INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Achieving equity in self-selected subsets of test items

Wang, Xiang Bo, Ph.D.

University of Hawaii, 1992
ACHIEVING EQUITY
IN SELF-SELECTED SUBSETS OF TEST ITEMS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF
UNIVERSITY OF HAWAI'I IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
EDUCATIONAL PSYCHOLOGY
AUGUST 1992

By
Xiang Bo Wang

Dissertation Committee
Peter Dunn-Rankin, Chairperson
Harold I. Ayabe
Ronald H. Heck
Francis M. Pottenger
Shuqiang Zhang
Copyright 1992
By
Xiang Bo Wang
All Rights Reserved
ACKNOWLEDGEMENTS

An endeavor of this magnitude is only successful with the generous support of many. My whole-hearted gratitude goes to all who guided and assisted me through this long journey.

My foremost gratitude goes to my dissertation committee members for their impromptu and unconditional assistance, without which this dissertation would not have been completed in time:

To Peter Dunn-Rankin for his immense encouragement and guidance of all these years during my Ph.D program. He has always made himself available for consultation and has been especially considerate of the needs of mine as a foreign student. I am especially grateful to him for patiently editing every chapter in this dissertation.

To Harold Ayabe for starting and maintaining my interests in testing and measurement and his special support of my work in this Department.

To Ronald Heck for his insightful recommendations to improve the structure and components of my research instrument as well as this dissertation.

To Frank Pottenger for his advice on how to carry out my data collection and on how to relate my dissertation study to the science education in the State of Hawai‘i.

To Shuqiang Zhang for his extensive in-depth comments to perfect my dissertation.

I wish to thank my friend John Thomas for his detailed description of the 1989 AP Chemistry Exam content from a teacher’s point of view.
The author owes the initiation of this dissertation research to Howard Wainer of
the Educational Testing Service, and David Thissen of the University of North Carolina.
Furthermore, I greatly appreciate their continued willingness to overcome the long-
distance barrier between Hawaii and the East Coast to freely provide their valuable time
and experiences in solving many of my problems in research design and IRT computer
implementation. I wish to thank the Educational Testing Service for supporting my study
and work on polytomous item response theory as a predoctoral fellow.

I sincerely thank the following people and institutions for the wide range of data
used in this dissertation:

Instrumental to the release of the 1989 AP Chemistry data were Walter
MacDonald and Rick Morgan of the College Board at the Educational Testing Service.
Valuable information on the structure and interpretation of the data was provided by
Behroz Manecshana, Joe Stevens and Hessie Taft. Their expert support and kind
cooperation is appreciated.

The author is deeply indebted to the following professors and chemistry teachers
for the dissertation data collected in the State of Hawai‘i: Gordon Barrow, John Gilje,
Randy Larsen, McDonald Ray of the Chemistry Department, University of Hawai‘i; Mel
Ishii of Aiea High School; Pam Fujinaka, Iolani School; Maria Mendoza and Amy
Watanabe from Kailua High School; Vickie Kirahira of Kaiser High School; Dennis
Otta, Kalani High School; Jene Kullingham and Jerry Johnson of Kamehameha High
School; Katherine Taniguchi of McKinley High School; Edward Ginoza from Maui
High School; Ruth Fukuchi of Mililani High School; Alice Kaneko, Moanalua High
School; Ted Brattstrom of Pearl City High School; Michael Hu from Punahou School;
Sharlene Higa of Roosevelt High School; and Nancy Rocheleau of the Sacret Hearts Academy. It is because of their full support and timely cooperation that the author succeeded in collecting data on over 600 subjects during a critical 20-day period after the students finished their AP curriculum and before they took the 1992 official AP Chemistry Exam.

Most importantly, I would like to thank my family, my dear wife, Liang and my lovely son, Louis for their love, understanding, warmth and encouragement necessary to finish this project. I am also particularly grateful to my parents in China for instilling deep in me the value and power of education. Graduation means the start of a life-time mission to contribute to the improvement of society.
On the basis of 18,462 students who took the 1989 Advanced Placement Chemistry Exam and the 618 students in Hawaii who replied to the AP Chemistry Survey and Test Kit, this dissertation has researched three major issues concerning students’ choices of and performance on self-selected test items, as well as on the technical aspects of equating the scores on chosen items.

First, most students tended to choose the items representing the "core" content of a AP chemistry. However, contrary to the common expectations that students would perform better on the items they choose to answer, the mean scores on the more frequently chosen "core" chemistry items were lower than the less frequently chosen "non-core" chemistry items. Interestingly, the choice patterns of the students of all ability levels were almost identical. Second, it can be concluded that students tended to choose items whose content was considered easier, more familiar and belonging to similar dimensions. They performed better on their chosen items. Finally, equating the scores on chosen items through IRT rather than using raw scores results in fine-tuned measures of student abilities and more accurate classifications of students.
<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
</tr>
<tr>
<td>ABSTRACT</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
</tr>
<tr>
<td>CHAPTER I</td>
</tr>
<tr>
<td>INTRODUCTION</td>
</tr>
<tr>
<td>CHAPTER II</td>
</tr>
<tr>
<td>LITERATURE REVIEW</td>
</tr>
<tr>
<td>Definition of Constructed Response Items</td>
</tr>
<tr>
<td>Advantages and Disadvantages of Constructed Response Items</td>
</tr>
<tr>
<td>Choices of Constructed Responses and Problems</td>
</tr>
<tr>
<td>CHAPTER III</td>
</tr>
<tr>
<td>RESEARCH ISSUES AND QUESTIONS</td>
</tr>
<tr>
<td>Investigating General Item Choice Tendencies</td>
</tr>
<tr>
<td>Investigating Effects of Curricular Experiences and Psychological Factors Behind Choice Behavior</td>
</tr>
<tr>
<td>Investigating the Consequences of Missclassification Using CR Items without Proper Statistical Adjustments</td>
</tr>
<tr>
<td>Chapter Summary</td>
</tr>
<tr>
<td>CHAPTER IV</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

ix
<table>
<thead>
<tr>
<th>Chapter Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Checking the Invariance of Item Parameter Estimates</td>
<td>45</td>
</tr>
<tr>
<td>Chapter Summary</td>
<td>49</td>
</tr>
<tr>
<td><strong>CHAPTER V</strong> REVIEW OF RELATED RESEARCH METHODOLOGY</td>
<td>51</td>
</tr>
<tr>
<td>Item Response Theory</td>
<td>52</td>
</tr>
<tr>
<td>Item Response Theory vs. Classical Measurement Theory</td>
<td>53</td>
</tr>
<tr>
<td>Invariance of Item Parameter and Examinee Ability Estimation</td>
<td>54</td>
</tr>
<tr>
<td>IRT Advantages over Classical Measurement Theory</td>
<td>56</td>
</tr>
<tr>
<td>Three Parameter and Nominal IRT Models</td>
<td>58</td>
</tr>
<tr>
<td>Test Score Equating</td>
<td>59</td>
</tr>
<tr>
<td>Definition and Assumption of Test Score Equating</td>
<td>60</td>
</tr>
<tr>
<td>Classical Equating Methods</td>
<td>61</td>
</tr>
<tr>
<td>Item Response Theory Linking Methods</td>
<td>63</td>
</tr>
<tr>
<td>Methods of Unidimensional and Multidimensional Scaling</td>
<td>64</td>
</tr>
<tr>
<td>Chapter Summary</td>
<td>65</td>
</tr>
<tr>
<td><strong>CHAPTER VI</strong> OVERALL ESSAY CHOICE TENDENCIES AND PERFORMANCE</td>
<td>66</td>
</tr>
<tr>
<td>Relative Difficulty Levels of the Five Essays</td>
<td>66</td>
</tr>
<tr>
<td>Essay Difficulty as Reflected by IRT Expected Scores</td>
<td>68</td>
</tr>
</tbody>
</table>
Relationship between Essay Dimensionality and Essay Choices .......... 131

Relationship between Students' Perceptions of Essay Difficulty and Their Choices .......... 132

Accounting for the Discrepancy between Choice Popularity and Mean Performance in Light of the Psychological Processes Underlying Multiple Choice Questions .......... 133

Dimensionality of the Four MC Items .......... 135

Perceptions of MC Item Difficulty Levels .......... 137

Comparison of Item Difficulties as Reflected by IRT, Student Perception and Mean Scores .......... 137

Accounting for the Relationship among Student Ability, Extent of Study, Item Choices and Performance through Multiple Choice Questions .......... 140

The Mean Ability of the Four Groups of Hawaii Students .......... 142

Student Judgment, Performance, Choice and Choice Change in Comparison I .......... 143

Student Judgment, Performance, Choice and Choice Change in Comparison II .......... 145

Student Judgment, Performance, Choice and Choice Change in Comparison III .......... 147

Summarizing the Relationship among Student Performance, Ability, Extent of Study, and the Item Choice Through Logistic Regression Models .......... 150

Choosing among More Than Two Items .......... 153

Chapter Summary .......... 156
<table>
<thead>
<tr>
<th>CHAPTER IX</th>
<th>EQUATING SCORES ON STUDENT SELECTED ITEMS</th>
<th>158</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investigating the Effects of Choice Through MC Items</td>
<td>........................................</td>
<td>161</td>
</tr>
<tr>
<td>Investigating the Relationship between the Raw Scores and IRT Equated Scores</td>
<td>........................................</td>
<td>164</td>
</tr>
<tr>
<td>Investigating the Accuracy of IRT Equating on Choice Items</td>
<td>........................................</td>
<td>171</td>
</tr>
<tr>
<td>T-Test of the Difference between Full and Partial IRT and Raw Scores</td>
<td>........................................</td>
<td>172</td>
</tr>
<tr>
<td>Intercorrelations Among the Full and Partial IRT and Raw Scores</td>
<td>........................................</td>
<td>173</td>
</tr>
<tr>
<td>Comprehensive Analysis of the Effects of Equating vs Not-equating Scores on Chosen Items</td>
<td>........................................</td>
<td>176</td>
</tr>
<tr>
<td>Comparison of Performance on the 20 vs 17 MC Items Based on Observed Raw Scores</td>
<td>........................................</td>
<td>177</td>
</tr>
<tr>
<td>Comparison of Performance on the 20 vs 17 MC Items Based on Standardized Scores</td>
<td>........................................</td>
<td>179</td>
</tr>
<tr>
<td>Comparison of Performance on the 20 vs 17 MC Items After IRT Equating</td>
<td>........................................</td>
<td>181</td>
</tr>
<tr>
<td>Illustration of the Effects of Equating</td>
<td>........................................</td>
<td>184</td>
</tr>
<tr>
<td>Chapter Summary</td>
<td>........................................</td>
<td>192</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER X</th>
<th>DISCUSSIONS AND RECOMMENDATIONS</th>
<th>194</th>
</tr>
</thead>
<tbody>
<tr>
<td>How to Reduce the Disparities of Students’ Choices</td>
<td>........................................</td>
<td>194</td>
</tr>
<tr>
<td>How to Control the Difficulty of CR Items</td>
<td>........................................</td>
<td>195</td>
</tr>
<tr>
<td>How to Improve Students’ General Performance on the Essay Parts</td>
<td>........................................</td>
<td>198</td>
</tr>
</tbody>
</table>
How to Encourage More Students to Choose Deliberately .................. 199

