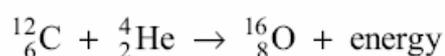
Recall Questions		Application Questions	
Question Number	% Correct	Question Number	% Correct
1	62.0	4	42.0
2	79.0	5	61.0
3	59.0	6	72.0
12	54.0	7	43.0
15	71.0	8	55.0
16	68.0	9	67.0
17	70.0	10	74.0
20	61.0	11	45.0
		13	54.0
		14	50.0
		18	41.0
		19	66.0
Mean mark	65.5	Mean mark	55.8
Standard deviation	8.0	Standard deviation	12.0

(VCAA, 2005b)

A similar analysis was performed on part B of the Unit 4, 2005 examination.

For example, a typical recall question from part B of the Examination 2 Unit 4 2005 is **Question 3b**, and part of the question previously mentioned on page 83, that is **Question 4a part I (Figure 4.7)**.

3b. Carbon is also a reactant in nuclear fusion reactions in some stars. One such reaction can be represented by the following equation.



For a given amount of carbon, significantly more energy is released in nuclear fusion reactions than in chemical reactions.

i. What is the source of the energy released in this nuclear fusion reaction?

*ii. Why is nuclear fusion **not** currently used as an energy source in our society?*

(1 + 1 = 2 marks)

Marks	0	1	2	Average
%	23	37	40	1.2

Comment:

3bi.

mass loss/nuclear binding energy/nuclear energy

3bii.

It is not yet practical.

(Too many students provided irrelevant information such as ‘dangerous nuclear waste’ and ‘possible nuclear disasters’. To obtain the mark for this question, students had to recognise that nuclear fusion is not yet a practical possibility. Students who identified this and then went on to provide additional material were still awarded the mark.)

(VCAA, 2005b, 2005d)

Figure 4.10: Question and mark distribution for Q 3b from Unit 4 2005 chemistry examination

This is a recall question. The course specifically states that students must know the principles of fusion reactions and the particular reaction shown is part of the required knowledge. Students would be expected to understand these equations and the related issues.

The students generally performed well on this question (3b) and as a simple recall this was to be expected. It would be interesting to speculate how they might have performed had this question been structured as a multiple-choice as it might easily have been. Whether students would have recognised the impracticality of fusion power if the option was presented in front of them is of course not possible to know, however, it is more likely that they would have as opposed to having to recall it from memory.

The solution to question 4a part i) (refer to page 84) is shown in Table 4.4

Table 4.4: Mark distribution for Q 4a from Unit 4 2005 chemistry examination

Marks	0	1	Average
%	65	35	0.4

Comment: It must have carboxyl and amino groups attached/bonded to the same carbon atom.

A structure that clearly showed the amino and carboxyl groups attached to the same C atom was also accepted as correct. Only 35 per cent of students received the mark for this simple question, indicating that the concept was not well understood. (VCAA, 2005b)

This question (4a) is interesting in that it demonstrates the issue again that when recall is needed students may perform poorly when they do not have the visual prompt in front of them. It is quite likely that had this been a multiple-choice question the performance may have been better.

Application questions from this paper is shown earlier on pages 84 and 85, they were Questions 3a and 4a part ii (Figures 4.2 and 4.3)

The recommended solutions (Figure 4.11) and marking guides (Table 4.5) are shown below for Question 3a.

Table 4.5: Mark distribution for Q 3a from Unit 4 2005 chemistry examination

Marks	0	1	2	3	4	Average
%	23	10	14	25	27	2.3

$$1.6 \times 10^6 \text{ g C} \rightarrow \text{CO}_2^* = 1.33 \times 10^5 \text{ mol}^* \rightarrow 1.33 \times 10^5 \times 393 = 5.24 \times 10^7 \text{ kJ}^*$$

$$0.40 \times 10^6 \text{ g C} \rightarrow \text{CO} = 0.33 \times 10^5 \text{ mol} \rightarrow 3.33 \times 10^3 \times \frac{232}{2} \rightarrow 0.39 \times 10^7 \text{ kJ}$$

$$\text{Total} = 5.63 \times 10^7 \text{ kJ}^*$$

One mark was given for correctly changing mass to moles; one mark for the 80:20 split; one mark for applying the energy per mol from ΔH ; and one mark for giving the correct answer consistent with the correct use of information from ΔH . Chemistry

This was not a particularly easy question but students performed reasonably well. A common error was to use 1 tonne of coke rather than 2 tonne of coke in the calculations. Another mistake was to forget the factor of 2 by using 232 rather than 232/2 in the CO calculation.

Although no marks were lost, too many students gave an unnecessary number of significant figures (up to 10!) in their final result. Students should be encouraged to use scientific notation when numbers are very large or very small.

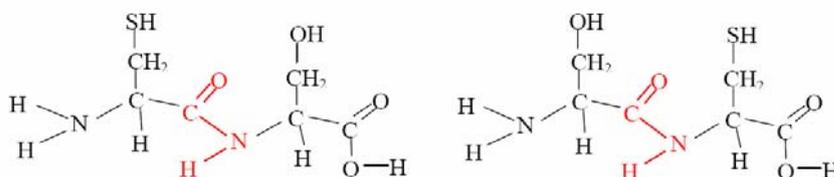
(VCAA, 2005b)

Figure 4.11: Question details for Q 3a from Unit 4 2005 chemistry examination

The marking guides and recommended solutions (Figure 4.12) and are shown below for Question 3a

4a.ii.

Marks	0	1	2	Average
%	42	16	43	1.1



One mark was given for each of the two possible links. In general, structures had to be drawn out as indicated, but O–H as –OH was accepted. If two fully correct semi-structural formulas were shown, one mark only was awarded.

This part was not very well done. It was surprising how many students could not join two amino acids to form a peptide link, rather preferring to produce acid anhydrides or ester-like links. –C–N–H– and –C–O–C– were also commonly provided.

(VCAA, 2005b)

Figure 4.12: Question solution details for Q 4a.ii from the Unit 4 2005 chemistry examination

The performance here was poor in the eyes of the examiners, however, it is interesting in light of the number of students who were unable to do part i mentioned earlier. It would seem that even though they did not know what an α -amino acid was, students were still able to perform the application aspect of joining the two acids together using a condensation pathway. However, it would appear that most of the students who could not identify an α -amino acid were unable to do the application question, assuming a direct correlation between the students who scored zero on question 4a part ii (42%) were also in the group that scored zero in question 4a part i

(65%), meaning that possibly one third of those who were unable to answer part i were able to answer part ii or at least some of it!

These questions also highlight the observation that a large number of students are unable to obtain any score at all in short-answer questions.

The full analysis of the part B questions is shown in Table 4.6.

Table 4.6: For the Unit 4 2005 part B summary results

Recall Questions	Average correct	Total marks	%	Application Questions	Average correct	Total marks	%
Question Number				Question Number			
1	4.1	5	82.0	3a	2.3	4	57.5
2a	0.8	1	80.0	5c	1.3	2	65.0
2ci	0.6	1	60.0	2b	2.2	3	73.3
2cii	1.1	2	55.0	4aii	1.1	2	55.0
2ciii	0.8	1	80.0	7b	1	2	50.0
2civ	0.8	1	80.0	7c	0.4	1	40.0
2cv	0.6	1	60.0	7d	0.5	1	50.0
3b	1.2	2	60.0	8a	0.8	2	40.0
3c	3.4	4	85.0	8b	1.4	4	35.0
4ai	0.4	1	40.0				
4b	2.7	4	67.5				
5a	2.3	3	76.7				
5b	0.9	1	90.0				
6a	0.3	1	30.0				
6b	0.3	1	30.0				
6c	1	2	50.0				
7a	0.8	1	80.0				
8c	1	2	50.0				
9a	1.4	2	70.0				
9b	0.4	1	40.0				
9c	0.3	1	30.0				
9d	0.8	2	40.0				
Mean mark			60.7	Mean mark			51.8
Standard deviation			19.7	Standard deviation			12.5

(VCAA, 2005d)

The summary statistics for student performance on the Unit 4 2005 examination 2 are shown in Table 4.7 along with the Unit 3 2005 examination 1 for comparative purposes.

Table 4.7: Summary for the 2005 Examination 1 and 2 papers

	Question type	
	Recall	Application
Examination	Mean (s.d.)	Mean (s.d.)
2005 Examination 1		
Part A ^a	68.6 (16.1)	50.7 (13.8)
Part B ^b	58.6 (15.0)	59.9 (14.2)
2005 Examination 2		
Part A ^a	65.5 (8.0)	55.8 (12.0)
Part B ^b	60.7 (19.7)	51.8 (12.5)

Note: a= multiple-choice questions

b= short-answer questions

The results here tend to show that students performed better in the multiple-choice section than in the comparable part B section of Examination 1, application questions being the exception. Students tended to perform more favourably in recall questions than application questions, again the part B section of Examination 1 application questions being the exception. The standard deviation of scores was also smaller in part A. This is probably a reflection on the lesser possibility for a zero score than in short-answer questions.

Similar analysis was performed on the Examination papers 1 and 2 of 2003, 2004, 2006, and 2007. In 2008 the VCAA changed the Year 12 Chemistry course that altered the content and style in which the course was to be taught and also altered the balance of the type of curriculum material studied (Learner, 2005). Where there had been an emphasis, intended or not, towards the application questions in examination 1 and more descriptive questions in examination 2, the balance was now much more even. Due to the difference in balance between these examinations and their predecessors, including them in the analysis of this study was only likely to cause ambiguity and a lack of certainty in the results.

Rather than go through each paper in detail, a summary of the results similar to the summaries for the 2005 papers follows. Any obvious differences and similarities are discussed. The results for the papers are outlined in Tables 4.10 and 4.11. Table 4.10 shows the summative results for the analysis of all papers from 2003 to 2007.

The results shown in Table 4.8 initially describe from an observational point of view that there are differences between the performances of students on the different examinations, the type of question and the material studied. Whether these differences are just observational or significant needed to be determined. Analysis of the above factors that is, Unit 3 or 4 examination, multiple-choice questions or short-answer questions and recall and application questions were analysed by ANOVA test measurements using the SPSS statistical package (Field, 2005; Pallant, 2007).

Table 4.8: Summary for the 2003, 2004, 2006 and 2007 Examination papers

Year	Unit	Part	Recall Questions	Application Questions
			Mean (s.d.)	Mean (s.d.)
2003	3	A	78.1(15.1)	61.5(12.3)
		B	62.4(14.4)	59.7(17.8)
	4	A	68.4(8.6)	62.7(17.5)
		B	60.1(10.2)	56.6(17.7)
2004	3	A	67.6(14.7)	56.1(12.5)
		B	60.1(16.4)	61.6(21.7)
	4	A	68.8(13.3)	57.4(9.4)
		B	61.3(17.1)	57.6(17.7)
2006	3	A	68.3(18.9)	58.5(17.6)
		B	62.2(20.0)	62.4(14.6)
	4	A	74.3(13.1)	56.3(7.3)
		B	61.8(15.9)	57.8(17.9)
2007	3	A	62.9(22.7)	56.1(17.5)
		B	71.0(13.2)	58.4(14.5)
	4	A	61.7(13.7)	56.7(13.6)
		B	60.1(11.1)	54.2(14.0)

(VCAA, 2003c, 2003d, 2004c, 2004d, 2006c, 2006d, 2007c, 2007d)

The summary results show that generally student performance on multiple-choice questions focussing on recall was superior to student performance on multiple-choice questions focussing on application. Also evident is that student performance on application questions was generally weaker than on recall questions though there is some variation from year to year. For example, in 2003 student performance on short-answer application questions was weaker than on recall questions, this situation was reversed in 2004.

Table 4.9: Overall Summary of Unit 3 and 4 performance (Years 2003-2007).

Examination 1 (Unit 3)	Mean (s.d.)		Mean (s.d.)
Part A recall	69.1(5.5)	Part A application	56.4(4.0)
Part B recall	62.9(4.8)	Part B application	60.4(1.6)
Examination 2 (Unit 4)			
Part A recall	67.7(4.6)	Part A application	57.8(2.8)
Part B recall	60.8(0.7)	Part B application	55.6(2.6)

Table 4.10: Combined Unit 3 and 4 performance (Years 2003-2007)

Combined examination results for units 3 and 4			
Examination 1&2	Mean		Mean
Part A recall	68.4	Part A application	57.2
Part B recall	61.8	Part B application	58.0
Recall A&B	65.1	Application A & B	57.6
Total Part A	62.5	Total Part B	59.9

Conclusions from the Analysis of the Chemistry Examinations (2003-2007)

The analysis by means and standard deviation analysis suggests the following conclusions (refer to Tables 4.8, 4.9 and 4.10):

- Performance on recall questions is better than on application questions.
- Performance on multiple-choice recall is better than on short-answer recall.
- Performance on application multiple-choice is less different from that on short-answer recall. The results here are distinctly unclear with performance trends reversing between Unit 3 and Unit 4 examinations.
- Performance on recall questions is generally better than on application questions regardless of question type although the distinction is less clear with short-answer questions.
- Performance in multiple-choice questions is generally better than short-answer questions.

Analysis Using ANOVA

The analysis of mean scores using ANOVA demonstrated some apparent trends in the data. To be able to reliably report on these with respect to the relevant research questions it was important to determine the significance of these differences (Field, 2005; Pallant, 2007).

An important part of the analysis in relation to answering the Research Questions was the determination of any patterns and trends associated with student performance on the VCE Chemistry examinations with respect to style and content of the questions. ANOVA testing allows a comparison between the performance means of the questions by their varying characteristics. Due to the very small sample size in each individual test, comparative measurements with ANOVA were unlikely to achieve any worthwhile results. To enable effective testing, the question breakdowns were combined to give overall distributions based on the above-mentioned factors. Consequently, the combined question breakdown for Unit 3 examinations, Unit 4 examinations and both examinations, respectively, provided the data for the analysis of the examinations (see Appendix E). The ANOVA analysis output (see Appendix F) also included Eta-squared, which provided information on the correlation between

the studied variables. Eta-squared is a measure of the effect size or magnitude of the effect (degree of association) between the independent and dependent variables. It is a squared measure of association and can be interpreted in a similar fashion other squared correlation values (Field, 2005). The results of the mean, standard deviation, ANOVA and Eta-squared are shown in Table 4.11.

Table 4.11: ANOVA and Eta-squared analysis of question performance by question type and classification VCE Chemistry 2003-2007

Comparison variables	Mean	Standard deviation.	Discriminating variable	N	ANOVA results						
					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

Response to Research Question 1: Do students perform more effectively on multiple-choice or short answer questions?

Comparison of performance on multiple-choice and short-answer; both application and recall questions

The ANOVA results show that there is a statistical or meaningful, though not strong, significance in the difference in performance on application questions regardless of whether they are multiple-choice or short-answer. Students' performance on recall multiple-choice questions was better than on recall short-answer questions. This was demonstrated by the means (short-answer) = 58.9, standard deviation = 16.0 compared to means (multiple-choice) = 62.2, standard deviation = 15.1, and statistically, the ANOVA test did show that the difference was significant $F(1,485) = 5.34; p < 0.05$.

The Eta-squared result demonstrates that only 1.1% of the variation in the multiple-choice score can be explained by variation in the short-answer score, so whilst the result is significant the variation is not strongly explained suggesting a large degree of independence between performance on multiple-choice questions and short-answer questions.

This result demonstrates that students do perform more effectively on multiple-choice type questions. A deeper examination of this result looks at the effect of performance on multiple-choice and short-answer on the basis of the content of question; namely multiple-choice or short-answer. The difference is not large when viewed on the basis of all questions, however, it can be shown that the difference is very much dependent on the questions being recall in content.

Comparison of performance on multiple-choice and short-answer application type questions

As presented in Table 4.11, students' performance on application short-answer questions (mean/standard deviation = 57.1/16.2) was slightly better than on application multiple-choice questions (56.2/13.6). However, the ANOVA results ($F(1,255) = 0.24; p > 0.05$) show that there is no statistical significance in the difference in performance on application questions whether they are multiple-choice or short-answer. Consistent with this finding, there was no meaningful difference

with the Eta-squared value showing that only 0.1% of the variation in the multiple-choice score can be explained by variation in the short-answer score.

Comparison of performance on multiple-choice and short-answer recall type questions

As presented in Table 4.11, students' performance on recall multiple-choice questions (mean/standard deviation = 69.3/13.8) was better than on recall short-answer questions (60.9/15.5). The ANOVA results show a statistically significant difference ($F(1,228) = 17.80; p < 0.001$) in performance on application questions whether they are multiple-choice or short-answer. However the Eta-squared result demonstrates that only 7.2% of the variation in the multiple-choice score can be explained by variation in the short-answer score, so whilst the result is statistically significant the variation is not strongly explained.

Response to Research Question 2: Do Students Perform More Effectively on Recall Type Questions or on Application Questions?

Comparison of performance on recall and application, all questions

As presented in Table 4.11, students' performance on recall questions (mean/standard deviation = 64.2.3/15.4) was better than on application questions (mean/standard deviation = 56.7 /15.2). The ANOVA results show a statistically significant difference ($F(1,485) = 29.37; p < 0.001$) in performance on all questions whether they are recall or application. However, the Eta-squared result demonstrates that only 5.7% of the variation in the recall score can be explained by variation in the application score, so whilst the result is statistically significant the variation is not strongly explained.

This result demonstrates that students do perform more effectively on recall type questions. A deeper examination of this result looks at the effect of performance on recall and application on the basis of the style of question; namely multiple-choice or short-answer.

Comparison of performance on recall and application multiple-choice questions

As presented in Table 4.11, students' performance on multiple-choice recall questions (mean/standard deviation = 69.3/13.8) was better than on multiple-choice application questions (mean/standard deviation = 56.2 /13.6). The ANOVA results

show a statistically significant difference $F(1,198) = 45.71$; $p < 0.001$ in performance on multiple-choice questions whether they are recall or application.

The Eta-squared result demonstrates that 18.8% of the variation in the recall multiple-choice score can be explained by variation in the application multiple-choice score, so whilst the result is significant the variation is not strongly explained. This result compared to the previous tests demonstrates that a stronger correlation exists between the performance of students on recall and application multiple-choice application questions. The result suggests that student performance on recall multiple-choice is matched to some degree by the same students' performance on application multiple-choice. An initial conclusion would be that multiple-choice performance is a factor in its own right regardless of the nature or content of the question.

Comparison of performance on recall and application short-answer questions

As presented in Table 4.11 students' performance on short-answer recall questions (mean/standard deviation = 60.9/15.5) was better than on short-answer application questions (mean/standard deviation = 57.1 /16.2). The ANOVA results show a statistically significant difference $F(1,285) = 3.96$; $p < 0.001$ in performance on multiple-choice questions whether they are recall or application. However the Eta-squared result demonstrates that only 1.4% of the variation in the recall score can be explained by variation in the application score, so whilst the result is statistically significant the variation is not strongly explained.

Put simply, students' ability to answer recall short-answer questions has only a limited influence on the same students' ability to answer application style short-answer questions. This result suggests that the prompting that the multiple-choice options give to the student influences the performance on recall questions. This prompting is not present in short-answer questions and this is likely to account for the smaller difference in the means.

As mentioned earlier the most compelling relationship appears to exist between students' ability to answer recall and application multiple-choice questions. Nearly all the relationships were significant; however, the Eta-squared results showed that most of the correlations were not strong.

Comparison of performance on all questions by Unit 3 and Unit 4

The relationship between the Unit 3 and Unit 4 performances on the different categories, recall or application and multiple-choice and short-answer were initially tested using ANOVA measurements (Field, 2005; Pallant, 2007). The results are shown in Table 4.12. The variations between students' performance were generally not significant, however, the differences in the means of the students' performance give some indicative results and are accordingly presented. The results show the observed differences between performances that could be deduced from the means, and also indicate the level of significance where any was found.

**Table 4.12: ANOVA of Unit 3 and Unit 4 performance by question type
Chemistry 2003-2007**

ANOVA					
Comparison variables	Mean	Standard deviation.	Discriminating variable	F	Sig (p)
Unit 3	60.9	16.5	Multiple-choice	1.502	0.222
Unit 4	63.5	13.6			
Unit 3	59.9	16.0	Short-answer	1.119	0.291
Unit 4	57.9	15.9			
Unit 3	57.3	15.7	Application	0.426	0.515
Unit 4	56.0	15.0			
Unit 3	64.5	16.0	Recall	0.052	0.819
Unit 4	64.0	15.0			
Unit 3	60.3	16.2	All Questions	0.052	0.954
Unit 4	60.2	15.2			

Comparison of performance on multiple-choice questions by Unit 3 and Unit 4

The ANOVA test results show that there is no statistical significance ($F(1,199) = 1.50$; $p > 0.05$) in the difference in performance on multiple-choice questions between the Unit 3 and Unit 4 examinations. Students' performance was better on

the Unit 4 multiple-choice questions than on Unit 3 multiple-choice questions. This was slightly supported by the means (Unit 3 mean/standard deviation = 60.9/16.5 compared to Unit 4 mean/standard deviation = 63.5/13.6) but was not supported as statistically significant.

Comparison of performance on short-answer questions by Unit 3 and Unit 4

The ANOVA test results show that there is no statistical significance ($F(1,286) = 1.12; p > 0.05$) in the difference in performance on short-answer questions between the Unit 3 and Unit 4 examinations. Students' performance on the Unit 4 short-answer questions was better than on Unit 3 short-answer questions. This was slightly supported by the means (Unit 3 mean/standard deviation = 59.9/16.0 compared to Unit 4 mean/standard deviation = 57.9/15.9) but not supported as statistically significant.

Comparison of performance on application question by Unit 3 and Unit 4

The ANOVA test results show that there is no statistical significance ($F(1,256) = 0.43; p > 0.5$) in the difference in performance on application questions between the Unit 3 and Unit 4 examinations. Students' performance on Unit 4 application questions was marginally better than on Unit 3 application questions. This was demonstrated by the means (Unit 3 mean/standard deviation = 57.3/15.7 compared to Unit 4 mean/standard deviation = 56.0/4.5) but not supported as statistically significant.

This result is somewhat surprising as it could be expected that with less calculation/application questions in Unit 4 Examination, that the examination might present as being a friendlier option from a student point of view. It is, however, worth reiterating that the differences in performance are not statistically significant.

Comparison of performance on recall questions by Unit 3 and Unit 4

The ANOVA test results show that there is no statistical significance ($F(1,229) = 0.03; p > 0.5$) in the difference in performance on short-answer questions between the Unit 3 and Unit 4 examinations. Students' performance on the Unit 4 recall questions was essentially the same as on Unit 3 recall questions. This was demonstrated by the means (Unit 3 mean/standard deviation = 64.5/16.0 compared to Unit 4 mean/standard deviation = 64.0/15.0). Essentially, there is almost no

difference in performance between the two examinations on this classification of question.

Comparison of performance on all questions by Unit 3 and Unit 4

The ANOVA test results show that there is little statistical or meaningful significance ($F(1,486) = 0.01$; $p > 0.5$) in the difference in performance on short-answer questions between the Unit 3 and Unit 4 examinations. Students' performance on the Unit 4 questions was essentially identical to that on the Unit 3 questions. This was demonstrated by the means (Unit 3 mean/standard deviation = 60.3/16.2 compared to Unit 3 mean/standard deviation = 60.2/15.2). Essentially there is almost no discernable difference in performance between the two examinations on all questions even though there were small differences on some of the more specific classifications mentioned above.

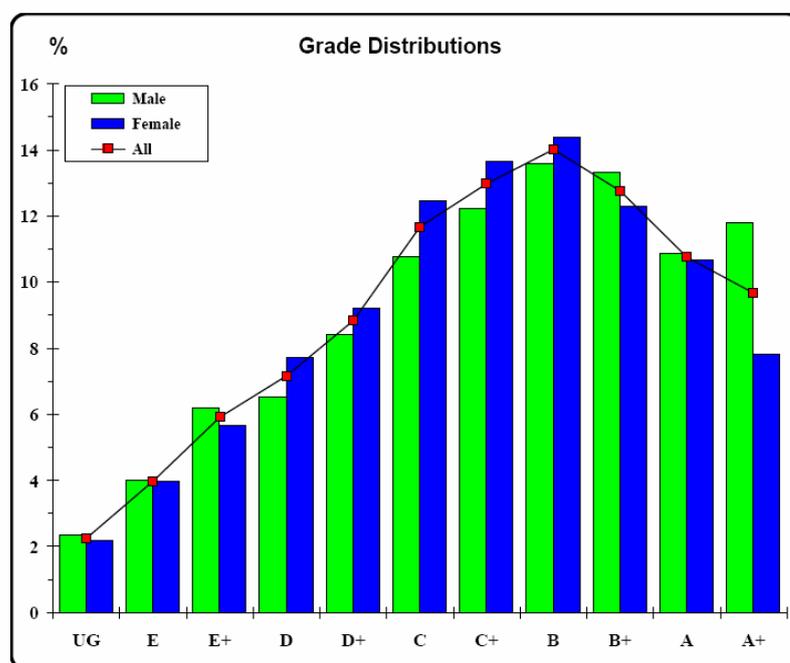
Essentially, there were no reportable differences in student performance on any category of question when the results are compared between the Unit 3 examination and the Unit 4 examination. This lack of difference is somewhat surprising considering the different emphasis in terms of subject content. The largest difference occurred with the multiple-choice questions where performance on the Unit 4 examination was higher. This could be due to the higher recall content that was present in the Unit 4 examinations. Overall, it is possible however, that some of the differences that may have existed could have been masked by the standardisation by the VCAA of student results. The standardisation procedure is designed to adjust the examination results of all students taking a particular examination where the examination was unusually easy or hard. This occurs after the examination where the VCAA statistically adjusts the grade distribution and absolute marks so that the overall appearance of the examinations in terms of grade allocation is essentially the same on both examinations (VCAA, 2010). Consequently it is difficult to detect which of the two examinations, Unit 3 or Unit 4, was actually more difficult because the results of two examinations are adjusted so that the overall mean and spread of both examinations are similar.

Response to Research Question 3: Does students' gender influence performance in chemistry examinations (or tests)?

Gender differences in performance

Further analysis of these papers highlights another important factor affecting performance on the examination. That is the discrimination of performance by the different genders.

As had been found by a number of researchers, the performance of males compared to females on the more theoretical physical sciences Physics and Chemistry was somewhat at odds to the performance on the less theoretical sciences such as Biology and Psychology (Beller & Gafni, 1991; Hamilton, 1998; Hedges & Howell, 1995).



(VCAA, 2005e)

Figure 4.13: Grade distributions for the 2005 Chemistry Examination 1

Data by the VCAA, whilst simplistic, does demonstrate this discrepancy. The grade distributions for all examinations published by the VCAA show the percentages of both genders that achieved certain grades. For example the graph in Figure 4.13 above shows the grade distributions for examination 1 2005.

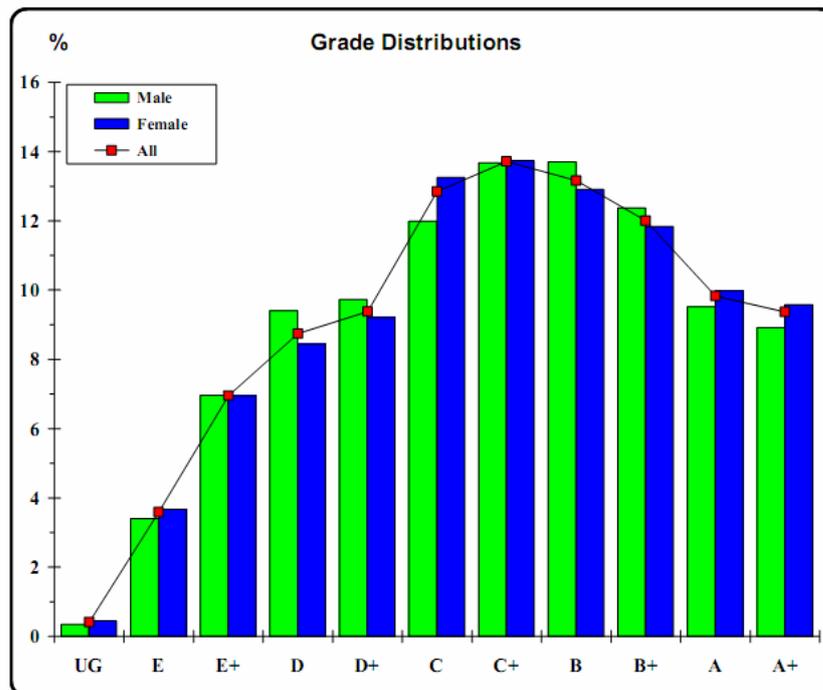
This graph (Figure 4.13) demonstrates that male students tend to outperform female students at the higher end of the grade distributions (A⁺, A, B⁺). The male students also “outperform” at the other end of the grade spectrum as well. Various interpretations can be supposed for this latter aspect of the results but it is not within the scope of this research. Performance in chemistry is the main focus rather than explaining poor performances. It is also worth noting the numbers of students doing chemistry. Across the five years being considered the numbers of males and females participating in Year 12 chemistry is shown in Table 4.13.

These data show that the number of females taking chemistry has been fairly constant over the five years, whereas the number of males, whilst lower, has steadily increased. That this is lower than the total females’ number is an interesting and perhaps surprising fact.

**Table 4.13: Participation numbers of male and female students in VCE
Chemistry 2003-2007**

Gender	2003	2004	2005	2006	2007
Males	4012	4104	4257	4350	4381
Females	4699	4705	4861	4812	4714

The grade distribution (Figure 4.13) is typical for all (2003-2007) the Chemistry Examination 1 distributions. Whilst there are some minor variations from year to year they generally show this pattern. When this is compared to the Biology examination for the same time period the results are somewhat different (see Figure 4.14).



(VCAA, 2005e)

Figure 4.14: Grade Distributions for the 2005 VCE Biology Examination 1

It is notable that in 2005 some 7982 females but only 3761 males took Biology. Clearly a 2:1 ratio in favour of the females again highlighted the observed preference for this more social of the sciences.

The performance in the examination also demonstrates this difference. Remembering that in Chemistry the male students performed more effectively than female students in the higher grades, this does not happen in Biology. The female students outperform the male students at the higher grades of A and A⁺ though not to the same extent that male students outperform female students in the same grades in Chemistry. This demonstrates the observations by other researchers (Beller & Gafni, 1991; Cox et al., 2004; Hamilton, 1998; Hedges & Howell, 1995).

Those female students that tend to perform better in this less physical science are also voting with their feet by choosing the subject in larger numbers than Chemistry. Conversely, the males choose Chemistry in larger numbers than they do Biology.

The observed trends so far commented on only provide a superficial view. Whilst more male students achieve at higher grades than female students, and this is

consistent across the years 2003 to 2007, it is important to determine whether this is a significant observation. The grade distributions (see Appendix G) across the years 2003 to 2007 (VCAA, 2003c, 2004c, 2005e, 2006c, 2007c) only shows three of the possible ten grades, and is intended to show the trend of the performance of the two genders across the range of possible grades. Thus only the grades of A⁺, C⁺ and E⁺ are examined. To act as a comparison for each grade and its distribution across the two genders, the estimated distributions have been calculated. As there are more females than males attempting chemistry it is reasonable to presume that the distribution of the grades would be proportional to the number from each gender attempting the subject. It is also a reasonable assumption that the VCAA has no deliberate intention to have examinations designed that will intentionally favour one gender over the other. Consequently the expected distributions reflect this. If, for example, 1000 students were to be awarded A⁺ grades and 60% were female students then 600 female students should receive A⁺ grades.

Chi-squared analysis of the VCE Chemistry examinations (2003-2007)

The Chi-squared analysis for each of the examinations from 2003 to 2007 (see Appendix E) analyses the Unit 3 and Unit 4 examinations for each of the A⁺, C⁺, and E⁺ grades.

For example, the Unit 3 2003 Examination is discussed as a demonstration of the information provided on gender performance on that particular examination.

Unit 3 Examination 2003, Chi-squared analysis of male and female performance

An important part of the analysis in relation to answering Research Question 3 (Does students' gender influence performance in chemistry examinations (or tests)?) was the determination of any patterns and trends associated with student performance on the VCE Chemistry examinations with respect to the gender of the students. Ideally, the expectation of awarding of the grades should be that there is no favouritism or bias present. Thus a reasonable expectation would be that 50% of the awarded A⁺ grades should go to males and the other 50% to the female candidates, assuming a 50/50 distribution of male and female candidates. Chi-squared analysis is most appropriate for this analysis as it allows a comparison between the actual means and

the expected means; in this case the numbers of students awarded each of the grades in question. A detected bias would have implications for the subject in terms of students' motivation towards the subject. Data provided by the VCAA describes the numbers of each gender awarded each grade A⁺, the highest grade, down to UG (ungraded), the lowest grade (Field, 2005; Pallant, 2007).

The total number of A⁺ grades awarded in the Unit 3 examination in 2003 was 869. The actual number of male students awarded this grade was 477 and female students 392. If distributed according to the number of females and males actually taking the examination and assuming an equitable distribution the number of male students receiving an A⁺ should have been 400 and female students 469 because more females (4653) compared to males (3973) studied Chemistry and sat this examination. That more male students were awarded A⁺ is common in all the examinations in Chemistry from 2003-2007, but is the difference significant? This trend alone is probably reason enough to question the assessment of the Chemistry as it apparently favours males over females particularly at the higher and arguably more important grades. The Chi-squared test of association (Table 4.14) shows that the difference is highly significant. That is, Chi-squared = 27.464 and the significance is $p < 0.001$. This suggests that the difference in performance is significant and could not reasonably be explained by chance.

Table 4.14: Frequency and Chi-squared analysis of the performance in terms of the A⁺, C⁺, E⁺ grades 2003 Unit 3

Gender	Grade	Observed N	Expected N	df	Chi-squared
Female	A ⁺	392	469	1	27.464*
Male	A ⁺	477	400		
Female	C ⁺	668	649	1	1.207
Male	C ⁺	536	555		
Female	E ⁺	217	230	1	1.361
Male	E ⁺	283	270		

*: $p < 0.001$

The total number of C⁺ grades awarded in the Unit 3 examination in 2003 was 1204. The actual number of male students awarded this grade was 536 and female students 668. If distributed according to the number of females and males actually taking the examination, and assuming an equitable distribution, the number of male students receiving a C⁺ should have been 555 and female students 649. At this grade the differences between the expected distribution and the actual distribution are much closer and this time the distribution reflects the actual distribution of males and females, that is, more female students were awarded C⁺ than were male students. Again this is generally the trend in all the Unit 3 examinations from 2003 to 2007. Superficially the Unit 3 examination would appear to be more able to accurately assess males and females at this level than is apparently the case at the higher grade. The main difference between this and the A⁺ grades is that the awarding of the C⁺ grade is slightly in favour of the female students whereas the A⁺ grades highly favoured the male students. The Chi-squared test of association shows this clearly and demonstrates that there is no significance in the distribution of the C⁺ grade between the males and the females. That is the Chi-squared = 1.207 and the significance is $p > 0.05$. This suggests that there is no discernable difference in performance of males and females with respect to their expected performance.

The total number of E⁺ grades awarded in the Unit 3 examination in 2003 was 500. The actual number of male students awarded this grade was 217 and female students 283. If distributed according to the number of females and males actually taking the examination and assuming an equitable distribution the number of male students receiving an E⁺ should have been 270 and female students 230. At this grade the differences between the expected distribution and the actual distribution are much closer than with the A⁺ distribution and much more similar to the C⁺ distribution and again the distribution reflects the actual distribution of males and females, that is more female students were awarded E⁺ than were male students. Again this is generally the trend in all the Unit 3 examinations from 2003 to 2007. Superficially the Unit 3 examination would appear to be more able to accurately assess males and females at this level than is apparently the case at the higher grade.

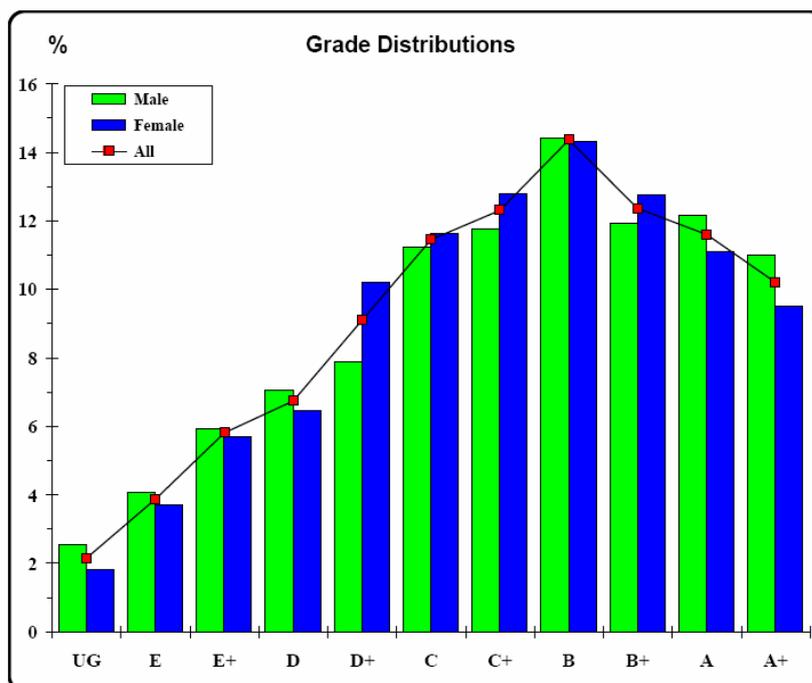
The Chi-squared test of association shows this clearly and demonstrates that there is little significance in the distribution of the E⁺ grade between the male students and the female students, that, is Chi-squared = 1.361 and the significance is $p > 0.05$. This

suggests that there is no discernable difference in performance of males and females with respect to their expected performance.

Overall the analysis of this particular examination shows that male students are significantly outperforming female students at the A⁺ level but the differences at the middle and lower end of the scale are not significant.

Unit 4 Examination 2003, Chi-squared analysis of male and female performance

The Unit 4 examination is one that observationally appears to favour the females over the males. As a teacher of chemistry I had noted that the females I taught were happier with the Unit 4 examination because it had far less calculations involved. It should be reiterated at this point that the earlier analysis has shown that on an overall basis there appears to be little difference on the performance on the Unit 3 examination compared to the Unit 4 examination (see Table 4.12). An examination of the Unit 4 examination results show that whilst the overall results may have been similar in Unit 3 compared to Unit 4 the distribution of the performance by grades awarded was different in Unit 4 compared to Unit 3. This is demonstrated by the comparison of the 2005 Unit 3 examination (see Figure 4.13) with the 2005 Unit 4 examination (see Figure 4.15) which shows that the skewness in the data is somewhat reduced compared to Unit3.



(VCAA, 2005e)

Figure 4.15: Grade distributions for the 2005 Chemistry Examination 2

The female students have performed more strongly (or the male students less so) on this examination than they did on the Unit 3 examination from the same year.

As shown in Table 4.15, the total number of A⁺ grades awarded in the Unit 3 examination in 2003 was 804. (Presumably about 60 students who attempted the Unit 3 examination did not attempt Unit 4). The actual number of male students awarded this grade was 408 and female students 396. If distributed according to the number of females and males actually taking the examination and assuming an equitable distribution the number of males receiving an A⁺ should have been 370 and females 434. That more male students than female students were awarded A⁺ is common in all the Unit 4 examinations in Chemistry from 2003-2007 as was the case with the Unit 3 examinations, but is the difference significant? The Unit 3 examination showed a significant absolute difference of 85, that is 85 more male students received an A⁺ than did female students. The Unit 4 examination shows a smaller absolute difference of 12 with just 12 more male students receiving an A⁺ grade than did female students. This suggests that this examination was better balanced in terms of its accessibility to females compared to the Unit 3 examination.

Table 4.15: Frequency and Chi-squared analysis of the performance in terms of the A⁺, C⁺, E⁺ grades 2003 Unit 4

Gender	Grade	Observed N	Expected N	df	Chi-squared
Female	A ⁺	408	370	1	7.230*
Male	A ⁺	396	434		
Female	C ⁺	431	475	1	7.545*
Male	C ⁺	602	558		
Female	E ⁺	241	227	1	1.361
Male	E ⁺	252	266		

*: $p < 0.01$

The Chi-squared test of association (see Table 4.15) shows again that the difference is significant. That is the Chi-squared = 7.23 and the significance is $p < 0.05$. These values are smaller than with the equivalent grade at Unit 3. Again this supported the premise that the more descriptive, less analytical content of Unit 4 improves the females' chances of performing well. This suggests that the difference in performance is significant and could not reasonable be explained by chance.

The total number of C⁺ grades awarded in the Unit 4 examination in 2003 was 1033. The actual number of male students awarded this grade was 431 and female students 602. If distributed according to the number of females and males actually taking the examination and assuming an equitable distribution the number of male students receiving a C⁺ should have been 475 and female students 558. At this grade the differences between the expected distribution and the actual distribution are much closer and this time the distribution reflects the actual distribution of males and females, that is, more females were awarded C⁺ than were males. Again this is generally the trend in all the Unit 4 examinations from 2003 to 2007, as it was with Unit 3 examinations. This time, however, the differences are larger than they were in Unit 3. That is more females were awarded a C⁺ than might have been expected. The Chi-squared test of association (see Table 4.15) shows this clearly and demonstrates that there is significance in the distribution of the C⁺ grade between the males and the

females. That is the Chi-squared = 7.55 and the significance is $p < 0.05$. This suggests that there is difference in performance of males and females with respect to their expected performance.

The total number of E^+ grades awarded in the Unit 4 examination in 2003 was 493. The actual number of male students awarded this grade was 241 and female students 252. If distributed according to the number of females and males actually taking the examination and assuming an equitable distribution the number of male students receiving an E^+ should have been 227 and female students 266. At this grade the differences between the expected distribution and the actual distribution are much closer than with the A^+ distribution and much more similar to the C^+ distribution and again the distribution reflects the actual distribution of males and females, that is more female students were awarded E^+ than were male students. Again this is generally the trend in all the Unit 3 examinations from 2003 to 2007. Superficially the Unit 4 examination (like the Unit 3 examination) would appear to be more able to fairly assess males and females at this level than is apparently the case at the higher grade. The Chi-squared test of association (see Table 4.15) shows this and demonstrates that there is little significance in the distribution of the E^+ grade between the males and the females. That is the Chi-squared = 1.361 and the significance is $p > 0.05$. This suggests that there is no discernable difference in performance of males and females with respect to their expected performance.

On the basis of these initial studies it appears that the performance by male students at the higher grades is greater than could be expected according to the number of males sitting. The differences in the lower grades appear to be less significant or not significant. The differences, when analysed by grade, appear less pronounced in the Unit 4 examinations than the Unit 3 examinations.

Table 4.16 shows the Chi-squared values for each examination from 2003 to 2007. For each grade A^+ , C^+ and E^+ the Chi-squared value relates to the difference in performance at that grade for male students compared to female students.

A⁺ grade analysis

These results, in Table 4.16, show a clear significance in the differences in performance of male students and female students at the A^+ level with male students gaining a significantly greater proportion of the A^+ grades. The significance was

mostly $p < 0.001$ the only exceptions being three Unit 4 examinations where that significance was $p < 0.05$. Again demonstrating that this examination more evenly awards A^+ grades to females and males. The Chi-squared values were consistently higher for Unit 3.

Table 4.16: 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*	5.931#	0.306
2004 Unit 4	7.150#	1.132	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

* : $p < 0.001$

: $p < 0.05$

C^+ grade analysis

These results, in Table 4.16, show a variable significance in the performance of male students and female students at the C^+ level. Essentially the differences were significant (at $p < 0.05$) in some years but generally the significance, if any, was weak. The Chi-squared values are much smaller than was the case with the A^+ values again demonstrating a much weaker relationship between the grade awarded to male students and female students. The reversing of the trend observed with the A^+ grades somewhat redresses the imbalance that was present with the A^+ grades and, to some extent, explains the apparent similarity in overall performances on the Unit 3 and Unit 4 examinations.

E⁺ grade analysis

The results, in Table 4.16, of the Chi-squared analysis of the grade distributions of the E⁺ grade show quite clearly that the grade distributions closely match the balance between the genders. . With the notable exceptions of the Unit 3 examinations from 2006 and 2007, which show a significant difference ($p < 0.05$) between the allocation of the E⁺ grade between male students and female students with more male than female being awarded E⁺ grades, the differences are not significant and the Chi-squared values themselves are usually very small.

Table 4.17 shows the Chi-squared analysis of the whole of Unit 3 examinations, the whole of Unit 4 examinations and all examinations.

Table 4.17: Chi-squared analysis of Male and Female performance on the Unit 3 and Unit 4 Examinations in Chemistry

	A ⁺		C ⁺		E ⁺	
<i>Examination</i>	<i>Chi-squared</i>	<i>Sig (p)</i>	<i>Chi-squared</i>	<i>Sig (p)</i>	<i>Chi-squared</i>	<i>Sig (p)</i>
All Unit 3	189.5	0.000	14.57	0.000	6.528	0.011
All Unit 4	62.58	0.000	11.896	0.001	3.210	0.073
All Examinations	234.6	0.000	26.25	0.000	0.298	0.585

With such large numbers of cases involved in the combined analysis it is to be expected that the significance levels be more pronounced than they are when considering a single examination. Consequently all the results, except those for E⁺, show significant differences ($p < 0.001$) between the males and females in the awarding of grades. Generally female students do not get the proportion of higher grades that is due to them in terms of the number of them that are attempting the examination. The most marked difference occurs in the distribution of the A⁺, demonstrated when considering the size of the Chi-squared values.

The A⁺ Chi-squared values are very high, showing the significance of the difference in the award of this grade to female students compared to male students with the males receiving the greater proportion of the A⁺ grades. It is also important to note that the difference is less marked in the Unit 4 examination than the Unit 3 examination. These trends were generally observed when considering each

examination individually but because of the smaller case size the difference was less evident.

The C⁺ grades, when considered as a whole, do show significance in the award of the grade when comparing males to females. The Chi-squared values show that the difference is less pronounced than was the case with the A⁺ grade. This also was reflected when each examination was considered separately.

The E⁺ results show even less difference in the allocation of grades between the males and females. Only Unit 3 shows a significant result and, when considering the case size, this is not a large or powerful significance.

Overall it appears that the Chemistry examinations do not award grades in proportion to the genders that attempt them and that this difference is most pronounced at the higher grades, significantly in favour of the males. The differences appear less pronounced in the Unit 4 examination as compared to the Unit 3 examination in spite of there appearing to be no overall differences between the performance levels of the Unit 3 and Unit 4 examination when they are compared to each other.

Summary of Analysis of the 2003-2007 Chemistry Examinations

Analysis of these examinations shows a number of significant trends or observations in the performance of students on these examinations.

The factors that were considered

- The type of question: Multiple-choice or Short-answer.
- The classification of question: Recall or Application
- Gender of student: Male or Female.

The following summarise the findings in terms of the observation being determined as significant using statistical analysis of ANOVA and Eta-squared or Chi-squared as appropriate to the type of data (Field, 2005; Pallant, 2007).

Significant results

- Students' responses on multiple-choice recall questions scored significantly better than short-answer recall questions (see Table 4.11).
- Students' responses on multiple-choice questions, generally, scored significantly better than short-answer questions (see Table 4.11).

- Students' responses on multiple-choice recall questions scored significantly better than multiple-choice application questions (see Table 4.11).
- Students' responses on short-answer recall questions scored significantly (but not strongly so) better than short-answer application questions (see Table 4.11).
- Students' responses on recall questions generally scored significantly better than application questions (see Table 4.11).
- Male students performed significantly better than female students in terms of the awarded A⁺ grades in all years 2003-2007. The difference most marked in the Unit 3 examinations (see Table 4.16 and 4.17).
- Female students out performed male students significantly in terms of the awarded C⁺ grades in all years 2003-2007. There was no particular trend between the Unit 3 and Unit 4 examinations (see Table 4.16 and 4.17).

Observed results that were not statistically significant

- Short-answer application questions scored slightly better than multiple-choice application questions (see Table 4.11).
- There was little difference in any aspect of the performance between the Unit 3 and Unit 4 examinations with regard to the type or classification of questions (see Table 4.12). They were essentially identical. There were, however, differences when gender was taken into consideration (see Table 4.16 and 4.17).
- There was little difference between the awarded E⁺ grades in all years 2003-2007. Neither gender showed a tendency to be awarded the grade above or below the expected distribution. The distribution varied from year to year and was rarely significant (see Table 4.11).
- None of the Eta-squared correlation tests showed, even when significant with respect to question type or classification, that the correlation was not very strong. The strongest Eta-squared showed only 18.8% causality between the measured variables (see Table 4.11).

Further discussion of the implications and meanings or interpretations of these findings will occur in Chapter 6.

Chapter 5: Results-School Trials and Interviews

This chapter investigates the results from the trial tests and student interviews. The first section of the chapter examines the results obtained from the trial tests. Rasch analysis (RUMM laboratory P/L, 2009) was performed on the data from the trial tests to determine the reliability and fit of the items and of the students taking part. The overall performance on the individual items and differential analysis examining the type of question (recall or application) (Research Question 1), the style of the question (multiple-choice or short-answer) (Research Question 2) and the gender of the student (Research Question 3) were analysed to see if any patterns or trends emerged. A comparison with the analyses performed in the previous chapter on the VCE Chemistry examinations is also discussed. The chapter then covers a description and analysis of the student and staff interviews (Research Questions 4 and 5) that were conducted. These interviews specifically looked at the students' and staff attitudes and preferences to the type and style of tests that they were experiencing.

School Trials of Chemistry Test Questions

The analysis of the past papers indicated some trends with regards to the performance of students and the type of question asked. For example, student performance on multiple-choice questions was better than on short-answer questions. The issue of whether this was a significant difference or simply a function of the complexity of the questions asked was partially addressed by the conduction of the trial tests. Some 192 students from Year 11 were involved with the trial tests. Five of these students were unable to complete more than half of the tests and were subsequently removed from the data. The number of students involved allowed a reasonable level of confidence in the reliability and validity of the questions in measuring performance on the questions. This was confirmed using Rasch analysis (RUMM Laboratory, 2009b).

The difficulty of the past papers mentioned above stems from the fact that, for example, a multiple-choice question assessing the recall of some aspect of chemistry was not going to be repeated in an identical (or nearly so) form on the same examination as a short-answer question. Consequently the observed result, that

performance on multiple-choice recall was significantly higher than short-answer recall, may have been a function of the difficulty of the content of the short-answer questions compared to the content of the multiple-choice questions rather than the multiple-choice mode being more able to draw out student understanding.

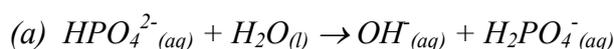
To attempt to determine an answer to this issue; nearly identical questions on a particular topic point were constructed so that the multiple-choice version and the short-answer version of a pair of questions were evenly matched in terms of the content difficulty.

Examples of paired difficulty questions

Acid-Base (recall) Question (See Appendix A)

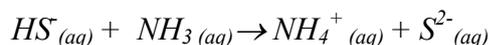
From the Short-answer Acid-Base Test, Question 3

3. *Identify the conjugate pairs in the following equation: (indicate which is the acid and which is the base for both conjugate pairs)*



From the Multiple-choice Acid-Base Test Question 3

3. *Consider the following equation:*



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

- A. *HS⁻ and NH₃*
- B. *HS⁻ and S²⁻*
- C. *NH₃ and S²⁻*
- D. *NH₄⁺ and HS⁻*

Both these questions examine essentially the same aspect of chemistry to a similar level of difficulty. Students in 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.

Stoichiometry (application) Question

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 $Zn_3(PO_4)_2$?*

Again each question is assessing student understanding of the same concept to a similar level of difficulty and again the hypothesis is that the same student should perform equally well on either question if the questions are assessing their ability equally well.

Consequential multiple-choice questions

A further unusual construct that was employed in these tests was the use of a consequential multiple-choice question. Typically many stoichiometry tests that involve answering a complex stoichiometric problem involve providing a consequential scaffold for the students when they attempt the question. These are very common in the VCE Chemistry examination papers as short-answer questions.

For example Question 7b from the 2005 Unit 3 examination

Question 7

Ethanoic acid (CH_3COOH) is a weak acid in water

7b. *A 0.100 M solution of ethanoic acid in water at 25°C has a pH of 2.88.*

- i.** *Calculate the hydrogen ion concentration in a 0.100 M solution of ethanoic acid.*
- ii.** *Calculate the acidity constant of ethanoic acid at 25°C.*

1 + 2 = 3 marks

(VCAA, 2005c)

To successfully answer this question and obtain full credit the student must correctly calculate the value of the hydrogen ion in the solution (part i.) and then use this value correctly in part ii. to obtain the required value for the acidity constant required. A

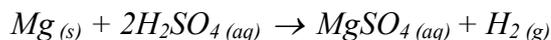
student who makes an error in part i cannot then obtain the correct value in part ii. To alleviate this double jeopardy problem examination markers are instructed to award marks for correct working even if the value obtained from those calculations is ultimately incorrect. In this case a student who obtains an erroneous value for the hydrogen ion concentration but then uses this value in a correct manner in part ii would be awarded 2 marks (from part ii.) and zero marks for part i. even though neither value was correct. Hence the student obtains 67% for the question.

This type of consequential system is not used in multiple-choice questions. Each multiple-choice question must stand alone even if more than one question arises from a particular stimulus material.