How to Minimize the Differential Gains in the Scores of Students of Different Ability Levels .................. 200

How to Equate Scores on the Choice items of Different Content .................. 203
APPENDICES

Appendix A  Advanced Placement Chemistry Survey and Test Kit  .......... 205
Appendix B  Chemistry Teacher Evaluation  .......... 219
Appendix C  Chemistry Teacher Expert Judgment Survey  .......... 225
Appendix D  Essential Chemistry Knowledge Components Required by the Multiple Choice Items in Part B and Part C of Advanced Placement Chemistry Survey and Test Kit  .......... 227

REFERENCES  .......... 228
LIST OF TABLES

Table

4.1 Major Characteristics of AP Chemistry Exam ........... 26
4.2 Examinee Composition of 1989 AP Chemistry Examination Classified in terms of Racial Groups and Gender Differences ........... 27
4.3 Estimated Item Parameters for Items 11-20 in Part B of the Survey and Test Kit ........... 30
4.4 Item Parameters for Comparison Items in Part C ........... 33
4.5 Summary of Demographic Information on Survey Subjects ........... 37
4.6 Comparison of Ability Estimates -- Descriptive Statistics ........... 44
6.1 Summary of Difficulty Rank of Five Essays Across Three Ability Ranges ........... 69
6.2 Frequencies of Essay Choice Combinations Based on the Total Population of 18,462 Examinees ........... 72
6.3 Mean Scores on Five Essays ........... 74
6.4 Relationship Between Rank of Essay Choice Preference and Mean Performance ........... 78
7.1 Summary of Teacher Ratings of Essay Difficulty Levels ........... 85
7.2 Intercorrelations Among Choice Preferences of Students of Ten Ability Levels ........... 102
7.3 Summary of ANOVA Results Based on Randomly Selected 11 Students across the Sequential and Selective Groups and Ten Ability Levels ........... 113
8.1 Descriptive Statistics of the Performance of Survey Subjects ........... 123
8.2 Rank Differences ........... 133
8.3 Comparisons of IRT Item Difficulty Estimates, RANKO Scale, Choice Frequencies and Mean Scores ........... 138

xvii
8.4 Mean Group Ability Estimates of Hawaii AP, College and Non-AP Chemistry Students............ 142
8.5 Relationship among Extent of Study, Judgment of Item Difficulty, Choices & Choice Change............ 144
8.6 Relationship among Extent of Study, Judgment of Item Difficulty, Choices and Choice Change............ 145
8.7 Relationship among Extent of Student, Judgment of Item Difficulty, Choices & Choice Change............ 148
8.8 Summary of Logistic Regressional Models............ 152
8.9 Summary Statistics for Comparison IV............ 155
9.1 Descriptive Statistics of the Scores on the 3 Chosen and 3 Un-chosen Items............ 161
9.2 Numbers of AP Chemistry Students Who Scored the Same, Higher or Lower After Choices............ 162
9.3 Descriptive Statistics of Raw and IRT Equated Scores on the 20 MC Items............ 165
9.4 Kendall’s Tau and Pearson Correlations between Raw and IRT Scores on the 20 MC Items............ 165
9.5 Number of Students Whose Rankings Moved Upward, Downward and Remained the Same from Raw Scores to IRT Equated Scores............ 166
9.6 Frequency Distributions of Rank Changes............ 167
9.7 Comparison of Students’ Rankings as Reflected by Raw Scores and IRT Scores............ 168
9.8 Illustrations of the Enhanced Accuracy of IRT Scores over Raw Scores............ 169
9.9 Summary Statistics and T-Statistics of IRT20 and IRT17 vs RAW20 and RAW17............ 172
9.10 Intercorrelations among Four Sets of Scores............ 173

xviii
9.11 Examples of the Relationship between IRT Scores and Item Difficulties

9.12 Illustration of the Effects of Choices and Choice Item Difficulties on IRT Ability Estimates and Rankings

........... 176

........... 188
LIST OF FIGURES

Figure

4.1 Comparison of Test Information Curves .................. 42
4.2 Scatterplot of Ability Estimates as Measured by the Original 75 MC Items and the 20 Selected MC Items .................. 44
4.3 Score Distributions of the Four Different Sized Random Samples .................. 46
4.4 Comparison of Item Difficulties of the 20 MC Items as Estimated by Four Random Samples .................. 47
4.5 Comparison of Item Discrimination of MC Items As Estimated by Four Random Samples .................. 48
4.6 Comparison of Item Guessing of the 20 MC Items as Estimated by Four Random Samples .................. 49
6.1 IRT Expected Scores of the Five Essays .................. 68
6.2 Essay Difficulties as Reflected by Conditional Mean Scores .................. 70
6.3 Frequencies of Essay Choice Combinations Based on the Total Population of 18,435 Examinees .................. 72
6.4 Popularity Rank of the Five Essays .................. 73
6.5 Mean Scores of Five Essays Conditioned on Ability Levels .................. 75
6.6 Mean Scores on Ten Essay Combinations .................. 77
6.7 Mean Scores on Three Chosen Essays Conditioned on Ability Levels .................. 79
7.1 Group Mean Abilities for 10 Essay Combinations .................. 95
7.2 Examinee Ability Profiles behind Ten Essay Choice Combinations .................. 97
7.3 Essay Choice Patterns of Students of Ten Ability Levels .................. 100

xx
7.4 Student Ability Composition in Essay Combination 5, 8 & 9 ................ 104
7.5 Mean Score Comparison of the Ten Levels of Students Across Four Essay Combinations ................ 106
7.6 Student Ability Composition in Essay Combination 5, 6 & 7 ................ 108
7.7 Performance Comparison Between Sequential vs. Selective Choosing of Level-1 Students Only ................ 111
7.8 Numbers of Sequentially vs Selectively Choosing Students Across Ten Ability Levels ................ 112
7.9 Performance Comparison Between Sequential vs Selective Choosing Across 10 Ability Levels ................ 114
7.10 Performance Comparison Between Sequential vs Selective Choosing Across 10 Ability Levels on Equal Group Size of 11 ................ 116
7.11 Mean Performance Comparison Between Sequential and Selective Choosing on Essay Combination 5, 6 & 8 ................ 117
8.1 Raw Score Distributions of National, Hawaii Total, AP and Non-AP Samples ................ 125
8.2 Comparison of the Overall Essay Choice Pattern of 554 Hawaii and 18,462 National Students ................ 126
8.3 Dimensionality of the Five Essay Questions ................ 130
8.4 Perceived Order of Essay Difficulties ................ 132
8.5 Dimensionality of the Four MC Items in Comparison IV ................ 136
8.6 Perceived Order of MC Items Difficulties ................ 137
9.1 Ability Distributions of AP Chemistry Students Who Scored the Same, Better, and Lower on the Basis of Chosen Items ................ 163
9.2 Scatterplot of the Raw Score20 and IRT Score20 ................ 166
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.3</td>
<td>Scatterplot of the Ability Estimates of 223 Hawaii AP Chemistry Students As Measured by Full and Partial Raw Scores</td>
<td>174</td>
</tr>
<tr>
<td>9.4</td>
<td>Scatterplot of the Ability Estimates of 223 Hawaii AP Chemistry Students As Measured by Full and Partial IRT Estimates</td>
<td>174</td>
</tr>
<tr>
<td>9.5</td>
<td>Scatterplot of the Ability Estimates of 223 Hawaii AP Chemistry Students As Measured by RAW20 and IRT20</td>
<td>175</td>
</tr>
<tr>
<td>9.6</td>
<td>Scatterplot of the Ability Estimates of 223 Hawaii AP Chemistry Students As Measured by RAW20 and IRT20</td>
<td>175</td>
</tr>
<tr>
<td>9.7</td>
<td>Relationship between RAW20 and RAW17 without Equating</td>
<td>178</td>
</tr>
<tr>
<td>9.8</td>
<td>Relationship between RAW20 and RAW17 without Equating in Z-scores</td>
<td>180</td>
</tr>
<tr>
<td>9.9</td>
<td>Relationship between IRT20 and IRT17 After Equating</td>
<td>183</td>
</tr>
<tr>
<td>9.10</td>
<td># of Students Whose Enhanced, Maintained and Lowered Their Ability Estimates from IRT20 to IRT17</td>
<td>185</td>
</tr>
<tr>
<td>9.11</td>
<td>Distribution of Rank Changes After IRT Equating</td>
<td>186</td>
</tr>
<tr>
<td>9.12</td>
<td>Comparison of Raw and IRT Equated Scores in Light of Mean Item Difficulties</td>
<td>190</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

There is increasing emphasis on assessing what is called the "generative or constructive process" of learning (Wittrock, 1991). Because of their ability to solicit deeper understandings, broader analyses, and higher levels of performance on the part of students, more and more testing agencies are beginning to incorporate constructed response (CR) items, such as essays or math proofs, into their standardized tests.

When an exam consists, in whole or in part, of CR items, it is a common practice to allow the examinee to choose a subset of CR questions from a larger pool so that a limited number of items would not unfairly affect the examinee. If choices were not allowed, the examinee might confront questions whose content may be unknown. However, allowing choices presents a serious problem in equating the scores of various subsets of CR items chosen by different examinees, because it results in the de facto administration of different test forms that are determined by examinees according to their own preferences. It is the purpose of this dissertation research to probe into the psychological processes underlying examinees' choices, to examine the consequences of allowing choices without proper adjustment of scores, and finally, to develop a procedure that will equate examinee self-selected subsets of items.

To understand this problem more concretely, Wainer, Wang and Thissen (1991) described two different ways to construct a 100-word vocabulary test to assess students' verbal abilities. Suppose we have an item bank of 100,000 words of varying difficulties.
The first method is to randomly select 100 words from this vocabulary corpus and to present them to an examinee. The proportion of correct answers on this 100-word test can be inferred as the student’s potential to answer the whole word population or an estimate of his vocabulary ability, except for sampling error. In other words, what we observe is a reasonable representation of what we do not observe. The items that are not selected can be viewed as "missing-at-random." Because of random missingness, a score from one random sample of words can be equated with another score from any other random sample of words, with known statistical confidence.

However, a different way to construct this vocabulary test is to let a student randomly select one word at a time from the item bank and simultaneously, to allow him to decide whether or not to try to answer it. He may deliberately omit the ones that are too difficult for him and continue to select other words until he encounters one that approximates his vocabulary ability. Suppose after he/she has screened 200 words and actually attempted 100 words, he has received a score of 90% correct after a number of deliberate omissions. It can be seen that this score of 90% has a completely different meaning from that as derived from the first method. All we know for sure from this score is that the student knows 90 words. It is impossible to make inferences about his potential or true vocabulary ability to answer the remaining words. In this instance, we cannot infer what we do not see. This situation is commonly referred to as "non-ignorable nonresponses." Due to this non-random missingness, an observed score CANNOT be equated to any other score.
Since the ultimate goal of a test is to provide fair, valid and accurate inferences about examinees' ability or competence as derived from test scores, the current practice of allowing choices of unequal items without proper adjustments has clearly violated a fundamental assumption of fairness or equity. Although there has been some research attempting to adjust scores on self-selected subsets of test items, mostly using regression and common-anchor equating of item response theory (IRT), there are still unsolved issues regarding the appropriateness of these methods. More specifically, the predicted scores of regression method conditional on the abilities of those students who choose an item does not take into account those who do not choose that item. In other words, the predicted scores for different chosen items are not based on the same group of subjects. On the other hand, although IRT equating method treats different groups of subjects choosing different subsets of items on a same scale, the effect of the violation of its basic assumption of missing-at-random is not known yet on its equating accuracy.

The position of this author is that we have to first understand what causes different choices and the score differences so that we know what is being adjusted or equated. As a result, there are three primary objectives for this dissertation research: (1) to find out students' general choices of test items, (2) their perceptions of item difficulties and dimensionalities, (3) their previous curricular experiences and, (4) their test performances. How do students choose in general? Why do students choose one subset of items over another? Which students choose appropriately or inappropriately for their actual ability levels? It is after the revelation of such a meaningful relationship that the relative accuracy of both regression and IRT equating methods will be examined. All these analyses will
ultimately shed light on a sound statistical procedure to estimate the upper-bound of equating accuracy, and ultimately, to achieve equity and fairness.
CHAPTER II

LITERATURE REVIEW

This chapter consists of three parts: (1) the definition of constructed response items; (2) a description of their advantages and disadvantages; (3) a summary of the research on the problem of allowing choices of constructed response items and the subsequent difficulty of equating the scores from the different choices. This chapter concludes with a discussion on approaches to equating different choices of CR items.

Definition of Constructed Response Items

Traditional tests fall into two categories: constructed-response (CR) tests and selected response tests (Powell & Gillespie, 1990). Constructed response items are generally defined as those items to which students must supply a proper response. They range from sentence-completion, short-answer items and cloze items, to open-ended essay questions, to written reports of tasks for performance assessment. Selected-response tests ask students to select an answer between or among alternatives. Since CR items are the main objective of research in this dissertation, the following discussion centers around the characteristics of CR items. The CR items referred to in this dissertation are either short-answer or essay questions.
Advantages and Disadvantages of Constructed Response Items

What are the main advantages of CR items on a test? In terms of Bloom’s taxonomy of educational objectives (1956), CR items such as short-answer and essay questions, if constructed well, tend to measure the cognitive domains of comprehension, application, analysis, synthesis and evaluation (Blosser & Mayer, 1983; Sax, 1989; Slem, 1981). Unlike multiple-choice (MC) items which tend to emphasize well-defined facts, concepts or constructs in isolation, CR items allow students freedom to respond within broader limits and with fewer restraints. This often results in divergent thinking on the part of students and allows them to generate unconventional and creative responses (Duchastel & Ungester, 1982). Furthermore, because they often require students to synthesize a topic with substantiated facts, guessing is reduced (Davey, 1987). The logistic advantage of CR items may be the reduced time to assemble a test with a few CR items, because less time is required for typing and mimeographing them.

Because of the characteristics cited, CR items remain as the most frequently used item type by teachers who find them sensitive to instruction and curriculum (Matter & Kevin, 1990). In fact, some research indicates that the evaluation of learning can be demonstrated as readily by essay tests as by multiple choice tests (Hogan and Mishler, 1980). Of course, neither essay nor MC tests can be effective if haphazardly constructed or scored. Sax and Collet (1968) found that college students who expected and prepared for a multiple-choice final examination did slightly better on essay finals and much better on multiple-choice final examination than did a comparable group of students who expected to take and studied for brief essays on their final examination.
Empirical evidence suggests that CR items seem to demonstrate higher construct validity for elementary students than MC items (Arrasmith, 1984). However, based on 432 students from 11 colleges who took the 1986 Advanced Placement Examination in Biology, Bridgeman (1989) found that CR scores appeared to be less closely correlated to first-semester biology course grades than multiple-choice scores. However, when the analyses were run separately according to the gender of the candidate, performance on both the MC and the essay items predicted grades equally well for males, but for females, predictions based on the essay portion were significantly less accurate.

However, the very advantages of essay questions have resulted, to a certain extent, in some of their disadvantages. The most serious disadvantage is the difficulty of scoring CR items objectively. It is also time-consuming. As early as the turn of this century, Starch and Elliott (1912, 1913a, 1913b) reported "... the range of marks given by different teachers to the same paper may be as large as 35 to 40 points."

Low levels of rater reliability may be caused by many factors, varying from teachers' expectations of students' performance (Chase, 1979), to teachers' dislike of students' spelling, grammar or punctuation errors (Marshall, 1967). Sometimes, even after experienced raters are instructed specifically to mark solely on content, there is still evidence of raters' subjective influences affecting their rating reliability (Ashburn, 1938; Hughes, Kelling and Tuck, 1983).

The second biggest disadvantage is that extended essays measure only limited aspects of student knowledge. Because they require time to write, the student can complete only a very limited number of CR items in allotted test time. As a result, tests of CR items may not always be a fair measure of what students actually know. This problem is especially
serious when a test consists of very few "big-unit" essay questions that tap deep understandings or syntheses of certain specialized topics. The limited number of items further compounds the potentially low rater reliability, because raters have relatively few reference points to consistently determine students’ true abilities. Coffman (1972) has shown that objectivity is improved more by increasing the number of short essay items than by allowing greater freedom in responding to fewer items.

Other disadvantages of CR tests are the possibilities of students’ bluffing and rote memorization. Although essay questions eliminate guessing, they do not prevent bluffing (Kaufman, 1964; Sullivan, 1987). This is because some prepared students may attempt to get a passing grade by answering something. Furthermore, if poorly designed, essay questions may require little more than rote memorization of a series of facts (Sax, 1989).

**Choices of Constructed Response Items and Problems**

The shortcoming of measuring a narrow range of knowledge is especially salient when CR items are used in large-scale high-stakes national tests, such as the Advanced Placement examinations administered by the College Board, Educational Testing Service. Because of the lack of national curricula in the US, the content and the depth students study in any discipline may vary substantially across the nation. With a limited number of CR items spanning a relatively narrow range of content domains on a national test, it is highly likely that some students may suffer a catastrophic loss or even failure due to the fact that they have never or only superficially studied the content represented by some of the CR
items. In order to avoid such dire consequences, a common compromise adopted by major test developers is to include a fairly large number of different CR items for students to choose from. In that way, a relatively large coverage of content is guaranteed. Another occasion where choices of items are allowed is computer based testing where students can skip the items which they deem too difficult (Johnson & et al, 1991). Nevertheless, such practices have caused a recalcitrant problem in test equating.

When should choices be allowed? Allowing choices among CR items is desired when the purpose of testing is to measure writing effectiveness rather than subject matter acquisition (Sax, 1989). Students can select those questions best suited to their writing skills (style, effectiveness, grammar and punctuation) and avoid the frustration of writing on an unfamiliar topic (Wiseman and Wrigley, 1958).

However, the situation is quite different if the purpose of CR items is to measure comprehension or understanding of subject matter. Because items differ in complexity, the least knowledgeable students could select the easier questions and in effect obtain higher scores or ratings than those who are willing to tackle more complex topics. If subject matter is important, all students should be expected to respond to the same items. Otherwise, the choice of topic rather than the degree of knowledge could account for the differences in students’ performances. If such differences are not properly adjusted, the tests would become psychometrically faulty (Stager & Mueller, 1991).

It should be obvious that nothing can be as dramatic as topical differences. For example, few would argue that writing about the treatment of AIDS patients would be similar to writing about ozone depletion. The quality of writing on such issues depends largely on the expertise of the writers. There is no psychometric way to equate the scores obtained on
the essays of such dramatically different topics. However, people would agree that the tasks of calculating the area of a circle vs. that of a trapezoid are more similar than different. A somewhat more relaxed example would be to explicate the causes of skin cancer vs. intestine cancer. It can be seen that the common characteristic of the latter two examples is that the tasks being asked in each example belong to one well defined discipline and they differ only in their relative difficulties as reflected in terms of depth of knowledge required to answer them. This feature is commonly referred to as unidimensionality which requires that items of a test or a subtest belong to one well-defined disciplinary area and only differ in their relative difficulties. Psychometrically, it is this type of test items that can be equated if they are randomly distributed among examinees.