To ensure that each test was as similar as possible and thus be able to match the typical stoichiometric scaffold question, one set of questions in the multiple-choice stoichiometry test was consequentially designed and marked.

Stoichiometry multiple-choice test question 4 (see Appendix A)

4. *Zinc reacts with dilute sulfuric acid according to the equation:*



In a certain experiment, 19.6g of magnesium is reacted with excess sulfuric acid.

- a. *How many mole of sulfuric acid are required to completely react with 19.6g of magnesium?*
 - A. 0.403 mol
 - B. 0.620 mol
 - C. 1.61mol
 - D. 2.48 mol
- b. *Calculate the volume of 1.50M sulfuric acid that would react with 19.6g of magnesium?*
 - A. 0.268 L
 - B. 0.413L
 - C. 1.08L
 - D. 1.65L
- c. *Calculate the mass of magnesium sulfate produced in the experiment.*
 - A. 24.2g
 - B. 37.2g
 - C. 97.0g
 - D. 149g

(1 + 1 + 1 = 3 marks)

To allow students who selected the incorrect answer to be able to successfully and fairly access parts b and c, the answer and distracter for parts b and c were directly linked to the four answer options in part a. The correct answer for 4a was option C (1.61 mol). This led to the correct answer for part b and part c. Students who made an error in part a (that is selecting A, B or D were not disadvantaged because each of these answers lead to one of the distracters in part b. For example, a student who selected A (by halving the moles of magnesium instead of doubling to get the moles of sulfuric acid) would, if they used correct procedures for part b, obtain an answer of 0.268 M (option A). Thus even though they were in error in part a they still had a fair opportunity to demonstrate their understanding for part b, as would happen had the question been asked in a short-answer format.

Results from the Trial Tests

Some 192 students participated in the class trials. The distribution of students was 99 males and 93 females. After the tests were collected from the trial schools they were marked and coded according to the coding framework (see Appendix H). 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, 2001; Connolly, 2009).

RUMM2030 Analysis of the Trial 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 P/L, 2009). The initial analysis using the RUMM2030 program showed that the trial 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

The RUMM2030 analysis showed that most items had satisfactory individual item fit residuals, that is, none were above 2.5 (see Appendix H-1) that accounted for the small average fit residual value mentioned above (Cavanagh et al., 2003; Connolly, 2009). Furthermore, only three items had individual probabilities that were likely to have impacted on the poor item-trait interaction value. These items were AB08 (Pr = 0.0093) and ST09 (0.00095) and ST05 (0.00045). Issues raised from the RUMM2030 item threshold maps further complicated concerns with these items.

The threshold summary (see Appendix H-2) demonstrated that several items had reverse thresholds indicating that the marking structure did not ideally match the question response provided by the student (Connolly, 2009). Reverse thresholds will be discussed in more detail on the next page.

Resolving Item and Person fit issues

Person fit

RUMM2030 Rasch analysis (RUMM laboratory P/L, 2009) was used to determine the fit of the trial test construct to the students taking part in the trials analysis (person fit) and also of each test item to all the other test items (item fit).

With over 190 students participating, it was likely that some students' responses would be of poor fit when compared to other students completing the tests. The poor fit may be due to a number of causes, for example, lack of understanding of some aspects of the material tested or lack of willingness to participate fully in the tests. This produces an erratic pattern when comparing the student responses to the remaining students' responses. As with poorly fitting items, students' with fit residual values of greater than 2.5 are considered to have poorly matched the test construct and the other students' responses (Andrich, 2005; Cavanagh et al., 2003).

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 $Pr(\text{Bonferroni}) = 0.05 / 24$ (number of items) = 0.00208. (Pallant, 2007)

The simplest expedient is to delete these students from the sample. An examination of the person fit characteristics showed three students with suspect fit residuals (over 2.0 and close to 2.5), namely students 45 (fit residual; 2.525), 53 (fit residual: 2.481) and 19 (fit residual: 2.462). These students were deleted from the sample set of students. This resulted in some improvement to the probability value (0.0001) and also resolved the reverse threshold issue with item AB12 negating the need for rescoring this item. This left 189 students in the sample. The RUMM2030 program deleted students with too few responses to be included in the analysis; these students were omitted from some analyses, e.g. ANOVA.

Item fit

The item fit analysis, which showed those questions with weak discrimination, appeared to only occur with the polytomous questions, which were all short-answer. The common issue with the five questions concerned (Acid-base questions AB10 and AB12 and Stoichiometry questions ST10, ST11 and ST12; see Appendix A) was that they did not discriminate effectively in the middle of the mark range. Essentially, students who were coded as a 1 for these questions were not matching the expected probability levels. Whilst the probability for the code 1 responses peaked in the middle of the score range as expected but the number of students achieving these grades was too low. Middle ability students were more likely to score either 0 or 2 not 1. That is, the probability of the student obtaining a correct response did not match the student's ability as determined by their performance on the remaining questions (Connolly, 2009). This issue is known as reverse thresholds.

On the advice of the statistician assisting with the analysis the items were rescored as either 0,0,1 or 0,1,1 and the analysis recommenced (Connolly, 2009). An example of this analysis is shown below with a comparison to a question that did discriminate effectively (Figures 5.1 and 5.2)

Example of a weakly discriminating question compared to a strongly discriminating question.

Question ST 10 demonstrates reverse thresholds (see Figure 5.1 and Appendix A). The probability lines for a zero score and for the score of 2 appear as might be expected. The students of low ability, as determined by performance on all the questions and shown on the student ability (logit) axis, have a high probability of

getting the zero score whereas students of high ability have a low probability of achieving the zero score. As would be expected, when the measured student ability rises, the probability of a zero score decreases and the probability of a 2 score increases. The difficulty with this question is the 1 score. This curve should start at a low probability rising in the middle and then dropping as the student ability rises so that as student ability increases the mark awarded moves from 0 to 1 and then 2. Whilst the curve does demonstrate some of this pattern it simply is not strong enough in the middle range of student ability. Students of middle ability (around a logit of 0.0) have a higher probability of getting a 2 or zero score than they do of getting a 1 score. To further emphasise the weakness of the distribution the Chi-square for this question is close to zero demonstrating that performance on this question correlates very poorly with performance on the remaining questions (Bond & Fox, 2001; Cavanagh et al., 2003; Connolly, 2009).

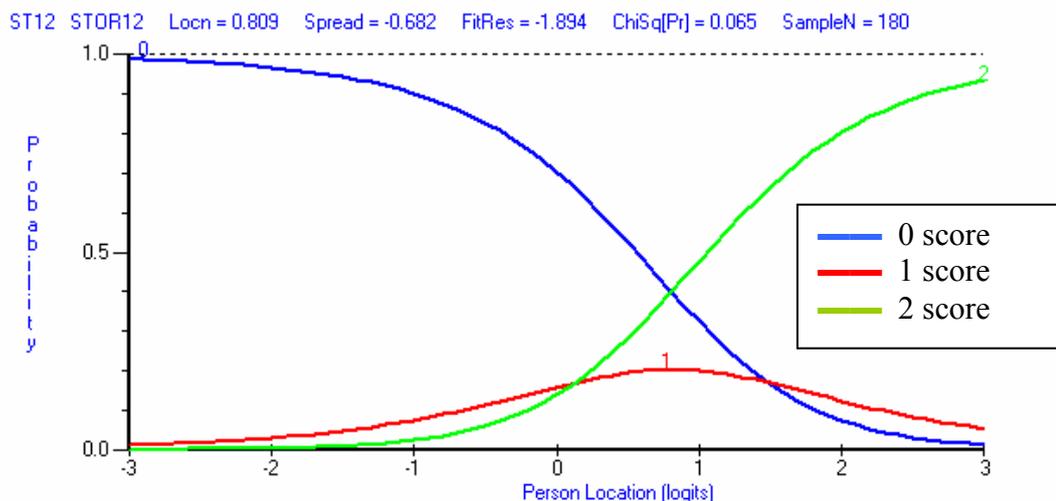


Figure 5.1: Rasch performance curves for question ST 12 using RUMM2030

A correctly performing polytomous situation is demonstrated in question AB07 (see Figure 5.2 and Appendix A). The curves here show that as ability increases the students' start with the highest probability score being 0, and then progressing to 1, and finally as ability reaches the higher levels the most probable score is 2. Note that the Chi-square probability is much higher with this question than with question ST12.

Essentially, ST12, instead of giving a polytomous description of student performance the question is only discriminating effectively on the basis of right and wrong. As

mentioned before the relatively small number of students on each question where reverse thresholds occurred were rescored as either 0,0,1 or 0,1,1 depending on which alternative produced the best item fit statistics (Cavanagh et al., 2003; Pallant, 2010).

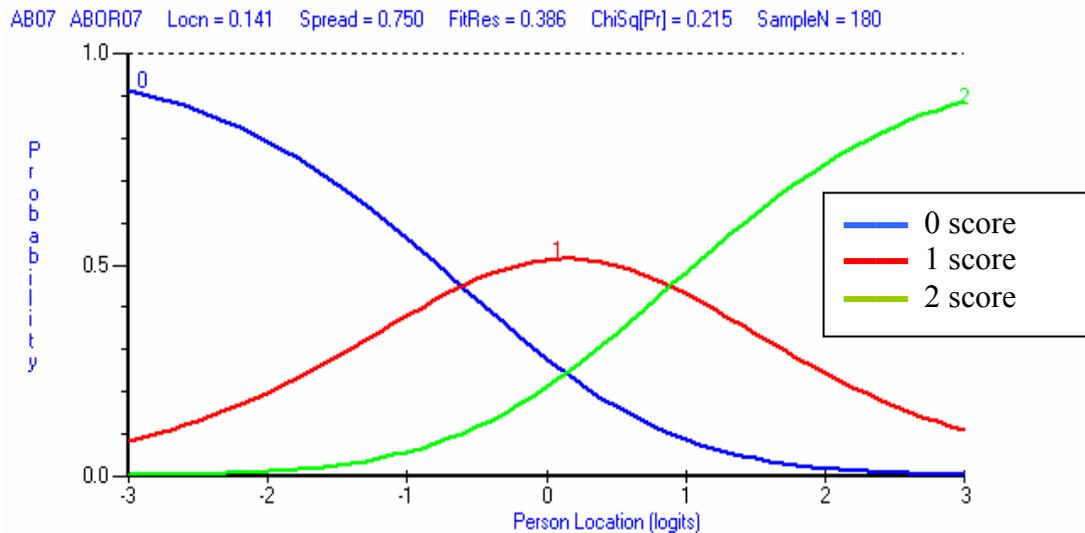


Figure 5.2: Rasch performance curves for question AB 7 using RUMM2030

Rescoring test items

Rescoring items such as ST12 involved two possible courses of action. The item can be rescored as 0,0,1 or as 0,1,1 instead of 0,1,2. For example, re-scoring as 001 would result in all students who were awarded 1 would be rescored as a 0 and all 2 scores would be rescored as 1. Re-scoring item ST012 as a 0,1,1 item produced an item probability of 0.00001 that was a weaker item fit value than previously existed for the item (0.065). Re-scoring as 0,0,1 produced a probability of 0.002 that was acceptable in terms of the Bonferroni adjusted value being sort for the whole sample of items, but it was lower than the original probability. This adjustment was still necessary, however, as it did resolve the reverse thresholds issue. Using a similar process items ST10, ST11 and AB10 were rescored to achieve maximum β probabilities. ST10 was rescored as 0,1,1; ST11 as 0,1,1 and AB10 as 0,1,1, each being rescored to give the maximum probability value for the item trait interaction. After completing the above rescoring item AB12 no longer exhibited reverse thresholds so was not rescored.

After the questions had been rescored to accommodate this issue the new graph for Question ST12 appeared as shown in Figure 5.3. As can be seen, this question now discriminates well between students of low ability and students of high ability.

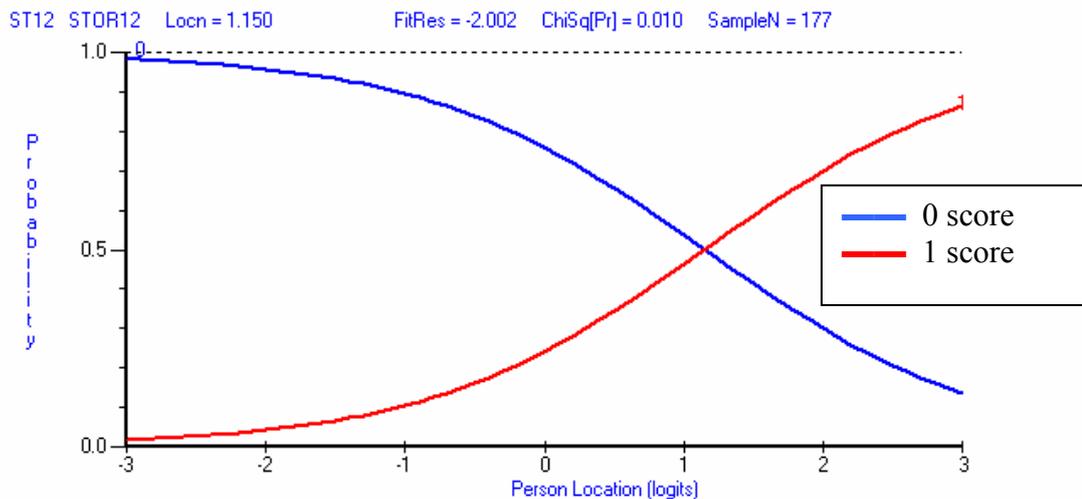


Figure 5.3: Rasch performance curves for question ST 12 (modified) using RUMM2030

Students of middle ability have an approximately 20 % chance of getting the question right. This suggests that ST12 was a difficult item on which to score well. It is important to note the Chi-squared value for this question (which was close to zero) is still low at 0.01 but is now closer to the individual item probability of 0.05.

After the four questions mentioned had been adjusted from being polytomous to single score items, all were found to function satisfactorily according to the Rasch model (Bond & Fox, 2007; Cavanagh et al., 2003). In view of the small sample size, the alternative of item deletion would have reduced the viability of the test question sample (Connolly, 2009).

Final trial test construct

Following these adjustments the statistics from RUMM2030 were as follows, item and person fit residuals (1.136 and 0.867), and the reliability index or Person Separation Index (PSI) was 0.785, all of which were good values (Cavanagh et al., 2003). The item-trait interaction measure (based on Chi-squared analysis was, however, still low (Pr= 0.00005), but a factor of 10 better than initially found. Analysis of the item fit characteristics showed that this was most likely due to the individual probability of item ST05 that 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. Item ST05 was somewhat experimental in nature in that it did have a degree of dependency on the student response to item ST04 (as mentioned earlier) so may not have been easily understood by the students.

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; Pallant, 2010). This meant that the adjusted test item structure gave a valid item set demonstrating strong unidimensionality in the items (Cavanagh et al., 2003; Connolly, 2009; Pallant, 2010), a necessary condition to validate Rasch analysis.

The item set of questions now performed satisfactorily and as mentioned above were included in the final analysis. A few questions performed very well and were excellent predictors of performance. For example, item ST03, a multiple-choice question on molarity was one of the better performing items. The item characteristic curves, Figures 5.3 and 5.4, show that the three person fit points (dots) match the shape of the fit curve indicating an excellent probability fit, demonstrating the success of the item (Cavanagh et al., 2003).

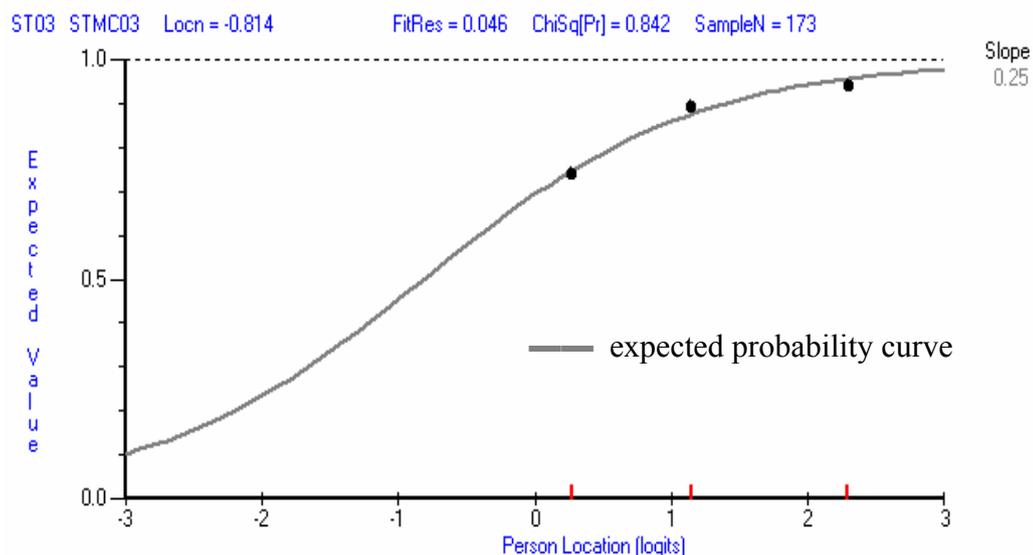


Figure 5.4: Rasch performance curve for question ST03 using RUMM2030
 (The three fit positions show closely they match the expected probability curve for the question).

The distracter analysis curves (Figure 5.5) for item ST03 also show the strength of this item. The solid red line following the curve shows the correct response pattern for the question. The three distracter curves are shown close to the x-axis.

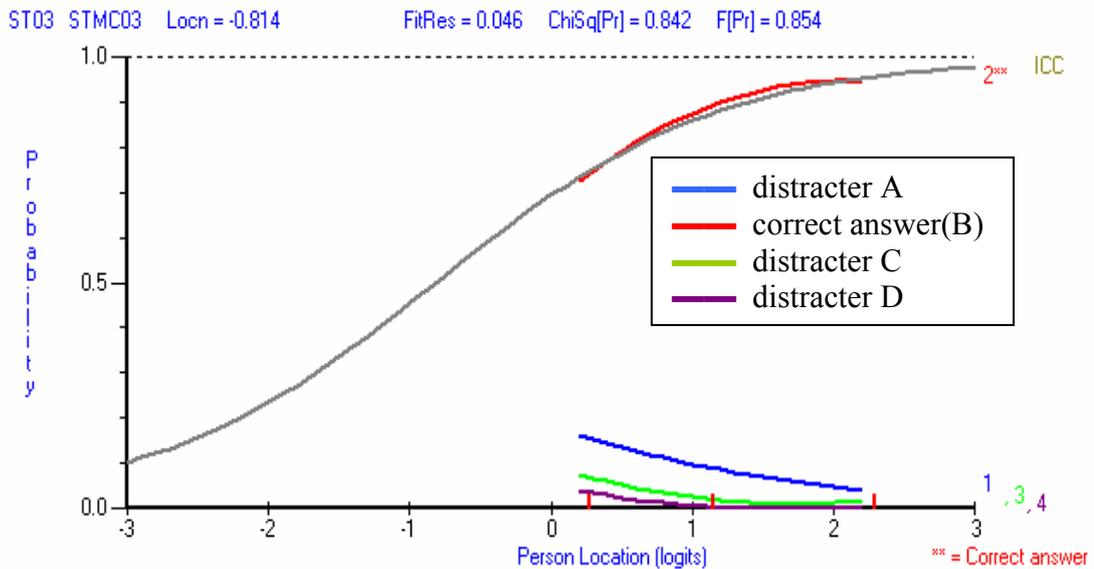


Figure 5.5: Rasch distracter curve for question ST03 using RUMM2030

The three distracters (1,3,4) have higher probability for lower person abilities whereas the correct response has higher probability correlating with higher student ability. The remaining items distracter and performance curves (see Appendix H-5) show the variety of the performances of each item. However, nearly all the items show good characteristics, which match the summary statistics previously mentioned.

Test validity

Correlation between multiple-choice and short-answer items

A significant factor in the use of the trial tests was the 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 (see page 119). As a measure of the performance on each item, the individual location measure of relative difficulty (item location) was used. This value indicates how difficult the item was, relative the other items in the tests (RUMM Laboratory, 2009b).

The mean values for location (see Table 5.1) 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.

When these ordered pairs (e.g. AB01:AB07) were plotted as a scatter plot, the relationship between the paired questions is evident (Figure 5.6). 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 good correlation between the variables.

Table 5.1 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

The r -squared value of 0.52 indicates that 52% of the variability of the short-answer responses could be explained by variation in the multiple-choice responses. (The

choice of using the multiple-choice values as the independent variable was an arbitrary one and the correlation values are the same if the short-answer values are used as the independent variable.) This correlation result is important as it demonstrates that the initial intention of writing questions that were able to test the students' ability on each type of presentation for the same content was essentially achieved.

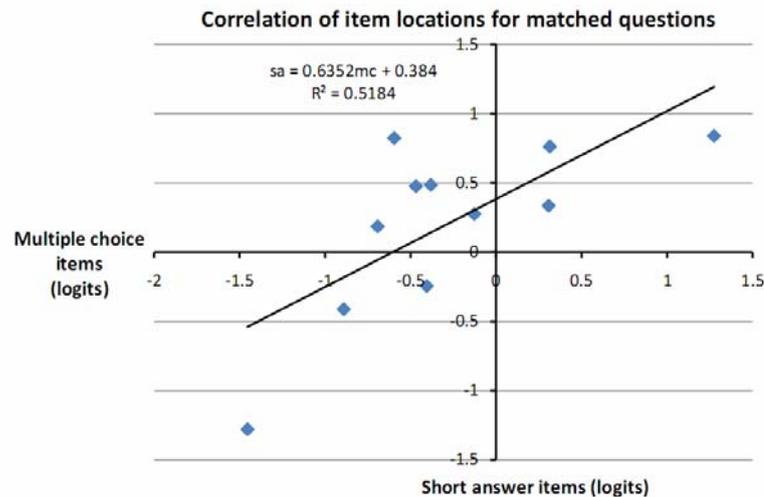


Figure 5.6 Correlation between matched item pairs of sample test items

Generally the order of difficulty of the two sets of items short-answer versions and multiple-choice versions were similar with the correlation value suggesting some dependency between the multiple-choice items and the short-answer items. This result appears to contradict the earlier results from the Rasch modelling which suggested a unidimensionality of the test items (page 127). This correlation result above simply shows that the items were matched as intended but each item still behaves independently, that is, performance on one item did not determine performance on another. In other words, the ranked difficulty of the multiple-choice items matched the rank order of difficulty of the short-answer versions of the items. The ANOVA results for this data $F(1,20), d.f.(2) = 2.82$ with $p > 0.05$ show that the difference between the means of the two sets (multiple-choice and short-answer) was not significant confirming the earlier mentioned analysis (see Appendix H-4).

The strength of this correlation adds credibility to the relationships that are discussed in the next section because, on a basic level, the tests have achieved what they were intended to do, that is compare performance on matched items. This initial analysis

helps to clarify Research Question 1: Do students perform more effectively on multiple-choice or short answer questions? The data shown above would appear to support this notion. Another aspect of the relationship between the multiple-choice test questions and the matching short-answer questions was the difference in the performance. It is evident from the scatter plot and the associated trend line that the multiple-choice questions proved slightly easier than the equivalent short-answer questions.

Targeting

The final measure of the effectiveness of the trial tests is an examination of the targeting of the items against the student ability. The item map (see Figure 5.7) shows how the item difficulties matched the student abilities on a common scale. 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.

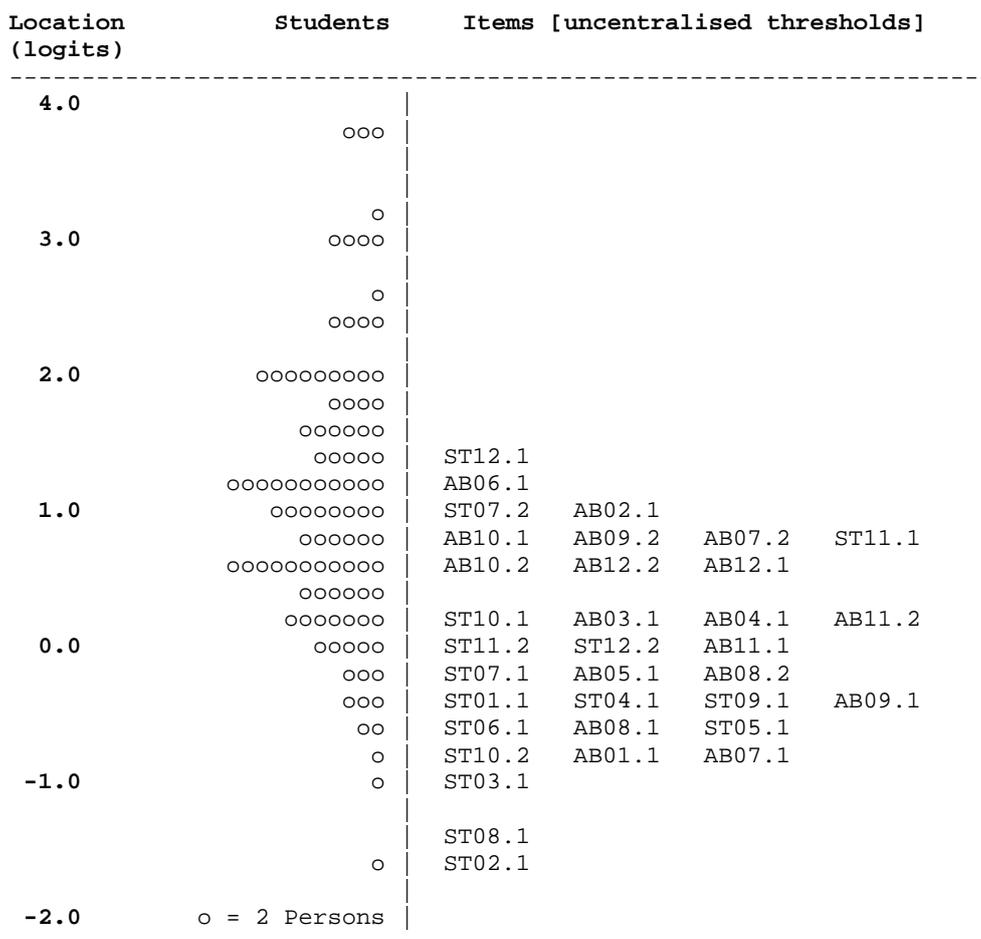


Figure 5.7: Item map showing distribution of items and students

As mentioned earlier in the methodology chapter, 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 did occur but the item distributions and person distributions both conformed to a normal distribution.

The test items show some weakness in properly assessing students of high ability. However, as mentioned earlier, this was expected, but the tests do match student ability well at the lower end of the ability range.

Overall it appears that the trial tests have (with some modification) produced a valid measure of student ability. The use of the Rasch analysis gives credibility to the internal validity of the test items (Andrich, 2005; Cavanagh et al., 2003).

Results from the RUMM2030 Analysis

The RUMM2030 program allowed detailed analysis of the tests and provided much useful information to assist in seeking answers to the research questions. In particular the analysis of the trial tests allowed comparison to the VCE examinations and also allowed a differential gender analysis of the items. The categories explored with RUMM2030 were:

1. multiple-choice or short-answer
2. recall (acid-base) and application (stoichiometry)

Response to Research Question 1: Do students perform more effectively on multiple-choice or short answer questions?

The RUMM2030 analysis provided information to allow a comparison of student performance on an overall basis (all students). This also allowed comparison with the findings of the VCE examinations. Whilst it is not expected that the measures (for example means) will be the same due to the limited nature of the trial test and the nature of the trial population (middle to upper class students), the patterns established in the VCE examination analysis should match to that of the trial test analysis. If the characteristics were substantially different to the whole population VCE examinations then the results obtained from the trial tests would have limited value in terms of transferability of the findings.

As an initial point of comparison the means and standard deviations of the two sets of tests were compared and are shown in Table 5.2

Table 5.2: Summary measures of the data from the trial tests

Measure	VCE examinations	Trial tests
Mean	60.3	73.6
Standard deviation	15.7	17.2

The trial test mean is substantially greater than that of the VCE examinations. Briefly there are two reasons that could account for this.

1. The VCE examination scores are standardised to match a predetermined allocation of grades so the mean percentage scores vary little from one year to the next. This is to normalise the results from one year to the next.
2. The student sample (trial tests), as previously mentioned, is not representative of the entire student population that normally takes part in the VCE examinations. To emphasise this point, the median study scores of the for schools involved were 31, 33, 34 and 36 compared to the median VCE study score for all schools and students which is 30 (VCAA, 2009b). This places the schools participating in the trial tests in the upper 50% of all Victorian schools. In simple terms, the participating schools have highly able students and typically feature well in the awarding of the top VCE grades. Consequently, it is a reasonable presumption that students' participating trial tests from these schools will likely score well and feature strongly in the awarding of the top grades in the subsequent Year 12 chemistry, should those students choose to take chemistry. Whilst this cannot be guaranteed, the researcher's extensive experience has shown that students who have performed well in Year 11 chemistry tend to perform well in Year 12 chemistry. This situation was appreciated before the study began, but, as previously stated one aspect of this study was to examine the discrepancy in the A⁺ and A grades that were awarded to male students and female students and by choosing students from high achieving schools would allow a better

opportunity to study this group as these are the students/schools most likely to achieve high grades in a VCE examination.

A point of comparison in the data that was encouraging was that the standard deviation value for both sets of tests (VCE examinations and trial tests) represent about 25% of the mean value, thus, it can be concluded that the spread of the results from the trial tests corresponds to the spread in the VCE examinations. This can also be concluded by comparing the distribution of marks by gender for the two sets of examinations (Figures 4.6 and 5.18).

Comparison of all Multiple-choice with all Short-answer questions

RUMM2030 analysis allowed the comparison of the trial 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).

As presented in Table 5.3, 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 (refer to Appendix H-6 and H-7 for details). This difference between the means is small.

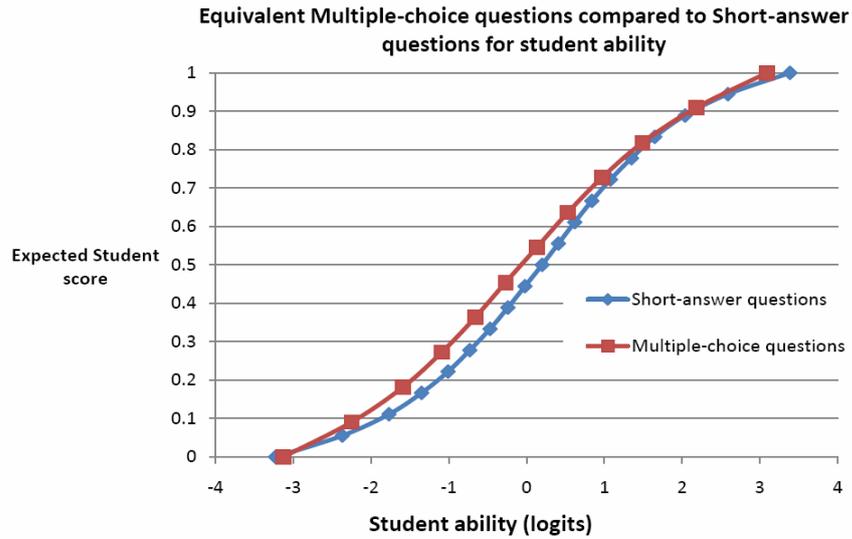


Figure 5.8: Multiple-choice compared to Short-answer response difference against expected score and student ability.

The RUMM2030 graphical analysis (Figures 5.8 and 5.9) shows several interesting aspects of students' abilities with respect to answering multiple-choice and short-answer questions.

Table 5.3: ANOVA analysis of question performance by question type and classification Chemistry trial tests

ANOVA results (N = 368, df = 1)							
Comparison variables	Mean	Standard deviation.	Discriminating variable	Sum of Squares	F	Sig (p)	Similarity to VCE analysis
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
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 VCE examinations
Recall Application	66.4 74.8	22.8 25.3	Multiple-choice and Short-answer questions	6450.3	11.01	0.001	Result was the reverse in the VCE examination
Recall Application	62.3 81.5	28.8 24.5	Multiple-choice Questions	33925	47.4	0.000	Result was the reverse in the VCE examination
Recall Application	68.6 69.9	26.0 30.7	Short-answer questions	162.0	0.20	0.65	Very similar result to VCE examination

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.

The t-test graph (Figure 5.9) from the RUMM2030 analysis shows that only a small portion of the students were in the 5% and 1% margins confirming that the results showed no significant differences between the multiple-choice scores and the short-answer scores.

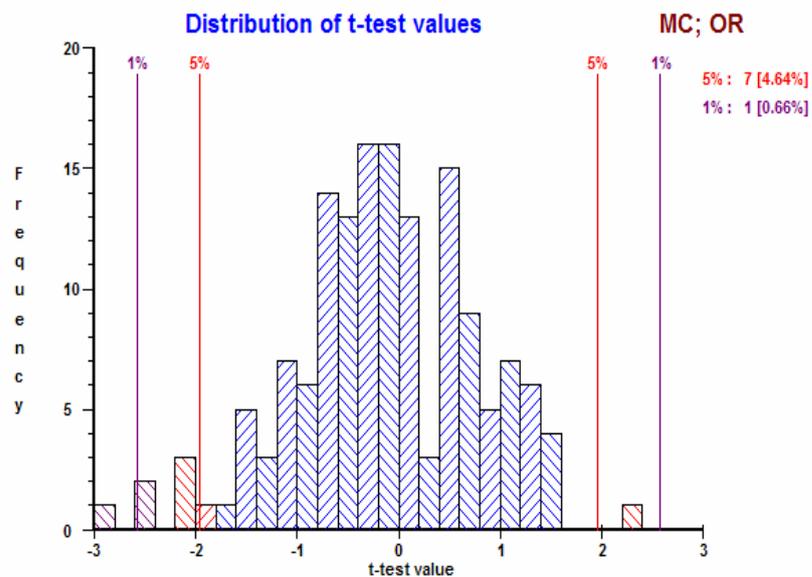


Figure 5.9: Multiple-choice (MC) compared to Short-answer (OR) responses t-test.

This result can be compared to that obtained in chapter 4 shows some similarity. The analysis of the VCE examinations (see Table 4.11) also showed that students performed better on multiple-choice questions than on short-answer, however, in that analysis the difference was significant. As with the VCE examination analysis, a closer interpretation of the differences in performance was performed.

Comparison of performance on multiple-choice and short-answer application type questions

As presented in Table 5.3, students' performance on application (stoichiometry) short-answer questions (mean/standard deviation = 70.0/30.7) was weaker than on application multiple-choice questions (81.5/24.5). The ANOVA results ($F(1,366) = 15.96$; $p < 0.01$) show that this is a statistically significant difference in performance on application questions.

The RUMM2030 analysis demonstrated this difference graphically as shown in Figure 5.10. As student ability increases, the stoichiometry short-answer questions become harder compared to the multiple-choice questions. This result supports the ANOVA findings that students find the multiple-choice stoichiometry easier than the short-answer versions of these questions and as student ability increases the difference becomes greater. This may reflect the more academic students being able to use the options to check their responses whereas less academic students may be unable to do this and have to resort to guessing and less able students are more likely to leave short-answer questions increasing the apparent difficulty of the item.

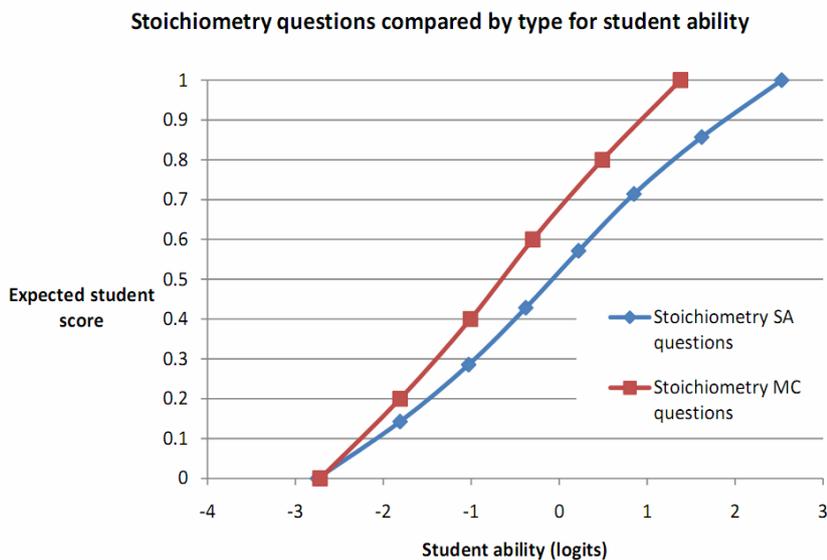


Figure 5.10: Multiple-choice (Stoichiometry) compared to Short-answer (Stoichiometry) response difference against expected score and student ability.

This result can be compared to that obtained in chapter 4 shows differences to the analysis of the VCE examination analysis (Table 4.11). In the VCE analysis there was almost no difference in performance on application multiple-choice compared to

application short-answer. The difference that has occurred in the trial sample may be due to the specific nature of the application questions chosen for the trial tests, that is, only stoichiometry questions were used to reflect application ability in the trial tests. The use of other less mathematical application questions in the VCE examinations, such as instrumental analysis, may have caused the difference between the trial test results and the VCE examination results.

Comparison of performance on multiple-choice and short-answer recall type questions

As presented in Table 5.3, students' performance on recall (Acid-Base) multiple-choice questions (62.3/28.8) was weaker than on recall short-answer questions (68.6/25.9). The ANOVA results show a statistically significant difference ($F(1,366) = 4.86; p < 0.05$) in performance on recall questions.

The RUMM2030 analysis demonstrated a somewhat different result when student ability was allowed for in the analysis (see Figure 5.11). At lower levels of student ability, there is little difference in student performance. However, as ability increases students find the multiple-choice questions slightly more difficult. Whilst the difference in raw score performance is quite marked, the difference once student ability is allowed for is much less, suggesting little real difference in performance on the two types of items.

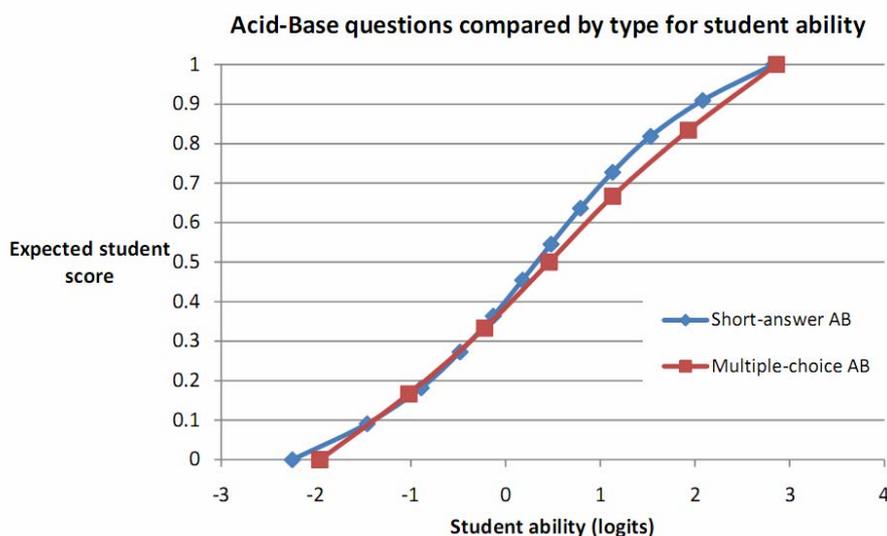


Figure 5.11: Multiple-choice (Acid-Base) compared to Short-answer (Acid-Base) response difference against expected score and student ability.

This result can be compared to that obtained in chapter 4 shows that the trial tests performed differently to the analysis of the VCE examinations. The analysis of the VCE examinations (see Table 4.11) showed that students performed better on multiple-choice questions than on short-answer.

Overall there are similarities in the students' performance on the multiple-choice questions being better than on short-answer questions which matches the result on the VCE examination analysis. Whilst the differences on the finer analysis, when comparing multiple-choice to short-answer on specific topic areas (application and recall), were different than in the VCE analysis, the overall comparison between the two sets of data appears to have validity.

Response to Research Question 2: Do students perform more effectively on recall type questions or on application questions?

As with the data addressing Research Question 1, analysis using the RUMM2030 program (RUMM laboratory P/L, 2009) was performed on the data relevant to this question. The recall questions were framed around the topic of Acid-Base chemistry and the application questions around stoichiometry.

Comparison of performance on recall and application, all questions

As presented in Table 5.3, 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 VCE 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 referred to in the previous section where the response to multiple-choice stoichiometry questions was unexpectedly high. Figure 5.12 shows the RUMM2030 analysis for the two question categories with adjustment for student abilities. Again 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.

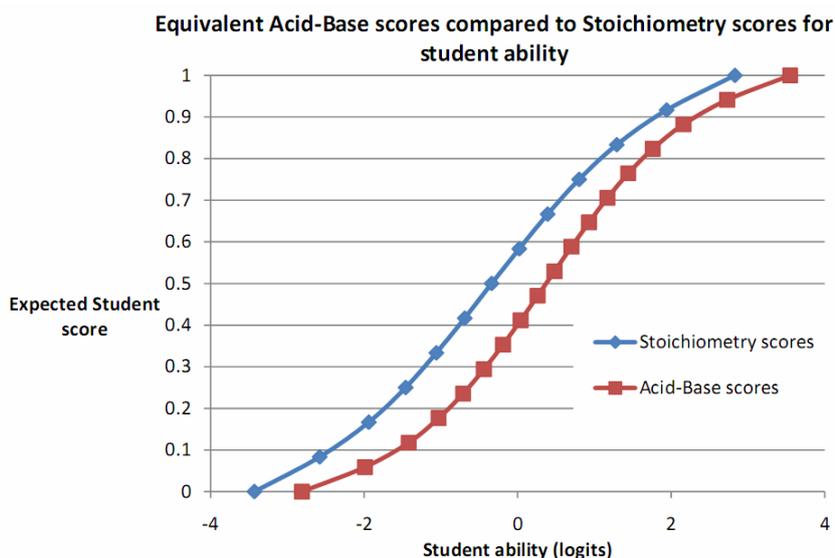


Figure 5.12: Recall (Acid-Base) compared to Application (Stoichiometry) response difference against expected score and student ability.

The difference on performance measured by the t-tests (Figure 5.13) was also shown to be significant with some 12.12% of results in the 5% limits portion of the distribution again showing the significance of this outcome (Pallant, 2010; RUMM Laboratory, 2009b).

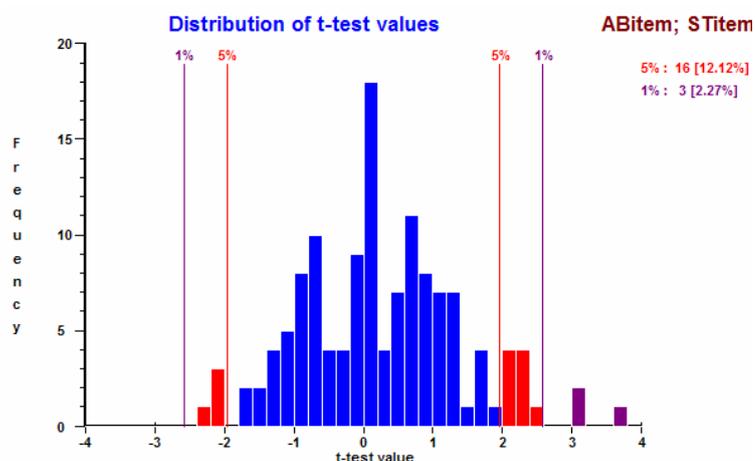


Figure 5.13: Acid-Base questions compared to Stoichiometry questions t-test distribution from RUMM2030.

The outcomes of this comparison were surprising when compared to the VCAA findings and introduce a point of discordance between the two sets of results. It would appear that the application questions may have been somewhat easier than anticipated and/or the recall questions harder than expected. However, the topic

(Acid-Base chemistry) chosen for the recall questions was one of the more abstract studied in Year 11 and this may go some way towards explaining the difference. The topic was chosen for its suitability within the existing teaching program at the schools taking part in the trial tests.

The differences in performance were studied in more detail by reducing the data (as was done for the multiple-choice to short-answer analysis) into more specific subtests.

Comparison of performance on recall and application multiple-choice questions

As presented in Table 5.3 students' performance on multiple-choice recall questions (62.3/28.8) was substantially weaker than on multiple-choice application questions (81.5 /24.5). The ANOVA results show a statistically significant difference $F(1,366) = 47.38$; $p < 0.001$ in performance on multiple-choice questions whether they are recall or application. As mentioned previously, the apparent ease of the multiple-choice application (stoichiometry) questions is very evident in these results.

When analysed with RUMM2030 to allow for student ability, the difference in performance is still obvious. The graph (Figure 5. 14) shows that students of similar ability found the stoichiometry multiple-choice easier than the acid-base multiple-choice.

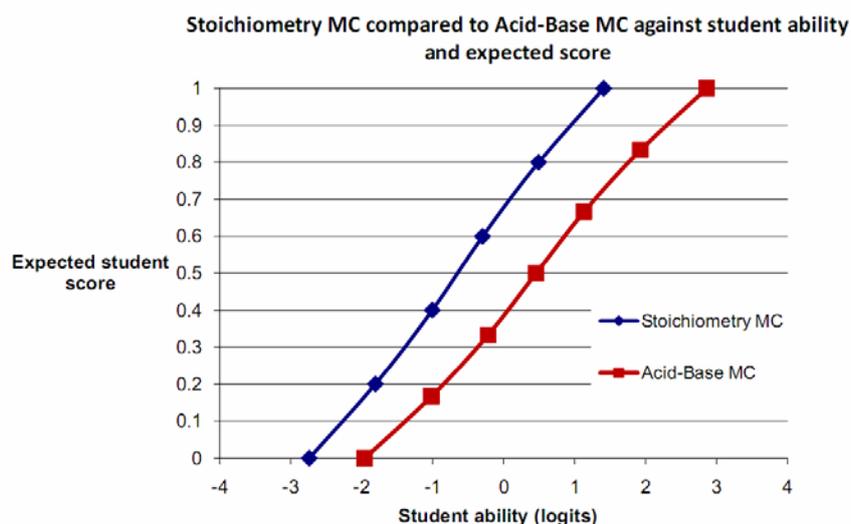


Figure 5.14: Multiple-choice (Stoichiometry) compared to Multiple-choice (Acid-Base) response difference against expected score and student ability.

However, when t-test analysis, using RUMM2030, was performed the difference did not prove to be significant with only 4.2 % of students in the 5% limit range (Figure 5.15). This outcome shows that whilst the raw score analysis shows the performance on stoichiometry multiple-choice was very good, the t-test analysis suggests that this difference is not as great as it at first appears.

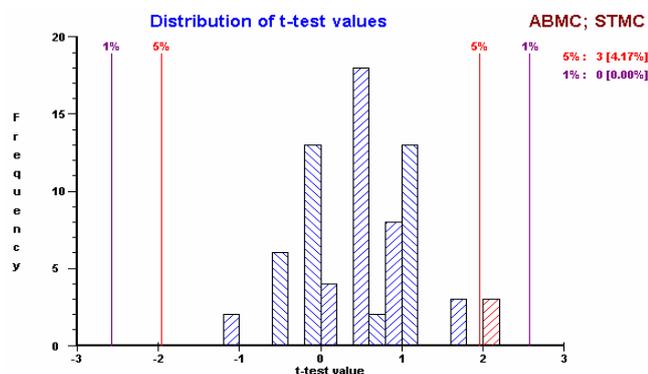


Figure 5.15: Acid-Base multiple-choice questions compared to Stoichiometry multiple-choice questions t-test distribution from RUMM2030.

Comparison of performance on recall and application short-answer questions

As presented in Table 5.3 students' performance on short-answer recall questions (68.6/26.0) was nearly the same as on the short-answer application questions (69.9/30.70). The ANOVA results show a statistically non-significant difference ($F(1,366) = 0.20; p > 0.05$) in performance on multiple-choice questions whether they are recall or application.

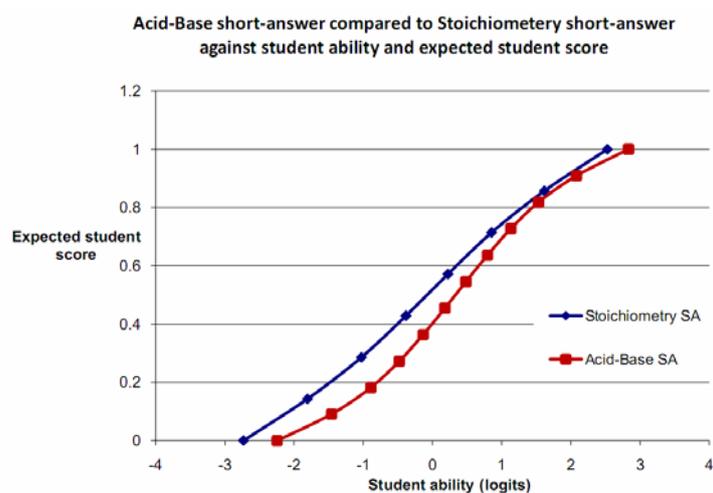


Figure 5.16: Short-answer (Stoichiometry) compared to Short-answer (Acid-Base) response difference against expected score and student ability.

When the comparison using RUMM2030 is made, the difference is somewhat more obvious with the stoichiometry short-answer questions being visibly easier than the acid-base short-answer questions (Figure 5.16). However, the t- test distribution again showed that the difference was not significant when the comparison allowed for student ability (Figure 5.17).

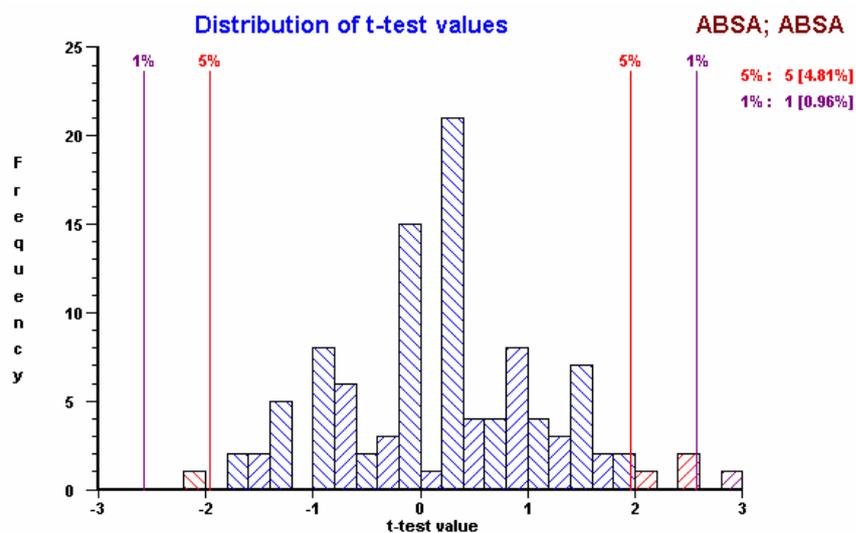


Figure 5.17: Acid-Base short-answer questions compared to Stoichiometry short-answer questions t-test distribution from RUMM2030.

Overall it would appear from this analysis that the multiple-choice stoichiometry results have had a substantial impact on the apparent performance of the students in the trial tests. The unusually high mean scores for these questions was above 80 and as all other categories had means around 60 to 70 this result does stand out. When reaching final conclusions and inferences care had to be taken to consider the impact of this result.

The results from the trial test analysis are summarised in Table 5.3. The final column in the table assesses how similar the results from the trial tests were to that of the VCE examination study, the results of which are shown in Table 4.11.

Response to Research Question 3: Does students' gender influence performance in chemistry examinations (or tests)?

In chapter 4, this question was addressed on the basis of the VCE chemistry examinations. That analysis showed that males performed significantly better in the

examinations than did females, particularly at the A⁺ end of the score range. An initial analysis of the student performance in the trial tests showed that the males again outperformed the females on the trial tests (see Appendices I-8 and I-10).

Table 5.4: Gender differences on the trial chemistry tests (means)

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>s.d.</i>
Male %	94	7364.2	78.3	15.8
Female %	90	6174.4	68.6	17.3

Table 5.5: Gender differences on the trial chemistry tests (ANOVA)

ANOVA					
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>
Between Groups	4360.3	1	4360.3	15.9	0.0001
Within Groups	49813.4	182	273.7		

The ANOVA results of the raw scores show that the males scored more highly (78.3%) than the females (68.9%) and that difference was significant ($F(1,181) = 15.9$; $p < 0.01$). The distribution of scores in the trial tests (Figure 5.18) also matches that of the VCE examination distributions (Figure 4.13).

Whilst the distribution is less uniform than the VCE distribution (the smaller sample size in the trial tests needs would 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 90 to 100% score range. 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 was relatively small compared to that of the VCE examinations.

The trial tests, however, allowed a finer examination of student performance on the various category types of question asked that was not possible on the VCE examination and was able to shed some light on where the males were outperforming the female students. The following analysis attempts to identify where, within the test structure, the males performed differently to the females.