The selection of unidimensional test items does not solve the problem completely, because of the nonrandom nature of students' choices. It seems mandatory to first understand why students choose particular items and what their score differences mean before proceeding to adjust their scores. There is no doubt that the relationship between students' choices and their resulting scores are interconnected and causal. Unfortunately, there has been virtually no study on this kind of relationship, and very limited research on how examinees choose items on a test, and how such choice differences can be properly adjusted, respectively.

As for students' item choosing behavior, Johnson (1991) found that of 148 college students taking a self-adapted computerized test of basic algebra skills, higher self-confidence and lower test anxiety were related to choosing more difficult first items but were not related to later choices. Overall, examinees chose items of moderate difficulty relative to their ability levels.
Ever since the birth of the College Board Advanced Placement examinations over 20 years ago, the scores from the CR items have been treated as if the items were of equal difficulty. However, various investigations in the past have provided strong evidence against such treatment (Fremer, Jackson and McPeek, 1968; Cowell, 1971; Livingston, 1988; Pomplun, Morgan and Nellikunnel, 1992). Numerous incidents have been reported that more able examinees often seem to be drawn to more difficult questions, and as a result, receive lower scores than less able examinees. For example, upon reviewing the psychometric characteristics of selected AP tests, Fremer, Jackson and McPeek (1968) reported that a higher ability group received a mean score of 2.73 on CR items lower than the mean score of 8.19 received by a lower ability group. In addition to finding out the mean score differences for groups of various abilities as reported above across four AP exams of 1991, Pomplun, Morgan and Nellikunnel (1992) also attempted to assess the effects of such differences on the classifications of examinees. They estimated that because examinee scores were not adjusted for item difficulties, almost as many as 4,000 examinees on the United States Government and Politics Examination might have been misclassified.

Based on polytomous IRT theory, Wainer, Wang and Thissen (1991) reported three major findings. First, MC questions and essay problems seem to yield the same ability trait estimates, although the authors admit that the two problem types may involve different cognitive processes. Second, under the assumption of missing-at-randomness, choosing an easier essay gives students of upper-middle ability as much as 1 point advantage in a range of 6 points over their counterparts who chose a hard essay. However, the choice of an easy or hard essay does not seem to affect the scores of those students who have either extremely
low or high abilities. Furthermore, they showed that the IRT equated ability estimates under the assumption of missing-at-randomness can serve as "upper bounds" of equating accuracy.

There are, however, many unknowns remaining to be investigated. To name just a few more important issues. First, what are examinees’ general tendencies of item choices on a large scaled standardized test? Although there are assumptions on how students choose, e.g., students tend to choose easier items, there has been virtually no systematic empirical evidence on the relationship between students’ abilities and their actual item choices. It is possible that a substantial proportion of students would choose the items that are appropriate for their levels, not just the easy ones. At the same time, some low-ability students may not know how to choose if most items are far beyond their abilities, while some extremely high-ability students may attempt some extremely difficult items.

The second question is, "Given that examinees’ general choice tendencies have been sorted out, why do they choose the way they did?" Although there may be numerous reasons behind examinees’ selections, such as limited time or mental stress, it is reasonable to hypothesize that some major factors are playing a dominant role in determining examinees’ choices, such as their actual abilities, their curricular experiences, their perceptions of the content familiarity, and the dimensionality and difficulty of the items. It is important to find out about the relationships among these factors so that one can understand in what respects students’ choice combinations differ. Therefore, when it comes to equating the different choice combinations, what to adjust for, not only in terms of total scores, but also in terms of experience and item content can be determined.

The third question is, "Given the information on students’ choice combinations according to their ability levels and their perceptions of item characteristics, how can their
final scores be adjusted so that a penalty would not be implicitly imposed on those students who choose harder items?" Up to now, there is no commonly agreed-upon method to solve the problem. The method of "upper bound of equating accuracy" used by Wainer, Wang and Thissen (1991) is still a hypothetical way out of the statistical impasse of non-randomness. Its accuracy has not been verified unequivocally, since there are no data that would provide the information on what scores examinees would have obtained if they had answered all the questions. Only when such data become available will we be able to empirically test the effects of the violation of missing-at-randomness on equating accuracy, and to verify the feasibility of the upper bound of equating accuracy.

The final question is, "What are the consequences if examinees’ scores are left unadjusted?" This relates to the current practice that treats all choice items equally. It is hypothesized that a considerable number of students, especially those of middle and upper-middle abilities, would suffer substantially because of their attempt to answer the items that are above their ability levels.

In summary, permitting choices of CR items in a test has taken current testing practice beyond our knowledge and capability to accurately describe student’s psychological processes when involved in item selection, and to statistically adjust for the differences in the scores among various subsets of chosen CR items. It is the purpose of this research to look into such unknowns.

This dissertation attempts to address the issue of CR item choices in a comprehensive way, ranging from finding out students’ general item choice tendencies, to assessing the psychological processes underlying their choices, to realizing the effect of the factors that underlie varying choices, to estimating significant differences among scores from different
sets of chosen items, to realizing the consequences of leaving choice scores untreated. The research questions, instrumentation and methods are to be discussed in detail in the next three chapters.
CHAPTER III

RESEARCH ISSUES AND QUESTIONS

In order to understand how students choose subsets of CR items and how their score differences can be properly adjusted, it is necessary to systematically investigate three main research issues:

1. examinees' general choice tendencies as a function of both their preferences and abilities;
2. the main psychological factors and curricular experiences behind examinee's choices;
3. the effects on ranking students in terms of their abilities due to the lack of statistical adjustment of their scores according to item difficulties.

The following is a discussion of specific research questions that can be asked in the process of investigating each of the three research issues.

Investigating General Item Choice Tendencies

Specific research questions can be asked to sort out students' general choice tendencies. The first question is "Do students choose CR items randomly?" The hypothesis is no, since a considerable number of people will choose the items that they feel confident in to boost their scores on a test. This hypothesis can be tested by a chi-square verification of the differences between the expected and observed frequencies across all possible choice
combinations. Based on the observed frequencies, items as well as item choice combinations can be ranked in terms of their popularity.

The second question is "What are the main characteristics of the most 'popular' items and item combinations?" The main item characteristics, such as item difficulties, discrimination indices and content domains, can be analyzed to deduce what item features attract most students. The hypothesis is that the popularity rank of items and item choice combinations correspond positively with the item difficulties, because test-smart students would naturally choose the items that are easiest to solve in order to maximize their scores, especially when easy and hard items are weighted equally.

The third question is "What is the relationship between the 'popularity' of items or item combinations and student abilities?" This question is intended to explore the extent to which students of varying abilities can accurately identify and choose the items that are best suited for their ability levels. Students will be divided into ten levels of abilities according to the deciles of their scores on the common items all examinees take. In the case of 1989 AP Chemistry, the items that were taken by all examinees were the 75 MC items in Section I. All 18,462 students will be divided into ten groups on the basis of their scores on these 75 MC items. The reason to classify students according to deciles is to make sufficiently fine distinctions among students' abilities, while guaranteeing approximately equal numbers of students in each group.

The choice frequencies of the students of each of the ten ability groups can be sorted out for all the items and all possible choice combinations. The "popularity" ranks of items and item combinations of the ten levels of students can be compared. Friedman (1937) two-way analysis of variance by ranks and associated multiple comparisons will be employed to
test the significance levels. The hypothesis is that there is an overall significance in the preference ranks of item choices among the ten levels of students. This significance is mainly attributed by the groups of extreme abilities, specifically, by level 1 or 2 vs. level 9 or 10.

With the information on the groups that exhibit statistically significantly different item choices, it will be interesting to look into the specific features of the items that are chosen by these groups, such as item selection orders and difficulty levels of chosen items. It is reasonable to assume that the item choices of students of level 10, the highest ability level, represent the optimum item choice patterns because they have the highest ability to accurately judge the relative difficulties of all items and choose the best items, most probably the easiest ones, to advantage their scores.

It is of equal importance to look into the item choices of the students of low abilities. To what extent do they choose items that are inappropriate to their abilities? How many of them fail to select or simply pick items on a "first-come-first-tried" basis?

Investigating Effects of Curricular Experiences and Psychological Factors on Choice Behavior

The information on students' general choice tendencies can be complemented by an investigation into the curricular and psychological factors behind students' choices in order to understand the relationship between what items they choose and why they choose these items. This task can be accomplished through two more research questions.
"Do students choose essays according to their familiarity, perceived difficulties, content dimensionalities, task differences and/or all of the four factors?" The hypothesis is that students tend to choose the kinds of items which they have studied before and deem less complex in terms of item difficulty and task involvement. After choosing one item, students would very likely choose another similar to their previous choice in as many content aspects as possible.

Although the above research question is essential to establishing the triangular relationship among students' curricular experiences, psychological processing and test performance, it is virtually impossible to investigate solely with test data. This is because no test solicits whether or not the respondents have studied each of the items. As an indispensable part of this dissertation research, a special survey and test instrument (to be described in the next chapter) is designed to tap into such information by asking three levels of students in Hawaii to assess their familiarity with some choice items, to compare their relative difficulties, to decide their preferences of choices, and finally to actually perform on them. Only in this way can this research question be properly investigated. Through a multiple regression analysis, it is possible to assess to what an extent student test performance is determined by their curricular experiences and their choices of items.

With the information on students' perceptions of item difficulties and content dimensionality, an investigation into the accuracy of test takers' judgment is possible. "Do perceived item difficulties and content dimensionalities correspond with subjects' empirical confirmations as estimated through item response theory and factor analysis?" This question relates directly to the relative accuracy of students' judgment, and ultimately, the appropriateness of their choices and the fairness of their scores. Students at the ten
different ability levels will be used. It is hypothesized that the higher the ability, the higher the correlation between the empirically estimated item difficulties and the item difficulties perceived by students. This should be true of content dimensionality as well.

**Investigating the Effects of Allowing Choices of Test Items**

**Before and After Proper Statistical Adjustment**

It is obvious that the information from the previous research questions will help cognitive scientists to realize what students' differential item choices symbolize. The natural next stage of inquiry is to find out the consequences of improper choices of items without proper statistical adjustment. Such an inquiry can be adroitly accomplished through a combination of IRT methodology and the information supplied by the previously mentioned survey and test instrument. If properly implemented, IRT has the unique flexibility and power of handling short, long and parallel forms of tests by "neutralizing" them into a common scale. IRT yields extremely similar ability estimates, commonly called "Θ", for these different forms.

The foremost question in score adjustment is, "**How do students' rankings differ if their scores are equated vs. not equated in general?**" This question relates to the current practice of not equating. Since the subjects of this study's survey and test instrument are to answer all the questions, the traditional equating methods can be used. Two measures will be used to assess the degrees of rank and score changes. The first one is a Pearson correlation between students' scores before equating and those after equating. The second measure is to compute the mean of the absolute score changes before and after equating. The
joint information from these two measures will reveal the extent of differences between equating and non-equating.

With the general effects of equating known, the next question is, "To what extent do students' item choices affect their ranks?" This time, we equate on the basis of only the items that the students say that they would have chosen. The correlation between students' chosen-item and full scores will indicate the impact of choices. If there is little change in students' rankings, we then will have verified two important statistical unknowns: first, the traditional equating method is robust to the violation of missing-at-randomness; second, the procedure of "upper-bound accuracy of equating" as proposed by Wainer, Wang and Thissen (1991) is statistically sound.

Note what is meant by "little" in the above paragraph is in comparison with the changes in the rankings of students in general investigated earlier. An effective summary statistic to compare the significant differences among the two obtained correlations is Fisher's z-transformed correlations (Fisher, 1921). Specifically, the two correlations will be transformed into z-scores using Fisher's r-to-z transformation and the probability levels of the difference between the two correlations can be consequently assessed through a z-table.

The above two equating processes are then to be repeated for each of the ten-level groups to assess, "What are the differential consequences for students of various abilities?" It is of particular interest to find out how much low-, middle- and high-ability students suffer in terms of their scores as a result of their inadequate choices. It is especially interesting to find the effects of "choosing-too-high" or "choosing-too-low" on students of various abilities.
It is hypothesized that on the average, it is the students ranging from upper-low (1st quartile) to lower-high ability (3rd quartile) students who are most vulnerable to the either positive or negative influences of item choices. This is because they have certain ability to choose but not all of them are sophisticated enough to choose accurately for their abilities. In addition, their abilities may not be wide enough to handle the range of problems they choose. If they happen to choose appropriate items, they maximize their scores. Otherwise, their scores suffer.

However, the situation may be quite different with extremely low- and high-ability students. Unlike the students of the inter-quartile range of abilities, the extremely high-ability students are expected to have sufficient leeway to correctly answer items of a much wider range of difficulties. Besides, they have higher abilities to accurately judge item difficulties in the first place. In other words, item choices do not matter that much to students of extremely high abilities, unless they happen to choose extremely easy items. However, it would be interesting to see how much the ceiling effect would be if students of extremely high-ability choose an easy item combination, especially when items are to be adjusted differentially according to their relative difficulties. Nevertheless, there will be significantly less change in the ranking orders and score changes in this group of students.

The scenario of extremely low-ability students will be the opposite to that of the extremely high-ability students. First of all, they have little ability to accurately judge item difficulties. As hypothesized earlier, a majority of the students in this group may answer the items on a first-come-first-tried basis. Second, since they can answer very few of the items correctly anyway, their choices really do not matter that much. As a result, there will be little change in their ranking orders and score changes after equating.
It should be pointed out that the valuable information on the differential effects of choices on students of various abilities will shed light on the reliability of the classification of students into various categories, such as letter grades. The direct measure of such reliability is a percentage of decision consistency (Popham, 1978) which offers the percentage of students that remain in the same category before and after equating. This gives a direct index of "How diverse the effects will be on the classifications of students if scores from choosing easy or hard items are equated vs. not equated?"

Chapter Summary

In this chapter, the author has laid out nine specific research questions that constitute three research issues, and the tactics on how to address the research questions. Although delineated separately, all these questions are inter-related. The success of the tactics depends on the information necessary to answer the ten research questions raised in this chapter. The next chapter illustrates how the instrument for this dissertation is designed and to whom it is administered to solicit the essential information to answer the above research questions, and ultimately, how equity and justice can be achieved out of choices.
CHAPTER IV

DEVELOPMENT OF RESEARCH INSTRUMENTS
AND DESCRIPTION OF RESEARCH SUBJECTS

The investigations of the three major issues and specific research questions described in Chapter II will be carried out on the basis of three data sets:


2. The data of "Advanced Placement Chemistry Survey and Test Kit" collected in twelve high schools throughout the state of Hawaii.

3. The data of "Chemistry Teacher Expert Judgment Survey" responded to by 12 high school chemistry teachers in the state of Hawaii.

This chapter intends to describe each of the three instruments, the data sets, as well as their respondents in detail. In addition, it will also describe the psychometric properties of the survey instruments to be used in this dissertation research.

The 1989 AP Chemistry Examination Data Set

Developed and administered by College Board, Educational Testing Service, the Advanced Placement Examination in Chemistry (AP Chemistry Examination) is intended to award college-equivalent credits in chemistry for the high school students who take their
second year of chemistry course in high school. Normally, students who receive a score of 4 or above are allowed to waive their first year of chemistry study at the college level.

**Structure of the 1989 AP Chemistry Examination:**

The first data set to be used in this dissertation research consists of the responses of 18,462 students to the 1989 AP Chemistry Examination (ETS, 1990). There are two major reasons to use this data set in this dissertation research. First, it is one of the first large-scale national standardized examinations which allows subsets of CR items to be chosen by examinees. The unique choice structure of this examination makes it ideal to sort out the general tendencies in students' subset selections. Second, since it contains a total population of 18,462 students who took AP Chemistry in 1989, students' choice patterns and item characteristics will be methodologically and statistically convincing.

The 1989 AP Chemistry examination has two sections. Section I consists of 75 MC questions. In this section, all examinees are expected to attempt all the questions in the allotted 90 minutes. Section II has four parts. Each part consists of different number of essay problems. Part A is a single essay problem which all the examinees are expected to answer. Therefore no free choice exists. Part B has two problems where examinees choose to answer one of the two problems. Part C is broken into eight parts from which examinees can choose five parts to answer. Therefore, there are 56 possible choice combinations ($\binom{8}{5}$) for examinees to select in Part C. Finally, Part D has five essay problems of which examinees answer three. This confronts students with 10 possible ways to select their essays.
($C_2$) in Part D. Viewed as a whole, Section II offers students 1,120 ways to organize their essay combinations ($1\times 2 \times 56 \times 10$).

Due to the complexity of possible essay choice patterns, 1989 AP Chemistry Exam becomes an ideal place to research students’ general tendencies of item choices described in Chapter III. All the research questions regarding students’ general tendencies of item choices can be satisfactorily answered with the help of this data set.

Furthermore, with the information on the total population of students, this data are in an unique position to serve three statistical purposes:

a. to calibrate item parameters of item difficulty, discrimination, and guessing for the MC items and the slopes and other parameters for the essay questions.

b. to provide accurate information on the relationship among essay choices, patterns, and student abilities.

c. to provide a subset of items which will be used to investigate the relationship among students’ perceptions of item characteristics and their choices.

**Summary of the Technical Characteristics of the 1989 AP Chemistry Examination:**

Based on a 3,000 random sample size, the 75 MC items in Section I of the 1989 AP Chemistry Exam has been fit satisfactorily with a three parameter logistic IRT model through computer program BILOG (Mislevy & Bock, 1983). The section’s marginal reliability is .91; the mean difficulty .49; the mean slope .73. Full information factor analysis (Bock, Gibbons, & Muraki, 1988) is also used to assess the dimensionality of the 75 MC items. Although it requires three dimensions to obtain an acceptable fit due to the large number of
items, these three dimensions are highly intercorrelated (.93). Such a high intercorrelation indicates that these three dimensions are essentially the same. It is also found that trimming off some of the late-appearing items that were not reached by a significant proportion of examinees can further strengthen the one dimensional solution.

As for Section II of CR questions, the average reliability of Cronbach Alpha is found to be .79. The correlation between Section I of MC items and Section II of CR items is to be .86. As for the validity of the AP Chemistry test, a general finding is that AP students generally do better in advanced college "courses than do the students who have taken the regular freshman-level courses at that institution." (ETS, 1989). Table 4.1 below summarizes some other descriptive statistics regarding the 1989 AP Chemistry.

Table 4.1

<table>
<thead>
<tr>
<th>Major Characteristics of AP Chemistry Exam</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Total Number of Students: 18,462)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section</th>
<th>I. Multiple Choice</th>
<th>II. Essay Question</th>
<th>Total Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum score</td>
<td>72</td>
<td>88</td>
<td>160</td>
</tr>
<tr>
<td>Mean Score</td>
<td>28.84</td>
<td>18.07</td>
<td>55.48</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>14.73</td>
<td>10.84</td>
<td>29.67</td>
</tr>
<tr>
<td>Reliability</td>
<td>.912 (K-20)</td>
<td>.791 (Cronbach Alpha)</td>
<td>.914 (ETS formula)</td>
</tr>
</tbody>
</table>
Examinee Composition of the 1989 AP Chemistry Examination:

Out of the 18,462 examinees who took the 1989 AP Chemistry Examination, 12,234 were male, while 6,228 were female. Therefore, there were twice as many male examinees as female examinees. Seven major racial/ethnic groups were identified in this data. Table 4.2 summarizes the examinee population in terms of gender and racial/ethnic groups:

Table 4.2
Examinee Composition of 1989 AP Chemistry Examination
Classified in terms of Racial Groups and Gender Differences

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>%</th>
<th>Male</th>
<th>%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>3,964</td>
<td>23.13</td>
<td>8,323</td>
<td>48.57</td>
<td>12,287</td>
</tr>
<tr>
<td>Asian</td>
<td>1,262</td>
<td>7.36</td>
<td>2,173</td>
<td>12.68</td>
<td>3,435</td>
</tr>
<tr>
<td>Black</td>
<td>319</td>
<td>1.86</td>
<td>271</td>
<td>1.58</td>
<td>590</td>
</tr>
<tr>
<td>Latin American</td>
<td>97</td>
<td>0.57</td>
<td>196</td>
<td>1.14</td>
<td>293</td>
</tr>
<tr>
<td>Mexican</td>
<td>60</td>
<td>0.35</td>
<td>122</td>
<td>1.58</td>
<td>182</td>
</tr>
<tr>
<td>American Indian</td>
<td>20</td>
<td>0.12</td>
<td>26</td>
<td>0.15</td>
<td>46</td>
</tr>
<tr>
<td>Pe Rican</td>
<td>16</td>
<td>0.09</td>
<td>27</td>
<td>0.16</td>
<td>43</td>
</tr>
<tr>
<td>Others</td>
<td>80</td>
<td>0.47</td>
<td>180</td>
<td>1.05</td>
<td>260</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>5,818</td>
<td>33.95</td>
<td>11,318</td>
<td>66.05</td>
<td>17,136</td>
</tr>
</tbody>
</table>

Note: 1. 1326 examinees were missing because they did not identify their races.