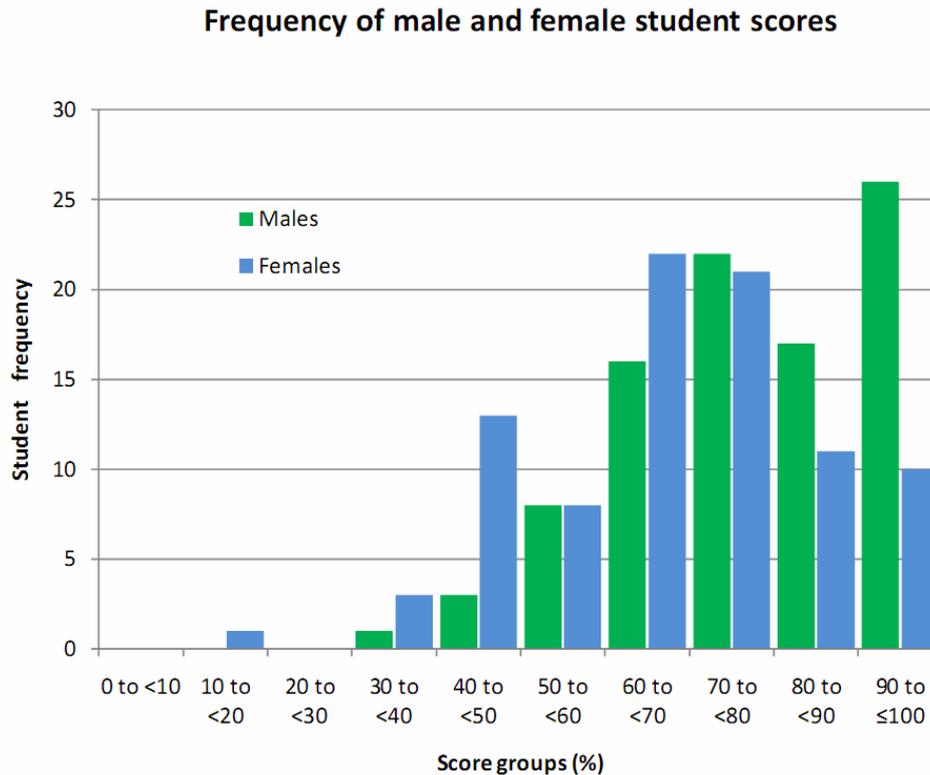


Figure 5.18: Distribution of male and female scores in the trial tests

Comparison of Multiple-choice and Short-answer questions by gender

The final trial test version for analysis in RUMM2030 was subdivided into two subtests, one containing all the multiple-choice questions and one containing the short-answer questions. The two subtests were then analysed for gender difference using RUMM2030 and ANOVA (RUMM Laboratory P/L, 2009).

Multiple-choice and gender

The ANOVA test results show that there is a significant statistical significance ($F(1,182) = 13.65; p < 0.001$) in performance on multiple-choice questions between the males and females. Male 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 (refer to Appendix H-6 and H-7 for details). The RUMM2030 graphical analysis difference shows, however, that the difference is less marked when the scores are adjusted for latent student ability as measured by RUMM2030.

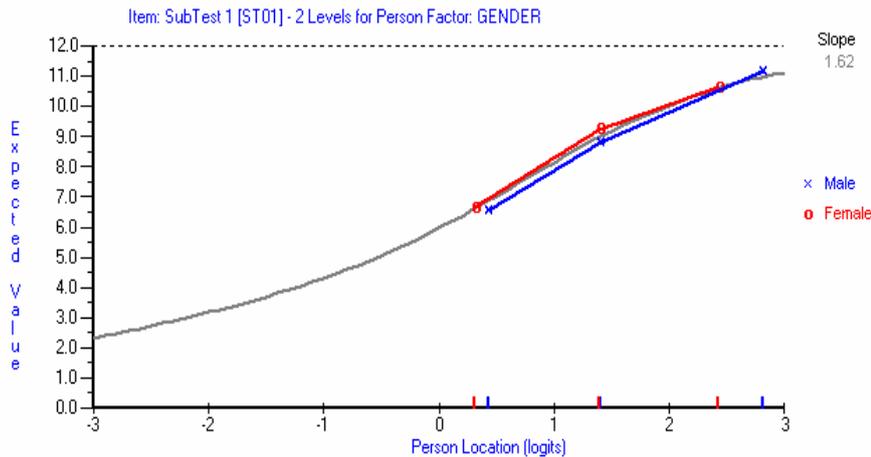


Figure 5.19: Multiple-choice questions showing gender difference against expected score and student ability.

The graph (Figure 5.19) shows that the male students, within their ability ranges, find the multiple-choice questions slightly more difficult than do the female students. In other words, in spite of the higher raw scores obtained by the males they find them somewhat more challenging than expected. Put simply, the female students found multiple-choice questions easier than did the male students for students of equal ability.

A significant inference is that it would appear that there may be more high ability male students than female students in the trial test sample which could account for the better performance of the males at the upper end of the performance scales in the VCE examinations (see Figures 4.5 and 4.7).

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 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 was 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 (63.5/23.2) (refer to Appendix H-6 and H-7 for details). The RUMM2030 graphical analysis shows that the difference is less marked when the scores are adjusted for latent student ability as measured by RUMM2030.

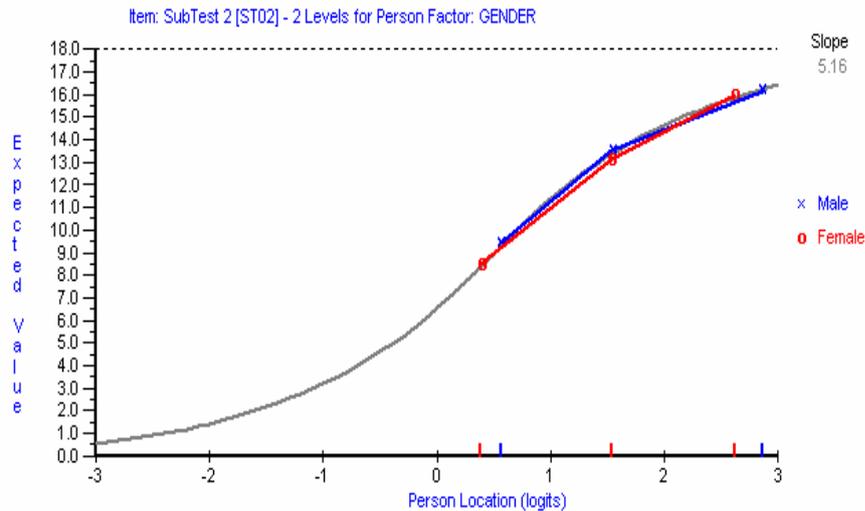


Figure 5.20: Short-answer questions: showing gender difference against expected score and student ability.

The RUMM2030 graphical analysis shows a different result than with the multiple-choice subtest. The graph (Figure 5.20) shows that the male students within their ability ranges find the short-answer questions slightly easier than do the 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.

However, in both instances the males outscored the females on both tests to a statistically significant degree. An interesting observation in Figure 5.20 is that the female students found the short-answer questions easier than did the male students at the very top end of the ability scale. This outcome could be explained as being expected in terms of the reportedly greater ability of female students in language expression and recall questions (Beller & Gafni, 2000). The very best female students may be more capable of expressing answers coherently than do the male students of equally high ability. If this were the case, it would provide some explanation for the differences in distribution between the top students in the Unit 3 VCE examination compared to the Unit 4 examination (see Figures 4.6 and 4.8).

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

The initial anecdotal observations from the researcher's teaching career suggested that the female students tended to find recall type questions easier than application questions. By creating subtests within RUMM2030 to separate the recall (Acid-Base) questions into one subtest and the application (Stoichiometry) questions into a second subtest allowed this idea to be tested.

Recall questions and gender

The ANOVA test results 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 was 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) (refer to Appendix H-6 and H-7 for details). This difference is relatively smaller and less significant than in the two previous analyses. This tends to show that the females were more able (or the boys less so) with these types of questions. Considering that the difference in the overall ability was 10 (mean difference on all items), the difference of 8 between the means here suggests that males do not have the same advantage when questions are recall based (see Table 5.2).

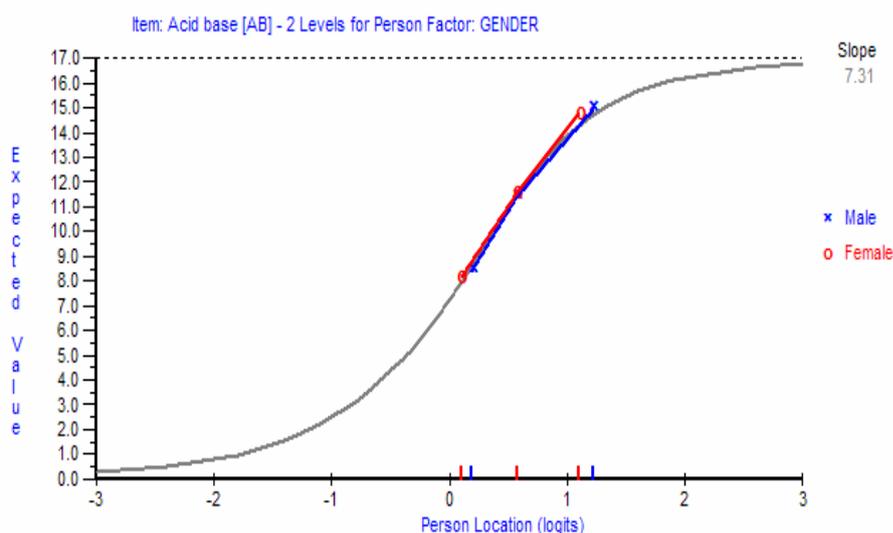


Figure 5.21: Recall (Acid-Base) questions: showing gender difference against expected score and student ability.

The RUMM2030 graphical analysis (Figure 5.21) however, 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 found the difficulty about the same.

Application questions and gender

The ANOVA test results 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 was 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 (67.2/27.1) (refer to Appendix H-6 and H-7 for details). This is the largest difference between the means of the different question classifications and is notable that the standard deviation of the female student scores was 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 some way to explaining the difference in performance between the female students in the Unit 3 examination and the Unit 4 examination (see Figures 4.5 and 4.7).

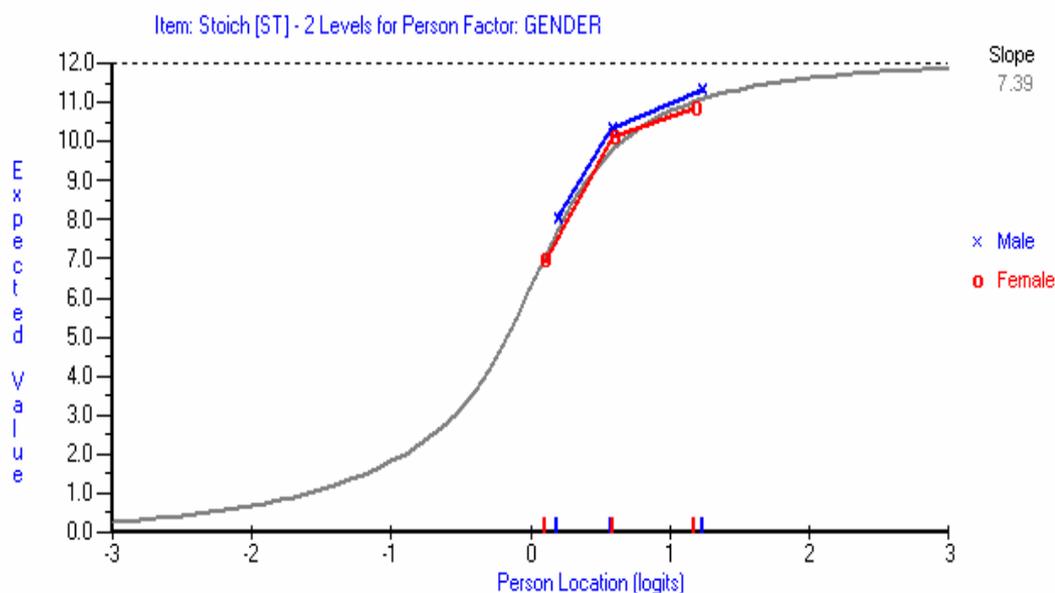


Figure 5.22: Application (Stoichiometry) questions: showing gender difference against expected score and student ability.

The RUMM2030 graphical analysis (Figure 5.22) shows that there is a notable 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. 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 5.22 is particularly important in terms of the relationship to the VCE examination results and may help explain the difference in performance of the male students and female students when comparing the Unit 3 examinations to the Unit 4 examinations. Whilst the male students outperformed the female students in both examinations, the differences were less pronounced than in Unit 4 (see Figures 4.5 and 4.7). As mentioned previously the content structure of Unit 3 was heavily loaded with stoichiometric calculations whereas Unit 4 had a much higher proportion of descriptive chemistry.

Given the observations demonstrated in Figure 5.22 it is a reasonable conclusion that the 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 Unit 3 and Unit 4 examinations. The difference is further emphasised when considering the effect for the recent examinations after the course content balanced changed in 2008. Inspection of the grade distribution report (VCAA, 2009a) demonstrates that the differences between the grade distributions have evened out (see Appendix H-8). Whereas the distribution for the years of this study (2003-2007) showed that males substantially outperformed females in the Unit 3 examination, the difference in Unit 4 was much less. The most recent examination, for which a complete year (2008) is available, shows that the male students are still outperforming the female students but the difference between Unit 3 and Unit 4 has been reduced in line with the more even distribution of content type in the Year 12 chemistry course. These data, whilst only based on one year's results, further support the observation that the female students perform more effectively on questions that are recall or content based and less effectively on stoichiometric application questions.

Whilst the RUMM2030 analysis has satisfied the internal validity of the trial tests, to allow an evaluation of the transferability of these gender analysis results to the wider

chemistry student population, the characteristics of the trial test needed to be compared to the characteristics of the previously analysed VCE examinations (Chapter 4). How well the results of the trial test findings match those of the analysis of the VCE examination analysis is important to assist in ascertaining the degree of transferability.

The trial test analysis in spite of some discordance with the VCE examination analysis was supportive of the VCE examination analysis and the trial test analysis enabled some possible explanations of the trends observed in both that was not possible to determine in the VCE analysis alone.

Response to Research Question 4: Do students have a preference for the type of question style in terms of: a) Multiple-choice or short-answer in general terms? and b) With respect to whether the question is assessing recall or application?

Results from Interviews

Conducting the interviews provided a first person opportunity to determine whether or not the students' preferences on question types matched the actual performance on the tests. To gain a larger number of results a number of students were interviewed by their teachers who recorded their responses either by using pen and paper notes or using an audio recording and transcribing the observations afterwards. All responses are transcribed in Appendix C-3.

The interview schedule

1. a. What type of questions have you experienced in your chemistry tests this year, multiple-choice, short-answer or both?
b. If yes have the test been all multiple-choice all short-answer or a mixture of both in the one test?
2. If you have responded to both what proportion of the test is multiple-choice and what proportion is short-answer? E.g. 50% multiple-choice 30% multiple-choice etc
3. Does the content of the question influence your decision?
 - a. For example if the question tests recall (e.g. name a strong acid) then is it better for the question to be multiple-choice or short-answer? Please explain your response.

- b. If the question is an application question (e.g. Calculate the number of mole of something) is it better for the question to be multiple-choice or short-answer? Please explain your response.
4. Do multiple-choice have any advantages compared to short-answer questions? Please explain your response.
 5. Do multiple-choice questions have any disadvantages compared to short-answer questions? Please explain your response.
 6. Do short-answer have any advantages compared to multiple-choice questions? Please explain your response.
 7. Do short-answer questions have any disadvantages compared to multiple-choice questions? Please explain your response.
 8. If you had the choice would you prefer chemistry tests to be multiple-choice only or short-answer only or a combination of both? Please explain your response.

Coding responses.

To aid in this interpretation, some student responses were coded according to whether they were positive or negative with respect to each question. This method enabled an assessment of the general trends of each gender and their preferences for particular types of questions. Whilst students responded either briefly or with some elaboration it was possible to deduce that they either preferred one type of question of the other. Students who generally favoured multiple-choice questions were coded 1 and students who preferred short-answer were coded 2. This process was applied to question 3. The coding is shown on the questions in Appendix C. The results of coding the responses of the 59 interviewed students are shown in Table 5.6.

Table 5.6: Responses to Research Question 4 (n = 59)

Question preference combination	Males	Females	Totals
Recall: MC and Application: MC	6	7	13
Recall: MC and Application: SA	19	22	41
Recall: SA and Application: MC	1	0	1
Recall: SA and Application: SA	1	3	4
Totals	27	29	59

Of the students involved, 70% preferred recall questions to be multiple-choice and the application questions to be short-answer. There were no discernable differences between the responses of the males and the females. Further description and analysis of the data for Question 3 follows later in this chapter. The evenness of the response shown in Table 5.6 was a little surprising, as it might have been expected that there would have been some gender differences evident in the responses. The important observation from this limited sample is that the number of students who preferred to answer recall questions as multiple-choice questions was very strong. Only 1 student out of the 56 showed a preference for recall short-answer questions. The number wanting only recall questions was also very small, 3 out of the 56. This compares with the number of students who preferred multiple-choice questions only, which was 11 out of 56, nearly three times the number. Responses to Question 8 provided little useful additional information with almost all students indicating a preference for the status quo of a combination of multiple-choice and short-answer. This response possibly reflects the experiences they have in chemistry assessment rather than the intended purpose of the question. The actual intention of Question 8 was to determine the students' preference for one type of question over the other if such a choice had to be taken.

To gain a better understanding of students' views regarding question type, the participating teachers were asked to put the following question to their classes: “*If a test was to be either all multiple-choice or all short-answer which would you generally prefer to do regardless of the test topic? That is which type of question do you like best?*” The entire class was asked this question and the teachers forwarded the simple poll results to the researcher. The outcome of this poll, shown in Table 5.7, included virtually all the students who took part in the trial tests depending on the individual attendance on that day. The raw data from each school is shown in Appendix C-4.

Table 5.7: Preferences of Students for Question Type (n=100)

Group	Percentage who favoured multiple-choice	Percentage who favoured short-answer
Males	66	34
Females	54	46

Chi-squared analysis of this data showed a value of 2.23, which is not statistically significant (see Appendix C-4). (The main reason for the sample not producing a significant result was most likely the sample size. Had the same trend continued in a sample just 50% larger the results would have been statistically significant). Whilst there are obvious differences between the data for males and females the difference is not statistically significant, the results do, however, support results found through the observations of other researchers (Beller & Gafni, 1991, 2000), though not strongly.

These results (Table 5.7) confirm the results demonstrated in Table 5.6. A strong preference was demonstrated for multiple-choice questions with males being the stronger supporters of the multiple-choice format.

Student responses by interview question

All the student responses to the interview questions are included in Appendix C-3. The first two questions aimed to clarify the experiences of the students with respect to the styles of questions they had experienced. The students' responses were quite uniform. All students indicated that they had commonly experienced tests that contained both multiple-choice and short-answer questions. The proportion of multiple-choice to short-answer questions in the tests experienced by the students was reported at between 30-50% multiple-choice. The most common response indicated a roughly one-third multiple-choice proportion in the tests. This was to be expected as all the students were taught by experienced teachers who were more than aware of the need to adequately prepare their students for the examination regime they would encounter in the following year.

Question 3 elicited a wide variety of responses. Students were asked if the type of question in terms of content influenced their choice of question style. The majority of students, around 90%, interviewed preferred content or recall questions to be multiple-choice. The most common reason given seemed to be based on the idea that multiple-choice gave the students an opportunity to be prompted by the options. A number of students indicated the possibility of eliminating incorrect options as an advantage with this type of question. Students who preferred short-answer questions for recall type questions did so mainly from the perspective of the confusion that sometimes occurred if the answer they thought was going to be the answer was not in

the options. Typical responses representative of these views were: Female student B8: School-A: “*multiple-choice as you have options to give you an idea but in short-answer you don’t.*” Male student B18: School-D “*Multiple-choice – if the question cannot be instantly answered, looking at available answers may remind you of the correct choice.*”

Question 3b (If the question is an application question (e.g. Calculate the number of moles of something) is it better for the question to be multiple-choice or short-answer? Please explain your response) focussed on application or calculation type questions. The views of the students were considerably different to the attitude towards multiple-choice questions. The students were generally of the view that short-answer responses gave them the best opportunity to perform well. There were, however, a number of students who preferred multiple-choice for both types of question. The percentage of students who opted for the short-answer application questions was about 70%. Typically the students indicated that the short-answer approach allowed them the opportunity to gain marks for correct working even if they could not provide the final correct answer. Typical examples of these responses were: Female student B4: School-A: “*short-answer, because I can get marks for showing working if I were to get the answer wrong*”-There were a number of responses similar to this, Male student A9: School-D “*short-answer – writing down my process of working out helps in checking my answers – less chance of making mistakes*”.

A view held by the small number of students who preferred multiple-choice for application questions focussed on the advantage of having the possible answer presented in the options and that having worked out an answer that was not an option meant that the students knew they had made an error and could then have a second chance at trying to calculate an answer that was presented in the options. A typical response was: Male student B2: School-C “*multiple-choice because that way when you work out the answer, it has to be one of the answers given, so that way you can work out if you have made a mistake or not.*” Of some significance was the observation that of the 59 interviewed students only one responded with short-answer for recall questions coupled with multiple-choice application questions.

Questions 5, 6, 7 and 8 (see Appendix C) from the interview schedule appear to ask the same questions. This was apparent on one occasion where a student exclaimed

that I was trying to trick him by asking the same question again. This strategy was employed with the knowledge that some students would respond in a nearly identical manner to questions 5 and 8 and, 6 and 7. However, the method was designed to ensure that students did have the opportunity to consider each type of question in the focus of their thoughts from both a positive and negative viewpoint. This tactic appears to have paid dividends with students giving new information in subsequent questions that was not in the previous questions. For example, whilst question 5 seeks advantages of multiple-choice over short-answer, question 8 seeks the opposite, that is, the disadvantages of short-answer over multiple-choice. (Questions 6 and 7 reinforce the students' ability to express their ideas about the two types of questions). By asking the questions, e.g. question 5 and question 8, about the same idea but from the reverse logic, that is one from a positive and the other a negative mindset, the questions encouraged the students to consider each perspective before answering. Nevertheless, this was not always the case with a number of students giving responses in the style of *"I can't add to what I said in the last question"*. The responses from questions 5 to 8 are summarised in Tables 5.8 and 5.9.

Table 5.8: Advantages and Disadvantages of Multiple-choice questions

Advantages	Disadvantages
Possible to eliminate or narrow down response.	Can be confusing or tricky with good options
Can work backwards from the answers	Can appear to have more than one correct answer
Prompting of answer from given options.	Penalised if you make a silly mistake or small error
Checking answer against options	Can't show your working out
Can guess the answer if it can't be worked out	Questions seemed designed to trick students
Usually easier questions	
Quicker to do	
Marking is more reliable.	

Table 5.9: Advantages and Disadvantages of Short-answer questions

Advantages	Disadvantages
Proves that the student really knows the work	Don't have an opportunity to check answer against any given options.
Can get marks for partially correct answers	If you don't know what to do you can't attempt the question at all.
Worth more marks	Usually worth a lot of marks each
Can see where you went wrong	Usually harder questions
Teacher bias can influence marking	No prompts from the question like multiple-choice

As could be expected from different students, opposing viewpoints were often expressed. For example, some students thought that the larger number of marks attributed to a short-answer question was an advantage whilst others thought it was a disadvantage.

Some of the comments from students that illustrated the results in Tables 5.8 and 5.9 were:

Multiple-choice advantage.

Female student B4: School-A “*Sometimes because the answer is on the page you just need to select it*”

Female student D10: School-A: “*Multiple-choice only. It is much easier and reassuring. If you don't come up with the right answer it forces you to look over your working and then try to work out where you have gone wrong. This cannot happen in short-answer questions, as you can't be sure you are right*”.

Female student B4: School-A “*Sometimes because the answer is on the page you just need to select it*”

Multiple-choice disadvantage.

Female student B4: School-A “*You don't get any marks if you do the working out method partially correct but get the wrong answer.*”

Short-answer advantage.

Male Student A3: School-D “*Short-answer, helps students show their working out step by step for future exams, and rewards points for you step, rather than lose all of the marks. Also once you finished the test, you can look at what went wrong in your steps thus correcting the mistakes. Also you can correct your method of approach. Also can help your vocabulary by writing.*”

Short-answer disadvantage

Male student A17: School-D “*Yes, if you don’t know how to do the question all you can do is leave it blank*”

Male student A4: School-D Q6:”*If you know your stuff there are more marks available and generally short-answer questions are more straight forward.*” Q7: “*Often it is easy to drop one mark (3/4 or 2/3) on short-answer questions with a minor mistake*”

In the last two quotes from the same student the advantage of asking the double questions is indicated as the student was able to reinforce his opinion as given in the first response with a second supporting argument.

The general opinion gained from the responses was that student views of the advantages and disadvantages of multiple-choice and short-answer were somewhat mixed but in keeping with the observations of other researchers (Braswell, 1990; Bridgeman, 1992; Burton, 1996; Haynie, 1994; Pressley et al., 1990; Simkin & Kuechler, 2005). Generally, the male students were more positive than were the female students about multiple-choice and females more positive about short-answer.

A particularly interesting observation is that the students who offered guessing as an advantage to multiple-choice questions were amongst the lower scoring students on the trial tests. Students who appreciated the advantages of being able to get marks for showing correct working also tended to be amongst the higher scoring students in the trial tests. For example two students who did not perform particularly well on the tests made the following similar observations about multiple-choice questions.

Male student C22: School-D: “*Help work through process of getting right answer, a choice of getting it right if you’re clueless*”. This student scored an average of 65.3% and was ranked 127th amongst all the students.

Female student A23: School-B: “ *If you are unsure of the answer you have a 1 in 4 chance even if you guess and you have something to work to*”. This student scored an average of 48.3% and was ranked 166th amongst all the students

This contrasted to the better scoring students.

Male student A3: School-D For Q5 multiple-choice disadvantages, “*It allows guessing most of the time rather than attempting having a go. Also some students can guess and get it right. It doesn’t show their strengths or weaknesses. Also skipping steps rather than showing how you got there. Also if you were taught the wrong method of working, it doesn’t let the teacher know or you. Thus not allowing you to correct an error.*”

and

Q6, short-answer advantages, “*helps students show their working out step by step for future exams, and rewards points for you step, rather than lose all of the marks. Also once you finished the test, you can look at what went wrong in your steps thus correcting the mistakes. Also you can correct your method of approach. Also can help your vocabulary by writing.*” This student scored an average of 93.1% and was ranked 18th amongst all the students

Post interview member checking

Several weeks after conducting the interviews a sample of six students from the original interview groups were asked to reflect upon the transcripts of their interviews to see if their viewpoints had changed. In each case the students indicated an affirmation of what they had originally said. One student said that she may have answered some of the questions a little differently having experienced the interview but it was unlikely to actually alter the way she viewed chemistry questions.

Overall results from the student interviews

The results from the student responses were inconclusive in terms of determining any particular favouritism for either short-answer or multiple-choice questions. In responses to Questions 3 (Does the content of the question influence your decision?) and 8 (If you had the choice would you prefer chemistry tests to be multiple-choice only or short-answer only or a combination of both?), that were the most likely to give some information to this issue the results were quite uniform. The majority of

students indicated a preference for multiple-choice for responding to recall type content and short-answer for responding to application questions.

Response to Research question 5- Do teachers consider multiple-choice or short-answer questions to be the most effective in demonstrating their students understanding of chemistry?

The interviews of four teachers reinforced the viewpoints of the students. The transcripts of these interviews are shown in Appendix C-4. None of the teachers offered new perspectives that had not already been covered by the students. The teachers seemed to quickly focus on the obvious student advantage of multiple-choice questions in that the options were in front of the students and they only had to pick correctly. For example, Teacher School A: *“Ideally the recall should be multiple-choice because the student has a chance of identifying the correct answer even when they can’t directly recall it”*. In terms of the disadvantages the teachers were able to express more clearly the point that students had some difficulty clearly expressing. Namely, that multiple-choice questions can involve a lot of work for usually one mark and that no benefit could be gained from showing any correct working out. For example, Teacher School C: *“Students tend to think that multiple-choice will be easier, but usually don’t consider that they are not all equally difficult and they won’t get a mark for being nearly right”*. The teachers perceived that short-answer questions enabled the student and teacher to see how the students arrived at the answers and this was a valuable tool for understanding student thinking and teaching practice. Pragmatically the teachers also commented that the multiple-choice tests were useful from the dual aspects of being able to quickly assess a wide variety of material and that they were easy to mark. For example, Teacher School D: *“Quicker to mark, they can be used to test outside subject area. Can test applied knowledge more easily”*.

The overall findings from the interview of the four teachers included the following. The teachers found multiple-choice questions easier to mark but harder to set and devise in the first place. All the teachers believed quite strongly that short-answer questions gave them, and their students, a far greater opportunity to demonstrate understanding of the concept being tested and also informed the teachers more about any possible short coming in the teaching program. They were all of the opinion that a mixture of both types of questions in a test was most suitable in that it gave a

greater breadth to the material tested and at the same time greater insight into student learning. The teachers' viewpoints supported the findings described in the literature (Becker & Johnson, 1999; Degado & Prieto, 2003; Walsted & Robson, 1997).

Summary of Chapter 5

In this chapter the reporting of the findings and results of the trial tests and student interviews was discussed. The trial tests were conducted with the cooperation of four schools and 192 students participated. The results showed that the trial test administered to the students was a reliable and valid test instrument. This was determined by analysis using the RUMM2030 program (RUMM Laboratory P/L, 2009). Further detailed analysis allowed a comparison with the data obtained from the analysis of the VCE Chemistry examinations. The analysis confirmed many of the findings from the VCE examination paper analysis; however, there were also some differences. These similarities and differences will be reviewed in the next chapter.

The opinions of 59 of these students were gained and the students' opinions about chemistry testing using multiple-choice and short-answer questions were recorded. The interviews showed a wide variety of opinions on issues related to multiple-choice and short-answer questions in the context of testing recall or application content. A number of recurring ideas were identified and reported. Four teachers were also interviewed on a similar range of questions. The opinions of the staff largely supported the views of the students.

Chapter 6: Discussion and Conclusions

Introduction

This final chapter draws together the findings from the literature and from the analysis and practical research performed. The main aim of the research is addressed along with the related research questions. The first section of the chapter presents a brief overview of the previous five chapters. The next section presents the major findings and their relationship to the particular research questions where they pertain. The third section addresses the overall findings and conclusions and the implications they may have for the teaching and assessment of VCE Chemistry. The limitations of the study are then addressed followed by suggestions for possible future research that evolves from this study. The chapter concludes with final closing comments.

Overview of Study

This study initially developed from the researcher's observations, along with those of a number of teaching colleagues, that there were apparent differences between the observed student performances in Year 12 VCE Chemistry at the researcher's school and those observed at other schools. In particular, female students achieved higher grades in the second semester examination whereas male students were more successful in the first semester examination. Why should this be occurring? A number of factors were hypothesised as being at the base of this issue and secondly was it particularly a problem. Were any observed differences actually significant?

A related issue was that VCE Chemistry was the only science subject where male students outperformed female students in the achievement of the higher grades (Cox et al., 2004) even though more female students attempt Year 12 Chemistry than male students, however, female students do not achieve as many of the highest grades as do the male students. One possibility was that Chemistry assessment favoured the males of the student cohort.

Consulting the literature on these issues, a number of pertinent observations were made that exemplify the breadth of this study. Apart from the gender of the students two other important factors influence the performance of the students. The two factors are: type of question, short-answer or multiple-choice, and the content type of the question, recall or application.

In addressing the differences between multiple-choice questions and short-answer questions, opinion was mixed, mainly due to the variety of perspectives that can be adopted when looking at these two question forms. The multiple-choice question has a number of advantages. They are easy and quick to score making them popular with both teachers and educational authorities (Dufresne et al., 2002; Holder & Mills, 2001). They allow a larger number of questions covering a wider variety of material to be covered in a given space of time (Becker & Johnson, 1999; Walstad & Becker, 1994). Multiple-choice questions do not disadvantage students with weaker writing and spelling skills (Zeidner, 1987) as much as do short-answer questions. They are easier to manage in terms of item storage and re-use of items through the formation of item banks (Bush, 2001; Haladyna & Downing, 1989). They reduce anxiety amongst students who see them as friendlier examinations (Fredericksen & Collins, 1989; Scouller, 1998; Snow, 1993). These brief examples demonstrate some of the advantages of multiple-choice over short-answer. Equally there are perceived disadvantages presented by a number of researchers. Multiple-choice questions take more time to correctly write (Brown et al., 1997) and such tests tend to favour recall learning over applied learning (Martinez, 1999). An important finding was that male students tended to be favoured by multiple-choice questions as opposed to short-answer questions where female students are generally considered to have higher literacy skills and are favoured by short-answer questions (Beller & Gafni, 2000; Bridgeman & Lewis, 1994; Lumsden & Scott, 1987). 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).

One method of addressing the difference between the two types of question is the use of two-tiered tests. These tests involve students initially attempting a multiple-choice question for part of the question credit and then attempting an explanation of how the answer selected was arrived at for the remaining credit in the question. This method can allow students the confidence of selecting an answer but also allows students to demonstrate their depth of understanding (Tamir, 1989; Treagust, 1995; Treagust & Chandrasegaran, 2007).

Another aspect of the research focussed on the type of content being examined and did the type of question (multiple-choice or short-answer) impact on the students'

ability to answer questions that were either application of learned material or recall of the material. Typically, as reported by a number of researchers, recall type questions are favoured by being asked in a multiple-choice format (Chan & Kennedy, 2002; Ercikan et al., 1998; Simkin & Kuechler, 2005). There is a general consensus that multiple-choice questions are not able to measure the same level of complexity of information as is possible with a short-answer question (Simkin & Kuechler, 2005).

The last variable to be considered was that of student gender. There was little doubt that male students tended to perform differently to female students in chemistry. The differences in gender performance have been much debated in the literature with various factors identified. 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. This finding is particularly relevant in terms of this study as it concurred with the finding that male students outperformed female students in the VCE Chemistry examination, particularly in the higher grades, which require good performance on the higher order questions.

Various forms of evidence were considered to examine the research questions, which were framed around providing an insight to the impact of the three variables of question type, content and gender.

The research topic was “Multiple-choice Questions Compared to Short-answer. Which Assesses Understanding of Chemistry More Effectively” and the research questions were:

1. Do students perform more effectively on multiple-choice or short answer questions?
2. Do students perform more effectively on recall type questions or on application questions?
3. Does students' gender influence performance in chemistry examinations (or tests)?
4. Do students have a preference for the type of question style in terms of
 - a. Multiple-choice or short-answer in general terms?

- b. With respect to whether the question is assessing recall or application?
5. Do teachers consider multiple-choice or short-answer questions to be the most effective in demonstrating their students understanding of chemistry?

The next chapter on research methodology described the theoretical basis of the research undertaken and outlined the practical methodology employed to address the three factors.

The evidence was provided by:

1. Analysing the last five years of the Chemistry study design (2003-2007) VCE Chemistry Examination papers (ten papers in all) conducted by the VCAA. The results published by the VCAA gives student achievement information on 400 multiple-choice questions and about 170 short-answer questions. 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 results were analysed with the statistical program SPSS for correlation differences between the various permutations of the variables (Field, 2005; Pallant, 2007).
2. Further evidence from the VCAA sources provided information on the performances of students by gender. The VCAA grade distribution reports provide information on the number of students of each gender that achieve each grade in each examination. This information was processed for significant differences between the various grades and the gender of the examinees.
3. Trial tests conducted by the researcher with 192 students provided information about all aspects of the three variables under consideration. The trial tests provided information about performance on two versions of essentially the same question but asked in different styles (multiple-choice or short-answer). This information was important because the analysis of past papers did not easily produce this sort of data, as it is rarely the practice of

examiners to ask the same question twice in the same paper! The trial papers also elicited information of the gender effect of performance. The use of Rasch analysis (RUMM laboratory P/L, 2009) allowed a much closer examination of the trial test data, particularly with respect to the comparison of the question type, question content and gender.

4. Finally, students were interviewed to ascertain their views about the types of questions they had experienced and their preferences for question type (if any) with respect to chemistry examinations.

Chapters 4 and 5 provided the results of the different types of research undertaken. Detailed data are contained in the attached appendices. Chapter 4 detailed the analysis of the three aspects of the VCE Chemistry papers. The results showed aspects of the chemistry examinations that were significantly different with respect of question type, question content and gender. Chapter 5 outlined the results of the trial tests and student interviews. These results also provided information on the aspects of question type, question content and gender performance. The interviews revealed that the students' own preferences on question type and it was possible to examine within the sample group whether or not the actual performance on the trial tests actually matched the preferences of the students for the different type of questions. This final chapter presents the conclusions and findings of the research.

Major Findings in Response to the Research Questions

Findings in response to Research Question 1: Do students perform more effectively on multiple-choice or short answer questions?

The research into this question focussed on student performance in the VCAA examinations and on the trial tests.

Past VCE examination papers (RQ 1)

The VCE examinations demonstrated that students performed better on multiple-choice-questions than on short-answer questions (Table 4.11). The mean for the multiple-choice questions was 62.2, (s.d. = 15.1) compared to that of the short-answer questions mean = 58.9 (s.d. = 16.0). This difference was statistically significant ($p < 0.05$). The difference in favour of multiple-choice questions was due to the responses on the questions classified as recall where there was a substantial difference in the means of some eight points (69.3 compared to 60.9). There was

almost no difference in performance on application questions. These results were within expectations and concurred with the findings of other researchers (Barnett-Foster & Nagy, 1996; Bridgeman & Lewis, 1994; Martinez, 1999). A possible explanation for this finding stems from the multiple-choice questions providing prompts for the correct answer in the offered responses, thus aiding the student in selecting the correct option. A further possibility could be that the difficulty level of the material tested in the multiple-choice test may be easier for the students to understand, that is, recall based questions do not require as higher level of sophistication in understanding as do multi-step application questions. Multiple-choice questions also offer the student an answer checking option that short-answer questions do not offer. These possibilities were suggested by the literature (Chan & Kennedy, 2002; Simkin & Kuechler, 2005) and by both the students and staff participating in the trial test study as an advantage with multiple-choice questions.

Trial tests (RQ1)

The trial tests were constructed in such a way as to attempt to offer students questions on the same topic and of equal difficulty in both multiple-choice and short-answer format. Initial Rasch analysis (RUMM laboratory P/L, 2009) of the trial tests demonstrated that this was generally achieved with the test meeting unidimensional status and at the same time showing a correlation between matched items of multiple-choice and short-answer content (Figure 5.6 and Appendix H-4). The results (Table 5.3) showed that again the multiple-choice questions (mean = 71.0; s.d. = 20.8) were more successfully answered than the short-answer questions (mean = 69.1; s.d.= 22.3), although the difference was not as great as in the VCE examination analysis (Table 4.11). This finding was not surprising considering the matching of the questions in terms of content and difficulty in the trial tests. More importantly, the differences at the finer analysis level of examining performance on recall only questions and application only questions was the opposite to what had been found in the VCE examination (Tables 4.11 and 5.3).

Two possible explanations are offered to explain these differences. Firstly, the recall questions on the trial tests were of an abstract nature in terms of the topic (Acid-Base chemistry) and as such may have been testing some aspects of application style thinking, making the questions somewhat more difficult than would have been ideal for a pure recall question. Secondly, the questions on the trial tests were matched so

that the topic covered in a pair multiple-choice / short-answer questions were of approximately equal difficulty. This is not the case in the VCE examinations where each question is unique. The possibility mentioned above that the material covered in the multiple-choice questions in the VCE examinations might be less complex than that of the VCE short-answer questions may explain some of the observed differences. Further analysis of the VCE questions is warranted to determine if the questions in themselves are of differing difficulty (that is the multiple-choice questions are inherently easier in terms of the content covered than short-answer questions).

Overall the results of the VCE examination and trial test analysis show that multiple-choice questions allowed students to score more highly than on short-answer questions.

Findings in response to Research Question 2: Do students perform more effectively on recall type questions or on application questions?

The initial expectation was that students would perform more effectively on recall questions than on application questions, a finding demonstrated in the literature (Chan & Kennedy, 2002; Ercikan et al., 1998; Martinez, 1999; Simkin & Kuechler, 2005).

Past VCE examination papers (RQ 2)

The analysis of the VCE examinations (Table 4.11) supported this view with the mean of the recall questions being 64.2 (s.d. = 15.4) compared to application questions with a mean of 56.7 (s.d. = 15.2). This trend was observed when the analysis was extended to comparing recall and application multiple-choice and short-answer questions. The difference was most marked with the multiple-choice question, that is, the recall multiple-choice questions had a much higher mean than the application multiple-choice questions (Table 4.11).

Trial tests (R.Q.2)

When the trial tests were analysed the opposite result was observed with the performance on application questions being stronger. This result (Table 5.3) was almost entirely due to the performance on multiple-choice questions where the mean of the application questions was 81.5 (s.d. = 24.5) compared to the mean of the recall multiple-choice questions 62.3 (s.d.= 28.8). The Rasch analysis showed this to be a

significant difference even when allowing for item and student abilities (Figure 5.14).

As the results appeared to be contradictory in comparing the VCE examination and the trial tests, reaching a conclusion with regard to recall compared to application was difficult. However, a number of factors need to be considered. The application multiple-choice questions appear to have been easier for the group of students taking part in the trial study. This may have been partly due to the recall questions being at the more abstract end of the recall spectrum of difficulty. As mentioned in chapter 5, the recall topic of Acid-Base chemistry was more difficult than would have been ideal for this type of question; however, it did fit in with the schools' teaching program. Ideally a topic such as the periodic table would have been a preferable concept area for the recall questions; however, this was not being taught at the time of the year of the trial tests. In the larger VCE examination the recall questions proved easier than the application questions (Table 4.11). This may well have been due to the recall questions being generally easier and less complex than the application questions. This notion was not supported, however, by the trial tests.

Whilst no definitive conclusion can be drawn for this research question, the balance of probability would favour students responding to recall questions being more likely to achieve higher scores than application questions. The fact that the trial tests contradict this finding is tempered by the unusually stronger multiple-choice application score which needs to be viewed with some suspicion as the mean was nearly 10 percentage points higher than the other mean scores in the analysis.

Findings in response to Research Question 3: Does students' gender influence performance in chemistry examinations (or tests)?

The analysis of gender performance on the VCE examinations was particularly interesting. The limited data supplied by the VCAA showed that males outperformed females at the highest grades of A⁺ and A, and the difference was substantial. What could not be determined from the analysis was where in the test structure the difference in performance was greatest (if any such effect existed). Overall the number of students achieving the highest grades was heavily skewed in favour of the males.

The trial tests sought to examine the difference in more detail. An important initial analysis, as a point of comparison between the two sets of data, showed that the

distribution of scores in the trial tests closely matched the distribution reported in the VCE examinations which added credibility to the comparison between the two sets of data (see Figures 4.6 and 5.18). This was in some part due to the selection of the students for the trial tests. The students taking part in the trial tests were from high achieving schools and as mentioned in the previous chapter were likely to perform well and as result the similarity in the performance (Tables 4.11 and 5.3) was not unexpected.

The trial test analysis supported strongly the findings of the analysis of the VCE examinations in terms of the mean scores. However, when gender difference analysis using RUMM2030 was performed the difference in performance was smaller. The Rasch analysis compared student performance according to gender but allowed for student ability as measured by the Rasch analysis. When this result was taken into account the performance by the students differed little by gender even though the mean scores were significantly in favour of male performance. In some instances, female performance (allowing for ability) was better than that of the male students even though the mean scores suggested otherwise. For example, in comparing multiple-choice question performance, the mean for males = 76.4 (s.d. = 19.8) was higher than that of the females, mean = 65.5 (s.d. = 20.4) (see Table 5.3). However, when analysed using Rasch, the gender difference analysis showed that females generally performed slightly better on multiple-choice than did males once student ability was allowed for (Figure 5.19). Another significant observation showed that males outperformed females on the stoichiometry questions (Figure 5.22). This observation when taken into view with the grade distributions for the VCE Chemistry examinations (see figures 4.5 and 4.7) may explain the fact that male students outperform female students much more so in the Unit 3 examination than in the Unit 4 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 on the examinations. Such an observation has been supported by numerous findings from the literature (Beller & Gafni, 2000; Cox et al., 2004; Hawkes, 2004). This observation was further emphasised when the most recent examination data (2008) was considered. In the 2008 examinations the distribution and balance of the two examinations was changed so that the stoichiometry was spread more evenly between the Unit 3 and Unit 4 examination.

The result (see Appendix H-8) shows that males still outperform the females but the difference between the two Units has evened out. However, the impact of the stoichiometric questions has now spread to Unit 4. That is, the domination of males in the allocation of the higher grades is less than it was in Unit 3 examination but greater than before in the Unit 4 examination.

Another important observation is implied in 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 (Tables 5.4 and 5.5). When the performance allows for student ability however, the differences are quite small (Figure 5.19), suggesting 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 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 section of 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 differences in the middle and bottom of the grade scales were generally small (Tables 4.16 and 4.17).

Findings in response to Research Question 4: Do students have a preference for the type of question?

The results for research question 4 and research question 5 were based on the outcomes of student and teacher information and responses. A sample (59) of the 192 students was interviewed and four teachers also provided responses.

With respect to the preferred type of question the students indicated a preference for multiple-choice (59%) over short-answer (Table 5.7). This preference was stronger amongst the males (66%) compared to females (54%), however, the differences in preferences were not great and the differences not significant ($p > 0.05$). The results support those previously found in the literature (Simkin & Kuechler, 2005).

The reason offered by students for their preferences were consistent with those found by other researchers (Simkin & Kuechler, 2005). Generally the strongest features reported by students in favour of multiple-choice were that the options offered prompted them towards the correct answers, allowed the possibility of cross checking results from calculations and finally offered the possibility of at least being able to make an informed guess rather than leaving a blank space. Students who favoured short-answer questions most often offered the response that it gave them the opportunity to gain partial credit for incomplete responses and also the opportunity to “*show what they knew*”. Overall the results of the interviews were indicative rather than conclusive and supported the findings of other researchers.

Findings in response to Research Question 5: Do teachers consider multiple-choice or short-answer questions to be the most effective in demonstrating their students understanding of chemistry?

The interviews of the teachers reinforced some of the viewpoints of the students. For example, both teachers and students felt that short-answer questions provided a better opportunity for students to demonstrate concept understanding. The teachers felt that the multiple-choice questions offered opportunities for students with weak English skills; however, the teachers also conceded that these questions required a greater degree of interpretive understanding. Teachers believed that the short-answer questions offered greater insights into how well students understood the material being tested. Teachers also commented that the multiple-choice tests were useful from the dual aspects of being able to quickly assess a wide variety of material and that they were easy to mark. Again the findings from this section of the research supported the findings of other researchers (Simkin & Kuechler, 2005).

Findings in Response to the Overall Thesis Question: Multiple-Choice Questions Compared to Short-Answer Response. Which Assesses Understanding of Chemistry More Effectively?

With respect to the question framing this thesis the results are not entirely conclusive. 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

(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 VCE Chemistry examinations are of a lesser difficulty than the short-answer questions (Table 4.11). A wider ranging and more extensive study of the VCE 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

An underlying methodology for 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. 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 highlighted through Research Question 3. Initial analysis showed that the males outperformed the females as may have been expected in terms of previous research. When the Rasch gender differential analysis (RUMM laboratory P/L, 2009) was included it showed that when latent student abilities are included for then there was little between the performance of the male and female students, that is, male and female students of equal ability perform very similarly in chemistry. This suggests that the male students choosing chemistry include a greater proportion of high ability males compared to the proportion of high ability females choosing chemistry. Whilst this proposition will need more testing there is some support when comparing the VCE Biology results to the VCE Chemistry results grade distributions (see Figures 4.5 and 4.6) where the females have achieved a greater proportion of the higher grades than the males in biology, implying that more high ability females choose biology than do high ability males.

Limitations of Study

Transference

The use of this study in terms of the transferability of the results is limited in some respects. Whilst there is good triangulation within the study (Denzin, 1997; Mathison, 1988) between the literature, VCE examination analysis, trial test analysis and student and staff surveys, there are three main limitations towards the application of the results to the wider community. Firstly, the limited size of the trial population (192 students) restricts the variety of possible responses and also limited the statistical analysis of the trial tests using the Rasch model. The ideal sample size for Rasch analysis is around 250 persons (Bond & Fox, 2007). Secondly, the student sample consisted of chemistry students only, selected from a number of middle class schools and as such were not representative of the wider community. That being said however, the sample did serve the purpose of being representative of higher performing chemistry students. Thirdly, the match between the VCE examination analysis and the trial test analysis was not perfect with a number of variations between the two sets of results (Table 5.3). For example, the means of the recall and application multiple-choice questions were in the reverse order in the trial tests when compared to the means of the VCE examinations, that is the mean of the recall multiple-choice questions was larger in the VCE examinations than the application multiple-choice mean but the mean of the application multiple-choice questions was higher in the trial tests than the mean of the recall multiple-choice questions. A similar pattern was repeated when comparing all recall questions to all application questions limiting the reliability of supportive statements based on the comparing the two sets of results.

Student interest

The VCE examinations are high stakes and generally receive maximum effort and interest from the participating students. Consequently, data from them can be reasonably assumed to fairly represent the genuine efforts and achievements of the students participating. The same cannot be said of the trial tests. The students are aware that they are not high stakes and so less than maximum effort can be assumed from the students. Consequently, a degree of caution should be exercised in any conclusions drawn directly from the trial test results. Part of the reason for choosing the sample from the particular schools used was to minimise this effect and the

Rasch analysis indicated only a handful of students whose individual results did not fit the model well. The generally good correlation between the paired items and the overall match with the VCE examination results show that a good degree of confidence can be placed in the trial tests despite the issue raised regarding student interest in the tests.

Practice element

The aim of the trial tests was to match items according to difficulty and content but have one version as multiple-choice and the other as short-answer. This introduced the possible problem of students using one test as practice for the other. From the analysis there does not appear to be any obvious evidence of this occurring. The teachers conducting the tests were asked to mix the order of delivery of the tests so that any practice element may be reduced in its impact on the results. However, the possibility of a practice effect cannot be ignored as being non-existent. The mixing of the delivery order of the tests will have minimised any statistical advantage that one form (that is multiple-choice compared to short-answer) of the test would have gained had the tests been delivered in a set order. For example if the multiple-choice version of the trial test had been completed always before the short-answer version then it would be reasonable to expect that the performance on the latter may have been advantaged.

Limited access to VCE results in detail

The analysis of the VCE results was based on the publicly released data from the VCAA. Ideally it would have been more useful to have the individual student data for each question so that better correlation between the trial tests and the VCE examinations could have been made.

Interviews

The interviews were inconclusive and did not provide much in the way of new insights into chemistry assessment. They did, however, confirm the views already established in the literature and so it is reasonable to assume that the students participating in the trial tests were not atypical of senior chemistry students thus adding to the reliability of the results of the trial test analysis.

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 VCE examinations. The most interesting 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 the actual VCE data may provide some useful insights into this issue.

The limited size of the trial sample and the low stakes nature of the trial 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 VCE examination analysis and the trial test analyses, the variations mentioned on the previous page suggest that a more detailed analysis and larger trials 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.

Summary

The analysis of the VCE examinations, trial tests and student interviews 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
- Male students outperformed female students, particularly at the higher grades.
- However, when individual student ability is taken into account the differences in performance between genders are much smaller.
- Students of both genders prefer, to a small extent, multiple-choice questions to short-answer questions.

References

- Aiken, R. L. (1987). Testing with multiple-choice items. *Journal of Research and Development in Education*, 20, 44-58.
- Anderson, G. (1998). *Fundamentals of educational research* (2nd. ed.). Bristol: PA: Falmer Press.
- Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., et al. (2001). *A taxonomy for learning, teaching, and assessing — A revision of Bloom's taxonomy of educational objectives*. New York: Addison Wesley Longman.
- 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.
- Beard, R. M., & Senior, I. J. (1980). *Motivating students*. London, U.K.: Routledge & Kegan Paul.
- 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.
- Bell, J. (1993). *Doing your research project*. Buckingham: Open University Press.
- Bell, R. C., & Hay, J. A. (1987). Differences and biases in English language examination formats. *British Journal of Educational Psychology*, 57, 212-220.
- 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.
- Bennet, R., Rock, D. A., & Wang, X. (1991). Equivalence of free response and multiple-choice items. *Journal of Educational Measurement*, 28, 77-92.
- Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement*, 28(1), 23-35.
- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32(3), 347-364.
- Biggs, J. B. (1973). Study behaviour and performance in objective essay formats. *Australian Journal of Education*, 17, 157-167.

- Blais, D. M. (1988). Constructivism: A theoretical revolution in teaching. *Journal of Developmental Education*, 11(2), 2-7.
- Bloom, B. S. (1956). *Taxonomy of educational objectives, Handbook I: The cognitive domain*. New York: David McKay Company Incorporated.
- Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement*, 27, 165-174.
- 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. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Bond, T., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- 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. (1991). Essays and multiple-choice tests as predictors of college freshman GPA. *Research in Higher Education*, 32(319-332).
- 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.
- Bridgeman, B., & Rock, D. A. (1993). Relationships among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement*, 30(4), 313-329.
- Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. London, UK: Routledge.
- Buehl, M. M., & Alexander, P. A. (2005). Motivation and performance differences in students' domain-specific epistemological belief profiles. *American Educational Research Journal*, 42(4), 697-726.
- Burns, R. (1994). *Introduction to research methods in education*. Cheshire: Longman.
- Burrell, G., & Morgan, G. (1979). *Sociological paradigms and organisational analysis*. London: Heinemann.
- Burton, L. (1996). A socially just pedagogy for the teaching of mathematics. In P. F. Murphy & C. V. Gipps (Eds.), *Equity in the classroom: Towards effective pedagogy for girls and boys* (pp. 136-146). London: Falmer Press.