The examinee population is predominantly white students (about 72%), while the second largest racial group is Asian (about 20%). The total of the remaining racial groups makes up about 4% of the total examinee population. In most racial groups, there are almost twice as many male examinees as female examinees. Only in the racial group of Blacks, there are more female examinees (319) than male examinees (271).
Since no information is available on students' perceived essay difficulty, content dimensionality and the relationship between students' choice and test performance, the attached instrument *Advanced Placement Chemistry Survey and Test Kit* (Kit) will be administered to collect such information (Refer to Appendix I for this instrument. The Kit is made up of four parts:

- **Part A:** General Information Survey
- **Part B:** Mini AP Chemistry Test
- **Part C:** AP Chemistry Item Comparison and Performance
- **Part D:** AP Chemistry Essay Comparison

All the items in this instrument are selected from 1989 AP Chemistry Exam in order to maintain the validity and continuity of the investigation. The following explains why and how each part is designed.

**Part A of General Information Survey:**

Part A is intended to collect some background information on the students who participated in this project. The information solicited ranges from students' gender and ethnic backgrounds, their current level of chemistry study, and their confidence in solving chemistry problems, to their potential future fields of specialization and their expected academic degrees. However, of all the questions asked in this part, Question 3--"Have you ever taken an AP Chemistry course?"-- is of particular importance. It is included to
distinguish between AP and non-AP students, and to check on the validity of AP Chemistry
items. It can help us find out, at least:

a. whether or not AP Chemistry students would do better than non-AP students.
b. whether or not the essay choices of AP Chemistry students would be
   significantly different from non-AP Chemistry students.
c. whether or not AP Chemistry students would more accurately differentiate the
difficulty levels of items.

Part B of Mini AP Chemistry Test:

Because it is difficult financially and logistically to hire and train personnel to score
essay questions, this study employs MC items to collect the information on students’
perceptions of item characteristics and preferences. The outcome of such a conceptualization
is Part B and Part C of the Kit.

Part B contains ten MC AP Chemistry items. These ten MC items are carefully
selected from Section I of MC items of the 1989 AP Chemistry Exam according to four
criteria:

(1) They, *in toto*, represent the content domain of the original Section I.
(2) Their difficulties span the original range of the examinee proficiency
distribution.
(3) They fit the IRT model well, possessing the highest possible discriminating
parameters and the lowest possible guessing parameters.
(4) They are situated in the principal dimension obtained from a factor analysis.
Previous research has shown that as long as subsets of items meet these four criteria, students’ ability estimates as measured by such subsets of items will be very similar to those as measured by the whole test, except for random errors (Hambleton, Swaminathan & Rogers, 1991; Wainer, Sireci & Thissen, 1991). The content knowledge required of these ten MC items is described in Appendix D.

The subset of the 10 selected MC items in Part B of the Kit serves as common anchor items (Holland and Rubin, 1992) for two types of equating: (1) to equate the scores of the participants in this survey with the scores of the examinees who took 1989 AP Chemistry test; and (2) to equate the scores of chosen items.

As shown below, these ten selected items exhibit ideal item parameters as estimated by the three-parameter IRT model (Birnbaum, 1968; Lord, 1974):

Table 4.3
Estimated Item Parameters for Items 11-20 in Part B of the Survey and Test Kit

<table>
<thead>
<tr>
<th>Item</th>
<th>Discrimination</th>
<th>Difficulty</th>
<th>Pseudo-chance</th>
<th>Chi-square</th>
<th>Degrees of Freedom</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>0.51</td>
<td>-2.27</td>
<td>0.17</td>
<td>6.7</td>
<td>7</td>
<td>0.47</td>
</tr>
<tr>
<td>12</td>
<td>0.86</td>
<td>0.48</td>
<td>0.16</td>
<td>4.5</td>
<td>8</td>
<td>0.81</td>
</tr>
<tr>
<td>13</td>
<td>1.15</td>
<td>-0.50</td>
<td>0.19</td>
<td>6.1</td>
<td>6</td>
<td>0.41</td>
</tr>
<tr>
<td>14</td>
<td>0.74</td>
<td>-0.14</td>
<td>0.13</td>
<td>5.3</td>
<td>7</td>
<td>0.63</td>
</tr>
<tr>
<td>15</td>
<td>0.72</td>
<td>3.34</td>
<td>0.21</td>
<td>5.1</td>
<td>9</td>
<td>0.83</td>
</tr>
<tr>
<td>16</td>
<td>0.88</td>
<td>0.62</td>
<td>0.13</td>
<td>3.6</td>
<td>7</td>
<td>0.83</td>
</tr>
<tr>
<td>17</td>
<td>0.82</td>
<td>1.71</td>
<td>0.19</td>
<td>4.4</td>
<td>9</td>
<td>0.89</td>
</tr>
<tr>
<td>18</td>
<td>1.03</td>
<td>1.12</td>
<td>0.21</td>
<td>7.2</td>
<td>8</td>
<td>0.52</td>
</tr>
<tr>
<td>19</td>
<td>0.95</td>
<td>0.97</td>
<td>0.17</td>
<td>5.9</td>
<td>8</td>
<td>0.67</td>
</tr>
<tr>
<td>20</td>
<td>0.97</td>
<td>1.80</td>
<td>0.19</td>
<td>5.9</td>
<td>9</td>
<td>0.76</td>
</tr>
</tbody>
</table>
According to Hambleton and Swaminathan (1985), although the normal range of the discrimination index can be theoretically defined as from \(-\infty\) to \(+\infty\), it is rare to find a discrimination index larger than 2. Negatively discriminating items are often discarded from ability tests. The higher the discrimination index, the better the item is in terms of its capability to pinpoint a student’s ability. It can be seen from Table 4.3 that all of the ten items display discrimination higher than 0.5.

The values of IRT item difficulties typically vary from -2.00 being extremely easy, through 0 being neither too easy nor too hard, and to +2.00 being extremely hard. In IRT, item difficulties are defined on the same scale as students’ abilities. As shown in Table 4.3, the ten items span a complete range starting from -2.2 for extremely low ability students, through 0 for middle-ability students, and all the way up to 1.79 for high-ability students. With an item difficulty as high as 3.3, Item 15 is a special item to target at participants of extremely high ability, such as college chemistry graduate students or AP Chemistry teachers.

Typically, the guessing parameter assumes values that are lower than the value that would result if examinees of low ability were to guess randomly on an item. Since a MC item of five alternatives would exhibit 20% chance for a low-ability examinee to guess, any item with its guessing parameter within the vicinity of .20 is considered sound, provided that it satisfies the other two conditions. Most of the ten items have less than .20 guessing parameter values.

It can be seen from their probability levels that all the ten MC items are fit satisfactorily with a three-parameter IRT model. None of the ten items has a significant \(X^2\) values.
Part C of AP Chemistry Item Comparison and Performance:

Part C is especially designed to collect information on the relationship among students’ perception of item difficulties, dimensionality and choice preferences. It contains four comparisons of MC items. Each comparison of MC items is followed by a series of questions and tasks. Comparisons I, II & III have two items each, while Comparison IV has four items. The item comparisons are specially designed, because they require participants to fulfil four tasks: recall, compare, choose and answer all the items. The main functions of these groups are to assess the effects of students’:

a. past curricular experiences on answering the items through the question "Do you remember ever studying a problem like …?"

b. perceptions of the content similarities through the question "How similar are these two questions in content?"

c. judgments of the difficulty levels of the items through the question "Which of the above two questions seems easier for you?"

d. choices of items on choosing and answering the items after their evaluation of content similarity and difficulty levels of the items through the questions "If you were allowed to answer only one question, which would you choose?"

Note that all participants still have to supply their answers to all items in the groups, although they are asked hypothetically about choosing one item over the other. This makes it possible to ACTUALLY see what students’ hypothetical scores would be if they did avoid the items. Through this kind of design, it is hoped that the data analyses will shed significant light on the difficult statistical problem of equating unobserved items on the basis of non-
ignorable non-responses. Furthermore, this design makes it possible to assess the influences of students’ curricular experiences, item content dimensionality, and difficulty on students’ choices.