- Bush, M. (2001). A multiple-choice test that rewards partial knowledge. *Journal of Further and Higher Education*, 25(2), 157-163.
- Carey, J. (1997). Everyone knows that $E=mc^2$ now, who can explain it? *Business Week*, 3457(66-68).
- Carver, R. P. (1993). The case against statistical significance revisited. *Journal of Experimental Education*, 21(4), 278-292.
- 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.
- Commons, C., Jarrett, S., McKenzie, C., Moseley, W., Porter, M., & Williamson, M. (1999). *Chemistry Two*. Melbourne: Heinemann.
- Condelli, L., & Wrigley, H. (2004). *Real world research: Combining qualitative and quantitative research for Adult ESL*. Paper presented at the National Research and Development Centre (NRDC) Second International Conference for Adult Literacy and Numeracy., Loughborough, England.
- Connolly, N. (2009). Rasch modelling advice. In R. Hudson (Ed.) (pp. Mr Connolly provided the author with advice regarding the Rasch model analysis of the Year 11 class tests.). Sydney: ACER.
- Cox, J., Leder, J., & Forgasz, H. (2004). Victorian Certificate of Education: Mathematics, science and gender. *Australian Journal of Education*, 48(1), 27-46.
- Crabtree, B., & Miller, W. (1999). *Doing qualitative research*. Thousand Oaks: SAGE.
- Creswell, J. (2008). *Research design: Qualitative, quantitative, and mixed methods approaches*. Los Angeles: SAGE.
- Dekkers, J., & De Laeter, J. (1997). The changing nature of upper secondary school science enrolments. *Australian Science Teachers Journal*, 43(4), 35-41.
- Dekkers, J., & De Laeter, J. (2001). Enrolment trends in school science education in Australia. *International Journal of Science Education*, 23(5), 487-500.
- Delgado, A. R., & Prieto, G. (2003). The effect of item feedback on multiple-choice test responses. *British Journal of Psychology*, 94(1), 73-85.
- Denzin, N., & Lincoln, Y. S. (2005). *The SAGE handbook of qualitative research*. Thousand Oaks, CA: SAGE.
- Denzin, N. K. (1997). Triangulation in educational research. In J. P. Reeves (Ed.), *Educational research, methodology and measurement: An international handbook*. (2nd. ed., pp. 318-322). Oxford CA: Elsevier.

- Dobson, I. R., & Calderon, A. J. (1999). *Trends in science education: Learning, teaching and outcomes 1987-1997*. Melbourne: Australian Council of Deans of Science.
- Driver, R., Asoko, H., Leach, J., Mortimer, E., & Scott, P. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher*, 23(7), 5-12.
- 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).
- Ebel, R., & Frisbie, D. A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs: Prentice Hall.
- Entwistle, A., & Entwistle, N. (1992). Experiences of understanding in revising for examinations. *Learning and Instruction*, 2, 1-22.
- Entwistle, D. R., Alexander, K. L., & Olsen, L. S. (1994). The gender gap in maths: Its possible origins in neighborhood effects. *American Sociological Review*, 59, 822-838.
- Epstein, M. L., & Brosvic, G. M. (2002). Immediate feedback assessment technique: Multiple-choice test that behaves like an essay examination. *Psychological Reports*, 90(1), 226.
- Epstein, M. L., Epstein, B. B., & Brosvic, G. M. (2001). Immediate feedback during academic testing. *Psychological Reports*, 88(3), 889-895.
- 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.
- Fenna, D. S. (2004). Assessment of foundation knowledge: Are students confident in their ability? *European Journal of Engineering Education*, 2, 307-313.
- Fennema, E. (1979). Women and girls in mathematics-equity in mathematics education. *Educational Studies in Mathematics*, 10(4), 389-401.
- Field, A. (2005). *Discovering statistics using SPSS (Introducing statistical methods)* (2nd. ed.). Thousand Oaks, CA: SAGE Publications Inc.
- Fielding, N. G., & Fielding, J. L. (1986). *Linking data*. Beverly Hills: SAGE.
- Fontana, A., & Frey, J. H. (1994). Interviewing: The art of science. In N. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research*. (pp. 361-376). Thousand Oaks: Sage Publications.
- 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.
- Fraser, B. (1994). Two decades of classroom environment research. In B. Fraser & H. J. Walberg (Eds.), *Handbook of research on science teaching and learning*. New York: Macmillan.
- Fraser, B. J. (Ed.). (1991). *Two decades of classroom environment research*. London: Pergamon.
- Fraser, B. J., & Tobin, K. G. (1991). Combining qualitative and quantitative methods in classroom research. In B. J. Fraser & H. J. Walberg (Eds.), *Educational*

- Environments: Evaluation, antecedents and consequences.* (pp. 271-291). London: Pergamon Press.
- Fraser, B. J., & Tobin, K. G. (1998). Qualitative and quantitative landscapes of classroom learning environments. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of Science Education* (pp. 623-640). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Fredericksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18(9), 27-32.
- Frederickson, N. (1984). Implications for cognitive theory for instruction in problem solving. *Review of Educational Research*, 54, 363-407.
- Gay, L. (1980). The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement*, 17(1), 45-50.
- Geertz, C. (1977). *The interpretation of cultures: Selected essays*. New York: Basic Books.
- Gibbs, G. R. (2002). *Qualitative analysis with Nvivo*. Buckingham: Open University Press.
- Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. CA: Sage: Newbury Park.
- Gunstone, R. F. (1995). Constructivist learning and the teaching of science. In B. Hand & V. Prain (Eds.), *Teaching and learning in science: The constructivist classroom* (pp. 3-20). Sydney: Harcourt Brace.
- Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, 2(1), 37-50.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items.* (3rd ed.) Mahwah, NJ: L Erlbaum Associates.
- Halasa, K. (1998). *Annotated bibliography-ethics in educational research*. Retrieved 9th April, 2004, from <http://www.aare.edu.au/ethics/aareethc.htm>
- 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.
- Hammersley, M., & Atkinson, P. (1995). *Ethnography: Principles in practice*. New York: Tavistock.
- Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. London and New York: Routledge.
- 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.

- Hildebrand, G. M. (1996). Redefining achievement. In P. F. Murphy & C. V. Gipps (Eds.), *Equity in the classroom: Towards effective pedagogy for girls and boys* (pp. 149-172). London: Falmer Press.
- Hobson, A., & Ghoshal, D. (1996). Flexible scoring for multiple-choice exams. *The Physics Teacher*, 34(5), 284.
- Holder, W. W., & Mills, C. N. (2001). Pencils down, computer up: The new CPA exam. *Journal of Accountancy*, 191(3), 57-60.
- Johnson, R. B. (1997). Examining the validity structure of qualitative research. *Educational Evaluation and Policy Analysis*, 118(2), 282–292.
- Kemmis, S., & McTaggart, R. (2000). Participatory action research. In N. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 555-604). Thousand Oaks: Sage Publications.
- Kniveton, B. H. (1996). A correlation analysis of multiple-choice and essay measures. *Research in Education*, 56, 73-84.
- Kreig, R. G., & Uyar, B. (2001). Student performance in business and economics statistics: Does exam structure matter? *Journal of Economics and Finance*, 25(2), 229-241.
- Kvale, S. (1996). *Interviews*. London: Sage publications.
- Learner, R. (Ed.). (2005). *Chemistry-study design*. East Melbourne: Victorian Curriculum and Assessment Authority.
- Lee, V.S. (1999). Creating a blueprint for the constructivist classroom. *The National Teaching and Learning Forum*, 8(4), 1-4.
- Levin, J. & James, A. F. (1991) *Elementary statistics in social research* (5th ed.). New York: Harper Collins
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage Publications.
- 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.
- Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, 31(3), 234-250.
- Lumsden, K. G., & Scott, A. (1987). The economics student reexamined: Male-female difference in comprehension. *Journal of Economic Education*, 18(4), 365-375.
- Maor, D., & Fraser, B. (1996). Use of classroom environment perceptions in evaluating inquiry based computer assisted learning. *International Journal of Science Education*, 18(4), 401-421.

- 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. (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.
- Mason, S. (1993). Employing qualitative and quantitative research in one study. *British Journal of Nursing, 2*(17), 869-872.
- Mathison, S. (1988). Why triangulate? *Educational Researcher, 17*(2), 13-17.
- McCormick, R., & James, M. (1988). *Curriculum evaluation in schools* (2nd ed.). London: Croon Helm, CA.
- Merriman, S. B. (1988). *Case study research in education: A qualitative approach*. San Francisco: Jossey-Bass.
- Mertens, D. (2005). *Research and evaluation in education and psychology: integrating diversity with quantitative, qualitative, and mixed methods*. Thousand Oaks, CA: SAGE.
- Middleton, J. A., & Spanias, P. A. (2005). Motivation for achievement in mathematics: Findings, generalizations, and criticisms of the research. *Journal for Research in Mathematics Education, 30*(1), 65-88.
- Miller, J., & Chamberlin, M. (2000). Women are teachers, men are professors: A study of student perceptions. *Teaching Sociology, 28*(4), 283-298.
- Myers, J. L., & Well, D. W. (2003). *Research design and statistical analysis*. (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- 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.
- Nungester, R. J., & Duchastel, P. C. (1982). Testing versus analysis: Effects on retention. *Journal of Educational Measurement, 74*(1), 18-22.
- O'Neill, K. A., & McPeck, 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.
- Patton, M. (1990). *Qualitative evaluation and research methods*. Newbury Park: SAGE.
- Paxton, M. (2000). A linguistic perspective on multiple-choice questioning. *Assessment & Evaluation in Higher Education, 25*(2), 109-120.
- Peshkin, A. (1993). The goodness of qualitative research. *Educational Researcher, 22*(2), 24-30.

- Petrie, H. (1986). Testing for critical thinking. In D. Nyberg (Ed.), *Philosophy of Education* (pp. 3-19). Normal, IL.: Philosophy of Education Society.
- Pilliner, A. (1973). *Experiment in educational research*. Milton Keynes: Open University Press.
- 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.
- Ramsden, P. (1988). Studying learning; improving teaching. In P. Ramsden (Ed.), *Improving learning: New perspectives* (pp. 13-31). London, U.K.: Kogan Page.
- Reichardt, C. S., & Cook, T. D. (Eds.). (1979). *Beyond qualitative and quantitative methods*. Beverly hills: SAGE.
- 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.
- Rogers, W. T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and test: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement*, 59(2), 234-247.
- Rolison, M. A., & Medway, F. J. (1985). Teachers' expectations and attributions for student achievement: Effects of label, performance pattern, and special education intervention. *American Educational Research Journal*, 22(4), 561-573.
- RUMM Laboratory. (2009a). Getting started with RUMM2030: RUMM Laboratory P/L.
- RUMM Laboratory. (2009b). Interpreting RUMM2030: RUMM Laboratory P/L.
- RUMM Laboratory P/L. (2009). RUMM2030.
- Russell, B. (1998, 20/9/1998). *Code of ethics*. Retrieved 9th April, 2004, from <http://www.aare.edu.au/ethics/ethcfull.htm>
- Ryan, A. M., & Patrick, H. (2001). The classroom social environment and changes in adolescents' motivation and engagement during middle school. *American Educational Research Journal*, 38(2), 437-460.
- Ryan, G. W., & Bernard, H. R. (2000). Data management and analysis methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 769-802). London: Sage Publications.
- Sax, G., & Collet, L. S. (1968). An empirical comparison of the effects of recall and multiple-choice tests on student achievement. *Journal of Educational Measurement*, 5(169-173).
- 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.

- Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple-choice question examinations versus assignment essay. *Higher Education, 35*, 453-472.
- Silverman, D. (1993). *Interpreting qualitative data*. London: Sage publications.
- 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.
- Simons, H. (1982). Conversation piece: the practice of interviewing in case study research. In R. McCormick (Ed.), *Calling education to account*. (pp. 239-246). London: Heinemann.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet based tests. *Journal of Educational Measurement, 25*, 15-29.
- Smith, J. K. (1983). Quantitative versus qualitative research: An attempt to clarify the issue. *Educational Researcher, 12*(3), 6-13.
- Snow, R. E. (1993). Construct validity and constructed-response tests. In E. B. Randy & C. W. William (Eds.), *Construction versus choice in cognitive measurement*. Hillsdale, NJ: Lawrence Erlbaum.
- Snyder, A. (2003). The new CPA exam: Meeting today's challenges. *Journal of Accountancy, 196*(6), 11-12.
- Stobart, G., Elwood, J., & Quinlan, M. (1992). Gender bias in examinations: How equal are the opportunities? *British Educational Research Journal, 18*(3), 261-276.
- Tamir, P. (1971). An alternative approach to the construction of multiple-choice test items. *Journal of Biological Education, 5*, 305-307.
- Tamir, P. (1989). Some issues related to the use of justifications to multiple-choice answers. *Journal of Biological Education, 23*(4), 285-292.
- Tashakkori, A., & Teddlie, C. (1998). *Combining qualitative and quantitative approaches*. Thousand Oaks, CA: SAGE.
- The Coalition of Americans for Research Ethics. (1999). *Do no harm*. Retrieved 8th April, 2004, from http://www.stemcellresearch.org/pr/pr_1999-08-25.htm
- Thissen, D. (1993). Repealing rules that no longer apply to psychological measurement. In N. Frederickson, R. J. Mislevy & I. Bejar (Eds.), *Test theory for a new generation of tests*. (pp. 79-97). Hillsdale: Erlbaum.
- Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*(2), 113-123.
- Thomson, S., Wernert, N., Underwood, C., & Nicholas, M. (2007). *Highlights from TIMSS 2007 from Australia's perspective.*, June 4th, 2009, from http://www.acer.edu.au/documents/TIMSS_2007-AustraliaHighlights.pdf
- Traub, R. E. (1992). On the equivalence of traits assessed by multiple-choice and constructed response questions. In R. Bennet & W. Ward (Eds.),

- Construction versus choice in cognitive measurement* (pp. 29-44). Hillsdale, N.J.: Lawrence Erlbaum and Associates.
- Treagust, D. F. (1988). The development and use of diagnostic instruments to evaluate students' misconceptions in science. *International Journal of Science Education*, 10, 159-169.
- Treagust, D. F. (1995). Diagnostic assessment of students science concepts. In *Learning science in schools: Research reforming practice* (pp. 327-346). Mahwah, NJ: Lawrence Erlbaum Associates.
- Treagust, D. F., & Chandrasegaran, A. L. (2007). The Taiwan national science concept learning study in an international perspective. *International Journal of Science Education*, 29(4), 391-403.
- VCAA. (2002). *Grade distribution report for 2002*. Retrieved 9th June 2008, from http://www.vcaa.vic.edu.au/vce/statistics/2002/section3/vce_chemistry_ga02.pdf
- VCAA. (2003a). *2003 Chemistry Exam 1*. Retrieved October 4th, 2008, from <http://www.vcaa.vic.edu.au/vce/studies/chemistry/exams/chemistry12003.pdf>
- VCAA. (2003b). *2003 Chemistry Exam 2*. Retrieved October 4th, 2008, from http://www.vcaa.vic.edu.au/vce/studies/chemistry/exams/2003_chem2.pdf
- VCAA. (2003c). *Grade distribution report for 2003*. Retrieved 9th June 2008, from http://www.vcaa.vic.edu.au/vce/statistics/2003/section3/vce_chemistry_ga03.pdf
- VCAA. (2003-2007). *VCE examinations and assessment reports*. Retrieved 8th July, 2008, from <http://www.vcaa.vic.edu.au/vce/studies/chemistry/exams.html#H2N10030>
- VCAA. (2004a). *2004 Chemistry exam 1*. Retrieved October 4th, 2008, from <http://www.vcaa.vic.edu.au/vce/studies/chemistry/exams/chem12004.pdf>
- VCAA. (2004b). *2004 Chemistry exam 2*. Retrieved October 4th, 2008, from <http://www.vcaa.vic.edu.au/vce/studies/chemistry/exams/chem12004.pdf>
- VCAA. (2004c). *Grade distribution report for 2004*. Retrieved 9th June 2008, from http://www.vcaa.vic.edu.au/vce/statistics/2004/section3/vce_chemistry_ga04.pdf
- VCAA. (2005a). *2005 Assessment report Chemistry exam 1*. Retrieved October 4th, 2008, from http://www.vcaa.vic.edu.au/vce/studies/chemistry/assessreports/2004/chemas_srepex204.pdf
- VCAA. (2005b). *2005 Assessment report Chemistry exam 2*. Retrieved October 4th, 2008, from http://www.vcaa.vic.edu.au/vce/studies/chemistry/assessreports/2005/chemist_ryassessreportnov05.pdf
- VCAA. (2005c). *2005 Chemistry exam 1*. Retrieved October 4th, 2008, from <http://www.vcaa.vic.edu.au/vce/studies/chemistry/exams/05chem1.pdf>
- VCAA. (2005d). *2005 Chemistry exam 2*. Retrieved October 4th, 2008, from <http://www.vcaa.vic.edu.au/vce/studies/chemistry/exams/2005chem2.pdf>

- VCAA. (2005e). *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. (2006a). *2006 Chemistry exam 1*. Retrieved October 4th, 2008, from http://www.vcaa.vic.edu.au/vce/studies/chemistry/exams/2006/2006chem1_w.pdf
- VCAA. (2006b). *2006 Chemistry exam 2*. Retrieved October 4th, 2008, from <http://www.vcaa.vic.edu.au/vce/studies/chemistry/exams/2006/2006chem2-w.pdf>
- VCAA. (2006c). *Grade distribution report for 2006*. Retrieved 9th June 2008, from http://www.vcaa.vic.edu.au/vce/statistics/2006/section3/vce_chemistry_ga06.pdf
- VCAA. (2006d). *VCE Chemistry Units 1-4*: Victorian Curriculum and Assessment Authority.
- VCAA. (2007a). *2007 Chemistry Exam 1*. Retrieved October 4th, 2008, from http://www.vcaa.vic.edu.au/vce/studies/chemistry/exams/2007/chem1_exam_07.pdf
- VCAA. (2007b). *2007 Chemistry Exam 2*. Retrieved October 4th, 2008, from <http://www.vcaa.vic.edu.au/vce/studies/chemistry/exams/2007/2007chem2.pdf>
- VCAA. (2007c). *Grade distribution report for 2007*. Retrieved 9th June 2008, from http://www.vcaa.vic.edu.au/vce/statistics/2007/section3/vce_chemistry_ga07.pdf
- VCAA. (2007d). *VCE statistics by subject*. Retrieved April 5th, 2009, from <http://www.vcaa.vic.edu.au/vce/statistics/2007/statssect2.html>
- VCAA. (2008). *General statistics about the VCE, VET and VCAL for 2008*. Retrieved 2nd July, 2009
- VCAA. (2009a). *Grade distribution report for Chemistry 2008*. Retrieved 28 February, 2010, from http://www.vcaa.vic.edu.au/vcaa/vce/statistics/2008/section3/vce_chemistry_ga08.pdf
- VCAA. (2009b). *Post compulsory completion and achievement information 2009*. Retrieved 7th Feb, 2010, from <http://www.vcaa.vic.edu.au/vce/statistics/schoolstats/postcompletiondata-2009schools.pdf>
- VCAA. (2010). *Statistical Moderation of VCE Coursework*. Retrieved May 17, 2010, from <http://www.vcaa.vic.edu.au/vce/exams/statisticalmoderation/statmod.html>
- Wainer, H., & Thissen, D. (1993). Combined multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, 6(2), 103-118.
- Walding, R., Fogliani, C., Over, R., & Bain, J. D. (1994). Gender differences in responses to questions on the Australian national chemistry quiz. *Journal of Research in Science Teaching*, 31(8), 833-846.

- Walstad, W. B. (1998). Multiple-choice tests for the economics course. In B. W. William & S. Phillip (Eds.), *Teaching undergraduate economics: A handbook for instructors*. (pp. 287-304). New York, NY: McGraw-Hill.
- Walstad, W. B., & Becker, W. E. (1994). Achievement differences on multiple-choice tests in economics. *American Economic Review*, 84(2), 193-196.
- Walstad, W. B., & Robson, D. (1997). Differential item functioning and male-female differences on multiple-choice tests. *Journal of Economic Education*, 28, 155-171.
- 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.
- Wesolowsky, G. O. (2000). Detecting excessive similarity in answers on multiple-choice exams. *Journal of Applied Statistics*, 27(7), 909-921.
- 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.
- Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student's perspective. *Journal of Educational Research*, 80(6), 352-358.
- Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement*, 27(5), 357-371.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

Appendices

Appendix A: Student trial tests

A-1: Stoichiometry Test: Short-answer Questions

Write your answers in the spaces provided

1. What is the percentage by mass of Zn in $\text{Zn}_3(\text{PO}_4)_2$?

(2 marks)

2. 7.3g of gaseous hydrogen chloride, HCl, is dissolved in 500ml of water.
 - a. Calculate the number of mole of HCl used.
 - b. Calculate the molarity (concentration) of the resulting solution.

(1 + 1 = 2 marks)

3. Zinc reacts with dilute hydrochloric acid according to the equation:
$$\text{Zn}_{(s)} + 2\text{HCl}_{(aq)} \rightarrow \text{ZnCl}_{2(aq)} + \text{H}_{2(g)}$$

In a certain experiment, 39.0g of zinc is reacted with excess hydrochloric acid.

 - c. How many mole of hydrochloric acid are required to completely react with 39.0g of zinc?
 - d. Calculate the volume of 0.50M hydrochloric acid that would react with 39.0g of zinc?
 - e. Calculate the mass of zinc chloride produced in the experiment.

(2 + 2 + 2 = 6 marks)

A-2: Acid- Base Test: Short-answer Questions

Write your answers in the spaces provided

1. Write down the name **and** formula of:

(a) a strong base

(b) a weak acid

(2 X 1 = 2 marks)

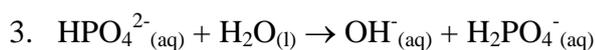
2. Write down one chemical property of:

(a) an acid

(b) a base

(2 X 1 = 2 marks)

2. Identify the conjugate pairs in the following equation: (indicate which is the acid and which is the base for both conjugate pairs)



(2 marks)

4. Write 2 equations to show the amphoteric nature of HCO_3^{-1}

acid: _____

base: _____

(2 marks)

5. (a) Write a balanced chemical equation, with state symbols, to show the reaction that occurs when dilute hydrochloric acid neutralises dilute potassium hydroxide solution.

(1 mark)

6. Referring to the concentration of the hydrogen ion $[\text{H}_3\text{O}^+]$ describe the difference between an acidic basic and neutral solutions.

(2 marks)

A-3: Stoichiometry Test: Multiple-choice Questions

You may use a periodic table to find rel. atomic masses

Circle the letter of the correct answer

(Please note: Consequential marking will occur in question 4)

1. The percentage by mass of oxygen in $\text{Mg}(\text{NO}_3)_2$ is closest to:

A. 11%
B. 48%
C. 65%
D. 78%

(1 mark)

2. 1.56g of solid calcium hydroxide, $\text{Ca}(\text{OH})_2$, is dissolved in 500ml of water.

- a. What is the number of mole of $\text{Ca}(\text{OH})_2$ used?

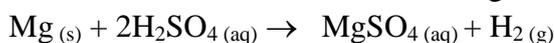
A. 0.021
B. 0.027
C. 36.53
D. 47.43

3. 0.23 mole of silver nitrate is dissolved in 400ml of water. The molarity (concentration) of the resulting solution is:

A. 0.092 M
B. 0.58 M
C. 1.74M
D. 92.0 M

(1 + 1 = 2 marks)

4. Zinc reacts with dilute sulfuric acid according to the equation:



In a certain experiment, 19.6g of magnesium is reacted with excess sulfuric acid.

- a. How many mole of sulfuric acid are required to completely react with 19.6g of magnesium?

A. 0.403
B. 0.620
C. 1.61
D. 2.48

- b. Calculate the volume of 1.50M sulfuric acid that would react with 19.6g of magnesium?

A. 0.268
B. 0.413
C. 1.08
D. 1.65

- c. Calculate the mass of magnesium sulfate produced in the experiment.

A. 24.2g
B. 37.2g
C. 97.0g
D. 149g

(1 + 1 + 1 = 3 marks)

A-4: Acid-Base Test: Multiple-choice Questions

Circle the letter of the correct answer

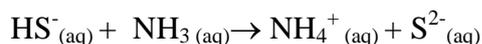
1. Select the group that contains strong acids

- A. HNO_3 , NH_3 , H_2SO_4
- B. NaOH , NH_3 , KOH
- C. HNO_3 , CH_3COOH , HCl
- D. HNO_3 , HCl , H_2SO_4

2. Which of the following is a property of an acid ?

- A. have a high pH
- B. accept protons from a base
- C. form the H_3O^{+1} ion in solution
- D. react with all metals to form hydrogen gas

3. Consider the following equation:



Which of the following are 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^-

4. Which of the following species is **not** amphoteric ?

- A. HS^-
- B. H_2O
- C. HF
- D. H_2PO_4^-

5. An acid reacting with a hydroxide will form

- A. A salt and carbon dioxide and water
- B. A salt and hydrogen
- C. A salt and water
- D. A salt and hydrogen and water.

6. Which is not true

- A. In acidic solutions $[\text{H}_3\text{O}^+] > [\text{OH}^-]$ and $[\text{H}_3\text{O}^+] > 10^{-7}\text{M}$
- B. In basic solutions $[\text{OH}^-] > [\text{H}_3\text{O}^+]$ and $[\text{H}_3\text{O}^+] > 10^{-7}\text{M}$
- C. In neutral solutions $[\text{H}_3\text{O}^+] = [\text{OH}^-] = 10^{-7}\text{M}$ (at 25°C)
- D. In basic solutions $[\text{H}_3\text{O}^+] < [\text{OH}^-]$ and $[\text{H}_3\text{O}^+] < 10^{-7}\text{M}$

(6 X 1 = 6 marks)

Appendix B: Permission letters for trials

B-1: Principal permission letter request form

Curtin University of Technology
School of Engineering, Science and Computing
(SMEC)



Participant Information Sheet

<<School>>

<<Principal>>

<<Address line 1>>

<<Address Line 2>> <<Post Code>>

Dear <<Name>>,

My name is Ross Hudson I am currently completing a piece of research for my Doctor of Science Education at Curtin University of Technology.

Purpose of Research

I am investigating the research topic : *Multiple-choice questions compared to short-answer. Which assesses understanding of Chemistry more effectively*

Your Role

I am seeking your permission to conduct research by asking for students to take part in a short test(s) on chemistry that will complement their learning. Students involved will undertake a number of short tests. The results of the tests will be given back to the students after the completion of the test.

I may also ask for the students' participation in a short interview (group) about their attitudes and opinions about assessment in chemistry. Again this participation will be voluntary and of short duration (10-15 mins)

Consent to Participate

The students and your school's involvement in the research is entirely voluntary. You have the right to withdraw at any stage without it affecting your rights or my responsibilities. When you have signed the consent form I will assume that you have agreed to participate and allow me to use the students data in this research.

Confidentiality

The information provided will be kept separate from the students' personal details, and only myself and my supervisor will only have access to this. The interview transcripts will not have student names or any other identifying information on them and in adherence to university policy, the interview tapes and transcribed information will be kept in a locked cabinet for at least five years, before a decision is made as to whether they should be destroyed.

Further Information

This research has been reviewed and given approval by Curtin University of Technology Human Research Ethics Committee (Approval Number SMEC20080044). If you would like further information about the study, please feel free to contact me on 0434-811-383- or by email r.d.hudson@optusnet.com.au Alternatively, you can contact my supervisor Dr. David Treagust on 08-9266-7924 or d.f.treagust@curtin.edu.au

Thank you very much for your involvement in this research.
Your participation is greatly appreciated.

PRINCIPAL'S CONSENT FORM

-
- I understand the purpose and procedures of the study.
 - I have been provided with the participation information sheet.
 - I understand that the procedure itself may not benefit me.
 - I understand that my schools involvement is voluntary and I can withdraw at any time without problem.
 - I understand that no personal identifying information will be used in any published materials.
 - I understand that all information will be securely stored for at least 5 years before a decision is made as to whether it should be destroyed.
 - I have been given the opportunity to ask questions about this research.
 - I agree to allow students form my school to participate in the study outlined to me.

Name: _____

Signature: _____

Date: _____

Curtin University of Technology
School of Engineering, Science and Computing (SMEC)

PARENT Information Sheet

My name is Ross Hudson I am currently completing a piece of research for my Doctor of Science Education at Curtin University of Technology.

Purpose of Research

I am investigating the research topic : *Multiple-choice questions compared to short-answer. Which assesses understanding of Chemistry more effectively*

Your Role

I will conduct research by asking for your Child to take part in short tests on chemistry that will complement their learning. Your Child's teachers and the College principal have already been contacted and have agreed in principle to the project. Students involved will undertake a number of short tests. The results of the tests will be given back to the students after the completion of the test. The tests will not in any way affect the students reported grades.

I may also ask for your Child's participation in a short interview (group) about their attitudes and opinions about assessment in chemistry. Again this participation will be voluntary and of short duration (10-15 mins)

Consent to Participate

Your Child's involvement in the research is entirely voluntary. You have the right to withdraw at any stage without it affecting your rights or my responsibilities. When you have signed the consent form I will assume that you have agreed to participate and allow me to use your data in this research.

Confidentiality

The information you provide will be kept separate from your personal details, and only myself and my supervisor will only have access to this. The interview transcript will not have your name or any other identifying information on it and in adherence to university policy, the interview tapes and transcribed information will be kept in a locked cabinet for at least five years, before a decision is made as to whether it should be destroyed.

Further Information

This research has been reviewed and given approval by Curtin University of Technology Human Research Ethics Committee (Approval Number SMEC20080044). If you would like further information about the study, please feel free to contact me on 0434-811-383- or by email r.d.hudson@optusnet.com.au Alternatively, you can contact my supervisor Dr. David Treagust on 08-9266-7924 or d.f.treagust@curtin.edu.au

Thank you very much for your involvement in this research.
Your participation is greatly appreciated.

PARENT CONSENT FORM

-
- I understand the purpose and procedures of the study.
 - I have been provided with the participation information sheet.
 - I understand that the procedure itself may not benefit my Child.
 - I understand that my and my Child's involvement is voluntary and I can withdraw at any time without problem.
 - I understand that no personal identifying information like my name and address will be used in any published materials.
 - I understand that all information will be securely stored for at least 5 years before a decision is made as to whether it should be destroyed.
 - I have been given the opportunity to ask questions about this research.
 - I agree to allow my Child to participate in the study outlined to me.
-

Name: _____

Student Name: _____

Signature: _____

Date: _____



**Curtin University of Technology
School of Engineering, Science and Computing (SMEC)
Participant Information Sheet**

My name is Ross Hudson I am currently completing research for my Doctor of Science Education at Curtin University of Technology.

Purpose of Research

I am investigating the research topic: *Multiple-choice questions compared to short-answer. Which assesses understanding of Chemistry more effectively*

Your Role

I will conduct research by asking for your Child to take part in short test on chemistry that will complement their learning. Your Child's teachers and the College principal have already been contacted and have agreed in principle to the project. Students involved will undertake a number of short tests. The results of the tests will be given back to the students after the completion of the test. The tests will not in any way affect the students reported grades.

I may also ask for your Child's participation in a short interview (group) about their attitudes and opinions about assessment in chemistry. Again this participation will be voluntary and of short duration (10-15 mins)

Consent to Participate

Your Child's involvement in the research is entirely voluntary. You have the right to withdraw at any stage without it affecting your rights or my responsibilities. When you have signed the consent form I will assume that you have agreed to participate and allow me to use your data in this research.

Confidentiality

The information you provide will be kept separate from your personal details, and only myself and my supervisor will only have access to this. The interview transcript will not have your name or any other identifying information on it and in adherence to university policy, the interview tapes and transcribed information will be kept in a locked cabinet for at least five years, before a decision is made as to whether it should be destroyed.

Further Information

This research has been reviewed and given approval by Curtin University of Technology Human Research Ethics Committee (Approval Number SMEC20080044). If you would like further information about the study, please feel free to contact me on 0434-811-383- or by email r.d.hudson@optusnet.com.au Alternatively, you can contact my supervisor Dr. David Treagust on 08-9266-7924 or d.f.treagust@curtin.edu.au

**Thank you very much for your involvement in this research.
Your participation is greatly appreciated.**

CONSENT FORM

- I understand the purpose and procedures of the study.
 - I have been provided with the participation information sheet.
 - I understand that the procedure itself may not benefit me.
 - I understand that my involvement is voluntary and I can withdraw at any time without problem.
 - I understand that no personal identifying information like my name and address will be used in any published materials.
 - I understand that all information will be securely stored for at least 5 years before a decision is made as to whether it should be destroyed.
 - I have been given the opportunity to ask questions about this research.
 - I agree to participate in the study outlined to me.
-

Name: _____

Signature: _____

Date: _____

Appendix C: Interview sheets, coding and transcripts

C-1: Student interview question list

1.
 - a. What type of questions have you experienced in your chemistry tests this year, multiple-choice, short-answer or both?
 - a. If yes have the test been all MC all SA or a mixture of both in the one test?
2. If you have responded to both what proportion of the test is MC and what proportion is SA? Eg 50% MC 30% MC etc
3. Does the content of the question influence your decision.
 - a. For example if the question tests recall(eg name a strong acid) then is it better for the question to be MC or SA? Please explain your response.
 - b. If the questions is an application question (eg Calculate the number of mole of something) is it better for the question to be MC or SA? Please explain your response.
4. Do MC have any advantages compared to SA questions? Please explain your response.
5. Do MC questions have any disadvantages compared to SA questions? Please explain your response.
6. Do SA have any advantages compared to MC questions? Please explain your response.
7. Do SA questions have any disadvantages compared to MC questions? Please explain your response.
8. If you had the choice would you prefer chemistry tests to be multiple-choice only or short-answer only or a combination of both.? Please explain your response.

C-2: Teacher interview question list

1. a. What type of questions have you use in your chemistry tests this year, multiple-choice, short-answer or both.?
 - a. If yes have the test been all MC all SA or a mixture of both in the one test?
2. If both, what proportion of the test is MC and what proportion is SA? Eg 50% MC 30% SA etc
3. Does the content of the question influence your decision?
 - a) For example if the question tests recall (eg name a strong acid) then is it better for the question to be MC or SA? Please explain your response.
 - b) If the question is an application question (eg Calculate the number of mole of something) is it better for the question to be MC or SA? Please explain your response.
4. Do MC have any advantages/disadvantages compared to SA questions? Please explain your response.
5. Do SA have any advantages/disadvantages compared to MC questions? Please explain your response.
6. If you had a completely free choice would you prefer chemistry tests to be multiple-choice only or short-answer only or a combination of both.? Please explain your response.

C-3 Transcripts of Student interviews

Interview transcript:

School: School-A		Student ID: D3	Sex: F
Q's	Responses		Code
1 a	Both		
b	If just a single test both		
2	40% MC, 60% SA		
3 a	MC – because it is easier to recognise an answer for example a strong acid than to have to recall it from memory		1
b	SA – because you can get marks for your working and still pick up marks even if you get the answer wrong		2
4	If you have rote learnt or fully committed something to memory, the process of elimination and knowledge you do have has an advantage in getting the question right		
5	If it takes a lot of work to work out the answer then you are not receiving a good amount of the marks for your effort		
6	If you fully understand something you are able to get multiple marks for your work		
7	It requires total recall and if you don't know or can't remember you have no chance of picking up marks deducing the possible answer		
8	Combination as this allows you to pick up marks no matter what stage your understanding is at		

School: School-A		Student ID: D6	Sex: F
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	30% MC, 70% SA		
3 a	SA – I think it should be put into a short-answer question because if there were options to choose from it would be easier pick out the correct answer		2
b	SA – because then you can have working out on the paper. If your method was right but the answer wrong you would still you would still get a mark for that.		2
4	There are going to be some things that you forget during a test. Having options to choose from can help refresh your memory		
5	Not really anything I can think of.		
6	You have to think about the questions and then think about ways to do the work to get the right answer.		
7	If you stuff your working at the start it can influence your whole answer and you could lose a lot of marks.		
8	I would have a combination of both because both help you think about the problems.		

School: School-A		Student ID: D10	Sex: F
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	40% MC, 60% SA		
3 a	MC – because you can be reassured that your answer may be right if you come up with one of the options listed		1
b	MC for the same reason as before		1
4	Yes, because the right answer will always be there and it is easier to have to select the right answers than have to come up with them yourself		
5	If you don't come up with the right answer you can't justify or explain anything that you did come up with.		
6	You can explain and justify a response. If your final answer is incorrect but your working out is correct you can gain marks for it		
7	If you have no idea of the answer it is harder to just guess.		
8	Multiple-choice only. It is much easier and reassuring. If you don't come up with the right answer it forces you to look over your working and then try to work out where you have gone wrong. This cannot happen in short-answer questions, as you can't be sure you are right.		

School: School-A		Student ID: B1	Sex: F
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	40% MC, 60% SA		
3 a	MC – the answer can be narrowed down if you are unsure		1
b	SA – it can be confusing having other numbers		2
4	Yes, if you don't know the answer you can guess		
5	If the question was asking for a definition of something you might have learnt a different one to what is being presented in front of you and it could confuse you.		
6	You can show all of your workings out in SA so even if you get the incorrect answer you might still get some marks		
7	The phrasing of the question can sometimes be confusing.		
8	Both but mostly SA		

School: School-A		Student ID: B6	Sex: F
Q's	Responses		Code
1 a	Both		
b	Mixture of both in one test		
2	30% MC, 70% SA		
3 a	MC – gives you the options in case you have forgotten or didn't know it.		1
b	MC - you get to check if the answer you find might be right if its on the MC list		1
4	Yes you have 25% chance at least at getting it right if you need to guess.		
5	Yes, sometimes can confuse you between different options		
6	It proves you really know it, with MC you could guess it all, get it right, yet have no idea what you are doing.		
7	You don't get the option of looking at which one could be your answer		
8	A combination of both, gives you a variety and a break from each type.		

School: School-A		Student ID: B4	Sex: F
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	60% MC, 40% SA		
3 a	MC, because I would be able to exclude the wrong answers and recall the correct one.		1
b	SA, because I can get marks for shouting working if I were to get the answer wrong.		2
4	Sometimes, because the answer is on the page, you just need to select it.		
5	Sometimes, more than one of the options seem correct and that can confuse people.		
6	Yes, because you can get marks for working.		
7	Yes, if you don't know the answer you cant really have a chance of getting it right, like MC		
8	I would like them to be 40 % MC and 60% SA because I seem to do better in short-answer questions.		

School: School-A		Student IDC4	Sex: F
Q's	Responses		Code
1 a	Both		
b	Both		
2	50% MC, 50% SA		
3 a	MC, because it uses recognition as opposed to just recalling the answer, In MC you have cues to the answer.		1
b	SA because there is room to do your working out and you don't feel you have to rush.		2
4	Yes because the right response is already there so you have a 25% chance of getting it right. It can also give you an idea of what you are supposed to do because you can sometimes work back from the answer		
5	They usually don't have any room to show your working out.		
6	They have room to work out and you don't feel pushed for time as much and you can see how you got your answer. Also you can get marks for the correct process even if you don't get the right answer		
7	Yes, if you don't know the answer you cant really have a chance of getting it right, like MC		
8	A mixture of both because they both have advantages and disadvantages		

School: School-A		Student ID: C8	Sex: F
Q's	Responses		Code
1 a	Both		
b	Both		
2	50% MC, 50% SA		
3 a	It is better for it to be MC because it is easier to pick the answer out of a group than it is if it is a SA where you have to try and think of the answer.		1
b	It is better if they are short-answer because there is often more than one mark for the questions so you can get marks for your working out.		2
4	They do because I am able to recognise the answer which is much easier than recall and if I run out of time in a test I can just circle any answer and I will have a chance of being correct.		
5	They are usually worth only 1 point and there never seems to be enough room for workings out. Also the answer might be worded differently to what I learnt and therefore I might not be able to recognise the right answer.		
6	I can just write my answer with my wording, if it is MC I may not recognise the right answer.		
7	They are more time consuming and if I run out of time in a test it is hard to just write anything in an attempt to get extra marks so I will probably end up leaving it blank		
8	Both they both have good points and disadvantages; they also test different methods of testing a student. SA uses recall and MC uses recognition. It is good to have a variety		

School: School-A		Student ID: C11	Sex: F
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	60% MC, 40% SA		
3 a	MC, because you know the answer is there, there is more cues and you are able to recognise the answer		1
b	MC because there are cues in the answers that might jog your memory and help you pick the right answer.		1
4	MC because they are generally easier because the answer is in front of you.		
5	Yes if you come up with the answer and its not one of the ones given then it means you have to do it again and again		
6	They are good at testing your real knowledge rather than just what you can memorise		
7	Because there are no clues you can't get any help from the question which means you can rally get stuck and end up not writing anything		
8	Combination of both because generally MC questions are really hard so it is good to have a mix of both		

School: School-A		Student ID: B5	Sex: F
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	40% MC, 50-60% SA		
3 a	MC as it triggers your memory as to which answer it is and you can eliminate questions		1
b	SA you can show your working and how you did it so you can still get marks if it is wrong		2
4	There is s set of options to choose from		
5	Often the questions are more difficult or one just general knowledge		
6	If you get the wrong final answer can you still show all your workings		
7	If you don't know what to do then there is nothing to choose from		
8	Combination of both but with the majority being SA as by doing this you can show your workings and the questions are generally easier to answer.		

School: School-A		Student ID: B8	Sex: F
Q's	Responses		Code
1 a	Both		
b	Mixture of both		
2	60% MC , 40% SA		
3 a	Multiple-choice as you have options to give you an idea but in short-answer you don't.		1
b	SA because there is more room to work it out and the rounding may be different to yours in and MC option.		2
4	MCs give you a choice of answers so you are more likely to remember the answers and you are able to eliminate improbable answers.		
5	They have less room and only the answer is worth any marks		
6	You can get marks for slightly right or almost right answers		
7	NOT ANSWERED		
8	A combination of both, or mostly SA		

School: School-A		Student ID: B15	Sex: F
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	60% MC , 40% SA		
3 a	MC – so if you have trouble recalling you can eliminate ones you definitely know are incorrect		1
b	MC – so if you make a slight miss calculation you can figure out which answer it is		1
4	Yes, because you can eliminate the answers one by one.		
5	No		
6	No		
7	Yes, if you are having trouble figuring out the questions you have nothing to help you answer		
8	I would prefer it to be mostly multiple-choice with a few short-answer		

School: School-C		Student ID: B11	Sex: M
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	25% MC 75% SA		
3 a	MC because one of the MC answers may jump out at you assisting your recall		1
b	SA because showing your working is a step you have to take anyway so it might as well contribute the answer to achieve the desired marks.		2
4	Yes, there is always a 25% chance of getting it right, which can be increased by process of elimination.		
5	The incorrect answer may seem correct through working or logic, so you circle and realise after that you missed a step for example		
6	Working through an application question is helpful in obtaining the correct answer		
7	If you don't understand the question you cannot obtain easy marks by just guessing		
8	Combination of both. You can get the best of both worlds and is the fairest way to mark a test		

School: School-C Student ID: B15 Sex: M			
Q's	Responses		Code
1 a	Both		
b	Mixture of both on major test		
2	25% MC 75% SA		
3 a	MC because you may not have memorised a specific strong acid, so the MC question gives you a fairer chance.		1
b	SA it is testing something you are supposed to strong on, and it is a step by step process.		2
4	Yes, because MC gives you (on average) a 25% chance to get the question right, however it doesn't test your knowledge as well.		
5	They don't thoroughly test your knowledge		
6	They are usually focused on topics you are strong on and are easy to complete if you have studied efficiently.		
7	If it is a complex/ discreet questions and your knowledge, is not so strong on that area, then it is easy to lose a lot of marks. They are time consuming.		
8	Combination, it tests you effectively on knowledge you are both strong and weak on.		

School: School-C Student ID: B14 Sex: M			
Q's	Responses		Code
1 a	Both		
b	A mixture in one test		
2	30% MC, 70% Short-answer		
3 a	MC – May not remember the answer but able to recognise it		1
b	Short-answer - Application of knowledge working should be shown		2
4	Yes in the aspect of theory questions, there can be no confusions, either right or wrong		
5	Yes, application questions (topic related) should be SA as working is needed, there is only one answer so MC does not have an advantage.		
6	Mathematical questions should be SA. The process of working should be part of the question as well as the answer.		
7	Harder to mark, SA can have multiple answers MC is either right or wrong		
8	Combination		

School: School-C Student ID: B12 Sex: M			
Q's	Responses		Code
1 a	Both		
b	A mixture in one test		
2	10% MC, 90% Short-answer		
3 a	If would be better if it was SA because you would be able to remember a strong acid that you've learnt in class		2
b	SA because you would be able to work through it.		2
4	Sometimes they are easier. They don't take long		
5	If they are hard they need a long time to work through it. Number of questions		
6	Yes, you can work throughout them. Have a lot of time		
7	Yes, hard		
8	Combination of both, because sometimes the multiple-choice helps you with short-answer and vice versa.		

School: School-C				Student ID: B2		Sex: M	
Q's	Responses					Code	
1	a	Both					
	b	Mixture usually					
2		30% MC , 70% SA					
3	a	MC because the answer is there on the page so it can trigger recall					1
	b	MC because that way when you work out the answer, it has to be one of the answers given, so that way you can work out if you have made a mistake or not.					1
4		Yes because if you don't know you can at least guess and because the answer is there when you read it, it triggers recall					
5		They're not worth as many marks					
6		Worth more marks					
7		If you don't know the answer, guessing is very difficult					
8		MC because it is easier to remember the answer if you read it and if you run out of time you can answer them quickly and when it comes to calculations you can check the answer.					

School: School-C				Student ID: 1		Sex: M	
Q's	Responses					Code	
1	a	Both					
	b	Mixture, usually more SA					
2		40% MC, 60% SA					
3	a	SA – don't have to solve or figure out the answer if the MC sometimes its really confusing and tricky					2
	b	MC – can use logic and guess the answer					1
4		Can guess, you at least have a correct chance of 25%					
5		Tricky answer and confusing to have to calculate each of it to make sure of the right answer					
6		Can explain the answer straight forward by ourselves freely					
7		If we don't now anything cant answer anything					
8		Both are needed MC and SA. MC sometimes has confusing answers and time consuming. Both types of questions are needed.					

School: School-C				Student ID: 3		Sex: M	
Q's	Responses					Code	
1	a	Both					
	b	Both					
2		45% MC, 55% SA					
3	a	MC - because the short-answers should have some room to explain					1
	b	SA – because it means					2
4		Can guess, you at least have a correct chance of 25% so you still have a 25% chance of getting it right.					
5		Yes, because there are always answers close to the answer but are not actually it					
6		If you do the right thing and then get it wrong, they can see that you know how to do it.					
7		If you have no idea then you can't do anything.					
8		Both it's a good mixture and good balance					

School: School-C Student ID: 2 Sex: F			
Q's	Responses		Code
1 a	Both		
b	Both		
2	40% MC, 60% SA		
3 a	If I have no idea I will just try and circle one that sounds vaguely correct but if I know the answer, I will answer it		1
b	MC so that way if its incorrect due to bad rounding I can just find the nearest answer		1
4	You have a one in four chance of getting it correct if you haven't a clue		
5	Not really they are usually OK		
6	I think I generally like SA better because I have a chance of getting some marks even if I don't fully know what to do, with MC I would have to rely on a guess.		
7	If you don't know then you have no chance of getting it correct		
8	A combination of both, because that way you have a chance to get some right but the SA you cant and it will determine the smart from the dumb.		

School: School-C Student ID: B20 Sex: F			
Q's	Responses		Code
1 a	Both		
b	Generally both		
2	25% MC, 75% SA		
3 a	MC because it gives us a range of certain answers		1
b	SA because it shows working		2
4	Yes – generally four choices = 25% of correct answer, quick and easy marks, no explanation required		
5	Yes – lose marks for not reading the question carefully, nor worth as much to overall mark, don't need to show working therefore not clear on how the answer is wrong (in some cases)		
6	Yes - shows understanding, gives more information which can help gain marks		
7	Yes, worth more marks so if unclear or don't know the answer, lots of marks can be lost. Explanations need to be precise and understood by others.		
8	MC only because I find it hard to explain myself.		

School: School-C Student ID: B18 Sex: F			
Q's	Responses		Code
1 a	Both		
b	Mixture of all		
2	30% MC , 70% SA		
3 a	MC because its something that's pretty easy and needs only 1 or ½ a mark		1
b	SA so you can get more marks for working out		2
4	Yes, because you already get the answers, it's a matter of choosing correctly		
5	No, because although they are easy they are worth about 25-30% of the exam		
6	No because answers are your own		
7	No cause if you work hard you should get the answer		
8	It would be good for a combinations of both otherwise it would be boring		

School: School-C Student ID: B24 Sex: F			
Q's	Responses		Code
1 a	Both		
b	Mixture of both		
2	30% MC and 70% SA		
3 a	MC because it just needs an answer and not much else so has a small mark		1
b	SA so you can show your working		2
4	Yes because if you do not know the answer but you know what it is not you can eliminate and have more chance		
5	Yes because you may get confused the by the answers even if you know the answer and may then pick the wrong one.		
6	Yes because you don't have similar answers		
7	Yes because if you do not know the answer you do not have a choice and can't guess		
8	Combination of both because different types of questions are more suited to MC or SA.		

School: School-C Student ID: B21 Sex: F			
Q's	Responses		Code
1 a	Both		
b	Mixtures		
2	30% MC , 70% SA, marks wise usually 10 MC questions and 8 SA questions		
3 a	MC		1
b	SA		2
4	Faster, more straight forward. Tests knowledge equally between students. No hazy marking		
5	Really similar answers can be confusing		
6	You may not know complete answers but you gave the chance to show what you do know		
7	Teacher can mark differently between students. They take ages, and if you get one answer wrong, you may stuff up the rest of the questions.		
8	Keep it how it is, it works very well.		

School: School-C Student ID: 12 Sex: F			
Q's	Responses		Code
1 a	Both		
b	Mixture of both		
2	30% MC, 70% SA		
3 a	SA because you are not given any prompts		2
b	SA because if the answer is wrong you can gain marks for formula		2
4	Yes, you are given options and they prompt your memory so it's easier to work out.		
5	They are often much harder		
6	You can gain marks for correct formula even if the answer is wrong		
7	Again, there are no prompts so less chance of guessing and getting it right		
8	Both, multiple-choice are hard but the answer is given. SA is not as hard.		

School: School-C		Student ID: B16	Sex: F
Q's	Responses		Code
1 a	Both		
b	Mixture of both		
2	30% MC and 70% SA		
3 a	MC because it is such a small but however broad questions and only worth one mark		1
b	SA shows that you know how to do calculations		2
4	Sometimes easier as answer is partially given to you, much quicker, good introduction to SA		
5	Cannot show workings Cannot show opinion / add points		
6	Allow you to show more knowledge, more in depth, allow for interpretation		
7	Take longer, can have broad answers, be misleading		
8	Mixture		

School: School-B		Student ID: B8	Sex: F
Q's	Responses		Code
1 a	Both		
b	Both		
2	50-50		
3 a	What does mean? Please reword I think this question would be more understandable if its MC ☺		1
b	SA as you have more room to work it out and you can see how you got your answer		2
4	Yes, because if your answer is not on the question choices, then obviously you're wrong		
5	Yes because some questions have more than one answer that is correct		
6	You can explain yourself more clearly and also back your answer up, just in case its wrong		
7	Yes, because if you get it wrong, you may not know		
8	Both because both have their advantages		

School: School-B		Student ID: B7	Sex: F
Q's	Responses		Code
1 a	Yes		
b	Both in one test		
2	40% MC, 60% SA		
3 a	The tests should be MC because I can then cancel out the wrong answers and choose the best answer		1
b	SA because it tests your knowledge more		2
4	Yes because you can cancel out your answers		
5	Yes because it does not test your strength its more luck than knowledge		
6	No, not really		
7	Yes, its time consuming		
8	Both, because it test your knowledge		

School: School-B Student ID: B23 Sex: F			
Q's	Responses		Code
1 a	Both		
b	Mixture of both		
2	About 30% MC		
3 a	MC the answers jog my memory		1
b	SA I get marks even if I get the final answer wrong, just for showing working out.		2
4	I can guess, if I am not sure		
5	MC sometimes does not give as many marks even when you have to do calculations		
6	You can show what you do understand and at least you might get some marks for the question		
7	Not able to match your final answer		
8	Combination, whether it is better to use MC or SA depends on the question		

School: Student ID: B13 Sex: F School-B			
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	40% MC , 60% SA		
3 a	MC		1
b	SA		2
4	You know that one of them is the answer and can guess		
5	Question word order confuses, and other answers distract		
6	Question is clear, no opportunity to be distracted by other answer		
7	If you muck up one part, you muck up the entire answer		
8	A combination of both, but more SA. SA is clearer to understand		

School: School-B Student ID: B15 Sex: F			
Q's	Responses		Code
1 a	Both		
b	Both on one test		
2	50%		
3 a	MC because your intuition answer is sometimes reconsidered when another answer you had not previously thought of is presented. I mean you get a chance to change your mind when you see the different choices in front of you		1
b	SA because you can get marks for working		2
4	You know if your on the right track (if your answer is even a possibility)		
5	Presenting more than one feasible answer which can change your mind		
6	You can at least get marks for working		
7	You don't definitely know whether or not your answer could be right. In a MC, you can get to the end and still choose the closet answer when dealing with calculations.		
8	Both, pros and cons.		