Another important feature about these four comparisons of MC items is the specially arranged item difficulty levels for the embedded items listed in Table 3:

Table 4.4

Item Parameters for Comparison Items in Part C

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Assignment</th>
<th>Difficulty</th>
<th>Discrimination</th>
<th>Pseudo-Chance</th>
<th>Chi-Square</th>
<th>Degrees of Freedom</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Very Easy</td>
<td>1</td>
<td>-2.63</td>
<td>0.31</td>
<td>0.17</td>
<td>6.7</td>
<td>7</td>
</tr>
<tr>
<td>I</td>
<td>Hard</td>
<td>2</td>
<td>1.59</td>
<td>0.84</td>
<td>0.13</td>
<td>7.9</td>
<td>9</td>
</tr>
<tr>
<td>II</td>
<td>Middle</td>
<td>1</td>
<td>-0.07</td>
<td>0.67</td>
<td>0.15</td>
<td>3.1</td>
<td>8</td>
</tr>
<tr>
<td>II</td>
<td>Middle</td>
<td>2</td>
<td>0.09</td>
<td>0.81</td>
<td>0.16</td>
<td>6.4</td>
<td>7</td>
</tr>
<tr>
<td>III</td>
<td>Very Hard</td>
<td>1</td>
<td>2.28</td>
<td>1.10</td>
<td>0.13</td>
<td>3.1</td>
<td>9</td>
</tr>
<tr>
<td>III</td>
<td>Very Hard</td>
<td>2</td>
<td>2.09</td>
<td>1.01</td>
<td>0.14</td>
<td>7.4</td>
<td>9</td>
</tr>
<tr>
<td>IV</td>
<td>Easy</td>
<td>1</td>
<td>-1.49</td>
<td>1.03</td>
<td>0.16</td>
<td>7.6</td>
<td>5</td>
</tr>
<tr>
<td>IV</td>
<td>Very Hard</td>
<td>2</td>
<td>2.56</td>
<td>0.81</td>
<td>0.16</td>
<td>4.4</td>
<td>9</td>
</tr>
<tr>
<td>IV</td>
<td>Middle</td>
<td>3</td>
<td>0.36</td>
<td>0.98</td>
<td>0.14</td>
<td>5.9</td>
<td>7</td>
</tr>
<tr>
<td>IV</td>
<td>Hard</td>
<td>4</td>
<td>1.09</td>
<td>0.93</td>
<td>0.09</td>
<td>5.8</td>
<td>8</td>
</tr>
</tbody>
</table>

Such alternations of test items of various difficulties are intended to assess the effects of the items of various difficulty levels on students’ choice patterns, choice appropriateness, losses or gains in actual scores and equating accuracy.

Furthermore, the reason for the four items in Group IV is to test the effects of choosing two or more items on test scores and choice patterns, specifically, the
differences between choosing three items and one item. Two of the major specific things to find out are:

a. whether or not the difficulties of selected items are close to one another;
b. whether or not the chosen items are of similar content dimensionality.

**Part D of AP Chemistry Essay Item Comparison:**

The essay comparisons contain the five original essay problems in Part D of 1989 AP Chemistry Exam which have been explicated earlier in this chapter. There is only one difference in the order that the five essay problems are presented: Essay 6, one of the most popular essays, is moved to the last position. Therefore, original Essays 8 and 9 are moved up one position. The reason for such a change is to test whether or not the low frequency of choosing the essay on nuclear chemistry is due to its terminal position.

Essay comparisons serve three functions:

1. to determine to what an extent participants of various chemistry abilities would replicate the choice patterns of the 18,462 examinees who actually took the Exam in 1989;
2. to determine the content dimensionality of the essay problems from the data based on the similarity comparison questions;
3. to discover the correlation between the perceived difficulties of the essays by participants in my survey and the actual essay choice frequency ranks of 18,462 students.
Note that participants are not asked to actually perform on these problems but only compared them. The maximum amount of allowable time with the participants was only 30 to 45 minutes. As a matter of fact, 45 minutes was hardly enough for low-ability students to finish Part C.

**Teacher Evaluation**

The teachers of all the classes were also surveyed with the "Teacher Evaluation" (See Appendix B). This instrument is designed to assess the extent to which the teachers have taught the various chemistry subjects as represented by the twenty MC and five essay questions in the Kit. All the evaluation questions in this instrument are parallel to those in the Kit except for the perspective. Instead of asking a student "To what extent have you studied a problem like...?", the teachers are asked "To what extent have you taught a problem like ...?" The information obtained in this instrument is to be correlated with their students' performance and choices so that instructional effects can be sorted out. This correlation also provides evidence for the sensitivity of the AP Chemistry Exam to instruction.
Subjects

With the enthusiastic support from over a dozen of administrators, chemistry professors and teachers, the Kit was administered to 678 students at four academic levels of chemistry studies: (1) one upper-division class of 15 students in the Department of Chemistry, University of Hawaii, who have completed, at least, four semesters of college chemistry studies; two chemistry classes of 81 students in the Chemistry Department, who have completed their first year of college chemistry; (3) thirteen high school AP Chemistry classes of 237 students, a majority of the registered AP Chemistry population throughout the state of Hawaii in 1992; (4) eleven high school general chemistry classes of 334 students.

The Kit was responded to on a voluntary basis and a certain number of students did not completely finish the survey Kit due to the lack of sufficient time, or from academic pressures such as examinations from other courses. Moreover, a small number of students supplied uniform answers in their answer sheets, such as bubbling all "B" options, and so on. To preclude undue influences of incomplete or random data, whoever did not finish through Part C of the Kit, or supplied uniform answers is left out of the analyses. Part C is the last part of comparison and evaluation on the specially designed multiple choice questions. As a result, sixty subjects have been deleted from the total sample pool, and all the analyses on Hawaii subjects to be reported in the remainder of the dissertation will be based on 618 cases. Table 4.5 summarizes the main demographic information of the 618 Hawaii subjects.
Table 4.5

Summary of Demographic Information on Survey Subjects

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>262</td>
<td>42.4</td>
<td>262</td>
<td>42.4</td>
</tr>
<tr>
<td>Female</td>
<td>314</td>
<td>50.8</td>
<td>576</td>
<td>93.2</td>
</tr>
<tr>
<td>Unidentified</td>
<td>42</td>
<td>6.8</td>
<td>618</td>
<td>100.0</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>211</td>
<td>34.1</td>
<td>211</td>
<td>34.1</td>
</tr>
<tr>
<td>Chinese</td>
<td>139</td>
<td>22.5</td>
<td>350</td>
<td>56.6</td>
</tr>
<tr>
<td>Caucasian</td>
<td>75</td>
<td>12.1</td>
<td>425</td>
<td>68.7</td>
</tr>
<tr>
<td>Filipino</td>
<td>74</td>
<td>12.0</td>
<td>499</td>
<td>80.7</td>
</tr>
<tr>
<td>Hawaiian/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part-Hawaiian</td>
<td>38</td>
<td>6.2</td>
<td>537</td>
<td>86.9</td>
</tr>
<tr>
<td>Portuguese/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>6</td>
<td>1.0</td>
<td>543</td>
<td>87.9</td>
</tr>
<tr>
<td>Black</td>
<td>3</td>
<td>0.5</td>
<td>546</td>
<td>88.4</td>
</tr>
<tr>
<td>Samoan</td>
<td>1</td>
<td>0.2</td>
<td>547</td>
<td>88.6</td>
</tr>
<tr>
<td>Other Asian</td>
<td>47</td>
<td>7.6</td>
<td>594</td>
<td>96.2</td>
</tr>
<tr>
<td>Others</td>
<td>21</td>
<td>3.4</td>
<td>615</td>
<td>99.6</td>
</tr>
<tr>
<td>Identified</td>
<td>3</td>
<td>0.4</td>
<td>618</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Total Hawaii Student # 618

Note: AP stands for Advanced Placement.

The main reason to survey the four academic levels mentioned above is to include students of a wide range of chemistry abilities so that the complete range of chemistry abilities reflected in the original 1989 AP Chemistry Exam can be approximated as closely as possible. There are some specific purposes to include the four levels of students. First of all, the reason to include the upper-division college class is to solicit expert judgment. It
is assumed that students at such a high level of chemistry studies should be proficient at solving the 20 MC items in the Kit. Their perceptions of item difficulties and dimensionality ought to be sufficiently accurate and reliable so that the perceptions of the rest of the participants can be compared with theirs.

The main reason to include the two first-year college chemistry classes is to see how similar their performance on the MC items, and their perceptions and judgment are to those of AP Chemistry students. Presumably, these first-year college students are equivalent to high AP Chemistry students in terms of their chemistry coverage, because they are supposed to have acquired the chemistry knowledge covered by the AP curriculum. However, it is admitted that there are some important differences between high school AP students and first-year university students. It is possible that high school AP students are generally more motivated for and more successful in their chemistry study than average first-year college students. This is because high school chemistry students take AP chemistry to enter competitive universities, whereas average freshmen take AP chemistry to meet some sort of academic requirement.

It is extremely fortunate for this dissertation project to have received full support and cooperation from all the AP chemistry teachers and students in both private and public schools throughout the state of Hawaii. Although the initial contacts with the teachers were established in late 1991, the Kit was administered to most AP students during their final three weeks of review before their May 18 1992 AP Chemistry Exam. This time frame was chosen, because by this time, the students had completed their intended AP Chemistry curricular, and were at a peak of readiness and motivation to take both the Kit and the actual AP Chemistry Examination. As part of the introduction to the Kit, almost all the AP
chemistry teachers emphasized the fact that the 20 MC items in Kit are real AP chemistry items, correlate highly with the full version of 75 MC items, and that it would be a good opportunity for the students to assess their true abilities.

Why were so many non-AP Chemistry students included in this study? The reason is that on Section I of the 75 MC items of the 1989 AP Chemistry Exam, 423 examinees scored between 0 and 6 points (ETS, 1990). It is possible that a number of students, highly likely non-AP students, took the Exam. As a result, it seems necessary to include some non-AP chemistry students. However, there is something more important to this inclusion. Since this group of students just completed their first year of chemistry, they would be faced with a certain number of questions completely unknown to them. It is interesting to find out how they would choose during such circumstances. Will they just pick items in their sequential order?

Chemistry Teacher Expert Judgment Survey -- A Follow-up Study

After a preliminary analysis of students' choices that revealed differential choice combinations of the essay items (to be discussed in detail in Chapter VI), a follow-up survey was conducted to ask the nine chemistry teachers their expert judgment on why some essay combinations were more favored than others. Called "Chemistry Teacher Expert Judgment Survey" (See Appendix C), this instrument inquired into three areas of teachers' judgment. First, they were asked to rank the five essays in their relative difficulties and to offer their explanations on the ranking. The average rank of these five essays will be correlated with students' mean scores on and the preference order of the five essays.
The second area of information requested was the likelihood of AP Chemistry textbooks, *in general*, to address the chemistry subjects reflected by the five essays. When asked "How likely do AP chemistry textbooks address such a subject as reflected by Essay ...?", the teachers rated the likelihood on a five point scale, from 1 meaning "hardly addressed" to 5 meaning "extensively addressed".