School: School-B Student ID: A 19 Sex: F			
Q's	Responses		Code
1 a	Both		
b	Both usually		
2	50% MC		
3 a	MC because you don't have to do any working out for it – it is either right or wrong therefore it is easier just to circle a letter		1
b	SA because you can show your working out and by doing this you can at least get some marks for working out if you made a silly mistake and got the wrong answer.		2
4	Yes, if you have no idea what the answer is, you can guess out of 4 possible answers. It also makes you feel more confident if you get an answer that is definitely printed on the sheet.		
5	Yes, you can't show working out which means you can't show that you at least had some idea of what the question was about		
6	Yes you can show working out		
7	Yes, there are no clues or prompts to guide you in the right direction to the correct response		
8	Combination of both, because it gives you a chance to do.		

School: School-B Student ID: A 17 Sex: F			
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	About 30% MC, 70% SA		
3 a	Multiple chance, because you can see if the option you're thinking to is a possible option		1
b	SA because you may get the wrong answer in the end but you may get some marks for the calculation		2
4	Yes because if you get an answer isn't one of the options you know its wrong		
5	Yes, you don't get method works		
6	Yes, you can get method marks, even if your final answer is incorrect		
7	Not really except that you don't have much chance of guessing.		
8	A combination		

School: School-B Student ID: A 13 Sex: F			
Q's	Responses		Code
1 a	Both		
b	Mixture of both		
2	50/50		
3 a	MC because you can eliminate incorrect responses		1
b	MC because you can check off against the answers they give you		1
4	Yes they are generally easier		
5	No		
6	Yes because sometimes they're worth a lot of easy marks.		
7	Yes if you have no idea what you're doing, but generally no.		
8	Combination because it's a good way of testing knowledge		

School: School-B		Student ID: A 5	Sex: F
Q's	Responses		Code
1 a	Both		
b	Mixture of both		
2	25% MC , 75% SA		
3 a	MC – I may recall what is the answer if I see it just in case I have a mental blank		1
b	I don't mind, it needed to be worked out either way, MC I may be able to be more sure of my answer but SA relies on myself and my instincts		1
4	Yes, I can check answer, I can work backwards		
5	No, I still know it or not MC I may be able to guess and a 1 in 4 or 1 in 5 chance of getting it correct numbers together		
6	SA makes me use my knowledge rather than connecting answers given to what I got from timings random		
7	SA – if I don't know it, I can't get it at all and I don't like not putting down an answer		
8	Combination, they both have advantages and disadvantages. I don't mind I am comfortable with both.		

School: School-B		Student ID: A 10	Sex: F
Q's	Responses		Code
1 a	I have experienced both multiple-choice and short-answers in all of my tests		
b	The tests have been a mixture of both MC and SA in the one test		
2	25% MC , 75% SA		
3 a	Probably bets as a MC because you just need to pick the correct answer and the ABCD might remind you		1
b	I prefer questions to be short-answer when I have to calculate something because then I can still get some marks if I do a mistake or an error on my calculator		2
4	MC are good for basic questions, like facts and in exams they are great when you are running out of time, you have a 25% chance of getting it right		
5	You don't get any marks if you do the working out method partially correct but the incorrect answer		
6	You can get some marks for your method even if your answer is not right		
7	NOT ANSWERED		
8	Combination of both		

School: School-B		Student ID: A 14	Sex: F
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	30% MC and 70% SA		
3 a	MC because it is an easier question and only requires about a mark		1
b	It would be better in short-answer so that you can see the working out		2
4	Yes because you can check if you have a similar answer to the MC answers		
5	Yes because you have to work out some answers and you only get 1 mark and the working out is not being corrected		
6	Yes, because you get marked on working out as well		
7	Yes because you don't get to compare your answers like MC answers to see if you have a similar / right answers		
8	Both because it is good to have a vary of MC and SA		

School: School-B		Student ID: A 7	Sex: F
Q's	Responses		Code
1 a	Combination of multiple-choice, short and extended answers		
b	As above		
2	40% MC , 60% SA		
3 a	SA with MC the other options may influence my answer		2
b	SA If it were MC I'd feel more pressure to quick through it quickly		2
4	You have the advantage of 4 possible answers , easier		
5	Prefer MC to SA, SA questions tend to be more difficult and makes it harder for me to answer them		
6	Tests individuals' ability to apply knowledge without getting any help from the question and you can show what you know and get some marks.		
7	Generally tougher		
8	Only MC, the chances of doing better increase		

School: School-B		Student ID: A 23	Sex: F
Q's	Responses		Code
1 a	Both		
b	Mixture of both		
2	About 50% MC , 50% SA		
3 a	MC because if it's a short-answer question you might become confused as to why the question is being asked and how to answer.		1
b	SA because if it is a difficult question and you make one small mistake you won't get any marks for it whereas in a SA you can get some marks for your working out		2
4	If you are unsure of the answer you have a 1 in 4 chance even if you guess and you have something to work to.		
5	If you think you have the right answer you aren't going to look over your shoulder		
6	You are more likely to work through your response step by step		
7	They take longer and you have very little idea what your answer should look like		
8	MC		

School: School-D		Student ID: C 7	Sex: M
Q's	Responses		Code
1 a	Both		
b	Mixture of both in one test		
2	30% MC 70% SA		
3 a	MC because if it is SA it would be too easy to just remember one or two strong acid		1
b	A SA question that requires working out should be best in SA where there is adequate space provided to show what you are doing.		2
4	Yes, MC has advantages compared to SA because if you do not know the answer you can still have a chance of getting it right. Furthermore, MC can refresh the student's memory with the choices given.		
5	Yes because the question are always harder than SA due to the advantages.		
6	Don't really know		
7	Only that you can show your working out and you might get some marks even if you stuff up the question.		
8	I would prefer a combination of both. Because it can balance the questions.		

School: School-D		Student ID: C 21	Sex: M
Q's	Responses		Code
1 a	Multiple-choice and short-answer		
b	Mixture		
2	20% MC		
3 a	MC because looking at the answers can prompt memory to recall the actual answer		1
b	SA because it is worth more marks		2
4	Process of elimination, and if the answer is unknown, there is a chance that the answer will be reached.		
5	Yes, because if the answer involves calculations then it is a lot of work just to get one mark.		
6	They can be worth more marks if one know the answer		
7	If the answer is unknown, then one will loose a lot of marks		
8	A bit of both would be good		

School: School-D		Student ID: C 8	Sex: M
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	40% MC 60% SA		
3 a	MC because it is easier to pull out the answer or remember depending if you remembered the word or studied it previously		1
b	MC, you can check and guess		1
4	You can do the elimination method and find an adequate response		
5	No, because it is easier to cancel knowledge or regurgitate knowledge from within		
6	Depends on the question, if you studied it previously, SA is easier		
7	If you haven't studied much or fail to remember then you can't write anything		
8	MC, easy ☺		

School: School-D		Student ID: B3	Sex: M
Q's	Responses		Code
1 a	All		
b	A mixture of both / test		
2	30% MC		
3 a	MC because correct answers are suggested rather than relying purely on recall		1
b	SA, it allows for a flow of calculation		2
4	A correct answer is present, it just has to be found		
5	There can be misleading answers designed to kill you		
6	They require more straight forward answers		
7	If you don't know the answer you can't work it out or guess.		
8	Combination. A broader means of testing		

School: School-D		Student ID: A 16	Sex: M
Q's	Responses		Code
1 a	Both		
b	Mixture of both in one test		
2	40% MC 60% SA		
3 a	MC because possible answers are provide, to give some clues		1
b	MC because you can always check with the choices to see if you're doing the question properly		1
4	Yes, as they provide given answers to check with		
5	Yes as they provided you with selected answer, and its frustrating to get an answer which doesn't match any of the answers		
6	Yes, they provide more time and space to calculate		
7	Yes, as you wont be sure whether you've gotten the correct answer or not		
8	Multiple-choice only, as the questions are shorter and less complicated, with answers to compare your calculated answer with.		

School: School-D		Student ID: A 3	Sex: M
Q's	Responses		Code
1 a	Both		
b	Mixture of both		
2	30% MC 70% SA		
3 a	MC, this is because with less options and choices it forces us to think, rather than just give a straight answer. Also no working out is needed.		1
b	I would say either, as both have their advantages and disadvantages. But mostly SA as it allows students to show steps.		2
4	For MC, if you accidentally make a calculation error, and then saw your answer was not on the choices, thus it helps you to go back and check kind of like a second chance.		
5	For MC, it influences guessing most of the time rather than attempting having a go. Also some students can guess and get it right. It doesn't show their strengths or weaknesses. Also skipping steps rather than showing how you got there. Also if you were taught the wrong method of working, it doesn't let the teacher know or you. Thus not allowing to you correct an error.		
6	SA, helps students show their working out step by step for future exams, and rewards points for you step, rather than lose all of the marks. Also once you finished the test, you can look at what went wrong in your steps thus correcting the mistakes. Also you can correct your method of approach. Also can help your vocabulary by writing.		
7	SA doesn't allow students to use methods such as the process of elimination, which influences some thinking. Also no second attempt is an option, once you read the question wrong, it's over. But this can help them learn not to repeat the mistake.		
8	Combinations of both, as some students are more skilled with MC and some with SA. Thus if a test is all MC, it can disadvantage the SA skilled students. Don't know as some know how to use various methods such as the process of elimination, don't know why, but I learned this of experience, possibly it's your DNA. Also It can test a variety of skills for the students.		

School: School-D		Student ID: A 7	Sex: M
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	30-40% MC 60-70% SA		
3 a	SA because more options, i.e. can list anything harder if multi-choice and don't know one of the option		2
b	SA because you can get marks for working if the answer is wrong also, too much working out required for not enough marks it is a MC questions, therefore should be SA		2
4	Multiple options, if you don't know the answer you can take a guess		
5	Risk of circling the wrong answer. Sometimes more than one answer is right, sometimes too much working is required for too little marks, wording of questions can be poorly worded to trip students up / multiple interpretations, some answers are ambiguous, some are chosen as the most correct answer which is subject or opinion		
6	Yes, space for working out to think through the problem, more marks for working. Even if you get the question wrong		
7	No, usually no ambiguity in the question, more marks available		
8	Short-answer only or combination of both (20-30% MC, 70-80% SA) short-answer questions test understanding more than multi choice, as in multi choice you can take a guess if you don't know the answer.		

School: School-D		Student ID: A 9	Sex: M
Q's	Responses		Code
1 a	Both		
b	Both in the one test		
2	40% MC 60% SA		
3 a	SA as recall questions are relatively simple so it is easier to get those marks in SA		2
b	MC as it is easier to make an educated guess with MC questions		1
4	Yes, because if I don't know the answer I usually have a 1 in 4 or 1 in 5 chance of getting it correct		
5	Yes, because it is easier to do working out with short-answer questions		
6	Easier to work through harder problems.		
7	If answer is unknown, you can guess		
8	I would prefer it to be mostly MC questions with some extension SA questions at the end		

School: School-D Student ID: A 17 Sex: M			
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	50% MC 50% SA		
3 a	MC may have a different but still right definition, so 1 choice of answers would help if that is the case		1
b	SA – writing down my process (i.e. working) helps in checking my answers – less chance of making mistakes		2
4	Yes – still have a chance of getting question right if you don't know, the answers may help in remembering how to do a question if you have forgotten how.		
5	Yes – working not as clear, easier to make mistakes		
6	Yes – forced to set work out more clearly, therefore make less mistakes		
7	Yes – if you don't know how to do a question all you can do is leave it blank		
8	MC only – if you don't know how to answer question, can still guess it correct		

School: School-D Student ID: B 2 Sex: M			
Q's	Responses		Code
1 a	Both		
b	Mixture of both or all SA		
2	78% SA and 23% MC		
3 a	Short-answer because people cant guess so more people get them wrong		2
b	As above		2
4	Yes, if you don't know the answer you have a chance of getting it right		
5	If you know the answer but other people don't then it can affect your chance of doing better as they can guess		
6	People that don't understand the question have less chance of getting it right		
7	If you don't understand the question you are stuffed		
8	Both – questions I don't understand should be MC, but ones I do know should be SA		

School: School-D Student ID: B 14 Sex: M			
Q's	Responses		Code
1 a	Both		
b	Use a mixture of both in the one test		
2	Approx 20% MC, 80% SA		
3 a	Better for the question to be MC, no that less time is devoted to fact recall – as this is not a good differentiate of ability		1
b	Short-answer allows for the question length to be more developed and sophisticated		2
4	Less time required, perhaps		
5	Possibility of guessing therefore knowledge not tested. More likely to misread question		
6	Allow for more sophisticated testing of student knowledge		
7	They are longer and often they are harder to follow so you can get confused		
8	A combination of both, as both fact recall and calculation questions are required for chemistry		

School: School-D		Student ID: B 18	Sex: M
Q's	Responses		Code
1	a Both types of questions		
	b Most tests have been a mixture of both multiple-choice and short-answer		
2	Please specify which test		
3	a Multiple-choice – if the question cannot be instantly answered looking at available answers may remind the examinee of the correct choice.		1
	b Short-answer – if the question requires working, as an application question, then marks can be amended for method as well as the correct answer		2
4	Yes - if there are questions that one cannot answer straight away in multiple-choice, then one may eliminate the more unlikely options and increase the probability of guessing the right answer		
5	Yes – multiple-choice questions do not, in general, give the student any chance to explain response which is marked as incorrect by general agreement.		
6	You can show your working and get some marks whereas in MC if you make a silly mistake and pick the wrong answer that's it.		
7	IF you stuck and can't think of what to do you end up writing nothing, at least in MC you can have a guess.		
8	If I have the choice, I would prefer chemistry test to be a combination of both multiple-choice and short-answers because, as I have already indicated above, the two types of questions test different parts of the brain, and therefore a combination of both question types is the best way of testing a range of learning of learning styles in a standardised exam.		

School: School-D		Student ID: C 10	Sex: M
Q's	Responses		Code
1	a Both		
	b Mixture		
2	40% MC 60% SA		
3	a Multiple-choice because you have a chance of picking the answer		1
	b SA – to prove how to calculate in order to get the result. If there's a mistake in one of the steps, probably get half of the mark given for the question.		2
4	Yes, i) enables you not to over explain your answers, ii) gets the point		
5	Yes - if the answer is wrong, you won't get half-mark but none of the marks given to the question		
6	Yes – let the examiner know why you use those calculations and how did we get the results		
7	Yes – takes time to write down the explanation		
8	Mixture because we would be able both types as there are pros and cons of each type		

School: School-D		Student ID: C 11	Sex: M
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	Generally 40% MC 60% SA		
3 a	MC – if you have a sound knowledge you need to rely less on it. As if you think of the answer before you see the answers you are more sure its correct		1
b	Short-answer – you will get more marks. Your working is not tested		2
4	If you don't know you at least have a certain percentage to get it right 25% in most cases		
5	Some can be translated to SA questions so in effect you only get one mark where as if it's SA you could get more.		
6	Same as 5. More marks generally		
7	If you don't know the answer and its worth 5 marks then you are screwed		
8	Both, we learn how to do working, for certain problems. It is better if we can put that to use.		

School: School-D		Student ID: C 14	Sex: M
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	40% MC 60% SA		
3 a	MC – gives different options, process of elimination can help choose the correct answer		1
b	SA – only one answer, easy to set out and usually short-answer gives more marks		2
4	Sometimes as obvious answer is with 3 other clearly incorrect answers		
5	Similar options can confuse		
6	There are marks awarded to working out even if the final answer is wrong		
7	If I have no idea I cant choose an answer from a list answers		
8	Both – the variety works well as it makes me have to think		

School: School-D		Student ID: C 17	Sex: M
Q's	Responses		Code
1 a	Both		
b	Yes all SA		
2	Mostly 30% MC, 70% SA		
3 a	MC because you have one in four chances of being correct		1
b	Short-answer - because you can get more marks from calculation steps		2
4	Yes, MC have one in four chance of getting correct, but SA only have one answer		
5	Yes, the mark are usually one marks but SA usually more		
6	Yes, can know step error when wrong, if step right, answer wrong, you can still get marks		
7	Yes, SA has one in infinite chance, but MC doesn't.		
8	Combination of both, cause I like the test setting now.		

School: School-D		Student ID: C 19	Sex: M
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	40% MC 60% SA		
3 a	MC – sometimes it helps to have a choice to help job the memory		1
b	MC - show working and see if youngest same answer, as a failsafe		1
4	If you have no idea on the question, there is still a chance at the mark		
5	You have to do more of them to get the same mark as a SA as they are only worth one.		
6	They give more marks for the same working out, just it is compulsory to show working		
7	They take more time and there is no 25% chance at a mark if you don't know the answer		
8	Combination – I like a variety of types of questions, otherwise I get bored.		

School: School-D		Student ID: C 22	Sex: M
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	35% MC, 65% SA		
3 a	Multi-choice – easier to recall with MC large amount learnt throughout year – so remembering recall questions (SA) can be hard		1
b	SA the opportunity arises for marks for working you can find out if your process is correct		2
4	Help work through process of getting right answer, a choice of getting it right if you're clueless		
5	Can over simplify some longer / harder questions, clueless people have a chance to get the question right		
6	Better for testing applications questions – chance for working marks		
7	Such a wide range of answers available – hard for recall questions		
8	The ideal chem. Test would use multiple-choice for recall questions (strong acid, bonding results) and application questions should be SA – (number mole, pH) this gives students the chance to show their knowledge of the processes they use.		

School: School-D Student ID: C 5 Sex: M			
Q's	Responses		Code
1 a	Both		
b	Each of these		
2	Some 100% MC / SA most are 40% MC / 60%		
3 a	Multi choice questions allow process of elimination and therefore are better off you know some but not all of the content		1
b	Calculation questions, unless short-answer, deserve more than the mark usually allotted to MC so probably SA		2
4	As question 4		
5	Often Mc questions can trick you (or you trick yourself doing them) even if you know the content		
6	If you know your stuff there are more marks available and generally SA questions are more straight forward		
7	Often it is easy to drop I mark (3/4 or 2/3) on SA questions with a minor mistake		
8	If its something I know about the content then MC 30%, SA 70% Otherwise MC.		

School: Student ID: C 20 Sex: M School-D			
Q's	Responses		Code
1 a	Both		
b	Mixture		
2	30% - 40% MC		
3 a	MC there are too many variables in SA		1
b	MC, if MC the question is worth 1 mark and I wont lose as much if it is a SA		1
4	Yes, it shows the student more about what the question is asking		
5	Yes, some answers are designed to trick the student		
6	Yes, if you know how to do them they can bring more marks then 1		
7	Yes, if you have no idea what you are doing it is impossible to start off.		
8	MC only, as there is a possibility for lesser gifted students to perform well.		

Teacher responses

School: School-C		Student ID: T	Sex: M
Q's	Responses		Code
1 a	Both		
b	Mixture of both in one test		
2	60% SA		
3 a	MC – they have to know most of them instead of 1 or 2 familiar, but they have the advantage of knowing that they only have to identify an answer		
b	SA – have to show working out which lets me know how well they are understanding the work		
4	Easy to mark, students may look at others answers. Students tend to think that MC will be easier but usually don't consider that they are not all equally difficult and they won't get a mark for being nearly right.		
5	Disadvantage: Time consuming in marking. Students often have difficulty articulating answers clearly.		
6	VCE exam format.		

School: School-D		Student ID: T	Sex: M
Q's	Responses		Code
1 a	Both		
b	Mixture of both in one test		
2	40% MC 60% SA		
3 a	No difference		
b	Students need to show equations and working out		
4	Quicker to mark, - can be used to test outside subject area. Can test applied knowledge more.		
5	SA ensures a student fully grasps the concepts, eliminates the “lucky” guess factor		
6	Combination of both		

School: School-A		Student ID: T	Sex: M
Q's	Responses		Code
1 a	Both		
b	Mixture of both in one test		
2	30-40% MC 60-70 SA sometimes however I'll use tests that are entirely one or the other		
3 a	Ideally the recall should be MC because the student has a chance of identifying the correct answer even when they can't directly recall it.		
b	SA can be either MC or SA because they have advantages in both. In MC they can work an answer out then check if their answer matches one offered. With SA they can get marks for working out and I can pick up any misconceptions that the class may have picked up.		
4	They are quicker to mark, and can test more ideas in the one test.		
5	Slow to mark but as mentioned earlier they do show where students' have misunderstandings and gives the students credit when they make a silly mistake.		
6	Combination of both, but if I felt that a test would be better as one or the other entirely I would continue to make that choice.		

School: School-B		Student ID: T	Sex: M
Q's	Responses		Code
1	a	Both	
	b	Mixture of both in one test,	
2		40% MC 60% SA	
3	a	Usually make recall MC if possible	
	b	Usually make them Short-answer so that the students can fully display their understanding	
4		Easy to mark, best for recall questions.	
5		Helps differentiate students by how well they go about answering the questions. You can see which students really know what they are doing and those who are struggling	
6		Combination of both	

C-4 Distribution of student responses question type preference

School	Question type	Males	Females	Total	MC	SA
School-D	MC	40		36		
	SA	23		27		
School-B	MC		20	20		
	SA		20	20		
School-C	MC	22	5	25		
	SA	11	6	19		
School-A	MC		26	26		
	SA		17	11		
Totals		96	94	190		
	MC	62	49		111	76
	SA	34	42			

Percentage distribution:

Group	Percentage who favoured multiple-choice	Percentage who favoured short-answer
Males	66	34
Females	54	46
All students	59	41

Chi Sq test

Actual	Males	Females	Total
Percentage who favoured multiple-choice	62	51	113
Percentage who favoured short-answer	34	43	77
Total	96	94	190

Expected	Males	Females
Percentage who favoured multiple-choice	57.0	54.0
Percentage who favoured short-answer	39.0	37.0

Chi ² calculations	Males	Females
	0.4415398	0.46580023
	0.6448805	0.68031349
Chi test	0.13513212	
Chi INV	3.84145534	
Chi ²	2.23253403	

Appendix D: VCE Chemistry question breakdown for Units 3 and 4

D-1: 2003 VCE Examination Unit 3 Question breakdown

2003 Unit 3 examination

Part A

Recall Questions		Application Questions	
Question Number	% correct	Question Number	% correct
1	80.0	3	45.0
2	90.0	4	86.0
6	77.0	5	62.0
7	65.0	8	49.0
9	94.0	10	64.0
11	80.0	13	70.0
12	79.0	14	67.0
16	46.0	15	57.0
18	92.0	17	67.0
		19	64.0
		20	46.0
Average mark and SD	78.1	Average mark and SD	61.5
SD	15.0	SD	12.0

2003 Unit 3 examination Part B

Recall Questions				Application Questions			
Average correct	Total marks		%	Average correct	Total marks		%
Question Number				Question Number			
1a	0.8	1	80.0	1b	1.25	2	62.5
1d	0.81	2	40.5	1c	1.57	6	26.2
2d	1.09	2	54.5	2a	0.81	1	81.0
3a	0.52	1	52.0	2b	0.87	1	87.0
4a	1.44	2	72.0	2c	2.15	3	71.7
5b	0.75	1	75.0	3b	1.37	3	45.7
5c	1.25	2	62.5	4b	0.89	1	89.0
				4c	1.25	2	62.5
				5a	1.59	3	53.0
				6a	1.81	3	60.3
				6b	1.33	3	44.3
				7a	1.35	3	45.0
				7b	1.1	2	55.0
				7c	1.04	2	52.0
Total	6.66	11	60.5		18.38	35	52.5
%	60				53		
Average mark and SD			62.4	Average mark and SD			59.7
SD			14.2	SD			17.8

D-2: 2003 VCE Examination Unit 4 Question breakdown

2003 Unit 4 examination

part A

Question Number	Recall Questions	Question Number	Application Questions
	% correct		% correct
7	65.0	1	82.0
8	62.0	2	88.0
10	59.0	3	73.0
11	71.0	4	67.0
12	79.0	5	49.0
13	75.0	6	61.0
14	70.0	9	76.0
15	58.0	17	50.0
16	83.0	18	48.0
20	62.0	19	33.0
Average mark and SD	68.4	Average mark and SD	62.7
SD	8.6	SD	17.5

2003 Unit 4 examination Part B

Recall Questions	Average correct	Total marks	%	Application Questions	Average correct	Total marks	%
Question Number				Question Number			
1a	0.63	1	63.0	1b	0.97	2	48.5
1c	1.23	2	61.5	2c	1.13	2	56.5
2a	0.68	1	68.0	2d	1.09	2	54.5
2b	0.67	1	67.0	3a	1.52	2	76.0
5a	0.55	1	55.0	3b	1.98	3	66.0
5b	0.6	1	60.0	3c	1.29	2	64.5
5c	1.31	2	65.5	4a	1.51	2	75.5
5d	1.39	3	46.3	4b	1.13	2	56.5
6a	1.51	2	75.5	4c	1.31	2	65.5
6b	1.25	2	62.5	7c	1.53	3	51.0
7a	1.36	4	34.0	7d	0.15	1	15.0
7b	1.1	2	55.0	7e	0.77	1	77.0
8a	0.62	1	62.0	8c	0.49	1	49.0
8c	1.98	3	66.0	8d	0.56	2	28.0
Total	14.88	26			15.43	27	
%	57.2				57.1		
Average mark and SD			60.1	Average mark and SD			56.0
SD			10.2	SD			17.7

D-3: 2004 VCE Examination Unit 3 Question breakdown

2004 Unit 3 examination

Part A

Recall Questions		Application Questions	
Question Number	% correct	Question Number	% correct
1	77.0	3	68.0
2	76.0	4	58.0
11	65.0	5	47.0
12	43.0	6	71.0
19	77.0	7	58.0
		8	64.0
		9	63.0
		10	55.0
		13	41.0
		14	24.0
		15	69.0
		16	59.0
		17	49.0
		18	66.0
		20	50.0
Average mark 67.6 and SD		Average mark 56.1 and SD	
SD	14.7	SD	12.5

2004 Unit 3 examination Part B

Recall Questions	Average correct	Total marks	%	Application Questions	Average correct	Total marks	%
Question Number				Question Number			
1bi	0.58	1	58.0	1a	0.81	1	81.0
1bii	0.33	1	33.0	1bv	1.61	2	80.5
1biii	0.34	1	34.0	2ai	0.72	1	72.0
1biv	0.64	1	64.0	2aii	0.79	1	79.0
5a	0.9	1	90.0	2aiii	0.75	1	75.0
5bi	0.6	1	60.0	2aiv	0.51	1	51.0
5bii	1.29	2	64.5	2b	1.36	2	68.0
5biii	0.72	1	72.0	2ci	0.57	2	28.5
5ci	0.71	1	71.0	2cii	0.6	1	60.0
5cii	0.52	1	52.0	2d	0.52	1	52.0
5d	0.63	1	63.0	3	2.58	5	51.6
				4ai	0.79	1	79.0
				4aii	0.39	1	39.0
				4aiii	0.91	1	91.0
				4bi	0.91	1	91.0
				4bii	0.39	1	39.0
				4biii	0.86	1	86.0
				4biv	0.18	2	9.0
				6a	0.96	2	48.0
				6b	2.22	4	55.5
				6c	1.17	2	58.5
Average mark and SD				Average mark and SD			
			60.1				61.6
SD				SD			
			16.4				21.7

D-4: 2004 VCE Examination Unit 4 Question breakdown

2004 Unit 4 examination part A

Question Number	Recall Questions	Question Number	Application Questions
	% correct		% correct
1	80.0	4	45.0
2	65.0	6	72.0
3	80.0	7	52.0
5	77.0	9	57.0
8	53.0	10	58.0
11	40.0	14	51.0
12	58.0	16	67.0
13	58.0		
15	82.0		
17	83.0		
18	67.0		
19	73.0		
20	78.0		
Average mark and SD	68.8	Average mark and SD	57.4
SD	13.3	SD	9.4

2004 Unit 4 examination Part B

Recall Questions	Average correct	Total marks		Application Questions	Average correct	Total marks	
			%				%
Question Number				Question Number			
1	3.2	5	64.0	2b	1.6	2	80.0
2a	0.9	1	90.0	2c	0.7	1	70.0
4a	0.4	1	40.0	3ai	2.2	3	73.3
4bii	0.9	1	90.0	3aii	1.2	3	40.0
4biii	0.5	1	50.0	3bi	1.5	2	75.0
5cii	0.5	1	50.0	3bii	0.3	1	30.0
6a	0.8	2	40.0	4bi	0.7	1	70.0
6b	2.7	4	67.5	5a	0.5	1	50.0
6ci	0.7	1	70.0	5bi	0.7	1	70.0
6cii	0.6	1	60.0	5bii	0.7	1	70.0
7a	0.6	2	30.0	5ci	0.4	1	40.0
7b	1.2	2	60.0	8b	1.5	3	50.0
8a	0.8	1	80.0	9ai	1.2	2	60.0
8c	1.1	2	55.0	9aii	0.4	1	40.0
9bi	1.4	2	70.0	9bii	0.9	2	45.0
9biii	1.3	2	65.0				
Average mark and SD			61.3	Average mark and SD			57.6
SD			17.1	SD			16.1

D-5: 2005 VCE Examination Unit 3 Question breakdown

2005 Unit 3 examination Part A

Recall Questions		Application Questions	
Question Number	% correct	Question Number	% correct
1	84.0	4	58.0
2	86.0	5	52.0
3	86.0	6	34.0
7	42.0	8	47.0
9	55.0	10	70.0
16	69.0	11	59.0
18	79.0	12	39.0
19	55.0	13	36.0
20	61.0	14	61.0
		15	33.0
		17	69.0
Average mark and SD	68.6	Average mark and SD	50.7
SD	16.1	SD	13.8

**2005 Unit 3
examination**

Part B

Recall Questions	Average correct	Total marks	%	Application Questions	Average correct	Total marks	%
Question Number				Question Number			
1a	0.6	1	60.0	1e	0.9	2	45.0
1b	0.7	2	35.0	2b	2.6	4	65.0
1c	0.5	1	50.0	3a	3.2	4	80.0
1d	0.5	1	50.0	3b	1.2	2	60.0
1e	0.9	2	45.0	3c	1.3	2	65.0
2a	0.6	1	60.0	4a	0.6	1	60.0
2c	3.3	4	82.5	4b	0.5	1	50.0
3d	1.8	3	60.0	6a	0.9	1	90.0
4c	1.1	2	55.0	6b	1.8	3	60.0
4d	1.4	2	70.0	6c	1	2	50.0
5a	0.6	1	60.0	7b	1.6	3	53.3
5b	2.6	4	65.0	8c	0.4	1	40.0
5c	1.1	2	55.0				
5d	0.5	1	50.0				
7a	0.7	1	70.0				
7c	0.8	3	26.7				
8a	0.8	1	80.0				
8b	0.8	1	80.0				
Total	19.3	33	58.5		16	26	61.5
%	58				62		
Average mark and SD			58.6	Average mark and SD			59.9
SD			15.0	SD			14.2

D-6: 2005 VCE Examination Unit 4 Question breakdown

**2005 Unit 4
examination**

part A

Recall Questions		Application Questions	
Question Number	% correct	Question Number	% correct
1	62.0	4	42.0
2	79.0	5	61.0
3	59.0	6	72.0
12	54.0	7	43.0
15	71.0	8	55.0
16	68.0	9	67.0
17	70.0	10	74.0
20	61.0	11	45.0
		13	54.0
		14	50.0
		18	41.0
		19	66.0
Average mark and SD	65.5	Average mark and SD	55.8
SD	8.0	SD	12.0

**2005 Unit 4
examination**

Part B

Recall Questions	Average correct	Total marks		Application Questions	Average correct	Total marks	
			%				%
Question Number				Question Number			
1	4.1	5	82.0	3a	2.3	4	57.5
2a	0.8	1	80.0	5c	1.3	2	65.0
2ci	0.6	1	60.0	2b	2.2	3	73.3
2cii	1.1	2	55.0	4aii	1.1	2	55.0
2ciii	0.8	1	80.0	7b	1	2	50.0
2civ	0.8	1	80.0	7c	0.4	1	40.0
2cv	0.6	1	60.0	7d	0.5	1	50.0
3b	1.2	2	60.0	8a	0.8	2	40.0
3c	3.4	4	85.0	8b	1.4	4	35.0
4ai	0.4	1	40.0				
4b	2.7	4	67.5				
5a	2.3	3	76.7				
5b	0.9	1	90.0				
6a	0.3	1	30.0				
6b	0.3	1	30.0				
6c	1	2	50.0				
7a	0.8	1	80.0				
8c	1	2	50.0				
9a	1.4	2	70.0				
9b	0.4	1	40.0				
9c	0.3	1	30.0				
9d	0.8	2	40.0				
	22.1	32	69.1		11	21	52.4
	69				52		
Average mark and SD			60.7	Average mark and SD			51.8
SD			19.7	SD			12.5

D-7: 2006 VCE Examination Unit 3 Question breakdown

**2006 Unit 3
examination**

Part A

		Recall Questions	Application Questions
Question Number	% correct	Question Number	% correct
1	73.0	2	63.0
3	86.0	5	28.0
4	71.0	6	78.0
9	45.0	7	90.0
15	86.0	8	69.0
16	78.0	10	50.0
17	39.0	11	71.0
		12	37.0
		13	55.0
		14	62.0
		18	59.0
		19	64.0
		20	35.0
Average mark and 68.3 SD		Average mark and SD	58.5
SD	18.9	SD	17.6

**2006 Unit 3
examination**

Part B

Recall Questions				Application Questions			
Question Number	Average correct	Total marks	%	Question Number	Average correct	Total marks	%
1	2.5	4	62.5	3a	0.9	1	90.0
2a	1.6	2	80.0	3b	0.4	1	40.0
2b	0.7	1	70.0	3c	0.6	1	60.0
2c	0.7	1	70.0	3d	1.8	3	60.0
4a	0.1	1	10.0	3e	3.1	5	62.0
5a	1.5	2	75.0	4b	2.9	7	41.4
7a	0.9	1	90.0	5b	1.8	3	60.0
7b	0.5	1	50.0	5c	1.3	2	65.0
7c	1.7	3	56.7	6a	6.4	8	80.0
8a	0.7	1	70.0	6b	0.7	1	70.0
8b	0.8	1	80.0	7d	2.3	4	57.5
8c	1.4	2	70.0				
8d	2.5	5	50.0				
Total	15.6	25	62.4		22.2	36	61.7
%	62.4				62		
Average mark and SD			64.2	Average mark and SD			62.4
SD			20.0	SD			14.6

D-8: 2006 VCE Examination Unit 4 Question breakdown

2006 Unit 4 examination

part A

		Recall Questions	Application Questions	
Question Number	% correct	Question Number	% correct	
1	93.0	3	67.0	
2	95.0	4	60.0	
5	75.0	7	53.0	
6	67.0	8	57.0	
10	57.0	9	56.0	
12	66.0	11	48.0	
13	85.0	20	45.0	
14	66.0			
15	78.0			
16	84.0			
17	73.0			
18	80.0			
19	73.0			
Average mark and SD	76.3	Average mark and SD	55.1	
SD	11.1	SD	7.4	

2006 Unit 4 examination Part B

Recall Questions	Average correct	Total marks	%	Application Questions	Average correct	Total marks	%
Question Number				Question Number			
1a	1.6	2	80.0	2a	0.9	1	90.0
1b	1.4	3	46.7	3a	1.2	2	60.0
2b	0.5	1	50.0	3b	2	3	66.7
2c	1.4	2	70.0	4a	0.8	1	80.0
2d	1.6	2	80.0	4b	0.8	1	80.0
3c	0.7	1	70.0	4c	0.9	2	45.0
5a	0.5	1	50.0	4d	1.2	2	60.0
5b	1.4	2	70.0	4e	1.6	3	53.3
6a	0.7	1	70.0	5c	0.5	2	25.0
6b	0.5	1	50.0	5d	2.6	5	52.0
6c	0.8	2	40.0	6d	0.4	1	40.0
7a	0.3	1	30.0	8a	1	2	50.0
7b	0.8	1	80.0	8b	0.6	2	30.0
8c	0.7	1	70.0	8d	0.6	1	60.0
9d	0.7	1	70.0	9a	1.8	3	60.0
				9b	0.8	2	40.0
				9c	1	2	50.0
Total	13.6	22			18.7	35	
%	61.8				53.4		
Average mark and SD			61.8	Average mark and SD			58.5
SD			15.9	SD			18.5

D-9: 2007 VCE Examination Unit 3 Question breakdown

**2007 Unit 3
examination**

Part A

Question Number	Recall Questions	Question Number	Application Questions
	% correct		% correct
1	77.0	3	56.0
2	78.0	4	66.0
5	66.0	6	62.0
9	48.0	7	49.0
11	71.0	8	71.0
13	82.0	10	70.0
18	30.0	12	77.0
		14	56.0
		15	39.0
		16	23.0
		17	51.0
		19	79.0
		20	30.0
Average mark and SD	64.6	Average mark and SD	56.1
SD	18.9	SD	17.5

2007 Unit 3 examination

Part B

Recall Questions	Average correct	Total marks		Application Questions	Average correct	Total marks	
Question Number			%	Question Number			%
1b	0.8	1	80.0	1a	1.2	2	60.0
1c	2.3	3	76.7	2bi	1.3	2	65.0
2a	0.9	1	90.0	2bii	0.3	1	30.0
4c	0.8	1	80.0	2biii	0.4	1	40.0
4d	1.9	3	63.3	3a	1.4	2	70.0
5ai	0.7	1	70.0	3b	1.7	3	56.7
5aii	0.8	1	80.0	3c	0.3	1	30.0
5ci	0.7	1	70.0	4ai	0.6	1	60.0
5cii	0.5	1	50.0	4aii	0.5	1	50.0
5ciii	0.5	1	50.0	4aiii	1.1	2	55.0
				4b	0.5	1	50.0
				5b	0.8	1	80.0
				5di	2.1	4	52.5
				5dii	1.0	2	50.0
				6a	0.5	1	50.0
				6b	0.6	1	60.0
				6c	0.8	2	40.0
				6d	0.8	2	40.0
				7a	2.3	3	76.7
				7bi	1.3	2	65.0
				7bii	0.7	1	70.0
				7biii	1.6	3	53.3
Total	9.9	14			21.8	39	
%	70.7				55.9		
Average mark and SD				Average mark and SD			
			71.0				54.7
SD			13.2	SD			13.5

D-10: 2007 VCE Examination Unit 4 Question breakdown

**2007 Unit 4
examination**

part A

Recall Questions		Application Questions	
Question Number	% correct	Question Number	% correct
2	58.0	1	65.0
5	71.0	3	73.0
6	78.0	4	51.0
7	77.0	9	41.0
8	50.0	10	51.0
12	66.0	11	67.0
13	56.0	15	44.0
14	64.0	16	44.0
20	35.0	17	58.0
		18	83.0
		19	42.0
Average mark and SD		Average mark and SD	
	61.7		56.3
SD	13.7	SD	14.1

2007 Unit 4 examination Part B

Recall Questions	Average correct	Total marks	%	Application Questions	Average correct	Total marks	%
Question Number				Question Number			
2a	1.4	2	70.0	1	3.4	6	56.7
2b	5.5	8	68.8	3a	1.6	3	53.3
4d	1.4	2	70.0	3b	0.7	1	70.0
5a	1.5	3	50.0	4a	2.1	4	52.5
7a	1.7	3	56.7	4b	0.7	2	35.0
7b	0.9	2	45.0	4c	1.4	3	46.7
7c	2.2	6	36.7	5b	3	5	60.0
				6a	2.7	5	54.0
				6b	0.7	2	35.0
				8a	0.7	1	70.0
				8b	1.9	3	63.3
				8c	1.2	2	60.0
				8d	0.3	1	30.0
Total	14.6	26			20.4	38	
%	56.2				53.7		
Average mark and SD			56.7	Average mark and SD			52.8
SD			13.4	SD			13.0

D-11: 2003-2007 VCE Examination Unit 3 Examinations combined Question breakdown

Part A Exams Semester 1					Part B Exams Semester 1				
Year	Q no.	Score	Class	Type	Year	Q no.	Score	Class	Type
2003	1	80	Recall	MC	2003	1a	80.0	Recall	SA
	2	90	Recall	MC		1d	40.5	Recall	SA
	6	77	Recall	MC		2d	54.5	Recall	SA
	7	65	Recall	MC		3a	52.0	Recall	SA
	9	94	Recall	MC		4a	72.0	Recall	SA
	11	80	Recall	MC		5b	75.0	Recall	SA
	12	79	Recall	MC		5c	62.5	Recall	SA
	16	46	Recall	MC		1b	62.5	Application	SA
	18	92	Recall	MC		1c	26.2	Application	SA
	3	45.0	Application	MC		2a	81.0	Application	SA
	4	86.0	Application	MC		2b	87.0	Application	SA
	5	62.0	Application	MC		2c	71.7	Application	SA
	8	49.0	Application	MC		3b	45.7	Application	SA
	10	64.0	Application	MC		4b	89.0	Application	SA
	13	70.0	Application	MC		4c	62.5	Application	SA
	14	67.0	Application	MC		5a	53.0	Application	SA
	15	57.0	Application	MC		6a	60.3	Application	SA
	17	67.0	Application	MC		6b	44.3	Application	SA
	19	64.0	Application	MC		7a	45.0	Application	SA
	20	46.0	Application	MC		7b	55.0	Application	SA
2004	1	77	Recall	MC	2004	7c	52.0	Application	SA
	2	76	Recall	MC		1bi	58.0	Recall	SA
	11	65	Recall	MC		1bii	33.0	Recall	SA
	12	43	Recall	MC		1biii	34.0	Recall	SA
	19	77	Recall	MC		1biv	64.0	Recall	SA
	3	68.0	Application	MC		5a	90.0	Recall	SA
	4	58.0	Application	MC		5bi	60.0	Recall	SA
	5	47.0	Application	MC		5bii	64.5	Recall	SA
	6	71.0	Application	MC		5biii	72.0	Recall	SA
	7	58.0	Application	MC		5ci	71.0	Recall	SA
	8	64.0	Application	MC		5cii	52.0	Recall	SA
	9	63.0	Application	MC		5d	63.0	Recall	SA
	10	55.0	Application	MC		1a	81.0	Application	SA
	13	41.0	Application	MC		1bv	80.5	Application	SA
14	24.0	Application	MC	2ai	72.0	Application	SA		

	15	69.0	Application	MC		2aii	79.0	Application	SA
	16	59.0	Application	MC		2aiii	75.0	Application	SA
	17	49.0	Application	MC		2aiv	51.0	Application	SA
	18	66.0	Application	MC		2b	68.0	Application	SA
	20	50.0	Application	MC		2ci	28.5	Application	SA
2005	1	84	Recall	MC		2cii	60.0	Application	SA
	2	86	Recall	MC		2d	52.0	Application	SA
	3	86	Recall	MC		3	51.6	Application	SA
	7	42	Recall	MC		4ai	79.0	Application	SA
	9	55	Recall	MC		4aii	39.0	Application	SA
	16	69	Recall	MC		4aiii	91.0	Application	SA
	18	79	Recall	MC		4bi	91.0	Application	SA
	19	55	Recall	MC		4bii	39.0	Application	SA
	20	61	Recall	MC		4biii	86.0	Application	SA
	4	58.0	Application	MC		4biv	9.0	Application	SA
	5	52.0	Application	MC		6a	48.0	Application	SA
	6	34.0	Application	MC		6b	55.5	Application	SA
	8	47.0	Application	MC		6c	58.5	Application	SA
	10	70.0	Application	MC	2005	1a	60.0	Recall	SA
	11	59.0	Application	MC		1b	35.0	Recall	SA
	12	39.0	Application	MC		1c	50.0	Recall	SA
	13	36.0	Application	MC		1d	50.0	Recall	SA
	14	61.0	Application	MC		1e	45.0	Recall	SA
	15	33.0	Application	MC		2a	60.0	Recall	SA
	17	69.0	Application	MC		2c	82.5	Recall	SA
2006	1	84	Recall	MC		3d	60.0	Recall	SA
	2	86	Recall	MC		4c	55.0	Recall	SA
	3	86	Recall	MC		4d	70.0	Recall	SA
	7	42	Recall	MC		5a	60.0	Recall	SA
	9	55	Recall	MC		5b	65.0	Recall	SA
	16	69	Recall	MC		5c	55.0	Recall	SA
	18	79	Recall	MC		5d	50.0	Recall	SA
	19	55	Recall	MC		7a	70.0	Recall	SA
	20	61	Recall	MC		7c	26.7	Recall	SA
	4	58.0	Application	MC		8a	80.0	Recall	SA
	5	52.0	Application	MC		8b	80.0	Recall	SA
	6	34.0	Application	MC		1e	45.0	Application	SA
	8	47.0	Application	MC		2b	65.0	Application	SA
	10	70.0	Application	MC		3a	80.0	Application	SA
	11	59.0	Application	MC		3b	60.0	Application	SA
	12	39.0	Application	MC		3c	65.0	Application	SA
	13	36.0	Application	MC		4a	60.0	Application	SA

2007	14	61.0	Application	MC	2006	4b	50.0	Application	SA
	15	33.0	Application	MC		6a	90.0	Application	SA
	17	69.0	Application	MC		6b	60.0	Application	SA
	1	77.0	Recall	MC		6c	50.0	Application	SA
	2	78.0	Recall	MC		7b	53.3	Application	SA
	5	66.0	Recall	MC		8c	40.0	Application	SA
	9	48.0	Recall	MC		1a	60.0	Recall	SA
	11	71.0	Recall	MC		1b	35.0	Recall	SA
	13	82.0	Recall	MC		1c	50.0	Recall	SA
	18	30.0	Recall	MC		1d	50.0	Recall	SA
	3	56.0	Application	MC		1e	45.0	Recall	SA
	4	66.0	Application	MC		2a	60.0	Recall	SA
	6	62.0	Application	MC		2c	82.5	Recall	SA
	7	49.0	Application	MC		3d	60.0	Recall	SA
	8	71.0	Application	MC		4c	55.0	Recall	SA
	10	70.0	Application	MC		4d	70.0	Recall	SA
	12	77.0	Application	MC		5a	60.0	Recall	SA
	14	56.0	Application	MC		5b	65.0	Recall	SA
	15	39.0	Application	MC		5c	55.0	Recall	SA
	16	23.0	Application	MC		5d	50.0	Recall	SA
	17	51.0	Application	MC		7a	70.0	Recall	SA
	19	79.0	Application	MC		7c	26.7	Recall	SA
	20	30.0	Application	MC		8a	80.0	Recall	SA
						8b	80.0	Recall	SA
						1e	45.0	Application	SA
						2b	65.0	Application	SA
						3a	80.0	Application	SA
				3b	60.0	Application	SA		
				3c	65.0	Application	SA		
				4a	60.0	Application	SA		
				4b	50.0	Application	SA		
				6a	90.0	Application	SA		
				6b	60.0	Application	SA		
				6c	50.0	Application	SA		
				7b	53.3	Application	SA		
				8c	40.0	Application	SA		
				1b	80.0	Recall	SA		
				1c	76.7	Recall	SA		
				2a	90.0	Recall	SA		
				4c	80.0	Recall	SA		
				4d	63.3	Recall	SA		
				5ai	70.0	Recall	SA		

5aii	80.0	Recall	SA
5ci	70.0	Recall	SA
5cii	50.0	Recall	SA
5ciii	50.0	Recall	SA
1a	60.0	Application	SA
2bi	65.0	Application	SA
2bii	30.0	Application	SA
2biii	40.0	Application	SA
3a	70.0	Application	SA
3b	56.7	Application	SA
3c	30.0	Application	SA
4ai	60.0	Application	SA
4aii	50.0	Application	SA
4aiii	55.0	Application	SA
4b	50.0	Application	SA
5b	80.0	Application	SA
5di	52.5	Application	SA
5dii	50.0	Application	SA
6a	50.0	Application	SA
6b	60.0	Application	SA
6c	40.0	Application	SA
6d	40.0	Application	SA
7a	76.7	Application	SA
7bi	65.0	Application	SA
7bii	70.0	Application	SA
7biii	53.3	Application	SA

D-12: 2003-2007 VCE Examination Unit 4 Examinations combined Question breakdown

Part A Exams					Part B Exams				
Year	Semester 2				Year	Semester 2			
	Q no.	Score	Class	Type		Q no.	Score	Class	Type
2003	7	65	Recall	MC	2003	1a	63.0	Recall	SA
	8	62	Recall	MC		1c	61.5	Recall	SA
	10	59	Recall	MC		2a	68.0	Recall	SA
	11	71	Recall	MC		2b	67.0	Recall	SA
	12	79	Recall	MC		5a	55.0	Recall	SA
	13	75	Recall	MC		5b	60.0	Recall	SA
	14	70	Recall	MC		5c	65.5	Recall	SA
	15	58	Recall	MC		5d	46.3	Recall	SA
	16	83	Recall	MC		6a	75.5	Recall	SA
	20	62	Recall	MC		6b	62.5	Recall	SA
	1	82	Application	MC		7a	34.0	Recall	SA
	2	88	Application	MC		7b	55.0	Recall	SA
	3	73	Application	MC		8a	62.0	Recall	SA
	4	67	Application	MC		8c	66.0	Recall	SA
	5	49.0	Application	MC		1b	48.5	Application	SA
	6	61.0	Application	MC		2c	56.5	Application	SA
	9	76.0	Application	MC		2d	54.5	Application	SA
	17	50.0	Application	MC		3a	76.0	Application	SA
	18	48.0	Application	MC		3b	66.0	Application	SA
	19	33.0	Application	MC		3c	64.5	Application	SA
2004	1	80	Recall	MC	4a	75.5	Application	SA	
	2	65	Recall	MC	4b	56.5	Application	SA	
	3	80	Recall	MC	4c	65.5	Application	SA	
	5	77	Recall	MC	7c	51.0	Application	SA	
	8	53	Recall	MC	7d	15.0	Application	SA	
	11	40	Recall	MC	7e	77.0	Application	SA	
	12	58	Recall	MC	8c	49.0	Application	SA	
	13	58	Recall	MC	8d	28.0	Application	SA	
	15	82	Recall	MC	2004	1	64.0	Recall	SA
	17	83	Recall	MC		2a	90.0	Recall	SA
	18	67	Recall	MC		4a	40.0	Recall	SA
	19	73	Recall	MC		4bii	90.0	Recall	SA
	20	78	Recall	MC		4biii	50.0	Recall	SA
	4	45.0	Application	MC		5cii	50.0	Recall	SA
	6	72.0	Application	MC		6a	40.0	Recall	SA

	7	52.0	Application	MC		6b	67.5	Recall	SA
	9	57.0	Application	MC		6ci	70.0	Recall	SA
	10	58.0	Application	MC		6cii	60.0	Recall	SA
	14	51.0	Application	MC		7a	30.0	Recall	SA
	16	67.0	Application	MC		7b	60.0	Recall	SA
2005	1	62	Recall	MC		8a	80.0	Recall	SA
	2	79	Recall	MC		8c	55.0	Recall	SA
	3	59	Recall	MC		9bi	70.0	Recall	SA
	12	54	Recall	MC		9biii	65.0	Recall	SA
	15	71	Recall	MC		2b	80.0	Application	SA
	16	68	Recall	MC		2c	70.0	Application	SA
	17	70	Recall	MC		3ai	73.3	Application	SA
	20	61	Recall	MC		3aii	40.0	Application	SA
	4	42.0	Application	MC		3bi	75.0	Application	SA
	5	61.0	Application	MC		3bii	30.0	Application	SA
	6	72.0	Application	MC		4bi	70.0	Application	SA
	7	43.0	Application	MC		5a	50.0	Application	SA
	8	55.0	Application	MC		5bi	70.0	Application	SA
	9	67.0	Application	MC		5bii	70.0	Application	SA
	10	74.0	Application	MC		5ci	40.0	Application	SA
	11	45.0	Application	MC		8b	50.0	Application	SA
	13	54.0	Application	MC		9ai	60.0	Application	SA
	14	50.0	Application	MC		9aii	40.0	Application	SA
	18	41.0	Application	MC		9bii	45.0	Application	SA
	19	66.0	Application	MC	2005	1	82.0	Recall	SA
2006	1	93.0	Recall	MC		2a	80.0	Recall	SA
	2	95.0	Recall	MC		2ci	60.0	Recall	SA
	5	75.0	Recall	MC		2cii	55.0	Recall	SA
	6	67.0	Recall	MC		2ciii	80.0	Recall	SA
	10	57.0	Recall	MC		2civ	80.0	Recall	SA
	12	66.0	Recall	MC		2cv	60.0	Recall	SA
	13	85.0	Recall	MC		3b	60.0	Recall	SA
	14	66.0	Recall	MC		3c	85.0	Recall	SA
	15	78.0	Recall	MC		4ai	40.0	Recall	SA
	16	84.0	Recall	MC		4b	67.5	Recall	SA
	17	73.0	Recall	MC		5a	76.7	Recall	SA
	18	80.0	Recall	MC		5b	90.0	Recall	SA
	19	73.0	Recall	MC		6a	30.0	Recall	SA
	3	67.0	Application	MC		6b	30.0	Recall	SA
	4	60.0	Application	MC		6c	50.0	Recall	SA
	7	53.0	Application	MC		7a	80.0	Recall	SA
	8	57.0	Application	MC		8c	50.0	Recall	SA

	9	56.0	Application	MC		9a	70.0	Recall	SA
	11	48.0	Application	MC		9b	40.0	Recall	SA
	20	45.0	Application	MC		9c	30.0	Recall	SA
2007	2	58.0	Recall	MC		9d	40.0	Recall	SA
	5	71.0	Recall	MC		3a	57.5	Application	SA
	6	78.0	Recall	MC		5c	65.0	Application	SA
	7	77.0	Recall	MC		2b	73.3	Application	SA
	8	50.0	Recall	MC		4a	55.0	Application	SA
	12	66.0	Recall	MC		7b	50.0	Application	SA
	13	56.0	Recall	MC		7c	40.0	Application	SA
	14	64.0	Recall	MC		7d	50.0	Application	SA
	20	35.0	Recall	MC		8a	40.0	Application	SA
	1	65.0	Application	MC		8b	35.0	Application	SA
	3	73.0	Application	MC	2006	1a	80.0	Recall	SA
	4	51.0	Application	MC		1b	46.7	Recall	SA
	9	41.0	Application	MC		2b	50.0	Recall	SA
	10	51.0	Application	MC		2c	70.0	Recall	SA
	11	67.0	Application	MC		2d	80.0	Recall	SA
	15	44.0	Application	MC		3c	70.0	Recall	SA
	16	44.0	Application	MC		5a	50.0	Recall	SA
	17	58.0	Application	MC		5b	70.0	Recall	SA
	18	83.0	Application	MC		6a	70.0	Recall	SA
	19	42.0	Application	MC		6b	50.0	Recall	SA
						6c	40.0	Recall	SA
						7a	30.0	Recall	SA
						7b	80.0	Recall	SA
						8c	70.0	Recall	SA
						9d	70.0	Recall	SA
						2a	90.0	Application	SA
						3a	60.0	Application	SA
						3b	66.7	Application	SA
						4a	80.0	Application	SA
						4b	80.0	Application	SA
						4c	45.0	Application	SA
						4d	60.0	Application	SA
						4e	53.3	Application	SA
						5c	25.0	Application	SA
						5d	52.0	Application	SA
						6d	40.0	Application	SA
						8a	50.0	Application	SA
						8b	30.0	Application	SA
						8d	60.0	Application	SA

	9a	60.0	Application	SA
	9b	40.0	Application	SA
	9c	50.0	Application	SA
2007	2a	70.0	Recall	SA
	2b	68.8	Recall	SA
	4d	70.0	Recall	SA
	5a	50.0	Recall	SA
	7a	56.7	Recall	SA
	7b	45.0	Recall	SA
	7c	36.7	Recall	SA
	1	56.7	Application	SA
	3a	53.3	Application	SA
	3b	70.0	Application	SA
	4a	52.5	Application	SA
	4b	35.0	Application	SA
	4c	46.7	Application	SA
	5b	60.0	Application	SA
	6a	54.0	Application	SA
	6b	35.0	Application	SA
	8a	70.0	Application	SA
	8b	63.3	Application	SA
	8c	60.0	Application	SA
	8d	30.0	Application	SA

D-13: 2003-2007 VCE Examination Unit 3 &4 All Examinations combined

Question breakdown

Q no.	Score	Class	Type
1	82	Application	MC
2	88	Application	MC
3	73	Application	MC
4	67	Application	MC
5	49.0	Application	MC
6	61.0	Application	MC
9	76.0	Application	MC
17	50.0	Application	MC
18	48.0	Application	MC
19	33.0	Application	MC
4	45.0	Application	MC
6	72.0	Application	MC
7	52.0	Application	MC
9	57.0	Application	MC
10	58.0	Application	MC
14	51.0	Application	MC
16	67.0	Application	MC
4	42.0	Application	MC
5	61.0	Application	MC
6	72.0	Application	MC
7	43.0	Application	MC
8	55.0	Application	MC
9	67.0	Application	MC
10	74.0	Application	MC
11	45.0	Application	MC
13	54.0	Application	MC
14	50.0	Application	MC
18	41.0	Application	MC
19	66.0	Application	MC
3	67.0	Application	MC
4	60.0	Application	MC
7	53.0	Application	MC
8	57.0	Application	MC
9	56.0	Application	MC
11	48.0	Application	MC
20	45.0	Application	MC
1	65.0	Application	MC

3	73.0	Application MC
4	51.0	Application MC
9	41.0	Application MC
10	51.0	Application MC
11	67.0	Application MC
15	44.0	Application MC
16	44.0	Application MC
17	58.0	Application MC
18	83.0	Application MC
19	42.0	Application MC
3	45.0	Application MC
4	86.0	Application MC
5	62.0	Application MC
8	49.0	Application MC
10	64.0	Application MC
13	70.0	Application MC
14	67.0	Application MC
15	57.0	Application MC
17	67.0	Application MC
19	64.0	Application MC
20	46.0	Application MC
3	68.0	Application MC
4	58.0	Application MC
5	47.0	Application MC
6	71.0	Application MC
7	58.0	Application MC
8	64.0	Application MC
9	63.0	Application MC
10	55.0	Application MC
13	41.0	Application MC
14	24.0	Application MC
15	69.0	Application MC
16	59.0	Application MC
17	49.0	Application MC
18	66.0	Application MC
20	50.0	Application MC
4	58.0	Application MC
5	52.0	Application MC
6	34.0	Application MC
8	47.0	Application MC
10	70.0	Application MC
11	59.0	Application MC

12	39.0	Application MC
13	36.0	Application MC
14	61.0	Application MC
15	33.0	Application MC
17	69.0	Application MC
4	58.0	Application MC
5	52.0	Application MC
6	34.0	Application MC
8	47.0	Application MC
10	70.0	Application MC
11	59.0	Application MC
12	39.0	Application MC
13	36.0	Application MC
14	61.0	Application MC
15	33.0	Application MC
17	69.0	Application MC
3	56.0	Application MC
4	66.0	Application MC
6	62.0	Application MC
7	49.0	Application MC
8	71.0	Application MC
10	70.0	Application MC
12	77.0	Application MC
14	56.0	Application MC
15	39.0	Application MC
16	23.0	Application MC
17	51.0	Application MC
19	79.0	Application MC
20	30.0	Application MC
1b	62.5	Application SA
1c	26.2	Application SA
2a	81.0	Application SA
2b	87.0	Application SA
2c	71.7	Application SA
3b	45.7	Application SA
4b	89.0	Application SA
4c	62.5	Application SA
5a	53.0	Application SA
6a	60.3	Application SA
6b	44.3	Application SA
7a	45.0	Application SA
7b	55.0	Application SA

7c	52.0	Application SA
1a	81.0	Application SA
1bv	80.5	Application SA
2ai	72.0	Application SA
2aii	79.0	Application SA
2aiii	75.0	Application SA
2aiv	51.0	Application SA
2b	68.0	Application SA
2ci	28.5	Application SA
2cii	60.0	Application SA
2d	52.0	Application SA
3	51.6	Application SA
4ai	79.0	Application SA
4aii	39.0	Application SA
4aiii	91.0	Application SA
4bi	91.0	Application SA
4bii	39.0	Application SA
4biii	86.0	Application SA
4biv	9.0	Application SA
6a	48.0	Application SA
6b	55.5	Application SA
6c	58.5	Application SA
1e	45.0	Application SA
2b	65.0	Application SA
3a	80.0	Application SA
3b	60.0	Application SA
3c	65.0	Application SA
4a	60.0	Application SA
4b	50.0	Application SA
6a	90.0	Application SA
6b	60.0	Application SA
6c	50.0	Application SA
7b	53.3	Application SA
8c	40.0	Application SA
1e	45.0	Application SA
2b	65.0	Application SA
3a	80.0	Application SA
3b	60.0	Application SA
3c	65.0	Application SA
4a	60.0	Application SA
4b	50.0	Application SA
6a	90.0	Application SA

6b	60.0	Application SA
6c	50.0	Application SA
7b	53.3	Application SA
8c	40.0	Application SA
1a	60.0	Application SA
2bi	65.0	Application SA
2bii	30.0	Application SA
2biii	40.0	Application SA
3a	70.0	Application SA
3b	56.7	Application SA
3c	30.0	Application SA
4ai	60.0	Application SA
4aii	50.0	Application SA
4aiii	55.0	Application SA
4b	50.0	Application SA
5b	80.0	Application SA
5di	52.5	Application SA
5dii	50.0	Application SA
6a	50.0	Application SA
6b	60.0	Application SA
6c	40.0	Application SA
6d	40.0	Application SA
7a	76.7	Application SA
7bi	65.0	Application SA
7bii	70.0	Application SA
7biii	53.3	Application SA
1b	48.5	Application SA
2c	56.5	Application SA
2d	54.5	Application SA
3a	76.0	Application SA
3b	66.0	Application SA
3c	64.5	Application SA
4a	75.5	Application SA
4b	56.5	Application SA
4c	65.5	Application SA
7c	51.0	Application SA
7d	15.0	Application SA
7e	77.0	Application SA
8c	49.0	Application SA
8d	28.0	Application SA
2b	80.0	Application SA
2c	70.0	Application SA

3ai	73.3	Application SA
3aii	40.0	Application SA
3bi	75.0	Application SA
3bii	30.0	Application SA
4bi	70.0	Application SA
5a	50.0	Application SA
5bi	70.0	Application SA
5bii	70.0	Application SA
5ci	40.0	Application SA
8b	50.0	Application SA
9ai	60.0	Application SA
9aii	40.0	Application SA
9bii	45.0	Application SA
3a	57.5	Application SA
5c	65.0	Application SA
2b	73.3	Application SA
4aii	55.0	Application SA
7b	50.0	Application SA
7c	40.0	Application SA
7d	50.0	Application SA
8a	40.0	Application SA
8b	35.0	Application SA
2a	90.0	Application SA
3a	60.0	Application SA
3b	66.7	Application SA
4a	80.0	Application SA
4b	80.0	Application SA
4c	45.0	Application SA
4d	60.0	Application SA
4e	53.3	Application SA
5c	25.0	Application SA
5d	52.0	Application SA
6d	40.0	Application SA
8a	50.0	Application SA
8b	30.0	Application SA
8d	60.0	Application SA
9a	60.0	Application SA
9b	40.0	Application SA
9c	50.0	Application SA
1	56.7	Application SA
3a	53.3	Application SA
3b	70.0	Application SA

4a	52.5	Application	SA
4b	35.0	Application	SA
4c	46.7	Application	SA
5b	60.0	Application	SA
6a	54.0	Application	SA
6b	35.0	Application	SA
8a	70.0	Application	SA
8b	63.3	Application	SA
8c	60.0	Application	SA
8d	30.0	Application	SA
7	65	Recall	MC
8	62	Recall	MC
10	59	Recall	MC
11	71	Recall	MC
12	79	Recall	MC
13	75	Recall	MC
14	70	Recall	MC
15	58	Recall	MC
16	83	Recall	MC
20	62	Recall	MC
1	80	Recall	MC
2	65	Recall	MC
3	80	Recall	MC
5	77	Recall	MC
8	53	Recall	MC
11	40	Recall	MC
12	58	Recall	MC
13	58	Recall	MC
15	82	Recall	MC
17	83	Recall	MC
18	67	Recall	MC
19	73	Recall	MC
20	78	Recall	MC
1	62	Recall	MC
2	79	Recall	MC
3	59	Recall	MC
12	54	Recall	MC
15	71	Recall	MC
16	68	Recall	MC
17	70	Recall	MC
20	61	Recall	MC
1	93.0	Recall	MC

2	95.0	Recall	MC
5	75.0	Recall	MC
6	67.0	Recall	MC
10	57.0	Recall	MC
12	66.0	Recall	MC
13	85.0	Recall	MC
14	66.0	Recall	MC
15	78.0	Recall	MC
16	84.0	Recall	MC
17	73.0	Recall	MC
18	80.0	Recall	MC
19	73.0	Recall	MC
2	58.0	Recall	MC
5	71.0	Recall	MC
6	78.0	Recall	MC
7	77.0	Recall	MC
8	50.0	Recall	MC
12	66.0	Recall	MC
13	56.0	Recall	MC
14	64.0	Recall	MC
20	35.0	Recall	MC
1	80	Recall	MC
2	90	Recall	MC
6	77	Recall	MC
7	65	Recall	MC
9	94	Recall	MC
11	80	Recall	MC
12	79	Recall	MC
16	46	Recall	MC
18	92	Recall	MC
1	77	Recall	MC
2	76	Recall	MC
11	65	Recall	MC
12	43	Recall	MC
19	77	Recall	MC
1	84	Recall	MC
2	86	Recall	MC
3	86	Recall	MC
7	42	Recall	MC
9	55	Recall	MC
16	69	Recall	MC
18	79	Recall	MC

19	55	Recall	MC
20	61	Recall	MC
1	84	Recall	MC
2	86	Recall	MC
3	86	Recall	MC
7	42	Recall	MC
9	55	Recall	MC
16	69	Recall	MC
18	79	Recall	MC
19	55	Recall	MC
20	61	Recall	MC
1	77.0	Recall	MC
2	78.0	Recall	MC
5	66.0	Recall	MC
9	48.0	Recall	MC
11	71.0	Recall	MC
13	82.0	Recall	MC
18	30.0	Recall	MC
1a	80.0	Recall	SA
1d	40.5	Recall	SA
2d	54.5	Recall	SA
3a	52.0	Recall	SA
4a	72.0	Recall	SA
5b	75.0	Recall	SA
5c	62.5	Recall	SA
1bi	58.0	Recall	SA
1bii	33.0	Recall	SA
1biii	34.0	Recall	SA
1biv	64.0	Recall	SA
5a	90.0	Recall	SA
5bi	60.0	Recall	SA
5bii	64.5	Recall	SA
5biii	72.0	Recall	SA
5ci	71.0	Recall	SA
5cii	52.0	Recall	SA
5d	63.0	Recall	SA
1a	60.0	Recall	SA
1b	35.0	Recall	SA
1c	50.0	Recall	SA
1d	50.0	Recall	SA
1e	45.0	Recall	SA
2a	60.0	Recall	SA
2c	82.5	Recall	SA

3d	60.0	Recall	SA
4c	55.0	Recall	SA
4d	70.0	Recall	SA
5a	60.0	Recall	SA
5b	65.0	Recall	SA
5c	55.0	Recall	SA
5d	50.0	Recall	SA
7a	70.0	Recall	SA
7c	26.7	Recall	SA
8a	80.0	Recall	SA
8b	80.0	Recall	SA
1a	60.0	Recall	SA
1b	35.0	Recall	SA
1c	50.0	Recall	SA
1d	50.0	Recall	SA
1e	45.0	Recall	SA
2a	60.0	Recall	SA
2c	82.5	Recall	SA
3d	60.0	Recall	SA
4c	55.0	Recall	SA
4d	70.0	Recall	SA
5a	60.0	Recall	SA
5b	65.0	Recall	SA
5c	55.0	Recall	SA
5d	50.0	Recall	SA
7a	70.0	Recall	SA
7c	26.7	Recall	SA
8a	80.0	Recall	SA
8b	80.0	Recall	SA
1b	80.0	Recall	SA
1c	76.7	Recall	SA
2a	90.0	Recall	SA
4c	80.0	Recall	SA
4d	63.3	Recall	SA
5ai	70.0	Recall	SA
5a _{ii}	80.0	Recall	SA
5ci	70.0	Recall	SA
5cii	50.0	Recall	SA
5c _{iii}	50.0	Recall	SA
1a	63.0	Recall	SA
1c	61.5	Recall	SA
2a	68.0	Recall	SA

2b	67.0	Recall	SA
5a	55.0	Recall	SA
5b	60.0	Recall	SA
5c	65.5	Recall	SA
5d	46.3	Recall	SA
6a	75.5	Recall	SA
6b	62.5	Recall	SA
7a	34.0	Recall	SA
7b	55.0	Recall	SA
8a	62.0	Recall	SA
8c	66.0	Recall	SA
1	64.0	Recall	SA
2a	90.0	Recall	SA
4a	40.0	Recall	SA
4bii	90.0	Recall	SA
4biii	50.0	Recall	SA
5cii	50.0	Recall	SA
6a	40.0	Recall	SA
6b	67.5	Recall	SA
6ci	70.0	Recall	SA
6cii	60.0	Recall	SA
7a	30.0	Recall	SA
7b	60.0	Recall	SA
8a	80.0	Recall	SA
8c	55.0	Recall	SA
9bi	70.0	Recall	SA
9biii	65.0	Recall	SA
1	82.0	Recall	SA
2a	80.0	Recall	SA
2ci	60.0	Recall	SA
2cii	55.0	Recall	SA
2ciii	80.0	Recall	SA
2civ	80.0	Recall	SA
2cv	60.0	Recall	SA
3b	60.0	Recall	SA
3c	85.0	Recall	SA
4ai	40.0	Recall	SA
4b	67.5	Recall	SA
5a	76.7	Recall	SA
5b	90.0	Recall	SA
6a	30.0	Recall	SA
6b	30.0	Recall	SA

6c	50.0	Recall	SA
7a	80.0	Recall	SA
8c	50.0	Recall	SA
9a	70.0	Recall	SA
9b	40.0	Recall	SA
9c	30.0	Recall	SA
9d	40.0	Recall	SA
1a	80.0	Recall	SA
1b	46.7	Recall	SA
2b	50.0	Recall	SA
2c	70.0	Recall	SA
2d	80.0	Recall	SA
3c	70.0	Recall	SA
5a	50.0	Recall	SA
5b	70.0	Recall	SA
6a	70.0	Recall	SA
6b	50.0	Recall	SA
6c	40.0	Recall	SA
7a	30.0	Recall	SA
7b	80.0	Recall	SA
8c	70.0	Recall	SA
9d	70.0	Recall	SA
2a	70.0	Recall	SA
2b	68.8	Recall	SA
4d	70.0	Recall	SA
5a	50.0	Recall	SA
7a	56.7	Recall	SA
7b	45.0	Recall	SA
7c	36.7	Recall	SA

Appendix E: SPSS ANOVA results and Eta-squared results

E-1: SPSS Anova results comparing Recall v Application Multiple-choice questions

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Score * Class	200	100.0%	0	.0%	200	100.0%

Report

Class	Mean	N	Std. Deviation	Median
Application	56.1852	108	13.55580	57.0000
Recall	69.3043	92	13.81728	71.0000
Total	62.2200	200	15.13529	62.5000

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Score * Class	Between Groups	(Combined)	8550.545	1	8550.545	45.713	.000
	Within Groups		37035.775	198	187.049		
	Total		45586.320	199			

Measures of Association

	Eta	Eta Squared
Score * Class	.433	.188

Notes: Significant F value sig 0.000 and a coefficient of determination of 18.8%

E-2: SPSS Anova results comparing Recall v Application Short-answer questions

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Score * Class	287	100.0%	0	.0%	287	100.0%

Report

Class	Mean	N	Std. Deviation	Median
Application	57.1255	149	16.24296	56.5000
Recall	60.8645	138	15.51527	60.7500
Total	58.9233	287	15.97946	60.0000

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Score * Class	Between Groups	(Combined)	1001.594	1	1001.594	3.963	.047
	Within Groups		72026.499	285	252.725		
	Total		73028.094	286			

Measures of Association

	Eta	Eta Squared
Score * Class	.117	.014

Note: result is significant but only to < than 0.05, Coefficient of determination is very small only 1.4%. Little difference between the scores.

E-3: SPSS Anova results comparing Short-answer v Multiple-choice Application questions

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Score * Type	257	100.0%	0	.0%	257	100.0%

Score

Type	Mean	N	Std. Deviation	Median
MC	56.1852	108	13.55580	57.0000
SA	57.1255	149	16.24296	56.5000
Total	56.7304	257	15.15094	56.7000

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Score * Type	Between Groups	(Combined)	55.364	1	55.364	.240	.624
	Within Groups		58709.679	255	230.234		
	Total		58765.043	256			

Measures of Association

	Eta	Eta Squared
Score * Type	.031	.001

Very small F value results not significant value (0.6). Coefficient of Det is very small.

E-4: SPSS Anova results comparing Short-answer v Multiple-choice Recall questions

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Score * Type	230	89.5%	27	10.5%	257	100.0%

Report

Score

Type	Mean	N	Std. Deviation	Median
MC	69.3043	92	13.81728	71.0000
SA	60.8645	138	15.51527	60.7500
Total	64.2404	230	15.39644	65.0000

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Score * Type	Between Groups	(Combined)	3931.960	1	3931.960	17.804	.000
	Within Groups		50352.594	228	220.845		
	Total		54284.554	229			

Measures of Association

	Eta	Eta Squared
Score * Type	.269	.072

Solid F value that is significant. sig 0.000. Coefficient of determination is only 7.2%

E-5: SPSS Anova results comparing Application v Recall All questions

Comparing Application vs recall all Qs

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Score * Class	487	100.0%	0	.0%	487	100.0%

Report

Class	Mean	N	Std. Deviation	Median
Application	56.7304	257	15.15094	56.7000
Recall	64.2404	230	15.39644	65.0000
Total	60.2772	487	15.70663	60.0000

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Score * Class	Between Groups	(Combined)	6845.760	1	6845.760	29.369	.000
	Within Groups		113049.597	485	233.092		
	Total		119895.357	486			

Measures of Association

	Eta	Eta Squared
Score * Class	.239	.057

Note: F value suggests significance. sig 0.000. Coefficient of Det is only 5.7%

E-6: SPSS Anova results comparing Short-answer v Multiple-choice All Questions

Comparing SA vs MC all Qs

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Score * Type	487	100.0%	0	.0%	487	100.0%

Report

Type	Mean	N	Std. Deviation	Median
MC	62.2200	200	15.13529	62.5000
SA	58.9233	287	15.97946	60.0000
Total	60.2772	487	15.70663	60.0000

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Score * Type	Between Groups	(Combined)	1280.943	1	1280.943	5.238	.023
	Within Groups		118614.414	485	244.566		
	Total		119895.357	486			

Measures of Association

	Eta	Eta Squared
Score * Type	.103	.011

Note: F value fairly small but result is significant with a sig of 0.023. the co eff of det is just 1.1%.

Appendix F: SPSS Chi-squared results comparing gender and reported grades

F-1: SPSS Chi-squared results comparing Males v Females for A⁺, C⁺ and E⁺ grades 2003 Unit 3 Examination

NPar Tests comparing Males v Females for A⁺ grades 2003 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Frequency A ⁺	869	438.6571	42.32056	392.00	477.00

Chi-Square Test

Frequencies

Frequency A⁺

A ⁺	Observed N	Expected N	Residual
Male	477	400.0	77.0
Female	392	469.0	-77.0
Total	869		

Test Statistics

	Frequency A ⁺
Chi-Square(a)	27.464
df	1
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 400.0.

NPar Tests comparing Males v Females for C⁺ grades 2003 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq C ⁺	1204	609.2359	65.62941	536.00	668.00

Chi-Square Test

Frequencies

Freq C⁺

C ⁺	Observed N	Expected N	Residual
Male	536	555.0	-19.0
Female	668	649.0	19.0
Total	1204		

Test Statistics

	Freq C ⁺
Chi-Square(a)	1.207
df	1
Asymp. Sig.	.272

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 555.0.

NPar Tests comparing Males v Females for E⁺ grades 2003 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq E ⁺	500	254.3560	32.74400	217.00	283.00

Chi-Square Test

Frequencies

Freq E⁺

E ⁺	Observed N	Expected N	Residual
Male	217	230.0	-13.0
Female	283	270.0	13.0
Total	500		

Test Statistics

	Freq E ⁺
Chi-Square(a)	1.361
df	1
Asymp. Sig.	.243

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 230.0.

F-2: SPSS Chi-squared results comparing Males v Females for A⁺, C⁺ and E⁺ grades 2003 Unit 4 Examination

NPar Tests comparing Males v Females for A⁺ grades 2003 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Frequency A ⁺	804	402.0896	6.00307	396.00	408.00

Chi-Square Test

Frequencies

Frequency A⁺

A ⁺	Observed N	Expected N	Residual
Male	408	370.0	38.0
Female	396	434.0	-38.0
Total	804		

Test Statistics

	Frequency A ⁺
Chi-Square(a)	7.230
df	1
Asymp. Sig.	.007

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 370.0.

NPar Tests comparing Males v Females for C⁺ grades 2003 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq C ⁺	1033	530.6534	84.36125	431.00	602.00

Chi-Square Test

Frequencies

Freq C⁺

C ⁺	Observed N	Expected N	Residual
Male	431	475.0	-44.0
Female	602	558.0	44.0
Total	1033		

Test Statistics

	Freq C ⁺
Chi-Square(a)	7.545
df	1
Asymp. Sig.	.006

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 475.0.

NPar Tests comparing Males v Females for E⁺ grades 2003 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq E ⁺	493	246.6227	5.50422	241.00	252.00

Chi-Square Test

Frequencies

Freq E⁺

E ⁺	Observed N	Expected N	Residual
Male	241	227.0	14.0
Female	252	266.0	-14.0
Total	493		

Test Statistics

	Freq E ⁺
Chi-Square(a)	1.600
df	1
Asymp. Sig.	.206

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 227.0.

F-3: SPSS Chi-squared results comparing Males v Females for A⁺, C⁺ and E⁺ grades 2004 Unit 3 Examination

NPar Tests comparing Males v Females for A⁺ grades 2004 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Frequency A ⁺	894	450.9463	41.83760	405.00	489.00

Chi-Square Test

Frequencies

Frequency A⁺

A ⁺	Observed N	Expected N	Residual
Male	489	417.0	72.0
Female	405	477.0	-72.0
Total	894		

Test Statistics

	Frequency A ⁺
Chi-Square(a)	23.300
df	1
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 417.0.

NPar Tests comparing Males v Females for C⁺ grades 2004 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq C ⁺	1139	580.5979	78.75616	490.00	649.00

Chi-Square Test

Frequencies

Freq C⁺

C ⁺	Observed N	Expected N	Residual
Male	490	531.0	-41.0
Female	649	608.0	41.0
Total	1139		

Test Statistics

	Freq C ⁺
Chi-Square(a)	5.931
df	1
Asymp. Sig.	.015

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 531.0.

NPar Tests comparing Males v Females for E⁺ grades 2004 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq E ⁺	472	238.0508	21.92744	214.00	258.00

Chi-Square Test

Frequencies

Freq E⁺

E ⁺	Observed N	Expected N	Residual
Male	214	220.0	-6.0
Female	258	252.0	6.0
Total	472		

Test Statistics

	Freq E ⁺
Chi-Square(a)	.306
df	1
Asymp. Sig.	.580

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 220.0.

F-4: SPSS Chi-squared results comparing Males v Females for A⁺, C⁺ and E⁺ grades 2004 Unit 4 Examination

NPar Tests comparing Males v Females for A⁺ grades 2004 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Frequency A ⁺	899	449.6607	8.50321	441.00	458.00

Chi-Square Test

Frequencies

Frequency

A ⁺	Observed N	Expected N	Residual
Male	458	418.0	40.0
Female	441	481.0	-40.0
Total	899		

Test Statistics

	Frequency A ⁺
Chi-Square(a)	7.154
df	1
Asymp. Sig.	.007

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 418.0.

NPar Tests comparing Males v Females for C⁺ grades 2004 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq C ⁺	1151	581.4466	58.22228	517.00	634.00

Chi-Square Test

Frequencies

Freq C⁺

C ⁺	Observed N	Expected N	Residual
Male	517	535.0	-18.0
Female	634	616.0	18.0
Total	1151		

Test Statistics

	Freq C ⁺
Chi-Square(a)	1.132
df	1
Asymp. Sig.	.287

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 535.0.

NPar Tests comparing Males v Females for E⁺ grades 2004 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq E ⁺	486	243.0658	4.00358	239.00	247.00

Chi-Square Test

Frequencies

Freq E⁺

E ⁺	Observed N	Expected N	Residual
Male	239	226.0	13.0
Female	247	260.0	-13.0
Total	486		

Test Statistics

	Freq E ⁺
Chi-Square(a)	1.398
df	1
Asymp. Sig.	.237

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 226.0.

F-5: SPSS Chi-squared results comparing Males v Females for A⁺, C⁺ and E⁺ grades 2005 Unit 3 Examination

NPar Tests comparing Males v Females for A⁺ grades 2005 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Frequency A ⁺	873	444.6105	58.97842	377.00	496.00

Chi-Square Test

Frequencies

Frequency A⁺

A ⁺	Observed N	Expected N	Residual
Male	496	407.0	89.0
Female	377	466.0	-89.0
Total	873		

Test Statistics

	Frequency A ⁺
Chi-Square(a)	36.460
df	1
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 407.0.

NPar Tests comparing Males v Females for C⁺ grades 2005 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq C ⁺	1182	601.0321	76.37599	514.00	668.00

Chi-Square Test

Frequencies

Freq C⁺

C ⁺	Observed N	Expected N	Residual
Male	514	551.0	-37.0
Female	668	631.0	37.0
Total	1182		

Test Statistics

	Freq C ⁺
Chi-Square(a)	4.654
df	1
Asymp. Sig.	.031

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 551.0.

NPar Tests comparing Males v Females for E⁺ grades 2005 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq E ⁺	534	267.1348	6.00411	261.00	273.00

Chi-Square Test

Frequencies

Freq E⁺

E+	Observed N	Expected N	Residual
Male	261	249.0	12.0
Female	273	285.0	-12.0
Total	534		

Test Statistics

	Freq E ⁺
Chi-Square(a)	1.084
df	1
Asymp. Sig.	.298

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 249.0.

F-6: SPSS Chi-squared results comparing Males v Females for A⁺, C⁺ and E⁺ grades 2005 Unit 4 Examination

NPar Tests comparing Males v Females for A⁺ grades 2005 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Frequency A ⁺	908	454.0088	2.00108	452.00	456.00

Chi-Square Test

Frequencies

Frequency A⁺

A ⁺	Observed N	Expected N	Residual
Male	456	423.0	33.0
Female	452	485.0	-33.0
Total	908		

Test Statistics

	Frequency A ⁺
Chi-Square(a)	4.820
df	1
Asymp. Sig.	.028

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 423.0.

NPar Tests comparing Males v Females for C⁺ grades 2005 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq C ⁺	1095	554.1854	60.15697	487.00	608.00

Chi-Square Test

Frequencies

Freq C⁺

C ⁺	Observed N	Expected N	Residual
Male	487	510.0	-23.0
Female	608	585.0	23.0
Total	1095		

Test Statistics

	Freq C ⁺
Chi-Square(a)	1.942
df	1
Asymp. Sig.	.164

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 510.0.

NPar Tests comparing Males v Females for E+ grades 2005 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq E+	517	259.1044	12.49747	246.00	271.00

Chi-Square Test

Frequencies

Freq E+

E+	Observed N	Expected N	Residual
Male	246	241.0	5.0
Female	271	276.0	-5.0
Total	517		

Test Statistics

	Freq E+
Chi-Square(a)	.194
df	1
Asymp. Sig.	.659

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 241.0.

F-7: SPSS Chi-squared results comparing Males v Females for A⁺, C⁺ and E⁺ grades 2006 Unit 3 Examination

NPar Tests comparing Males v Females for A⁺ grades 2006 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Frequency	864	448.7245	83.38668	347.00	517.00

Chi-Square Test

Frequencies

Frequency A⁺

A ⁺	Observed N	Expected N	Residual
Male	517	411.0	106.0
Female	347	453.0	-106.0
Total	864		

Test Statistics

	Frequency A ⁺
Chi-Square(a)	52.142
df	1
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 411.0.

NPar Tests comparing Males v Females for C⁺ grades 2006 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq C ⁺	1246	625.6982	40.92754	582.00	664.00

Chi-Square Test

Frequencies

FreqC⁺

C ⁺	Observed N	Expected N	Residual
Male	582	592.0	-10.0
Female	664	654.0	10.0
Total	1246		

Test Statistics

	Freq C ⁺
Chi-Square(a)	.322
df	1
Asymp. Sig.	.571

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 592.0.

NPar Tests comparing Males v Females for E⁺ grades 2006 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq E ⁺	467	241.2355	41.83490	191.00	276.00

Chi-Square Test

Frequencies

Freq E⁺

E ⁺	Observed N	Expected N	Residual
Male	191	222.0	-31.0
Female	276	245.0	31.0
Total	467		

Test Statistics

	Freq E ⁺
Chi-Square(a)	8.251
df	1
Asymp. Sig.	.004

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 222.0.

F-8: SPSS Chi-squared results comparing Males v Females for A⁺, C⁺ and E⁺ grades 2006 Unit 4 Examination

NPar Tests comparing Males v Females for A⁺ grades 2006 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Frequency A ⁺	897	457.2096	61.92470	386.00	511.00

Chi-Square Test

Frequencies

Frequency A⁺

A ⁺	Observed N	Expected N	Residual
Male	511	425.0	86.0
Female	386	472.0	-86.0
Total	897		

Test Statistics

	Frequency A ⁺
Chi-Square(a)	33.072
df	1
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 425.0.

NPar Tests comparing Males v Females for C⁺ grades 2006 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq C ⁺	1070	537.5589	36.92867	498.00	572.00

Chi-Square Test

Frequencies

Freq C⁺

C ⁺	Observed N	Expected N	Residual
Male	498	507.0	-9.0
Female	572	563.0	9.0
Total	1070		

Test Statistics

	Freq C ⁺
Chi-Square(a)	.304
df	1
Asymp. Sig.	.582

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 507.0.

NPar Tests comparing Males v Females for E⁺ grades 2006 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq E ⁺	549	275.3752	15.48938	259.00	290.00

Chi-Square Test

Frequencies

Freq E⁺

E ⁺	Observed N	Expected N	Residual
Male	259	260.0	-1.0
Female	290	289.0	1.0
Total	549		

Test Statistics

	Freq E ⁺
Chi-Square(a)	.007
df	1
Asymp. Sig.	.932

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 260.0.

F-9: SPSS Chi-squared results comparing Males v Females for A⁺, C⁺ and E⁺ grades 2007 Unit 3 Examination

NPar Tests comparing Males v Females for A⁺ grades 2007 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Frequency A ⁺	885	462.2565	91.44057	349.00	536.00

Chi-Square Test

Frequencies

Frequency A⁺

A ⁺	Observed N	Expected N	Residual
Male	536	426.0	110.0
Female	349	459.0	-110.0
Total	885		

Test Statistics

	Frequency A ⁺
Chi-Square(a)	54.765
df	1
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 426.0.

NPar Tests comparing Males v Females for C⁺ grades 2007 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq C ⁺	1182	597.7157	62.66755	528.00	654.00

Chi-Square Test

Frequencies

Freq C⁺

C ⁺	Observed N	Expected N	Residual
Male	528	569.0	-41.0
Female	654	613.0	41.0
Total	1182		

Test Statistics

	Freq C ⁺
Chi-Square(a)	5.697
df	1
Asymp. Sig.	.017

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 569.0.

NPar Tests comparing Males v Females for E⁺ grades 2007 Unit 3 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq E ⁺	468	238.9402	33.67518	200.00	268.00

Chi-Square Test

Frequencies

Freq E⁺

E ⁺	Observed N	Expected N	Residual
Male	200	225.0	-25.0
Female	268	243.0	25.0
Total	468		

Test Statistics

	Freq E ⁺
Chi-Square(a)	5.350
df	1
Asymp. Sig.	.021

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 225.0.

F-10: SPSS Chi-squared results comparing Males v Females for A⁺, C⁺ and E⁺ grades 2007 Unit 4 Examination

NPar Tests comparing Males v Females for A⁺ grades 2007 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Frequency A ⁺	930	469.7505	46.78446	418.00	512.00

Chi-Square Test

Frequencies

Frequency A⁺

A ⁺	Observed N	Expected N	Residual
Male	512	446.0	66.0
Female	418	484.0	-66.0
Total	930		

Test Statistics

	Frequency A ⁺
Chi-Square(a)	18.767
df	1
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 446.0.

NPar Tests comparing Males v Females for C⁺ grades 2007 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq C ⁺	1181	596.0991	57.25099	533.00	648.00

Chi-Square Test

Frequencies

Freq C⁺

C+	Observed N	Expected N	Residual
Male	533	567.0	-34.0
Female	648	614.0	34.0
Total	1181		

Test Statistics

	Freq C ⁺
Chi-Square(a)	3.922
df	1
Asymp. Sig.	.048

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 567.0.

NPar Tests comparing Males v Females for E⁺ grades 2007 Unit 4 Examination

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq E ⁺	377	188.6074	4.50470	184.00	193.00

Chi-Square Test

Frequencies

Freq E⁺

E ⁺	Observed N	Expected N	Residual
Male	193	181.0	12.0
Female	184	196.0	-12.0
Total	377		

Test Statistics

	Freq E ⁺
Chi-Square(a)	1.530
df	1
Asymp. Sig.	.216

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 181.0.

F-11: SPSS Chi-squared results comparing Males v Females for A⁺, C⁺ and E⁺ grades All Unit 3 Examinations (2003-2007)

NPar Tests comparing Males v Females for A⁺ grades Unit 3 Examination (2003-2007)

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Frequency A ⁺	4385	2239.9373	319.02847	1870.00	2515.00

Chi-Square Test

Frequencies

Frequency A⁺

A ⁺	Observed N	Expected N	Residual
Male	2515	2060.0	455.0
Female	1870	2325.0	-455.0
Total	4385		

Test Statistics

	Frequency A ⁺
Chi-Square(a)	189.541
df	1
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 2060.0.

NPar Tests comparing Males v Females for C⁺ grades Unit 3 Examination (2003-2007)

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq C ⁺	5953	3012.3146	324.55702	2650.00	3303.00

Chi-Square Test

Frequencies

Freq C⁺

C ⁺	Observed N	Expected N	Residual
Male	2650	2797.0	-147.0
Female	3303	3156.0	147.0
Total	5953		

Test Statistics

	Freq C ⁺
Chi-Square(a)	14.573
df	1
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 2797.0.

NPar Tests comparing Males v Females for E⁺ grades Unit 3 Examination (2003-2007)

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq E ⁺	2441	1235.9906	136.65263	1083.00	1358.00

Chi-Square Test

Frequencies

Freq E⁺

E ⁺	Observed N	Expected N	Residual
Male	1083	1146.0	-63.0
Female	1358	1295.0	63.0
Total	2441		

Test Statistics

	Freq E ⁺
Chi-Square(a)	6.528
df	1
Asymp. Sig.	.011

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 1146.0.

F-12: SPSS Chi-squared results comparing Males v Females for A⁺, C⁺ and E⁺ grades All Unit 4 Examinations

NPar Tests comparing Males v Females for A⁺ grades Unit 4 Examination (2003-2007)

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Frequency A ⁺	4438	2226.1546	125.81088	2093.00	2345.00

Chi-Square Test

Frequencies

Frequency A⁺

A ⁺	Observed N	Expected N	Residual
Male	2345	2082.0	263.0
Female	2093	2356.0	-263.0
Total	4438		

Test Statistics

	Frequency A ⁺
Chi-Square(a)	62.581
df	1
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5.
The minimum expected cell frequency is 2082.0.

NPar Tests comparing Males v Females for C⁺ grades Unit 4 Examination (2003-2007)

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq C ⁺	5530	2797.3331	297.27353	2466.00	3064.00

Chi-Square Test

Frequencies

Freq C⁺

C ⁺	Observed N	Expected N	Residual
Male	2466	2594.0	-128.0
Female	3064	2936.0	128.0
Total	5530		

Test Statistics

	Freq C ⁺
Chi-Square(a)	11.896
df	1
Asymp. Sig.	.001

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 2594.0.

NPar Tests comparing Males v Females for E⁺ grades Unit 4 Examination (2003-2007)

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq E ⁺	2422	1211.8993	32.99456	1178.00	1244.00

Chi-Square Test

Frequencies

Freq E⁺

E ⁺	Observed N	Expected N	Residual
Male	1178	1134.0	44.0
Female	1244	1288.0	-44.0
Total	2422		

Test Statistics

	Freq E ⁺
Chi-Square(a)	3.210
df	1
Asymp. Sig.	.073

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 1134.0.

F-13: SPSS Chi-squared results comparing Males v Females for A⁺, C⁺ and E⁺ grades All Unit 3 & 4 Examinations

NPar Tests comparing Males v Females for A⁺ grades Unit 3 & 4 Examinations (2003-2007)

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Frequency A ⁺	8823	4457.0972	446.20142	3963.00	4860.00

Chi-Square Test

Frequencies

Frequency A⁺

A ⁺	Observed N	Expected N	Residual
Male	4860	4142.0	718.0
Female	3963	4681.0	-718.0
Total	8823		

Test Statistics

	Frequency A ⁺
Chi-Square(a)	234.594
df	1
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 4142.0.

NPar Tests comparing Males v Females for C⁺ grades Unit 3 & 4 Examinations (2003-2007)

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq C ⁺	11483	5809.6443	621.80405	5116.00	6367.00

Chi-Square Test

Frequencies

Freq C⁺

C ⁺	Observed N	Expected N	Residual
Male	5116	5390.0	-274.0
Female	6367	6093.0	274.0
Total	11483		

Test Statistics

	Freq C ⁺
Chi-Square(a)	26.250
df	1
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 5390.0.

NPar Tests comparing Males v Females for E⁺ grades Unit 3 & 4 Examinations (2003-2007)

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Freq E ⁺	4863	2443.4557	170.09780	2261.00	2602.00

Chi-Square Test

Frequencies

Freq E⁺

E ⁺	Observed N	Expected N	Residual
Male	2261	2280.0	-19.0
Female	2602	2583.0	19.0
Total	4863		

Test Statistics

	Freq E ⁺
Chi-Square(a)	.298
df	1
Asymp. Sig.	.585

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 2280.0.

Appendix G: Grade distributions for VCE Chemistry 2003–2007- A⁺, C⁺ and E⁺

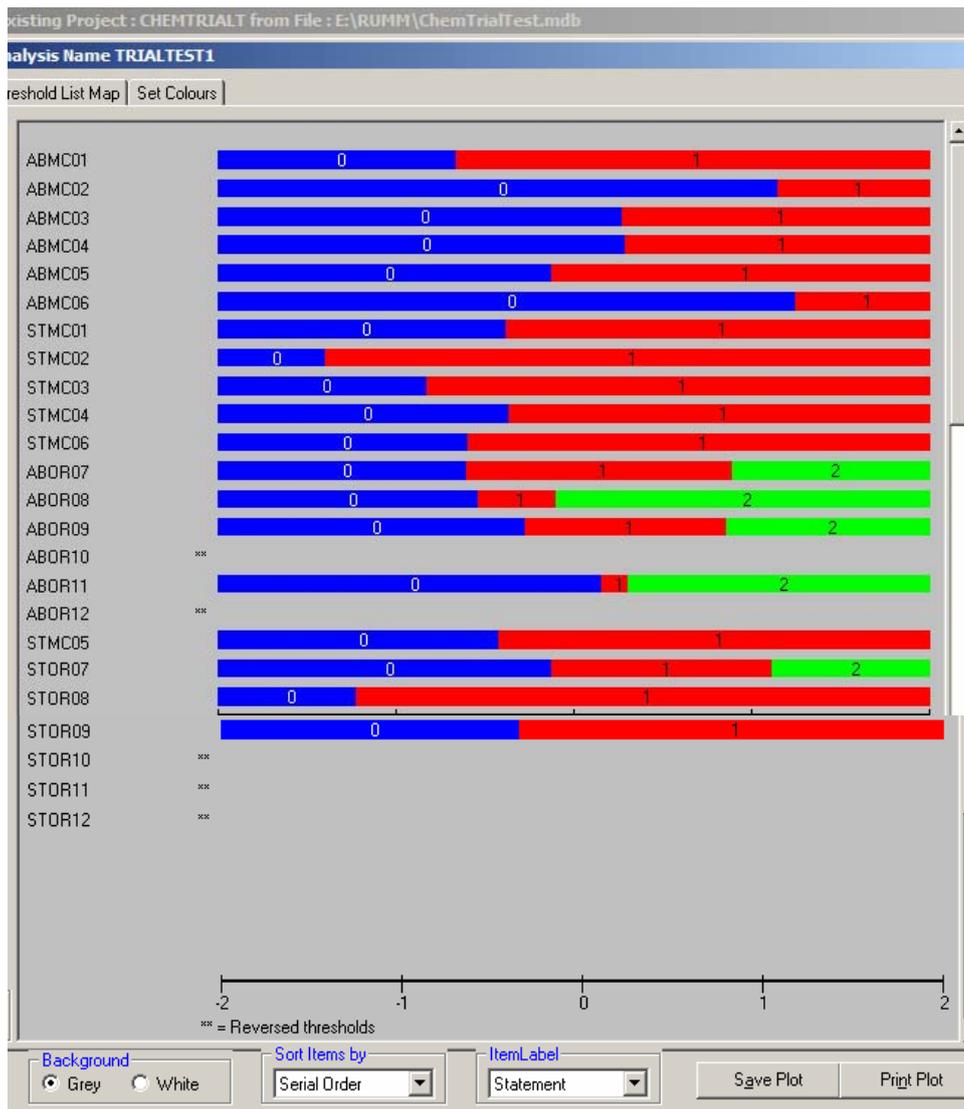
Year	2003 Unit 3	2003 Unit 4	2004 Unit 3	2004 Unit 4	2005 Unit 3	2005 Unit 4	2006 Unit 3	2006 Unit 4	2007 Unit 3	2007 Unit 4	Totals	Unit 3 Total	Unit 4 Total
Male A+	477	408	489	458	496	456	517	511	536	512	4860	2515	2345
Female A+	392	396	405	441	377	452	347	386	349	418	3963	1870	2093
Total A+	869	804	894	899	873	908	864	897	885	930	8823	4385	4438
Male C+	536	431	490	517	514	487	582	498	528	533	5116	2650	2466
Female C+	668	602	649	634	668	608	664	572	654	648	6367	3303	3064
Total C+	1204	1033	1139	1151	1182	1095	1246	1070	1182	1181	11483	5953	5530
Male E+	217	241	214	239	261	246	191	259	200	193	2261	1083	1178
Female E+	283	252	258	247	273	271	276	290	268	184	2602	1358	1244
Total E+	500	493	472	486	534	517	467	549	468	377	4863	2441	2422
Total students	8626	8539	8733	8607	9030	8897	9063	8939	8982	8864	88280	44434	43846
Total Male	3973	3930	4069	3998	4207	4142	4307	4236	4321	4254	41437	20877	20560
Total Female	4653	4609	4664	4609	4823	4755	4756	4703	4661	4610	46843	23557	23286
Expected male A+	400	370	417	418	407	423	411	425	426	446	4142	2060	2082
Expected female A+	469	434	477	481	466	485	453	472	459	484	4681	2325	2356
Expected Male C+	555	475	531	535	551	510	592	507	569	567	5390	2797	2594
Expected Female C+	649	558	608	616	631	585	654	563	613	614	6093	3156	2936
Expected Male E+	230	227	220	226	249	241	222	260	225	181	2280	1146	1134
Expected Female E+	270	266	252	260	285	276	245	289	243	196	2583	1295	1288

Appendix H: Trial test data

H-1 Fit residuals and probabilities for the initial item set

Seq	Item	Type	Item Location	Fit Residual	DF	ChiSq	DF	Prob
20	ST08	Poly	-1.226	-0.418	159	0.273	2	0.8721
7	ST01	MC	-0.383	-0.057	165.	0.634	2	0.72825
17	AB12	Poly	0.77	-0.661	164	1.012	2	0.60277
1	AB01	MC	-0.663	-0.581	160	1.228	2	0.54122
11	ST06	MC	-0.597	0.002	165	1.578	2	0.45439
14	AB09	Poly	0.289	0.431	166	1.625	2	0.44379
16	AB11	Poly	0.229	0.139	166	1.803	2	0.4058
3	AB03	MC	0.273	-0.375	160	2.102	2	0.34953
5	AB05	MC	-0.124	-0.841	160	2.27	2	0.32149
9	ST03	MC	-0.832	0.121	165	2.297	2	0.31705
4	AB04	MC	0.284	0.902	159	2.312	2	0.31480
8	ST02	MC	-1.4	-0.909	165	2.696	2	0.25979
2	AB02	MC	1.144	2.009	160	3.077	2	0.21468
19	ST07	Poly	0.492	1.563	159	3.27	2	0.19499
23	ST11	Poly	0.505	-1.378	158	3.48	2	0.17549
6	AB06	MC	1.243	0.998	160	4.26	2	0.11881
24	ST12	Poly	0.809	-1.894	157	4.406	2	0.11045
22	ST10	Poly	-0.214	-1.255	158	5.839	2	0.05397
15	AB10	Poly	0.729	1.811	166	6.08	2	0.04783
12	AB07	Poly	0.141	0.386	166	6.146	2	0.04629
10	ST04	MC	-0.367	-1.067	166	8.487	2	0.01435
13	AB08	Poly	-0.323	1.003	166	9.338	2	0.00938
21	ST09	Poly	-0.352	-1.728	159	13.921	2	0.0009
18	ST05	Poly	-0.426	-2.094	166	15.405	2	0.00045

H-2 Threshold map for initial item set



H-3 Person-fit characteristics for original data sample

Serial	Extm	Location	SE	FitResid	Data Pts	PF_Gender
45		0.042	0.338	2.525	24	2
53		0.133	0.339	2.481	23	2
19		0.173	0.362	2.462	18	2
26		0.37	0.336	1.825	24	2
113		0.479	0.337	1.806	24	1
80		0.564	0.343	1.801	23	2
84		0.261	0.335	1.691	24	1
14		1.203	0.382	1.689	24	2
44		0.261	0.335	1.654	24	2
11		0.082	0.382	1.324	21	2
49		1.215	0.421	1.321	18	2
188		-0.423	0.357	1.268	24	1
36		-0.55	0.365	1.257	24	2
13		-0.302	0.35	1.233	24	2
81		0.764	0.379	1.228	18	2
75		0.764	0.379	1.228	18	2
59		-0.216	0.418	1.216	18	2
50		0.6	0.418	1.109	18	2
56		0.77	0.427	1.107	18	2
71		0.703	0.345	1.105	24	2
63		-0.184	0.344	1.093	24	2
41		-0.07	0.34	1.076	24	2
76		0.941	0.359	1.053	24	2
66		1.851	0.541	0.876	18	2
165		1.203	0.382	0.824	24	1
46		0.703	0.345	0.815	24	2
77		2.524	0.647	0.815	24	2
18		0.261	0.335	0.77	24	2
17		1.35	0.399	0.754	24	2
190		0.941	0.359	0.724	24	2
24		0.703	0.345	0.633	24	2
25		-0.826	0.389	0.603	24	2
68		-0.425	0.388	0.594	18	2
170		0.398	0.495	0.591	12	1
23		0.703	0.345	0.566	24	2
100		1.912	0.491	0.559	24	1
102		3.046	0.828	0.499	24	1
128		-0.07	0.34	0.497	24	1
168		1.697	0.451	0.484	24	1
78		1.203	0.382	0.462	24	2
27		-1.542	0.48	0.456	24	2
185		0.82	0.351	0.452	24	1

9	0.82	0.351	0.436	24	2
39	0.703	0.345	0.435	24	2
122	0.703	0.345	0.431	24	1
144	0.703	0.345	0.431	24	1
79	1.35	0.399	0.424	24	2
67	1.394	0.448	0.424	18	2
8	1.35	0.399	0.423	24	2
60	1.513	0.422	0.413	24	2
172	2.524	0.647	0.404	24	1
155	0.152	0.336	0.391	24	1
166	1.068	0.369	0.375	24	1
174	0.218	0.426	0.358	18	1
189	-0.302	0.35	0.352	24	2
167	1.203	0.382	0.337	24	1
121	1.068	0.369	0.31	24	1
143	1.068	0.369	0.31	24	1
73	1.426	0.537	0.304	12	2
180	1.35	0.399	0.258	24	1
103	0.6	0.418	0.228	18	1
52	1.513	0.422	0.214	24	2
183	2.176	0.552	0.214	24	1
191	0.6	0.418	0.199	18	1
98	1.912	0.491	0.18	24	1
150	-0.07	0.34	0.178	24	2
82	1.513	0.422	0.177	24	2
55	0.941	0.359	0.162	24	2
20	0.949	0.44	0.157	18	2
173	0.261	0.335	0.144	24	1
64	1.068	0.369	0.124	24	2
62	1.35	0.399	0.113	24	2
148	1.203	0.382	0.073	24	1
29	1.697	0.451	0.068	24	2
37	-0.07	0.34	0.063	24	2
171	1.697	0.451	0.045	24	1
1	2.524	0.647	0.031	24	2
175	0.946	0.461	0.026	18	1
131	0.703	0.345	0.019	24	1
169	0.479	0.337	-0.003	24	1
88	3.046	0.828	-0.012	24	1
157	1.068	0.369	-0.053	24	2
33	1.35	0.399	-0.056	24	2
101	0.738	0.467	-0.062	12	1
138	0.59	0.34	-0.089	24	1
151	0.042	0.338	-0.096	24	1
110	1.912	0.491	-0.132	24	1

31	-0.302	0.35	-0.132	24	2
57	2.176	0.552	-0.141	24	2
74	0.59	0.34	-0.164	24	2
30	0.274	0.41	-0.165	18	2
129	2.176	0.552	-0.173	24	1
95	1.068	0.369	-0.189	24	1
7	1.35	0.399	-0.218	24	2
141	0.59	0.34	-0.22	24	1
181	0.941	0.359	-0.243	24	1
162	3.046	0.828	-0.247	24	2
111	2.524	0.647	-0.261	24	1
65	1.912	0.491	-0.267	24	2
58	0.941	0.359	-0.3	24	2
160	1.513	0.422	-0.323	24	1
28	0.152	0.336	-0.351	24	2
156	1.35	0.399	-0.361	24	2
164	1.203	0.382	-0.362	24	1
154	0.82	0.351	-0.376	24	1
127	-0.826	0.389	-0.419	24	1
178	1.068	0.369	-0.427	24	1
161	1.35	0.399	-0.432	24	1
147	2.176	0.552	-0.435	24	1
6	0.261	0.335	-0.463	24	2
146	2.176	0.552	-0.518	24	1
187	1.697	0.451	-0.521	24	2
86	2.176	0.552	-0.528	24	1
54	1.601	0.486	-0.547	18	2
133	2.176	0.552	-0.562	24	1
124	2.176	0.552	-0.562	24	1
184	1.697	0.451	-0.592	24	2
43	0.59	0.34	-0.606	24	2
35	0.941	0.359	-0.611	24	2
136	1.697	0.451	-0.628	24	1
15	1.068	0.369	-0.654	24	2
10	1.697	0.451	-0.657	24	2
12	1.097	0.393	-0.667	23	2
134	2.176	0.552	-0.669	24	1
125	2.176	0.552	-0.669	24	1
177	0.261	0.335	-0.672	24	1
137	0.941	0.359	-0.708	24	1
38	1.068	0.369	-0.71	24	2
186	1.068	0.369	-0.718	24	2
72	1.912	0.491	-0.751	24	2
48	1.35	0.399	-0.771	24	2
163	0.63	0.372	-0.771	18	1

85	1.697	0.451	-0.824	24	1
83	1.203	0.382	-0.836	24	2
119	3.046	0.828	-0.842	24	1
120	3.046	0.828	-0.842	24	1
107	3.046	0.828	-0.842	24	1
159	1.513	0.422	-0.856	24	1
158	1.35	0.399	-0.859	24	1
112	2.796	0.851	-0.87	18	1
123	0.703	0.345	-0.876	24	1
145	0.703	0.345	-0.876	24	1
114	2.524	0.647	-0.903	24	1
87	3.046	0.828	-0.923	24	1
179	0.59	0.34	-0.93	24	2
99	1.068	0.369	-0.959	24	1
21	1.513	0.422	-0.964	24	2
115	1.697	0.451	-0.974	24	1
92	0.703	0.345	-0.975	24	1
182	0.479	0.337	-0.985	24	2
142	0.204	0.34	-0.999	23	1
104	3.046	0.828	-1.005	24	1
32	0.479	0.337	-1.005	24	2
118	2.524	0.647	-1.016	24	1
140	1.697	0.451	-1.025	24	1
109	2.524	0.647	-1.039	24	1
153	1.513	0.422	-1.048	24	1
16	2.176	0.552	-1.069	24	2
42	-0.302	0.35	-1.078	24	2
61	-0.649	0.547	-1.079	11	2
117	2.176	0.552	-1.085	24	1
149	-0.302	0.35	-1.086	24	1
96	2.524	0.647	-1.131	24	1
192	1.157	0.485	-1.131	18	1
47	0.37	0.336	-1.165	24	2
40	0.152	0.336	-1.196	24	2
108	2.176	0.552	-1.2	24	1
176	1.35	0.399	-1.204	24	1
105	2.176	0.552	-1.205	24	1
106	2.176	0.552	-1.272	24	1
34	1.068	0.369	-1.312	24	2
116	2.176	0.552	-1.313	24	1
5	0.574	0.353	-1.328	23	2
89	1.068	0.369	-1.371	24	1
139	1.912	0.491	-1.373	24	1
3	2.089	0.557	-1.378	23	2
69	0.738	0.467	-1.408	12	2

130	0.764	0.379	-1.445	18	1
70	0.339	0.458	-1.58	12	2
152	0.042	0.338	-1.898	24	1
51extm	3.828	1.19 ...			2
132extm	3.224	1.218 ...			1
126extm	3.828	1.19 ...			1
90extm	3.828	1.19 ...			1
4extm	3.828	1.19 ...			2
2extm	1.556	1.584 ...			2
135extm	3.828	1.19 ...			1

H-4- Correlation analysis of multiple-choice compared to short-answer responses

Anova: Single Factor

SUMMARY

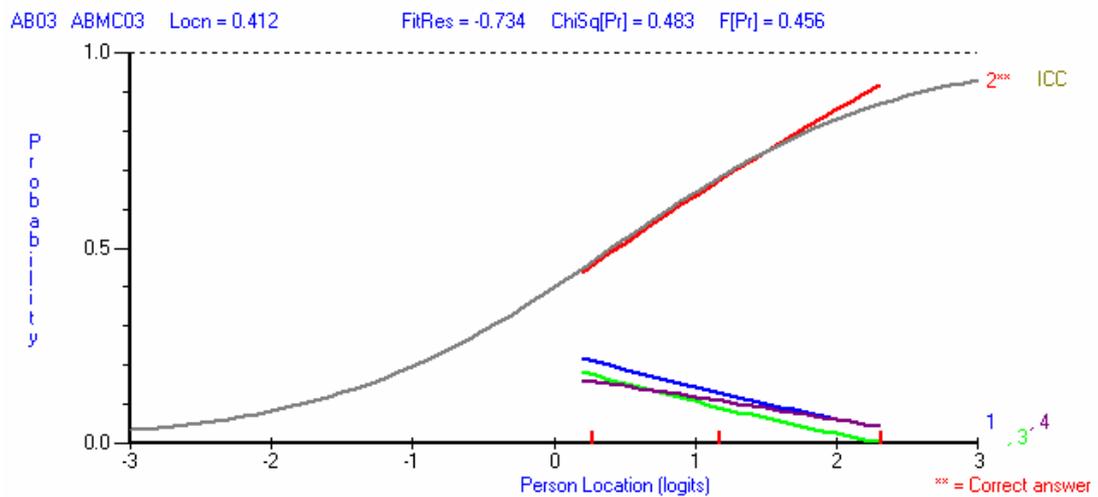
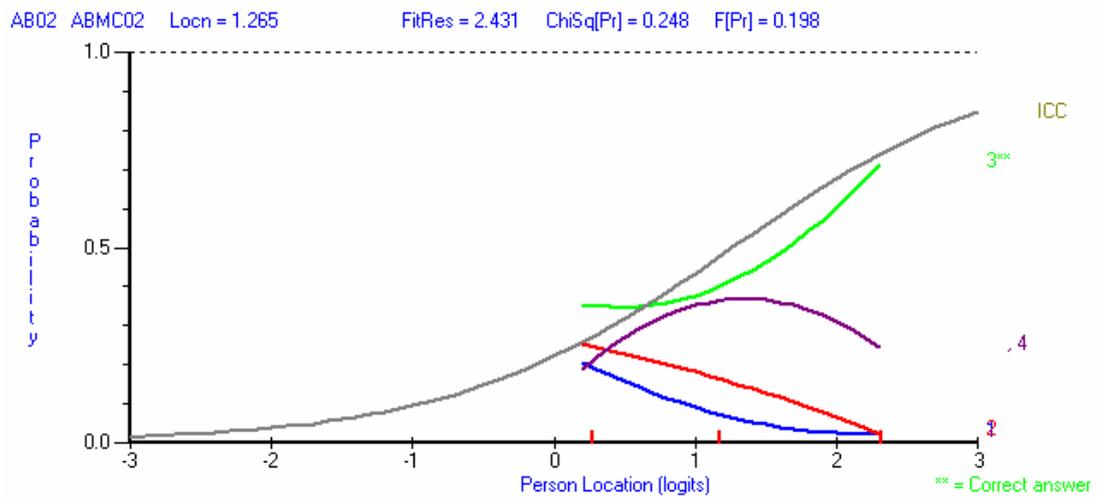
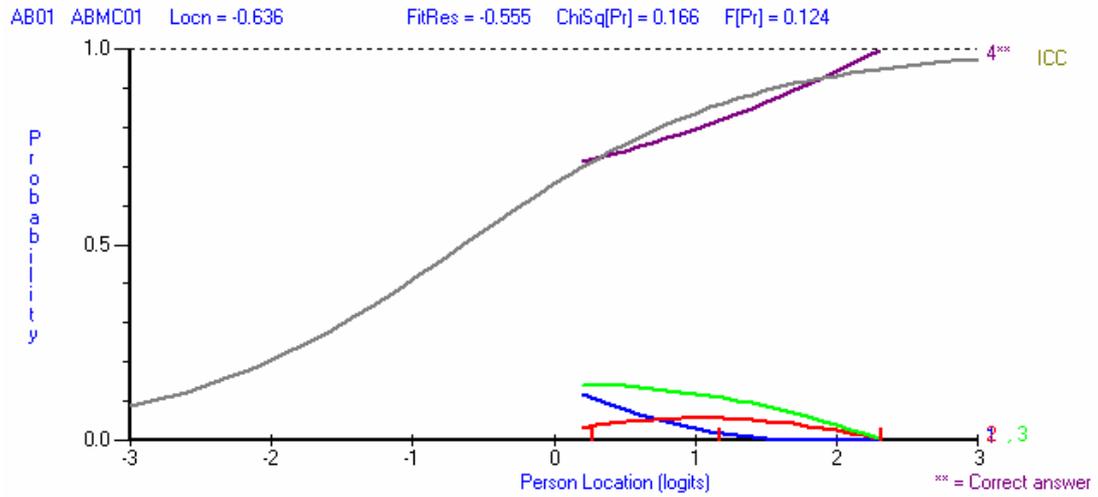
<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Multiple-choice questions	11	3.124	-0.284	0.521927
Short-answer questions	11	2.24	0.203636	0.406167

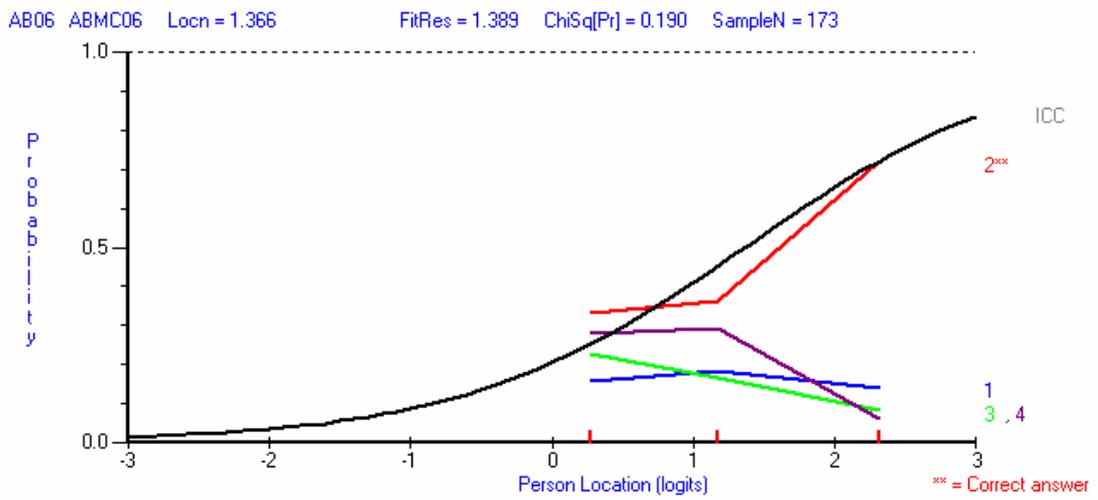
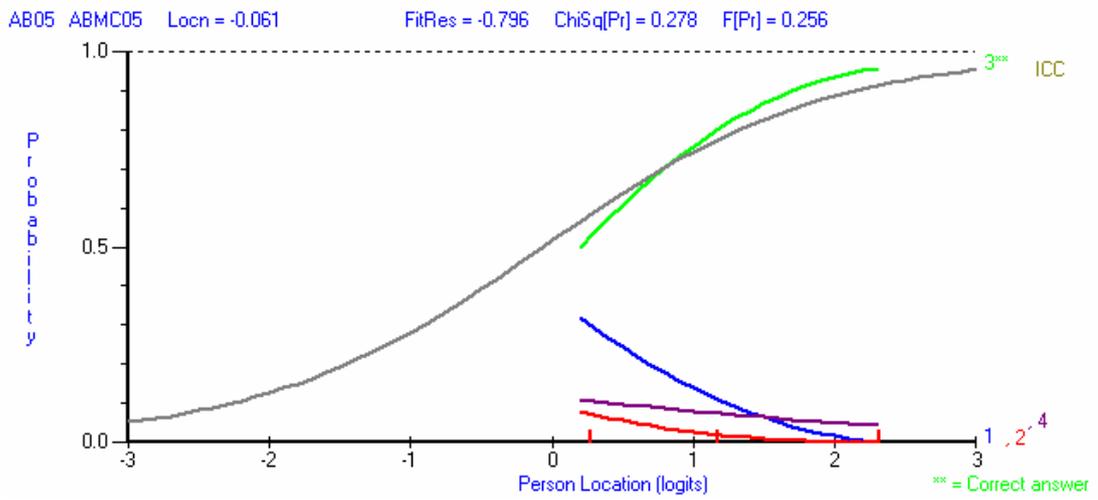
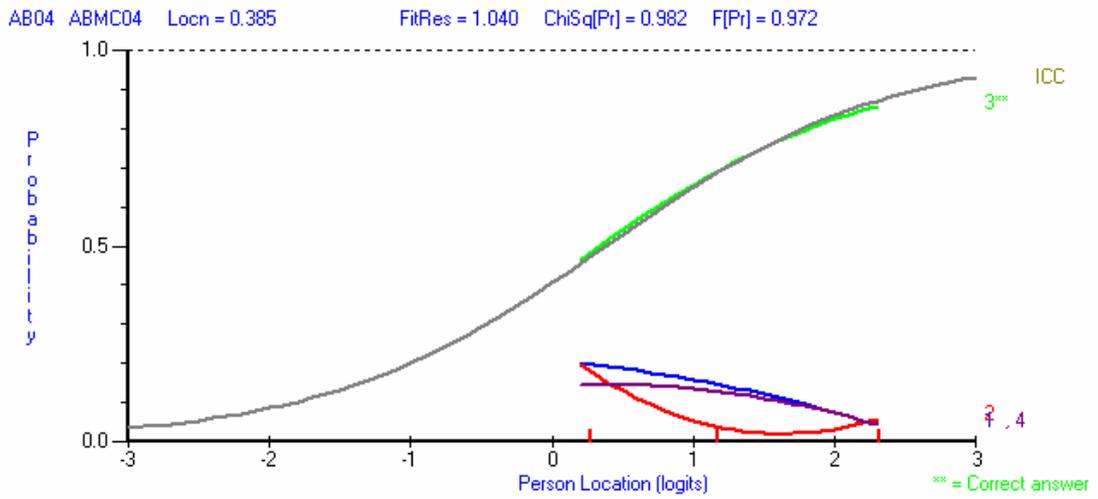
ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	1.307841	1	1.307841	2.818336	0.10875	4.351243
Within Groups	9.280943	20	0.464047			
Total	10.58878	21				

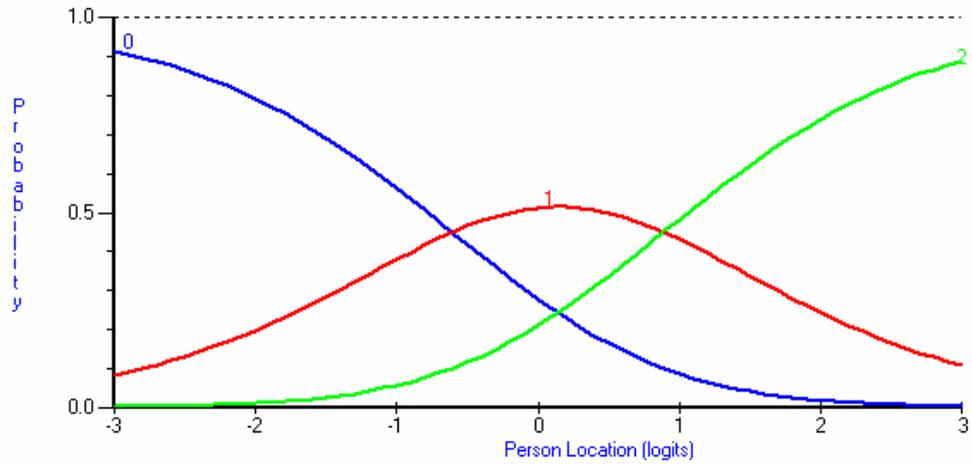
H-5 Item analysis curves

Original item analysis curves before rescoring.

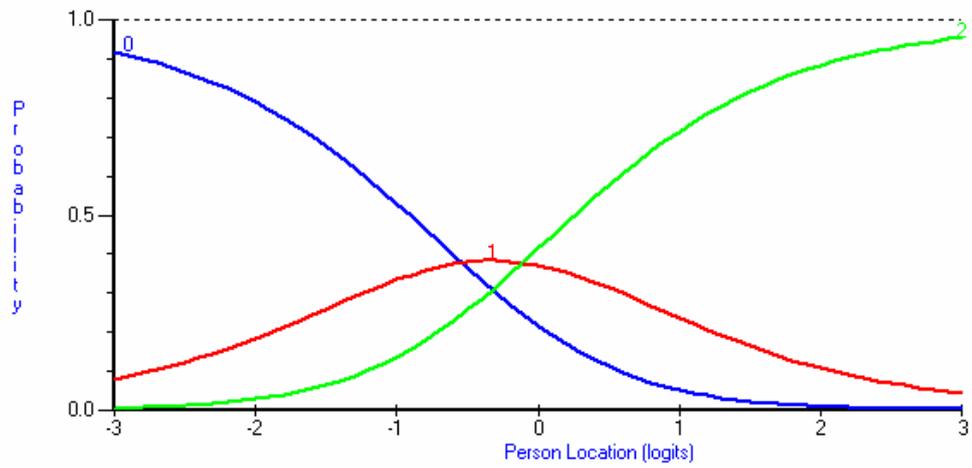




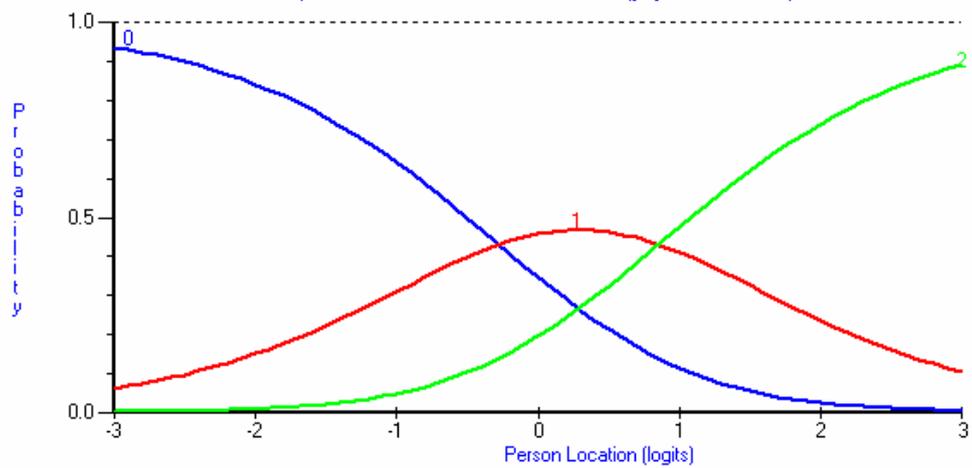
AB07 ABOR07 Locn = 0.141 Spread = 0.750 FitRes = 0.386 ChiSq[Pr] = 0.215 SampleN = 180



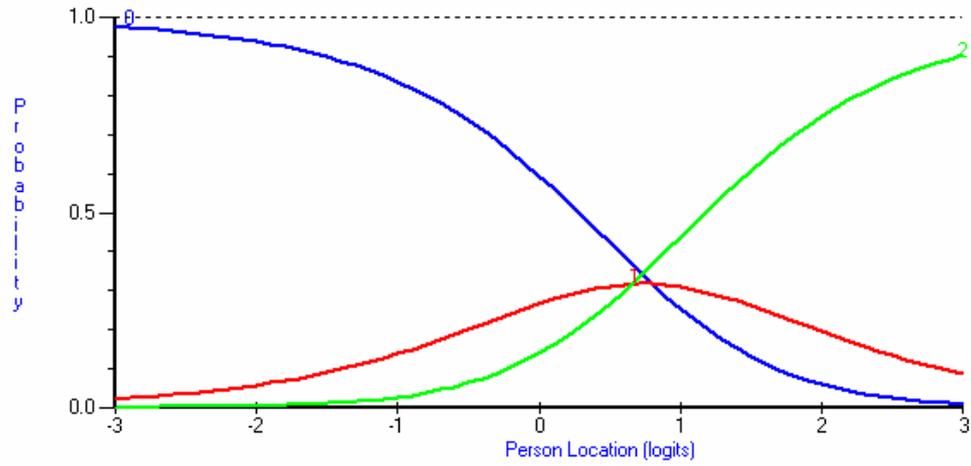
AB08 ABOR08 Locn = -0.323 Spread = 0.217 FitRes = 1.003 ChiSq[Pr] = 0.012 SampleN = 180



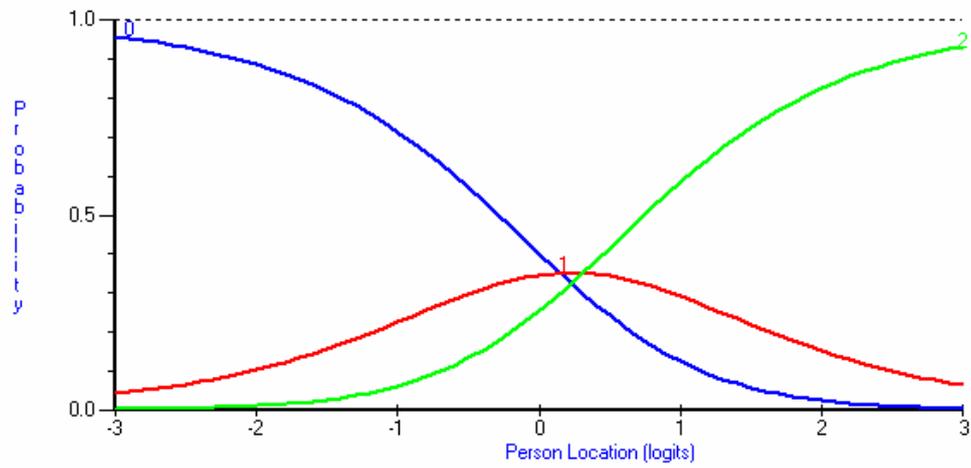
AB09 ABOR09 Locn = 0.289 Spread = 0.565 FitRes = 0.431 ChiSq[Pr] = 0.064 SampleN = 180



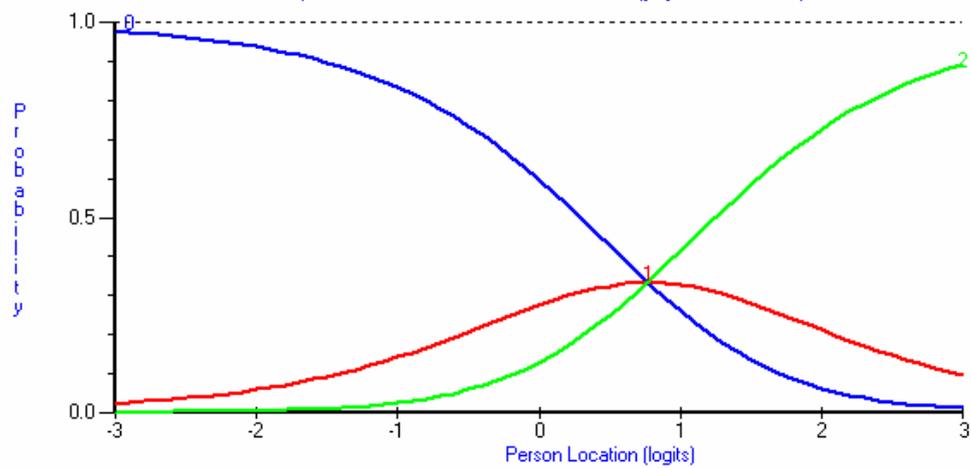
AB10 ABDR10 Locn = 0.729 Spread = -0.076 FitRes = 1.811 ChiSq[Pr] = 0.001 SampleN = 180

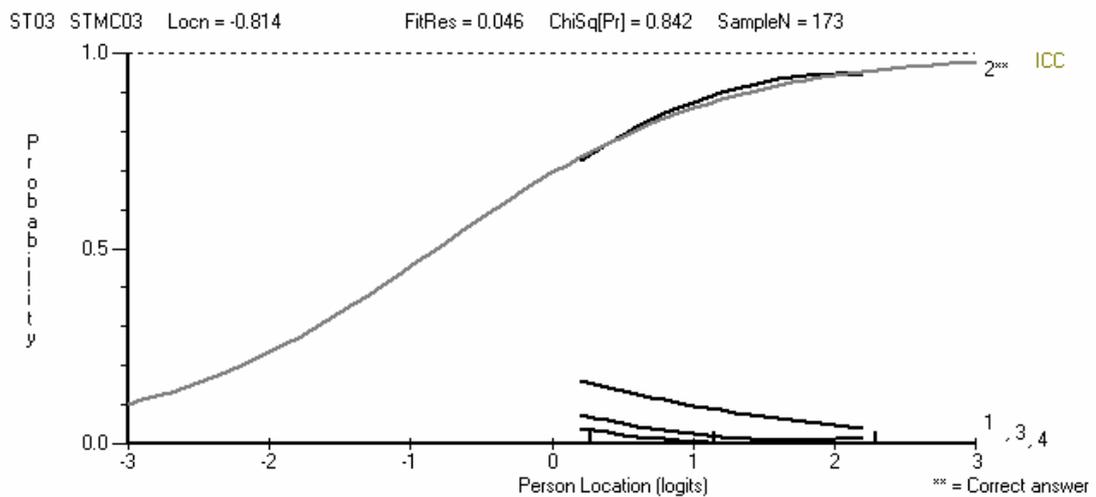
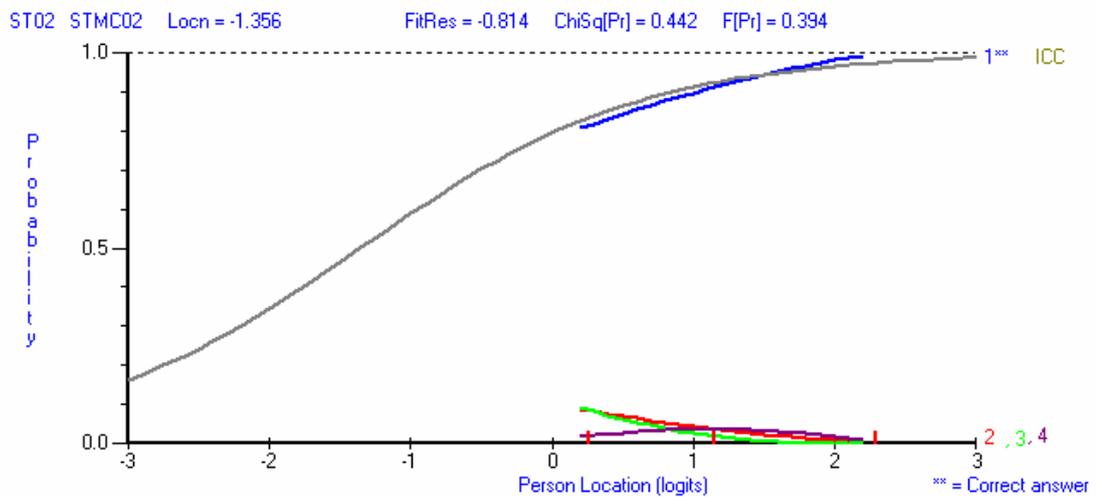
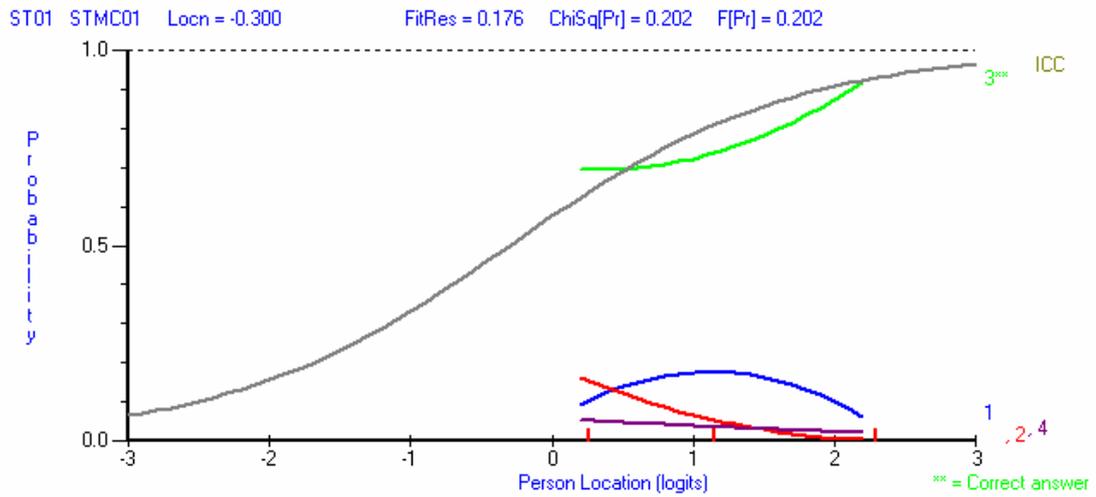


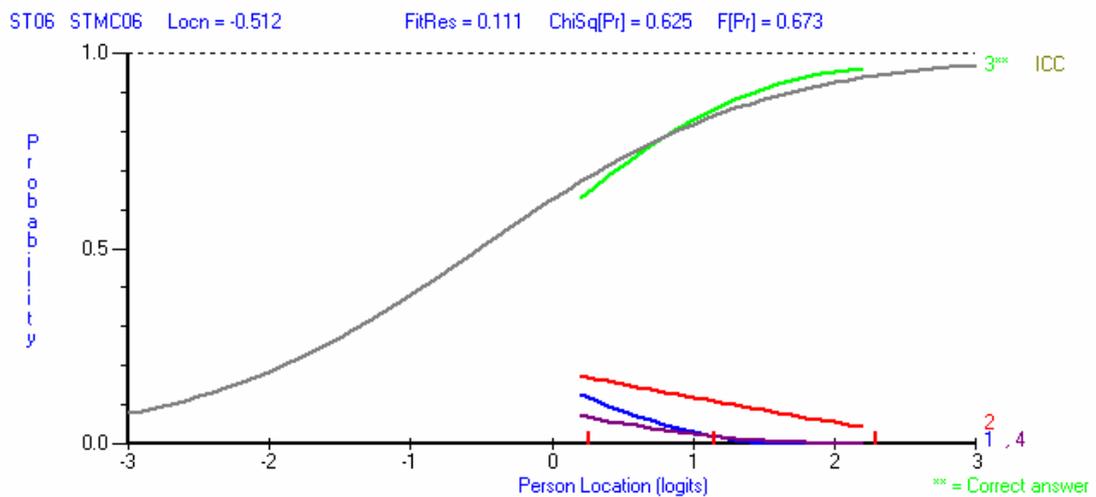
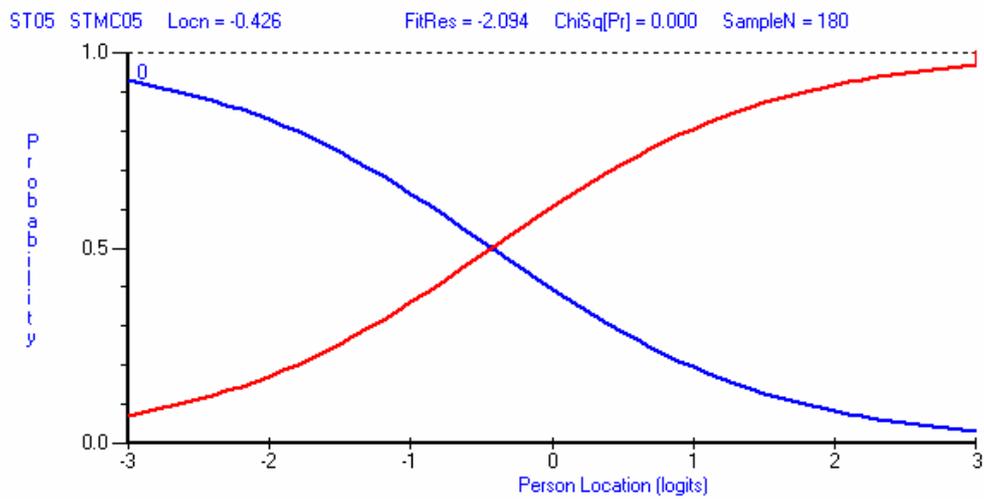
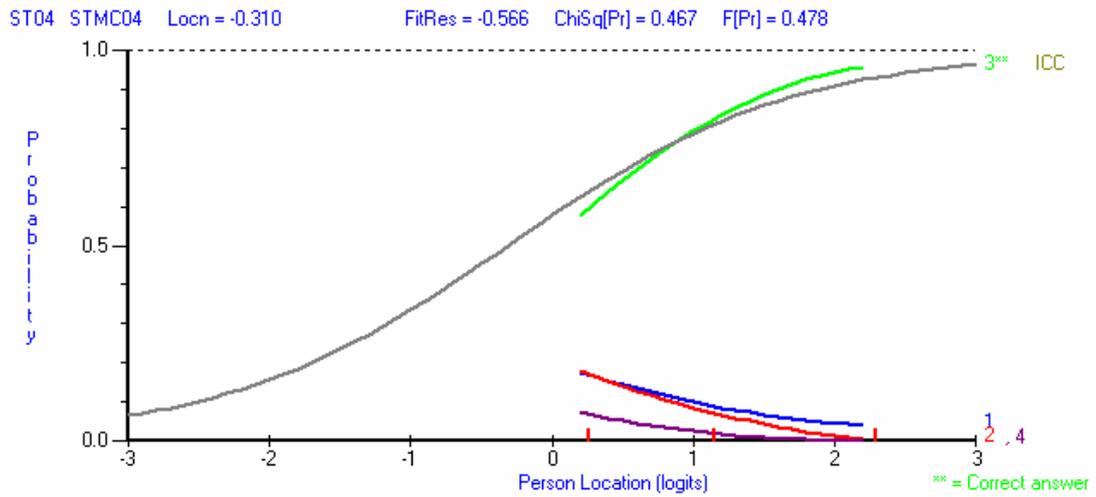
AB11 ABDR11 Locn = 0.229 Spread = 0.077 FitRes = 0.139 ChiSq[Pr] = 0.726 SampleN = 180



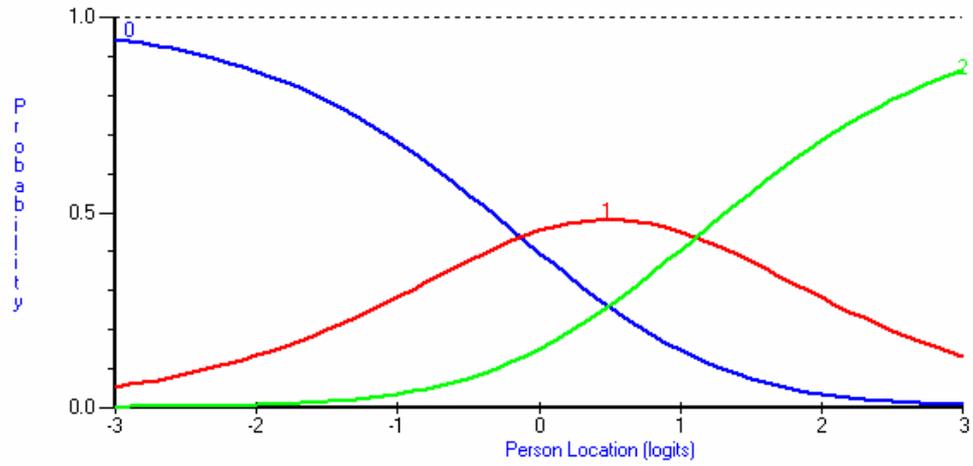
AB12 ABDR12 Locn = 0.770 Spread = -0.003 FitRes = -0.661 ChiSq[Pr] = 0.763 SampleN = 180



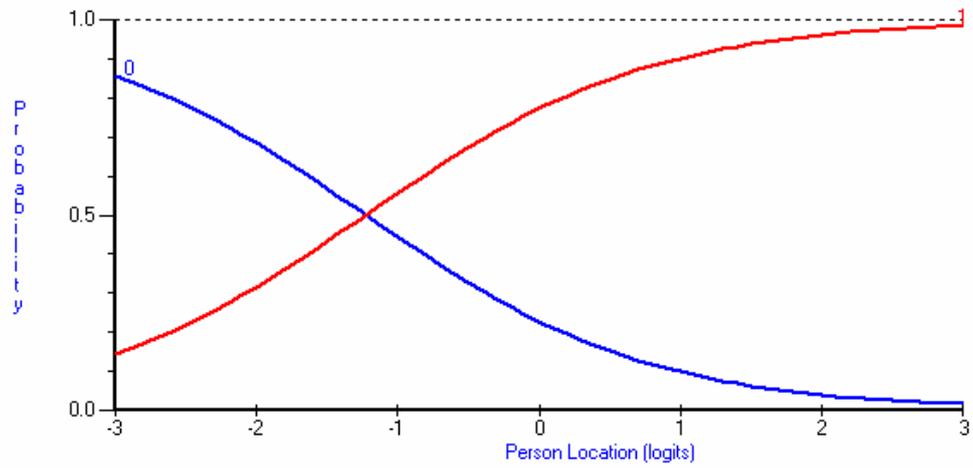




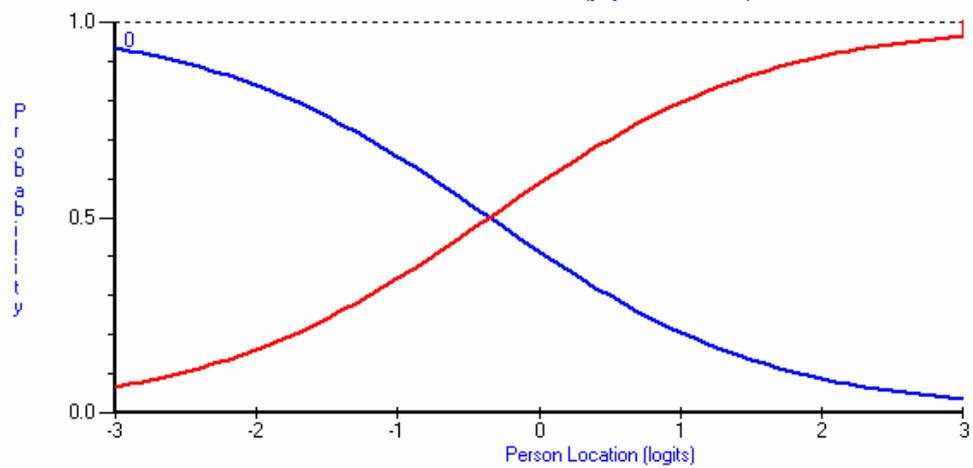
ST07 STOR07 Locn = 0.492 Spread = 0.618 FitRes = 1.563 ChiSq[Pr] = 0.266 SampleN = 180

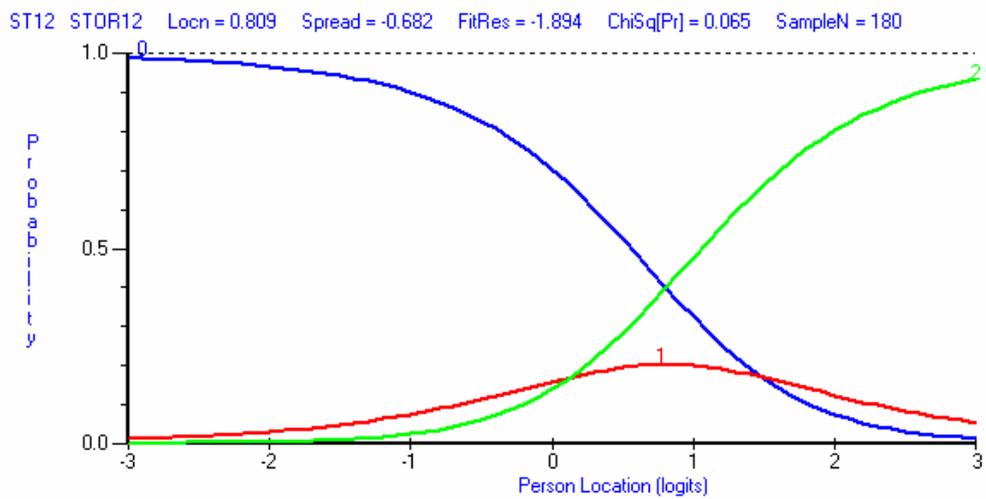
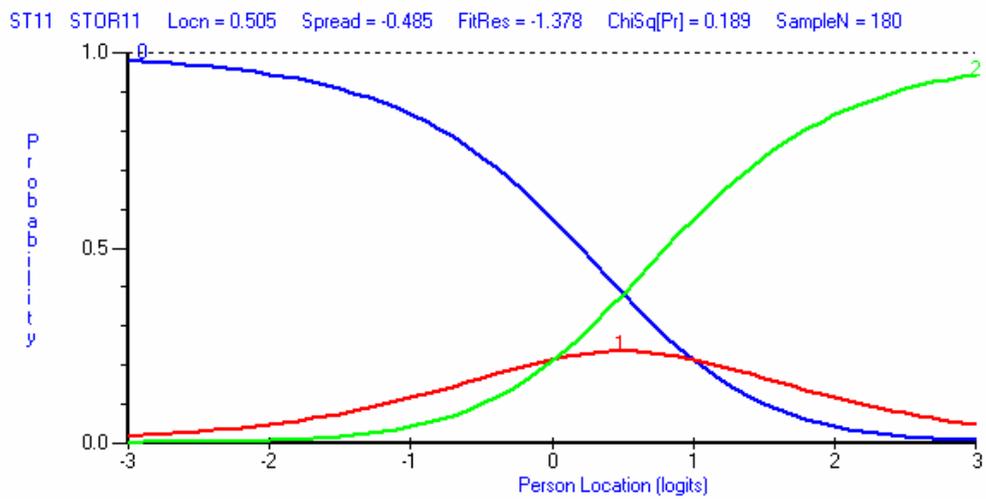
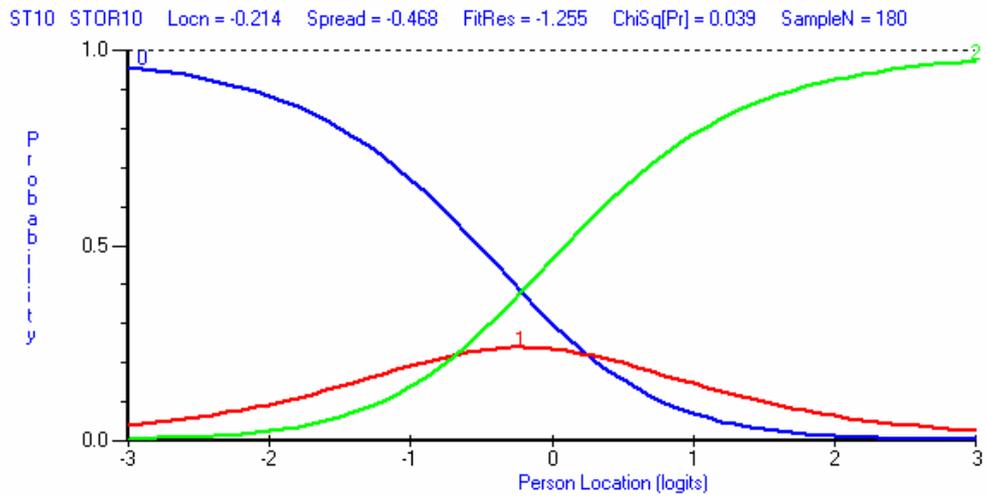


ST08 STOR08 Locn = -1.226 FitRes = -0.418 ChiSq[Pr] = 0.565 SampleN = 180



ST09 STOR09 Locn = -0.352 FitRes = -1.728 ChiSq[Pr] = 0.030 SampleN = 180





H-6: Student responses to trial test, location and fit values

ID	Total	Max	Miss	AB01	AB02	AB03	AB04	AB05	AB06	AB07	AB08	AB09	AB10	AB11	AB12	ST01	ST02	ST03	ST04
1	14	18	17	1	0	0	0	1	1							1	1	1	1
2	14	22	17	1	0	0	0	0	0	2	2	1	1	2	2	1	1	1	0
3	20	29	23	1	0	1	1	0	1	2	0	2	0	1	0	1	1	1	1
4	13	29	23	1	0	0	0	1	1	1	2	0	0	0	0	1	1	1	0
5	12	29	23	0	0	0	1	0	1	2	2	0	1	0	0	1	0	1	0
6	26	29	23	1	1	1	1	1	1	2	2	2	1	1	2	1	1	1	0
7	22	29	23	1	1	1	1	0	0	2	2	1	1	0	1	1	1	1	1
8	19	29	23	0	0	0	0	1	1	1	2	2	0	0	2	1	1	1	0
9	25	29	23	1	1	0	0	1	1	2	2	1	1	1	2	1	1	1	1
10	27	29	23	1	1	1	1	1	1	2	2	2	1	0	2	1	1	1	1
11	17	29	23	1	0	1	0	1	0	1	2	2	1	0	0	1	1	1	0
12	22	29	23	0	1	1	1	1	1	1	2	1	0	1	2	1	1	1	1
13	23	29	23	1	0	0	1	1	1	2	2	0	1	2	2	1	1	0	1
14	19	29	23	1	1	1	1	1	0	1	2	1	0	0	2	1	1	1	0
15	22	29	23	1	1	1	0	1	0	1	2	1	1	1	1	1	1	1	0
16	16	29	23	0	0	0	0	0	0	2	0	1	1	1	0	1	1	1	1
17	23	29	23	1	1	0	1	1	1	2	2	1	1	1	1	1	1	1	1
18	13	18	17	1	1	0	0	1	0							1	1	1	0
19	11	18	17	1	1	1	0	0	0							1	1	1	0
20	15	29	23	1	1	0	0	0	1	1	2	0	1	0	1	1	1	1	0
21	27	29	23	1	1	1	1	1	1	2	2	2	1	0	2	1	1	1	1
22	25	29	23	1	1	1	0	1	1	2	2	0	0	2	2	1	1	1	1

23	9	12	11													1	1	1	0
24	18	29	23	1	0	0	0	0	1	1	2	2	1	2	0	1	1	1	0
25	25	29	23	1	0	1	1	1	1	1	2	2	1	2	2	1	1	0	1
26	22	29	23	1	1	1	1	1	0	1	0	1	1	0	2	1	1	1	1
27	21	29	23	1	1	0	0	1	0	1	2	2	1	2	1	1	1	1	1
28	21	29	23	0	0	0	0	1	0	2	2	2	1	2	1	0	1	1	1
29	22	29	23	0	0	1	1	1	0	1	2	2	1	2	1	1	1	1	1
30	15	23	17							1	2	2	1	0	0	1	1	1	1
31	28	29	23	1	1	1	0	1	1	2	2	2	1	2	2	1	1	1	1
32	23	29	23	1	0	1	1	1	1	0	2	2	1	2	1	1	1	1	1
33	23	29	23	1	0	1	1	1	0	1	2	2	1	2	2	1	1	0	1
34	23	29	23	1	0	0	0	1	0	2	2	1	1	2	1	1	1	1	1
35	24	29	23	1	0	1	1	1	0	2	2	1	1	1	2	1	1	1	1
36	21	29	23	0	0	1	1	0	1	2	2	1	1	1	0	1	1	1	1
37	22	29	23	1	0	0	0	1	1	1	2	1	1	2	1	1	1	1	1
38	14	29	23	0	1	0	0	0	0	2	0	0	1	0	0	1	1	1	1
39	20	29	23	1	0	0	1	0	0	2	2	2	1	1	1	1	1	1	1
40	23	29	23	1	0	1	1	0	0	1	2	2	1	2	1	1	1	1	1
41	15	29	23	1	0	1	1	0	0	1	0	1	1	1	0	1	1	1	1
42	16	29	23	1	0	1	0	1	0	2	2	2	1	0	0	1	1	0	1
43	13	29	23	0	0	0	0	0	1	1	2	1	1	1	0	1	1	0	1
44	12	29	23	0	0	1	1	0	0	1	2	1	0	0	0	0	1	0	1
45	23	29	23	1	0	0	0	1	1	2	2	1	0	2	2	1	1	1	0
46	27	29	23	1	0	1	1	1	1	2	2	2	0	2	2	1	1	1	1
47	26	29	23	1	1	1	1	1	1	2	2	1	1	2	0	1	1	1	1
48	19	29	23	1	0	0	0	1	0	0	1	1	0	1	2	1	1	1	1
49	19	29	23	1	0	1	1	1	0	2	1	2	0	1	0	1	1	1	1

50	21	29	23	1	0	1	1	1	0	2	0	2	0	0	1	1	1	1	1
51	16	28	22	0	0	0		0	0	0	2	1	1	2	1	1	1	1	1
52	20	29	23	1	0	1	1	1	0	2	1	2	0	2	0	0	1	1	0
53	25	29	23	1	1	1	1	1	0	2	2	1	0	2	1	1	1	1	1
54	25	29	23	1	1	1	1	1	0	1	2	1	1	2	1	1	1	1	1
55	17	29	23	1	0	0	0	1	0	2	2	0	0	1	1	0	1	0	1
56	22	29	23	1	0	1	0	1	0	1	1	2	1	2	2	1	1	1	1
57	25	29	23	1	1	1	0	1	1	1	2	2	0	2	1	1	1	1	1
58	29	29	23	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1
59	26	29	23	1	1	0	1	1	1	1	2	2	1	2	2	1	1	1	1
60	27	29	23	1	0	1	1	1	1	2	1	2	1	2	2	1	1	1	1
61	13	13	12	1	1	1	1	1	1										
62	21	29	23	1	0	0	1	1	0	1	2	2	0	2	2	1	1	1	1
63	16	23	17							1	1	1	1	2	1	1	1	1	0
64	26	29	23	1	1	1	0	1	1	1	1	2	1	2	2	1	1	1	1
65	15	29	23	1	1	0	0	1	1	2	1	1	0	1	0	1	1	1	0
66	9	29	23	1	0	0	1	0	0	1	1	1	0	1	1	1	0	0	0
67	29	29	23	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1
68	26	29	23	1	1	0	1	1	1	1	2	2	1	2	2	1	1	1	1
69	27	29	23	1	0	1	1	1	1	2	1	2	1	2	2	1	1	1	1
70	19	29	23	1	0	0	0	1	0	0	1	1	0	1	2	1	1	1	1
71	19	29	23	1	0	1	1	1	0	2	1	2	0	1	0	1	1	1	1
72	21	29	23	1	0	1	1	1	0	2	0	2	0	0	1	1	1	1	1
73	29	29	23	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1
74	29	29	23	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1
75	27	29	23	1	1	1	1	1	1	2	2	2	1	2	1	1	1	1	1
76	27	29	23	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1

77	27	29	23	1	1	1	1	1	0	2	2	2	1	2	2	1	1	1	1
78	25	29	23	1	0	1	1	1	1	1	2	2	0	2	1	1	1	1	1
79	27	29	23	1	1	1	1	1	1	2	2	1	1	2	2	1	1	1	1
80	19	29	23	1	0	1	1	1	1	2	2	2	1	2	2	1	1	1	0
81	22	22	17	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1
82	27	29	23	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1
83	25	29	23	1	0	1	1	1	0	2	2	2	1	2	2	1	1	1	1
84	27	29	23	1	1	1	1	1	1	1	2	2	1	2	2	1	1	1	1
85	26	29	23	1	1	1	1	1	0	1	2	2	1	2	2	1	1	1	1
86	29	29	23	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1
87	27	29	23	1	0	1	1	1	1	2	2	2	1	2	1	1	1	1	1
88	26	29	23	1	0	1	1	1	0	2	2	2	1	2	2	1	1	1	1
89	28	29	23	1	1	1	1	1	0	2	2	2	1	2	2	1	1	1	1
90	15	22	17	1	0	1	1	1	1	1	2	1	1	2	1	0	1	1	0
91	28	29	23	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1
92	11	17	12	1	0	1	0	1	1	0	2	1	1	1	2				
93	25	29	23	1	0	1	1	1	0	2	2	2	1	2	1	1	1	1	1
94	21	29	23	1	1	0	1	1	0	2	2	1	1	1	1	0	1	1	1
95	25	29	23	1	1	1	1	1	1	2	2	1	1	1	2	1	1	0	1
96	28	29	23	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1
97	21	29	23	1	0	0	1	1	1	1	2	0	1	2	0	1	1	1	0
98	19	29	23	1	0	1	0	1	0	1	2	1	1	1	1	0	1	1	1
99	29	29	23	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1
100	21	29	23	1	0	1	0	1	0	2	1	2	1	1	0	1	1	1	1
101	28	29	23	1	1	1	1	1	1	2	1	2	1	2	2	1	1	1	1
102	28	29	23	1	1	1	1	1	1	2	2	2	1	2	1	1	1	1	1
103	27	29	23	1	1	1	0	1	0	2	2	2	1	2	2	1	1	1	1

104	25	29	23	1	0	1	1	1	1	1	2	2	1	2	2	1	1	1	1
105	16	29	23	1	1	1	1	1	1	1	1	0	0	2	0	1	0	0	0
106	23	29	23	1	1	1	0	0	1	1	2	2	1	1	1	1	1	1	1
107	24	29	23	1	1	1	1	0	1	2	2	1	1	2	2	1	1	1	1
108	16	23	17							1	2	2	1	2	2	0	0	1	1
109	19	28	22	1	1	1	1	1	0	1	2	2	1	2	2	1	0	1	1
110	22	29	23	1	1	1	1	1	0	0	0	2	1	2	1	1	1	1	1
111	21	29	23	1	1	1	0	1	1	0	2	2	1	2	1	1	1	0	1
112	27	29	23	1	1	1	1	0	1	2	2	2	1	2	2	1	1	0	1
113	19	29	23	1	1	1	0	1	1	0	2	2	1	1	1	1	1	1	1
114	16	23	17							1	2	2	1	2	2	0	0	1	1
115	18	29	23	1	0	0	1	0	1	1	2	1	1	1	1	1	1	1	1
116	13	17	12	1	1	1	1	1	1	0	2	1	1	1	2				
117	25	29	23	1	0	0	1	1	1	1	2	1	1	2	2	1	1	1	1
118	19	29	23	1	0	1	1	0	0	1	2	2	1	2	2	0	0	1	1
119	9	17	12	1	0	1	1	0	0	2	1	1	1	1	0				
120	11	17	12	1	0	1	1	1	1	1	1	1	0	2	1				
121	10	23	17							2	2	2	0	0	0	1	1	0	0
122	20	23	17							2	2	1	1	2	1	1	1	1	1
123	21	23	17							2	1	2	1	2	2	1	1	1	1
124	26	29	23	1	0	1	1	1	0	2	2	2	1	2	2	1	1	1	1
125	20	29	23	1	0	0	0	1	0	2	1	1	1	2	0	1	1	0	1
126	14	29	23	1	1	1	1	0	0	2	2	2	0	0	0	1	1	0	0
127	24	29	23	1	0	1	1	1	0	2	2	1	1	2	1	1	1	1	1
128	5	11	10												0	1			1
129	23	29	23	0	0	1	1	1	0	2	2	2	1	2	0	0	1	1	1
130	8	22	17	1	0	0	0	0	1	1	1	0	1	1	0	0	0	1	1

131	20	29	23	1	1	0	1	1	0	2	1	0	0	2	1	0	1	1	1
132	27	29	23	1	1	1	1	1	1	2	2	1	1	2	2	0	1	1	1
133	14	22	17	1	1	0	1	1	1	1	1	0	1	2	0	1	0	1	1
134	21	29	23	1	1	0	1	1	0	2	2	1	1	1	2	1	1	1	1
135	20	23	17							2	2	2	1	2	2	0	1	1	1
137	23	29	23	0	0	1	1	0	1	2	1	1	1	2	2	1	1	1	1
138	29	29	23	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1
139	15	22	17	0	0	1	1	1	0	2	2	1	1	2	2	0	1	1	0
140	17	23	17							1	1	2	1	2	2	1	1	0	1
141	22	29	23	1	0	0	0	1	0	2	1	1	1	2	1	1	1	1	1
142	16	29	23	1	1	1	0	0	0	1	1	1	1	0	0	1	1	1	1
143	18	29	23	0	1	1	0	1	0	1	2	1	1	2	0	0	1	1	1
145	17	29	23	0	0	1	1	0	1	2	2	1	0	1	1	0	0	1	1
146	17	29	23	1	1	0	0	1	0	1	1	1	1	1	0	0	1	0	1
147	12	29	23	0	0	0	1	1	0	1	1	1	1	0	0	1	1	1	1
148	12	29	23	0	0	0	0	1	1	0	1	0	0	1	0	0	1	0	1
149	14	29	23	0	0	0	1	1	0	0	1	1	0	0	0	1	1	1	1
150	20	29	23	1	1	1	0	0	0	1	2	2	1	2	2	1	1	1	1
151	22	29	23	1	0	1	1	1	0	1	2	0	1	2	1	1	1	1	1
152	12	29	23	0	0	0	1	1	0	1	1	0	0	0	0	0	1	0	1
153	11	29	23	1	1	0	1	1	0	0	1	0	1	1	0	0	0	1	0
154	19	29	23	1	0	0	0	0	1	1	2	0	1	1	1	1	1	1	1
155	21	29	23	1	1	0	1	1	0	1	2	1	0	2	0	1	1	1	1
156	22	29	23	1	1	1	1	1	0	0	1	0	1	2	2	1	1	1	1
157	18	29	23	1	0	0	1	1	1	0	1	0	0	1	1	1	1	1	1
158	13	29	23	1	0	1	0	1	0	1	2	2	1	0	1	0	1	1	0
159	12	22	17	1	1	1	1	1	0	1	2	1	0	0	0	0	1	0	1

160	25	29	23	1	1	1	1	1	1	2	2	2	1	2	2	0	1	1	1
161	17	29	23	1	1	1	0	1	0	0	1	1	1	1	2	0	1	1	1
162	5	29	23	1	1	0	0	0	0	0	0	1	0	1	0	0	1	0	0
163	18	29	23	1	1	0	1	1	0	2	2	2	0	2	2	0	0	1	0
164	9	29	23	1	1	1	1	0	0	0	0	0	0	0	0	1	1	0	0
165	20	29	23	1	0	1	0	1	1	1	1	2	1	1	2	1	0	1	1
166	19	29	23	0	1	1	1	0	0	1	2	1	1	2	1	1	0	1	0
167	24	29	23	1	1	1	0	1	0	1	1	2	1	2	2	1	1	1	1
168	16	22	17	0	1	1	1	0	0	1	2	2	1	2	1	0	1	1	1
170	15	29	23	1	0	1	1	0	1	1	2	1	1	1	0	1	1	0	0
171	23	29	23	0	0	1	1	1	1	2	1	2	1	2	2	0	1	1	1
172	27	29	23	1	1	1	1	1	0	2	2	2	1	2	2	1	1	1	1
173	20	29	23	1	1	1	1	1	0	1	1	2	1	0	1	1	1	1	0
174	21	29	23	1	0	1	1	0	0	2	2	2	1	2	2	0	0	0	1
175	13	29	23	1	1	1	1	1	0	0	2	2	0	0	0	1	0	1	0
176	21	28	22	1	0	1	1	1	0	2	2	1	1	2	1	0	1	1	1
177	13	26	20	0	0	0	0	0	1	2	2	1	1	1	0	1	1	1	0
178	25	29	23	1	1	1	1	1	1	1	1	2	1	2	0	1	1	1	1
179	19	29	23	0	0	1	1	1	0	1	0	0	0	2	2	1	1	1	1
180	23	29	23	0	1	1	1	1	0	2	2	0	1	2	2	0	1	1	1
181	23	29	23	1	1	1	1	0	1	2	2	1	0	2	2	0	1	1	1
182	14	29	23	1	0	0	0	0	0	1	2	0	0	0	0	0	1	1	1
183	18	27	22	1	0	0	0	0	0	1	2	1	0	2		1	1	1	1
184	29	29	23	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1
185	25	27	22	1	1	1	1	1	1	2	2	1	1	2		1	1	1	1
186	5	5	5													1	1	1	1
187	27	29	23	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	0

ST06	ST07	ST08	ST09	ST10	ST11	ST12	Locn	SE	Residua I	DegFree	DataPts	Serial	Gender	School	Score %	ID
1	2	1	1	1	1	0	1.282	0.59	-0.812	16.1	17	192	male	4	77.78	1
0							0.704	0.45	0.491	16.1	17	191	male	4	63.64	2
1	2	1	0	1	1	1	-0.258	0.39	1.857	21.8	23	188	male	4	41.38	3
1	2	1	0	0	0	0	0.757	0.4	0.461	21.8	23	185	male	4	65.52	4
1	1	0	1	0	0	0	2.577	0.68	0.374	21.8	23	183	male	4	93.10	5
1	1	1	1	1	1	1	1.253	0.44	0.278	21.8	23	181	male	4	75.86	6
1	2	1	1	1	1	0	1.445	0.46	0.428	21.8	23	180	male	4	79.31	7
1	1	1	1	1	1	1	1.253	0.44	-0.926	21.8	23	178	male	4	75.86	8
1	2	1	1	1	1	1	0.317	0.38	-0.459	21.8	23	177	male	4	55.17	9
1	2	1	1	1	1	1	1.445	0.46	-1.534	21.8	23	176	male	4	79.31	10
1	1	1	0	1	1	0	0.971	0.56	-0.611	16.1	17	175	male	4	72.22	11
1	2	1	0	1	1	0	0.426	0.52	0.302	16.1	17	174	male	4	61.11	12
1	2	1	1	1	1	0	0.174	0.38	0.093	21.8	23	173	male	4	51.72	13
1	1	1	0	1	1	0	2.577	0.68	0.374	21.8	23	172	male	4	93.10	14
1	2	1	1	1	1	1	1.903	0.53	0.222	21.8	23	171	male	4	86.21	15
1	2	1	1	1	1	0	0.801	0.69	0.484	10.4	11	170	male	4	75.00	16
1	1	1	1	1	1	0	0.606	0.39	-0.435	21.8	23	169	male	4	62.07	17
1	1	1	1	1	1	1	1.903	0.53	0.127	21.8	23	168	male	4	86.21	18
1	2	1	0	1	0	0	1.253	0.44	0.247	21.8	23	167	male	4	75.86	19
1	1	1	0	1	0	0	1.077	0.42	0.284	21.8	23	166	male	4	72.41	20
1	2	1	1	1	1	1	1.077	0.42	0.608	21.8	23	165	male	4	72.41	21
1	2	1	1	1	1	1	1.253	0.44	-0.299	21.8	23	164	male	4	75.86	22

1	2	1	0	1	0	1	0.64	0.44	-1.012	16.1	17	163	male	4	65.22	23
1	0	1	1	1	1	0	1.445	0.46	-0.289	21.8	23	161	male	4	79.31	24
1	1	1	1	1	1	1	1.445	0.46	-0.339	21.8	23	160	male	4	79.31	25
1	2	1	1	1	1	1	1.445	0.46	-1.1	21.8	23	159	male	4	79.31	26
1	0	1	1	0	1	1	1.658	0.49	-1.382	21.8	23	158	male	4	82.76	27
0	2	1	1	1	1	1	0.032	0.38	0.029	21.8	23	155	male	4	48.28	28
1	1	1	0	1	1	1	0.913	0.41	-0.251	21.8	23	154	male	4	68.97	29
1	1	1	0	1	1	0	1.445	0.46	-1.273	21.8	23	153	male	4	79.31	30
1	2	1	1	1	1	1	0.174	0.38	-1.611	21.8	23	152	male	4	51.72	31
1	1	1	1	1	0	1	0.317	0.38	0.274	21.8	23	151	male	4	55.17	32
1	1	1	1	1	1	0	-0.258	0.39	-0.655	21.8	23	149	male	4	41.38	33
1	2	1	1	1	1	1	1.445	0.46	0.153	21.8	23	148	male	4	79.31	34
1	2	1	1	1	1	0	2.577	0.68	-0.171	21.8	23	147	male	4	93.10	35
1	2	1	1	1	1	0	2.198	0.58	-0.663	21.8	23	146	male	4	89.66	36
1	2	1	0	1	1	1	0.757	0.4	-0.629	21.8	23	145	male	4	65.52	37
1	1	1	1	1	1	0	0.757	0.4	0.429	21.8	23	144	male	4	65.52	38
1	2	1	0	1	0	0	1.077	0.42	0.28	21.8	23	143	male	4	72.41	39
1	1	1	1	1	1	1	0.388	0.39	-0.767	20.8	22	142	male	4	57.14	40
1	1	1	0	1	0	0	0.913	0.41	-0.211	21.8	23	141	male	4	68.97	41
0	1	1	0	1	0	0	1.903	0.53	-0.744	21.8	23	140	male	4	86.21	42
0	1	1	0	1	0	0	1.903	0.53	-1.527	21.8	23	139	male	4	86.21	43
1	1	1	0	1	0	0	0.46	0.39	-0.345	21.8	23	138	male	4	58.62	44
1	2	1	1	1	1	1	1.253	0.44	-0.5	21.8	23	137	male	3	75.86	45
1	2	1	1	1	1	1	1.903	0.53	-0.383	21.8	23	136	male	3	86.21	46
1	2	1	1	1	1	1	3.937	1.21				135	male	3	100.00	47
1	2	1	1	1	1	1	2.198	0.58	-0.724	21.8	23	134	male	3	89.66	48
0	2	0	1	1	1	0	2.577	0.68	-0.297	21.8	23	133	male	3	93.10	49

1	2	1	1	1	1	1	3.414	1.28					132	male	3	100.00	50
0	1	1	1	1	1	0	1.077	0.42	0.159	21.8	23	131	male	3	72.41	51	
1	1	1	1	1	1	1	0.83	0.45	-0.903	16.1	17	130	male	3	69.57	52	
1	2	1	1	1	1	1	2.198	0.58	-0.223	21.8	23	129	male	3	89.66	53	
1	2	1	1	1	1	1	0.174	0.38	1.019	21.8	23	128	male	3	51.72	54	
1	1	1	1	1	1	1	-0.725	0.41	-0.159	21.8	23	127	male	3	31.03	55	
1	2	1	1	1	0	0	3.937	1.21					126	male	3	100.00	56
1	2	1	1	1	1	1	2.198	0.58	-0.724	21.8	23	125	male	3	89.66	57	
1	2	1	1	1	1	1	2.577	0.68	-0.297	21.8	23	124	male	3	93.10	58	
1	1	1	1	1	1	1	0.757	0.4	-0.629	21.8	23	123	male	3	65.52	59	
1	2	1	1	1	1	1	0.757	0.4	0.429	21.8	23	122	male	3	65.52	60	
	2	1	1	1	1	1	1.077	0.42	0.28	21.8	23	121	male	3	72.41	61	
0	2	1	1	1	0	0	3.937	1.21					120	male	3	100.00	62
1	1	1	1	1	0	1	3.937	1.21					119	male	3	100.00	63
1	2	1	1	1	1	1	2.577	0.68	-1.053	21.8	23	118	male	3	93.10	64	
1	1	0	1	0	0	0	2.577	0.68	-1.004	21.8	23	117	male	3	93.10	65	
0	0	0	0	1	0	0	2.577	0.68	-1.104	21.8	23	116	male	3	93.10	66	
1	2	1	1	1	1	1	1.903	0.53	-0.703	21.8	23	115	male	3	86.21	67	
1	1	1	1	1	1	1	2.577	0.68	-0.897	21.8	23	114	male	3	93.10	68	
1	2	1	1	1	1	1	0.757	0.4	2.225	21.8	23	113	male	3	65.52	69	
1	2	1	1	1	1	1	3.658	1.2					112	male	3	100.00	70
0	2	0	1	1	1	0	2.577	0.68	-0.234	21.8	23	111	male	3	93.10	71	
1	2	1	1	1	1	1	1.903	0.53	-0.19	21.8	23	110	male	3	86.21	72	
1	2	1	1	1	1	1	2.577	0.68	-1.03	21.8	23	109	male	3	93.10	73	
1	2	1	1	1	1	1	2.198	0.58	-1.234	21.8	23	108	male	3	89.66	74	
1	2	1	1	1	1	0	3.937	1.21					107	male	3	100.00	75
1	1	1	1	1	1	0	2.577	0.68	-1.103	21.8	23	106	male	3	93.10	76	

1	1	1	1	1	1	1	2.198	0.58	-1.22	21.8	23	105	male	3	89.66	77
1	2	1	1	1	1	1	3.13	0.86	-0.957	21.8	23	104	male	3	96.55	78
1	2	1	1	1	1	0	0.899	0.46	0.181	16.1	17	103	male	3	68.18	79
0	0	0	0	0	0	0	3.13	0.86	0.519	21.8	23	102	male	3	96.55	80
1							0.927	0.5	0.134	11.4	12	101	male	3	64.71	81
1	0	1	1	1	1	1	1.903	0.53	0.545	21.8	23	100	male	3	86.21	82
0	1	1	1	1	1	1	1.077	0.42	-1.35	21.8	23	99	male	3	72.41	83
1	1	1	1	1	1	1	1.903	0.53	0.234	21.8	23	98	male	3	86.21	84
1	1	1	1	1	1	1	3.13	0.86	-0.955	21.8	23	96	male	3	96.55	85
1	2	1	1	1	1	1	1.077	0.42	-0.233	21.8	23	95	male	3	72.41	86
1	2	1	1	1	1	1	0.757	0.4	-1.529	21.8	23	92	male	3	65.52	87
1	1	1	1	1	1	1	3.937	1.21				90	male	3	100.00	88
1	2	1	1	1	1	1	1.077	0.42	-1.259	21.8	23	89	male	3	72.41	89
0							3.13	0.86	0.002	21.8	23	88	male	3	96.55	90
0	2	1	1	1	1	1	3.13	0.86	-0.971	21.8	23	87	male	3	96.55	91
							2.577	0.68	-0.465	21.8	23	86	male	3	93.10	92
1	2	0	1	1	1	1	1.903	0.53	-0.637	21.8	23	85	male	3	86.21	93
1	1	1	1	1	1	0	0.317	0.38	1.358	21.8	23	84	male	3	55.17	94
1	2	1	1	1	1	0	0.913	0.41	0.713	21.8	23	190	female	3	68.97	95
1	1	1	1	1	1	1	-0.112	0.39	0.788	21.8	23	189	female	3	44.83	96
1	2	1	1	1	1	1	2.198	0.58	-0.064	21.8	23	187	female	3	89.66	97
1	1	1	1	1	0	1	1.253	0.44	-0.617	21.8	23	186	female	3	75.86	98
1	2	1	1	1	1	1	1.903	0.53	-0.431	21.8	23	184	female	3	86.21	99
1	2	1	1	1	1	0	0.46	0.39	-1.134	21.8	23	182	female	3	58.62	100
1	2	1	1	1	1	1	0.757	0.4	-0.323	21.8	23	179	female	3	65.52	101
1	2	1	1	1	1	1	3.13	0.86	-0.212	21.8	23	162	female	3	96.55	102
1	2	1	1	1	1	1	1.077	0.42	-0.289	21.8	23	157	female	3	72.41	103

1	0	1	1	1	1	1	1.253	0.44	-0.584	21.8	23	156	female	3	75.86	104
0	1	1	1	1	1	0	-0.112	0.39	-0.175	21.8	23	150	female	3	44.83	105
1	2	1	1	1	1	0	1.445	0.46	-0.713	21.8	23	83	female	2	79.31	106
1	0	1	0	1	1	1	1.658	0.49	0.29	21.8	23	82	female	2	82.76	107
0	0	1	1	1	1	0	0.83	0.45	1.2	16.1	17	81	female	2	69.57	108
	0	1	0	0	0	0	0.882	0.41	1.802	20.8	22	80	female	2	67.86	109
1	1	1	1	1	1	1	1.253	0.44	0.12	21.8	23	79	female	2	75.86	110
1	0	1	1	1	1	0	1.077	0.42	0.347	21.8	23	78	female	2	72.41	111
1	2	1	1	1	1	1	2.577	0.68	0.874	21.8	23	77	female	2	93.10	112
0	0	0	1	1	1	0	0.757	0.4	0.828	21.8	23	76	female	2	65.52	113
0	0	1	1	1	1	0	0.83	0.45	1.2	16.1	17	75	female	2	69.57	114
0	1	1	0	1	1	0	0.606	0.39	-1.107	21.8	23	74	female	2	62.07	115
							1.443	0.56	0.137	11.4	12	73	female	2	76.47	116
1	2	1	1	1	1	1	1.903	0.53	-0.865	21.8	23	72	female	2	86.21	117
0	0	1	1	1	1	0	0.757	0.4	1.085	21.8	23	71	female	2	65.52	118
							0.478	0.48	-1.262	11.4	12	70	female	2	52.94	119
							0.927	0.5	-0.937	11.4	12	69	female	2	64.71	120
1	0	1	0	0	0	0	-0.246	0.43	1.266	16.1	17	68	female	2	43.48	121
1	2	1	1	0	1	1	1.8	0.59	0.547	16.1	17	67	female	2	86.96	122
1	2	0	1	1	1	1	2.176	0.68	0.996	16.1	17	66	female	2	91.30	123
1	2	1	1	1	0	1	2.198	0.58	-0.119	21.8	23	65	female	2	89.66	124
1	2	1	1	1	1	1	0.913	0.41	-0.249	21.8	23	64	female	2	68.97	125
1	0	1	0	0	0	0	0.032	0.38	1.653	21.8	23	63	female	2	48.28	126
1	2	1	1	0	1	1	1.658	0.49	0.252	21.8	23	62	female	2	82.76	127
0	1	1	0	1	0	0	-0.483	0.64	-1.057	9.5	10	61	female	2	45.45	128
1	2	1	1	1	1	1	1.445	0.46	0.279	21.8	23	60	female	2	79.31	129
0							-0.401	0.45	0.433	16.1	17	59	female	2	36.36	130

1	1	1	1	1	1	1	0.913	0.41	-0.332	21.8	23	58	female	2	68.97	131
1	2	1	1	1	1	1	2.577	0.68	0.11	21.8	23	57	female	2	93.10	132
1							0.704	0.45	0.979	16.1	17	56	female	2	63.64	133
1	1	0	1	1	0	0	1.077	0.42	0.329	21.8	23	55	female	2	72.41	134
1	1	1	1	1	1	0	1.8	0.59	-0.238	16.1	17	54	female	2	86.96	135
1	1	1	1	1	1	1	1.445	0.46	0.1	21.8	23	52	female	2	79.31	137
1	2	1	1	1	1	1	3.937	1.21				51	female	2	100.00	138
0							0.899	0.46	1.084	16.1	17	50	female	2	68.18	139
1	2	0	0	1	1	0	1.032	0.47	1.25	16.1	17	49	female	2	73.91	140
1	2	1	1	1	1	1	1.253	0.44	-1.116	21.8	23	48	female	2	75.86	141
0	1	1	1	1	1	0	0.317	0.38	-1.474	21.8	23	47	female	2	55.17	142
1	0	1	1	0	1	1	0.606	0.39	0.666	21.8	23	46	female	2	62.07	143
1	2	0	1	1	0	0	0.46	0.39	1.872	21.8	23	44	female	2	58.62	145
1	1	1	1	1	1	1	0.46	0.39	-0.916	21.8	23	43	female	1	58.62	146
1	1	0	0	0	0	0	-0.258	0.39	-0.481	21.8	23	42	female	1	41.38	147
0	2	1	1	1	1	0	-0.258	0.39	0.725	21.8	23	41	female	1	41.38	148
1	1	1	1	1	1	0	0.032	0.38	-1.83	21.8	23	40	female	1	48.28	149
1	0	1	1	0	0	0	0.913	0.41	0.791	21.8	23	39	female	1	68.97	150
0	2	1	1	1	1	1	1.253	0.44	-0.437	21.8	23	38	female	1	75.86	151
1	1	1	0	1	1	1	-0.258	0.39	-0.237	21.8	23	37	female	1	41.38	152
0	2	0	1	0	0	0	-0.407	0.4	1.749	21.8	23	36	female	1	37.93	153
1	1	1	1	1	1	1	0.757	0.4	-1.195	21.8	23	35	female	1	65.52	154
1	1	1	1	1	1	1	1.077	0.42	-1.262	21.8	23	34	female	1	72.41	155
1	1	1	1	1	1	1	1.253	0.44	-0.32	21.8	23	33	female	1	75.86	156
1	2	1	1	1	1	0	0.606	0.39	-0.734	21.8	23	32	female	1	62.07	157
0	0	1	0	0	0	0	-0.112	0.39	0.399	21.8	23	31	female	1	44.83	158
1							0.335	0.43	0.16	16.1	17	30	female	1	54.55	159

1	1	1	0	1	1	0	1.903	0.53	0.289	21.8	23	29	female	1	86.21	160
0	1	1	1	1	0	0	0.46	0.39	-0.135	21.8	23	28	female	1	58.62	161
0	0	0	0	0	0	0	-1.511	0.5	0.78	21.8	23	27	female	1	17.24	162
1	0	1	0	1	0	0	0.606	0.39	1.834	21.8	23	26	female	1	62.07	163
1	1	0	0	1	0	0	-0.725	0.41	0.955	21.8	23	25	female	1	31.03	164
1	0	1	1	1	1	0	0.913	0.41	0.394	21.8	23	24	female	1	68.97	165
1	1	1	1	1	1	0	0.757	0.4	0.585	21.8	23	23	female	1	65.52	166
1	2	1	1	1	1	0	1.658	0.49	-0.691	21.8	23	21	female	1	82.76	167
1							1.107	0.48	0.455	16.1	17	20	female	1	72.73	168
1	0	1	0	0	1	0	0.174	0.38	0.455	21.8	23	18	female	1	51.72	170
1	0	1	1	1	1	1	1.445	0.46	0.594	21.8	23	17	female	1	79.31	171
1	1	1	1	1	1	1	2.577	0.68	-1.104	21.8	23	16	female	1	93.10	172
1	1	1	1	1	1	0	0.913	0.41	-1.054	21.8	23	15	female	1	68.97	173
1	0	1	1	1	1	1	1.077	0.42	1.586	21.8	23	14	female	1	72.41	174
0	0	1	1	0	0	0	-0.112	0.39	1.777	21.8	23	13	female	1	44.83	175
0	1	1	1	1	1		1.161	0.44	-0.612	20.8	22	12	female	1	75.00	176
1	1	0	0				0.124	0.4	0.786	18.9	20	11	female	1	50.00	177
1	2	1	1	1	1	1	1.903	0.53	-0.524	21.8	23	10	female	1	86.21	178
1	1	1	1	1	1	1	0.757	0.4	0.35	21.8	23	9	female	1	65.52	179
1	1	1	1	1	1	1	1.445	0.46	0.461	21.8	23	8	female	1	79.31	180
1	1	1	1	1	1	0	1.445	0.46	0.058	21.8	23	7	female	1	79.31	181
1	1	1	1	1	1	1	0.032	0.38	-1.148	21.8	23	6	female	1	48.28	182
1	2	1	1	1	1	0	0.733	0.42	-1.234	20.8	22	5	female	1	66.67	183
1	2	1	1	1	1	1	3.937	1.21				4	female	1	100.00	184
1	1	1	1	1	1	1	2.474	0.69	-1.124	20.8	22	3	female	1	92.59	185
1							1.376	0				2	female	1	100.00	186
1	1	1	1	1	1	1	2.577	0.68	0.076	21.8	23	1	female	1	93.10	187

H-7 :All student scores broken down by type and gender and ANOVA analysis

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)
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
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	

ANOVA outputs for gender analysis

Anova: Single Factor: Comparing male and female overall score %

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
Male %	94.00	7364.22	78.34	248.48	15.8
Female %	90.00	6174.40	68.60	300.05	17.3

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	4360.31	1.00	4360.31	15.93	0.00010	3.89
Within Groups	49813.35	182.00	273.70			
Total	54173.66	183.00				

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

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
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

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
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				

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

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
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

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
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				

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

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
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

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
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				

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

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
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

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
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				

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

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
Male AB MC scores %	94.00	6333.33	67.38	811.92	28.5
Female AB MC scores %	90.00	5133.33	57.04	805.10	28.4

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	4914.70	1.00	4914.70	6.08	0.0146	3.89
Within Groups	147162.60	182.00	808.59			
Total	152077.29	183.00				

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

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
Male AB SA scores %	94.00	6781.82	72.15	667.93	25.8
Female AB SA scores %	90.00	5845.45	64.95	662.20	25.7

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	2381.86	1.00	2381.86	3.58	0.0600	3.89
Within Groups	121052.83	182.00	665.13			
Total	123434.69	183.00				

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

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
Male Stoichiometry MC scores %	94.00	8200.00	87.23	463.23	21.5
Female Stoichiometry MC scores %	90.00	6800.00	75.56	681.15	26.1

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	6270.84	1.00	6270.84	11.01	0.0011	3.89
Within Groups	103703.07	182.00	569.80			
Total	109973.91	183.00				

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

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
Male Stoichiometry SA scores %	94.00	7357.14	78.27	705.41	26.5
Female Stoichiometry SA scores %	90.00	5514.29	61.27	1051.34	32.4

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	13284.02	1.00	13284.02	15.19	0.00014	3.89
Within Groups	159172.72	182.00	874.58			
Total	172456.74	183.00				

ANOVA analysis all question types

Anova: Single Factor: Comparing all stoichiometry with all acid-base responses.

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
Acid-base scores %	184.00	12217.65	66.40	520.30	22.80
Stoichiometry scores %	184.00	13758.33	74.77	641.65	25.30

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	6450.31	1.00	6450.31	11.10	0.0009	3.87
Within Groups	212636.82	366.00	580.97			
Total	219087.13	367.00				

Anova: Single Factor: Comparing all multiple-choice with all short-answer responses.

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
MC scores %	184.00	13072.73	71.05	431.61	20.80
SA scores %	184.00	12722.22	69.14	498.12	22.30

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	333.84	1.00	333.84	0.72	0.40	3.87
Within Groups	170140.97	366.00	464.87			
Total	170474.82	367.00				

Anova: Single Factor: Comparing all acid-base multiple-choice and short-answer responses.

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
Acid-base MC scores %	184.00	11466.67	62.32	831.02	28.80
Acid-base SA scores %	184.00	12627.27	68.63	674.51	25.90

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	3660.34	1.00	3660.34	4.86	0.03	3.87
Within Groups	275511.99	366.00	752.76			
Total	279172.33	367.00				

Anova: Single Factor: Comparing all stoichiometry multiple-choice and short-answer responses.

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
Stoichiometry MC scores %	184.00	15000.00	81.52	600.95	24.50
Stoichiometry SA scores %	184.00	12871.43	69.95	942.39	30.70

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	12312.00	1.00	12312.00	15.96	0.00008	3.87
Within Groups	282430.66	366.00	771.67			
Total	294742.66	367.00				

Anova: Single Factor: Comparing all stoichiometry multiple-choice and acid-base multiple-choice responses

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
Acid-base MC scores %	184.00	11466.67	62.32	831.02	28.80
Stoichiometry MC scores %	184.00	15000.00	81.52	600.95	24.50

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	33925.12	1.00	33925.12	47.38	0.00000000003	3.87
Within Groups	262051.21	366.00	715.99			
Total	295976.33	367.00				

Anova: Single Factor: Comparing all stoichiometry multiple-choice and acid-base multiple-choice responses

SUMMARY

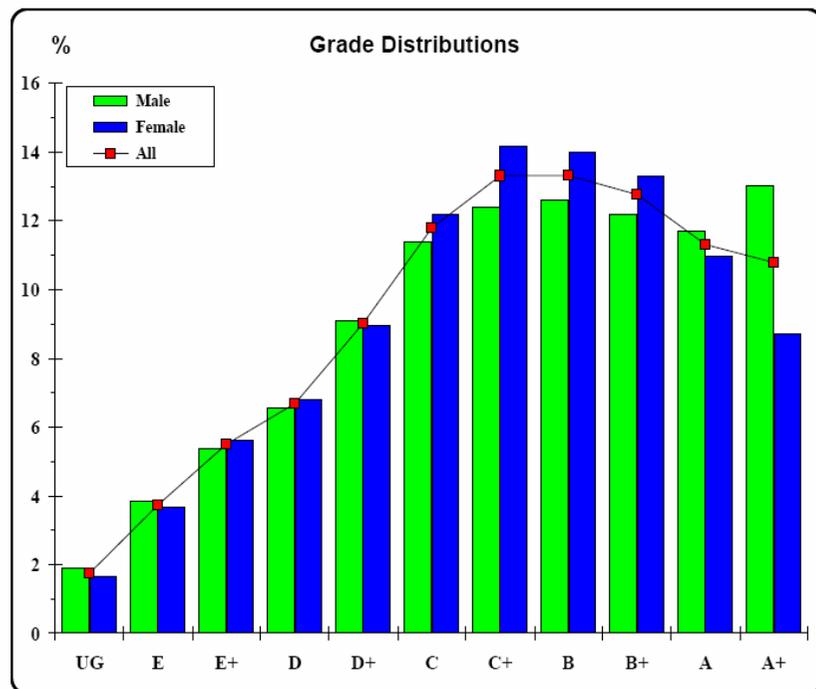
<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>	<i>s.d.</i>
Acid-base SA scores %	184.00	12627.27	68.63	674.51	26.00
Stoichiometry SA scores %	184.00	12871.43	69.95	942.39	30.70

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	161.99	1.00	161.99	0.20	0.65	3.87
Within Groups	295891.44	366.00	808.45			
Total	296053.43	367.00				

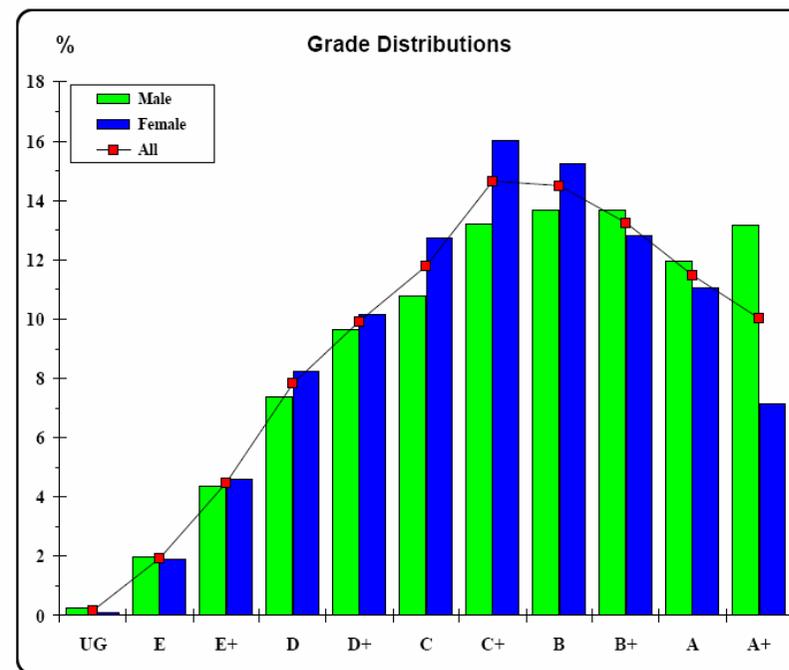
H-8: Grade distribution graphs for Chemistry 2008

Grade distribution for Unit 3 examination Chemistry 2008



(VCAA, 2009a)

Grade distribution for Unit 4 examination Chemistry 2008



(VCAA, 2009a)

Appendix I: Unit 3 examination results compared to Unit 4 examination results

I-1 Raw data comparing the Unit 3 and Unit 4 examinations.

UNIT 3				Unit 4			
<i>Q no.</i>	<i>Score</i>	<i>Type</i>	<i>Type</i>	<i>Q no.</i>	<i>Score</i>	<i>Type</i>	<i>Type</i>
1	80	Recall	MC	7	65	Recall	MC
2	90	Recall	MC	8	62	Recall	MC
6	77	Recall	MC	10	59	Recall	MC
7	65	Recall	MC	11	71	Recall	MC
9	94	Recall	MC	12	79	Recall	MC
11	80	Recall	MC	13	75	Recall	MC
12	79	Recall	MC	14	70	Recall	MC
16	46	Recall	MC	15	58	Recall	MC
18	92	Recall	MC	16	83	Recall	MC
3	45.0	Application	MC	20	62	Recall	MC
4	86.0	Application	MC	1	82	Application	MC
5	62.0	Application	MC	2	88	Application	MC
8	49.0	Application	MC	3	73	Application	MC
10	64.0	Application	MC	4	67	Application	MC
13	70.0	Application	MC	5	49.0	Application	MC
14	67.0	Application	MC	6	61.0	Application	MC
15	57.0	Application	MC	9	76.0	Application	MC
17	67.0	Application	MC	17	50.0	Application	MC
19	64.0	Application	MC	18	48.0	Application	MC
20	46.0	Application	MC	19	33.0	Application	MC
1	77	Recall	MC	1	80	Recall	MC
2	76	Recall	MC	2	65	Recall	MC
11	65	Recall	MC	3	80	Recall	MC
12	43	Recall	MC	5	77	Recall	MC
19	77	Recall	MC	8	53	Recall	MC
3	68.0	Application	MC	11	40	Recall	MC
4	58.0	Application	MC	12	58	Recall	MC
5	47.0	Application	MC	13	58	Recall	MC
6	71.0	Application	MC	15	82	Recall	MC
7	58.0	Application	MC	17	83	Recall	MC
8	64.0	Application	MC	18	67	Recall	MC
9	63.0	Application	MC	19	73	Recall	MC
10	55.0	Application	MC	20	78	Recall	MC
13	41.0	Application	MC	4	45.0	Application	MC
14	24.0	Application	MC	6	72.0	Application	MC
15	69.0	Application	MC	7	52.0	Application	MC
16	59.0	Application	MC	9	57.0	Application	MC
17	49.0	Application	MC	10	58.0	Application	MC

18	66.0	Application	MC	14	51.0	Application	MC
20	50.0	Application	MC	16	67.0	Application	MC
1	84	Recall	MC	1	62	Recall	MC
2	86	Recall	MC	2	79	Recall	MC
3	86	Recall	MC	3	59	Recall	MC
7	42	Recall	MC	12	54	Recall	MC
9	55	Recall	MC	15	71	Recall	MC
16	69	Recall	MC	16	68	Recall	MC
18	79	Recall	MC	17	70	Recall	MC
19	55	Recall	MC	20	61	Recall	MC
20	61	Recall	MC	4	42.0	Application	MC
4	58.0	Application	MC	5	61.0	Application	MC
5	52.0	Application	MC	6	72.0	Application	MC
6	34.0	Application	MC	7	43.0	Application	MC
8	47.0	Application	MC	8	55.0	Application	MC
10	70.0	Application	MC	9	67.0	Application	MC
11	59.0	Application	MC	10	74.0	Application	MC
12	39.0	Application	MC	11	45.0	Application	MC
13	36.0	Application	MC	13	54.0	Application	MC
14	61.0	Application	MC	14	50.0	Application	MC
15	33.0	Application	MC	18	41.0	Application	MC
17	69.0	Application	MC	19	66.0	Application	MC
1	84	Recall	MC	1	93.0	Recall	MC
2	86	Recall	MC	2	95.0	Recall	MC
3	86	Recall	MC	5	75.0	Recall	MC
7	42	Recall	MC	6	67.0	Recall	MC
9	55	Recall	MC	10	57.0	Recall	MC
16	69	Recall	MC	12	66.0	Recall	MC
18	79	Recall	MC	13	85.0	Recall	MC
19	55	Recall	MC	14	66.0	Recall	MC
20	61	Recall	MC	15	78.0	Recall	MC
4	58.0	Application	MC	16	84.0	Recall	MC
5	52.0	Application	MC	17	73.0	Recall	MC
6	34.0	Application	MC	18	80.0	Recall	MC
8	47.0	Application	MC	19	73.0	Recall	MC
10	70.0	Application	MC	3	67.0	Application	MC
11	59.0	Application	MC	4	60.0	Application	MC
12	39.0	Application	MC	7	53.0	Application	MC
13	36.0	Application	MC	8	57.0	Application	MC
14	61.0	Application	MC	9	56.0	Application	MC
15	33.0	Application	MC	11	48.0	Application	MC
17	69.0	Application	MC	20	45.0	Application	MC
1	77.0	Recall	MC	2	58.0	Recall	MC
2	78.0	Recall	MC	5	71.0	Recall	MC

5	66.0	Recall	MC	6	78.0	Recall	MC
9	48.0	Recall	MC	7	77.0	Recall	MC
11	71.0	Recall	MC	8	50.0	Recall	MC
13	82.0	Recall	MC	12	66.0	Recall	MC
18	30.0	Recall	MC	13	56.0	Recall	MC
3	56.0	Application	MC	14	64.0	Recall	MC
4	66.0	Application	MC	20	35.0	Recall	MC
6	62.0	Application	MC	1	65.0	Application	MC
7	49.0	Application	MC	3	73.0	Application	MC
8	71.0	Application	MC	4	51.0	Application	MC
10	70.0	Application	MC	9	41.0	Application	MC
12	77.0	Application	MC	10	51.0	Application	MC
14	56.0	Application	MC	11	67.0	Application	MC
15	39.0	Application	MC	15	44.0	Application	MC
16	23.0	Application	MC	16	44.0	Application	MC
17	51.0	Application	MC	17	58.0	Application	MC
19	79.0	Application	MC	18	83.0	Application	MC
20	30.0	Application	MC	19	42.0	Application	MC
1a	80.0	Recall	SA	1a	63.0	Recall	SA
1d	40.5	Recall	SA	1c	61.5	Recall	SA
2d	54.5	Recall	SA	2a	68.0	Recall	SA
3a	52.0	Recall	SA	2b	67.0	Recall	SA
4a	72.0	Recall	SA	5a	55.0	Recall	SA
5b	75.0	Recall	SA	5b	60.0	Recall	SA
5c	62.5	Recall	SA	5c	65.5	Recall	SA
1b	62.5	Application	SA	5d	46.3	Recall	SA
1c	26.2	Application	SA	6a	75.5	Recall	SA
2a	81.0	Application	SA	6b	62.5	Recall	SA
2b	87.0	Application	SA	7a	34.0	Recall	SA
2c	71.7	Application	SA	7b	55.0	Recall	SA
3b	45.7	Application	SA	8a	62.0	Recall	SA
4b	89.0	Application	SA	8c	66.0	Recall	SA
4c	62.5	Application	SA	1b	48.5	Application	SA
5a	53.0	Application	SA	2c	56.5	Application	SA
6a	60.3	Application	SA	2d	54.5	Application	SA
6b	44.3	Application	SA	3a	76.0	Application	SA
7a	45.0	Application	SA	3b	66.0	Application	SA
7b	55.0	Application	SA	3c	64.5	Application	SA
7c	52.0	Application	SA	4a	75.5	Application	SA
1bi	58.0	Recall	SA	4b	56.5	Application	SA
1bii	33.0	Recall	SA	4c	65.5	Application	SA
1biii	34.0	Recall	SA	7c	51.0	Application	SA
1biv	64.0	Recall	SA	7d	15.0	Application	SA
5a	90.0	Recall	SA	7e	77.0	Application	SA

5bi	60.0	Recall	SA	8c	49.0	Application	SA
5bii	64.5	Recall	SA	8d	28.0	Application	SA
5biii	72.0	Recall	SA	1	64.0	Recall	SA
5ci	71.0	Recall	SA	2a	90.0	Recall	SA
5cii	52.0	Recall	SA	4a	40.0	Recall	SA
5d	63.0	Recall	SA	4bii	90.0	Recall	SA
1a	81.0	Application	SA	4biii	50.0	Recall	SA
1bv	80.5	Application	SA	5cii	50.0	Recall	SA
2ai	72.0	Application	SA	6a	40.0	Recall	SA
2aii	79.0	Application	SA	6b	67.5	Recall	SA
2aiii	75.0	Application	SA	6ci	70.0	Recall	SA
2aiv	51.0	Application	SA	6cii	60.0	Recall	SA
2b	68.0	Application	SA	7a	30.0	Recall	SA
2ci	28.5	Application	SA	7b	60.0	Recall	SA
2cii	60.0	Application	SA	8a	80.0	Recall	SA
2d	52.0	Application	SA	8c	55.0	Recall	SA
3	51.6	Application	SA	9bi	70.0	Recall	SA
4ai	79.0	Application	SA	9biii	65.0	Recall	SA
4aii	39.0	Application	SA	2b	80.0	Application	SA
4aiii	91.0	Application	SA	2c	70.0	Application	SA
4bi	91.0	Application	SA	3ai	73.3	Application	SA
4bii	39.0	Application	SA	3aii	40.0	Application	SA
4biii	86.0	Application	SA	3bi	75.0	Application	SA
4biv	9.0	Application	SA	3bii	30.0	Application	SA
6a	48.0	Application	SA	4bi	70.0	Application	SA
6b	55.5	Application	SA	5a	50.0	Application	SA
6c	58.5	Application	SA	5bi	70.0	Application	SA
1a	60.0	Recall	SA	5bii	70.0	Application	SA
1b	35.0	Recall	SA	5ci	40.0	Application	SA
1c	50.0	Recall	SA	8b	50.0	Application	SA
1d	50.0	Recall	SA	9ai	60.0	Application	SA
1e	45.0	Recall	SA	9aii	40.0	Application	SA
2a	60.0	Recall	SA	9bii	45.0	Application	SA
2c	82.5	Recall	SA	1	82.0	Recall	SA
3d	60.0	Recall	SA	2a	80.0	Recall	SA
4c	55.0	Recall	SA	2ci	60.0	Recall	SA
4d	70.0	Recall	SA	2cii	55.0	Recall	SA
5a	60.0	Recall	SA	2ciii	80.0	Recall	SA
5b	65.0	Recall	SA	2civ	80.0	Recall	SA
5c	55.0	Recall	SA	2cv	60.0	Recall	SA
5d	50.0	Recall	SA	3b	60.0	Recall	SA
7a	70.0	Recall	SA	3c	85.0	Recall	SA
7c	26.7	Recall	SA	4ai	40.0	Recall	SA
8a	80.0	Recall	SA	4b	67.5	Recall	SA

8b	80.0	Recall	SA	5a	76.7	Recall	SA
1e	45.0	Application	SA	5b	90.0	Recall	SA
2b	65.0	Application	SA	6a	30.0	Recall	SA
3a	80.0	Application	SA	6b	30.0	Recall	SA
3b	60.0	Application	SA	6c	50.0	Recall	SA
3c	65.0	Application	SA	7a	80.0	Recall	SA
4a	60.0	Application	SA	8c	50.0	Recall	SA
4b	50.0	Application	SA	9a	70.0	Recall	SA
6a	90.0	Application	SA	9b	40.0	Recall	SA
6b	60.0	Application	SA	9c	30.0	Recall	SA
6c	50.0	Application	SA	9d	40.0	Recall	SA
7b	53.3	Application	SA	3a	57.5	Application	SA
8c	40.0	Application	SA	5c	65.0	Application	SA
1a	60.0	Recall	SA	2b	73.3	Application	SA
1b	35.0	Recall	SA	4aii	55.0	Application	SA
1c	50.0	Recall	SA	7b	50.0	Application	SA
1d	50.0	Recall	SA	7c	40.0	Application	SA
1e	45.0	Recall	SA	7d	50.0	Application	SA
2a	60.0	Recall	SA	8a	40.0	Application	SA
2c	82.5	Recall	SA	8b	35.0	Application	SA
3d	60.0	Recall	SA	1a	80.0	Recall	SA
4c	55.0	Recall	SA	1b	46.7	Recall	SA
4d	70.0	Recall	SA	2b	50.0	Recall	SA
5a	60.0	Recall	SA	2c	70.0	Recall	SA
5b	65.0	Recall	SA	2d	80.0	Recall	SA
5c	55.0	Recall	SA	3c	70.0	Recall	SA
5d	50.0	Recall	SA	5a	50.0	Recall	SA
7a	70.0	Recall	SA	5b	70.0	Recall	SA
7c	26.7	Recall	SA	6a	70.0	Recall	SA
8a	80.0	Recall	SA	6b	50.0	Recall	SA
8b	80.0	Recall	SA	6c	40.0	Recall	SA
1e	45.0	Application	SA	7a	30.0	Recall	SA
2b	65.0	Application	SA	7b	80.0	Recall	SA
3a	80.0	Application	SA	8c	70.0	Recall	SA
3b	60.0	Application	SA	9d	70.0	Recall	SA
3c	65.0	Application	SA	2a	90.0	Application	SA
4a	60.0	Application	SA	3a	60.0	Application	SA
4b	50.0	Application	SA	3b	66.7	Application	SA
6a	90.0	Application	SA	4a	80.0	Application	SA
6b	60.0	Application	SA	4b	80.0	Application	SA
6c	50.0	Application	SA	4c	45.0	Application	SA
7b	53.3	Application	SA	4d	60.0	Application	SA
8c	40.0	Application	SA	4e	53.3	Application	SA
1b	80.0	Recall	SA	5c	25.0	Application	SA

1c	76.7	Recall	SA	5d	52.0	Application	SA
2a	90.0	Recall	SA	6d	40.0	Application	SA
4c	80.0	Recall	SA	8a	50.0	Application	SA
4d	63.3	Recall	SA	8b	30.0	Application	SA
5ai	70.0	Recall	SA	8d	60.0	Application	SA
5aii	80.0	Recall	SA	9a	60.0	Application	SA
5ci	70.0	Recall	SA	9b	40.0	Application	SA
5cii	50.0	Recall	SA	9c	50.0	Application	SA
5ciii	50.0	Recall	SA	2a	70.0	Recall	SA
1a	60.0	Application	SA	2b	68.8	Recall	SA
2bi	65.0	Application	SA	4d	70.0	Recall	SA
2bii	30.0	Application	SA	5a	50.0	Recall	SA
2biii	40.0	Application	SA	7a	56.7	Recall	SA
3a	70.0	Application	SA	7b	45.0	Recall	SA
3b	56.7	Application	SA	7c	36.7	Recall	SA
3c	30.0	Application	SA	1	56.7	Application	SA
4ai	60.0	Application	SA	3a	53.3	Application	SA
4aii	50.0	Application	SA	3b	70.0	Application	SA
4aiii	55.0	Application	SA	4a	52.5	Application	SA
4b	50.0	Application	SA	4b	35.0	Application	SA
5b	80.0	Application	SA	4c	46.7	Application	SA
5di	52.5	Application	SA	5b	60.0	Application	SA
5dii	50.0	Application	SA	6a	54.0	Application	SA
6a	50.0	Application	SA	6b	35.0	Application	SA
6b	60.0	Application	SA	8a	70.0	Application	SA
6c	40.0	Application	SA	8b	63.3	Application	SA
6d	40.0	Application	SA	8c	60.0	Application	SA
7a	76.7	Application	SA	8d	30.0	Application	SA
7bi	65.0	Application	SA				
7bii	70.0	Application	SA				
7biii	53.3	Application	SA				

I-2: Comparison of Unit 3 and Unit 4 multiple-choice questions

Oneway ANOVA

Descriptives

Score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
3.00	100	60.9100	16.51818	1.65182	57.6324	64.1876	23.00	94.00
4.00	100	63.5300	13.57021	1.35702	60.8374	66.2226	33.00	95.00
Total	200	62.2200	15.13529	1.07023	60.1096	64.3304	23.00	95.00

ANOVA

Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	343.220	1	343.220	1.502	.222
Within Groups	45243.100	198	228.501		
Total	45586.320	199			

I-3: Comparison of Unit 3 and Unit 4 short-answer questions

Oneway ANOVA

Descriptives

Score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
3.00	145	59.9103	16.00695	1.32931	57.2829	62.5378	9.00	91.00
4.00	142	57.9155	15.94462	1.33804	55.2703	60.5607	15.00	90.00
Total	287	58.9233	15.97946	.94324	57.0668	60.7799	9.00	91.00

ANOVA

Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	285.493	1	285.493	1.119	.291
Within Groups	72742.600	285	255.237		
Total	73028.094	286			

I-4: Comparison of Unit 3 and Unit 4 application questions

Oneway ANOVA

Descriptives

Score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
3.00	142	57.2859	15.72281	1.31943	54.6775	59.8943	9.00	91.00
4.00	115	56.0443	14.45226	1.34768	53.3746	58.7141	15.00	90.00
Total	257	56.7304	15.15094	.94509	54.8692	58.5915	9.00	91.00

ANOVA

Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	97.948	1	97.948	.426	.515
Within Groups	58667.096	255	230.067		
Total	58765.043	256			

I-5: Comparison of Unit 3 and Unit 4 recall questions

Oneway ANOVA

Descriptives

Score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
3.00	103	64.4990	15.96725	1.57330	61.3784	67.6197	26.70	94.00
4.00	127	64.0307	14.97799	1.32908	61.4005	66.6609	30.00	95.00
Total	230	64.2404	15.39644	1.01521	62.2401	66.2408	26.70	95.00

ANOVA

Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	12.474	1	12.474	.052	.819
Within Groups	54272.080	228	238.035		
Total	54284.554	229			

I-6: Comparison of Unit 3 and Unit 4 all questions

Oneway ANOVA

Descriptives

Score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
3.00	245	60.3184	16.19140	1.03443	58.2808	62.3559	9.00	94.00
4.00	242	60.2355	15.23361	.97925	58.3065	62.1645	15.00	95.00
Total	487	60.2772	15.70663	.71174	58.8787	61.6757	9.00	95.00

ANOVA

Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.835	1	.835	.003	.954
Within Groups	119894.522	485	247.205		
Total	119895.357	486			