The third area of information was the teachers’ qualitative judgment on what characteristics were shared by the essays in some of the most and least favored essay combinations shared. Teachers’ insights into why so many or so few students chose the most or least favored combinations were sought.

The final area of information was the teachers’ holistic reflections on the overall quality of essay questions and the observed choice patterns of students. The teachers were asked to offer unrestricted comments on the quality of the essay problems, content coverage, and allowing choices, etc.

**Verifying the Invariance properties of Ability Estimates, Item Parameters and Test Information of the 20 Selected MC Items**

So far, detailed descriptions have been given regarding the structure, purpose and content of the three data sets as well as the composition of the three groups of research subjects yielding the three data sets. Special discussion has also been given to how the 20 MC items in the KIT were selected and arranged into the Kit. These 20 MC items have a special mission -- to measure the abilities of the respondents of this survey Kit so that their scores can be equated to those of the 1989 AP Chemistry examinees and the
psychological factors of the two groups of research subjects that underlie the item choices can be bridged together.

However, in order to perform this equating accurately, it is necessary to first verify three essential properties of the 20 MC items (Hambleton & Swaminathan, 1985). The first property is "Does the 20-item curtailed test resemble the original 75-item exam in terms of general test information pattern?" This is an important question, because substantial differences in test information may imply fundamental changes in the test function. The second property is "Will the 20 selected MC items in the Kit yield the same or, at least, similar ability estimates as the 75 MC items in the original AP Chemistry Exam?" This is a question of ability invariance of IRT. If this invariance does not hold, the foundation of the equating no long exists. The third question is "Will the 20 items exhibit the same, or at least, similar item parameters across different samples of examinees?" This is a question of item parameter invariance. If this invariance does not exist, there will not be consistent measurement of students abilities. It is important that all the three questions be confirmed before this instrument can be employed with confidence, because almost every application of item response theory capitalizes on test information and properties of ability and item parameter invariance.

Comparing Test Information Similarity:

The most efficient way to check whether or not the 20 selected MC items maintain similar measurement attributes is through what is called "test information". As one of the powerful features of IRT, test information function serves to demonstrate the
measurement efficacy of a test. Mathematically, test information is the sum of the item information that is independently contributed by the individual items in a test (Hambleton & Swaminathan, 1985).

Figure 4.1: Comparison of Test Information Curves
Original 75 (+) vs. Selected 20 (*) MC Questions
Figure 4.1 shows that the curtailed version of 20 MC items displays a parallel, although shorter, test information pattern as the original 75 items. The parallelism indicates that the curtailed version has retained the main characteristics of the original form. The shorter height of the information curve of the 20 MC items is expected because of the much smaller number of items. Note that the original version of 75 items is more than three times longer than the curtailed version of 20 items. As a result, the test information curve of the original test can be more than three times higher than that of the curtailed version.

Checking the Invariance of Ability Estimates:

In order to check whether or not the 20 MC items yield same or similar ability estimates, one thousand students were randomly selected from the population of 18,462 examinees who took the 1989 AP Chemistry Examination. Using BILOG (Mislevy & Bock, 1990), two analyses were carried out to estimate the ability estimates of the selected 1,000 students. The first analysis was based on the complete 75 items of the 1989 AP Chemistry Exam, while the second, on the 20 items of the Kit. As indicated by Table 4.6, the ability estimates of the 1,000 students from these two versions are extremely alike.

Specifically, for the two analyses, the means differ only by 0.030; the standard deviations, by 0.085; the minimum score, by 0.312 and the maximum score, by .056. The difference of 0.156 between the two reliabilities is expected since there are 3.75 (75/20) times more items in the original AP Chemistry Test than in the Kit.
Table 4.6
Comparison of Ability Estimates
Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
<th>KR-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>I20</td>
<td>1000</td>
<td>0.02214</td>
<td>0.87655</td>
<td>-2.29310</td>
<td>2.80000</td>
<td>.768</td>
</tr>
<tr>
<td>I75</td>
<td>1000</td>
<td>0.05233</td>
<td>0.96187</td>
<td>-2.60580</td>
<td>2.74430</td>
<td>.924</td>
</tr>
</tbody>
</table>

The correlation between the ability estimates as measured by the two versions is 0.93. Figure 4.2 displays a narrow scatterplot of the ability estimates of the 1,000 students as measured by the 20-item Kit and 75-item AP Chemistry Exam, respectively. It can be concluded beyond reasonable doubt that the 20 items in the Kit have maintained its ability invariance. It can also be concluded that the abilities of the potential participants in the Kit will be estimated comparably as well.

---+---+---+---+---+---+---+---+---+---+---+---+---+---+._-+---
| 3.5 | + |
| 3.0 | + A B AA + |
| 2.5 | + AAABACB AA + |
| 2.0 | + ADBDDFFJFHCCBB A + |
| 1.5 | + AAEDEKQRLOADCAC + |
| 1.0 | + AAACDBKJXUUMJECDBA + |
| 0.5 | + AEEFROUVZQ2V01BD + |
| 0.0 | + AFGKTOV2XQ0HFCGA A + |
| -0.5 | + BDBFFPJJHNNJEMGDA B A + |
| -1.0 | A CBDBBFRIFFEIDXAC + |
| -1.5 | A AAAA AA A + |
| -2.0 | + |
| -2.5 | + |
| -3.0 | + |
| -3.5 | + |
| -4.0 | + |
---+---+---+---+---+---+---+---+---+---+---+---+---+---+._-+---

* Note: 1. Plot of I20*I75. Legend: A = 1 obs, B = 2 obs, etc.
2. 17 obs hidden.

Figure 4.2: Scatterplot of Ability Estimates
As Measured by the Original 75 MC Items and the 20 Selected MC Items
Checking the Invariance of Item Parameter Estimates:

In order to test whether or not the three item parameters of the 20 selected MC items hold relatively constant regardless of the sample, four random samples are selected from the 18,462 students of the 1989 AP Chemistry. The four sample sizes are 1,000, 300, 100 and 50 students, respectively. Figure 4.3 indicates that the score distributions of the four groups of students cover the complete range of possible scores, namely, from 0 to 20 points and as the sample size increases, they gradually approach normal distributions.
The responses of the four groups of students are used to estimate the three parameters of threshold (difficulty), slope (discrimination) and guessing for each of the 20 items. Figure 4.4 below indicates the threshold (difficulty) estimates for the 20 items are virtually identical across the four random samples of students.
Figure 4.4: Comparison of Item Difficulties of the 20 MC Items
As Estimated by Four Random Samples

As for the parameter of item discrimination, Figure 4.5 below shows that this invariance decreases slightly as the sample sizes become smaller. Well documented (Swaminathan, 1983; Swaminathan & Gifford, 1983), this phenomenon is due to the fact that the calibration of item discrimination takes a large number of examinees to be stable. However, it can be observed that the amount of fluctuation is minimal as long as the sample size is bigger than 50.
Figure 4.5: Comparison of Item Discrimination of 20 MC Items As Estimated by Four Random Samples

Similar to the case of item discrimination estimation, there is a slight amount of fluctuation in the guessing parameters of the 20 MC items across the four random samples, due to the same reason that it takes a substantial number of examinees to reach a stable calibration. However, the amount of variation is still minimal, given the magnitudes of the absolute differences among the estimates.
Chapter Summary

This chapter has reviewed the structure and subject composition of the two data sets to be analyzed and compared in this dissertation research --- the 1989 AP Chemistry Examination and the AP Chemistry Survey and Test Kit. Discussion has been made on the rationale and development of the three research instruments -- AP Chemistry Survey and Test Kit, Teacher Evaluation and Chemistry Teacher Expert Judgment Survey. Three major aspects have been emphasized. First, both the MC and Essay sections of the 1989 AP
Chemistry Examination have been shown to possess sound psychometrical properties, such as high levels of internal consistency and inter-correlation. This provides a sound basis to extract items for the development of the survey instrument for this dissertation.

Second, the twenty MC items that are built into the Kit demonstrate high degrees of item parameter invariance with negligible fluctuations across four random samples of different sizes. What is more important is that these twenty MC items have been shown to yield virtually identical ability estimates for the examinees across different random samples of subjects. As a result, it is reasonable to assume that as long as the participants of the dissertation research span a wide range of abilities as the national population did, it is highly probable that these 20 items will function consistently in the same way as they did with the original 1989 AP Chemistry Exam examinees.

Finally, it has been shown that in addition to being carefully selected, the twenty MC items were specially arranged in the Kit with a special mission -- to evaluate the respondents' item choice patterns and the psychological processes underlying them.
CHAPTER V

REVIEW OF RELATED RESEARCH METHODOLOGY

Although many of the research questions of the dissertation are investigated through commonly used statistical methods, such as frequency distributions, conditional means, correlations, analysis of variance and regression, there are three other major analyses --- item response theory (IRT), test score equating, and multidimensional scaling, that play an essential role in the investigation of the psychological processes underlying examinees' choices of items and the accuracy of equating different chosen items.

IRT is adopted to (1) estimate item parameters of both the MC items and essay questions; (2) to equate scores; and (3) to reflect the effects of the lack of proper adjustment for item difficulties on scores. The unidimensional and multidimensional scaling is employed to analyze the comparison data on item similarities as well as to discern students' psychological processes underlying their choices, perceived item difficulties and content dimensionalities.

Although it takes volumes to explicate fully the above mentioned methods, this chapter will summarize each of the three methodologies with the emphasis on their applications in this dissertation research. More discussion is to be devoted to item response theory because it is less familiar to the general audience outside the field of educational and psychological measurement.

51
Item Response Theory

IRT can be defined as a theoretical and mathematical framework that seeks to explain and predict the relationship between examinees' abilities in a certain discipline and their test performance (Lord & Novick, 1968). Since abilities are not directly measurable, IRT, as one of various latent theories, assumes that an examinee's ability consists of a set of unobservable or latent traits that impact on his/her test performance. This set of traits forms a k-dimensional space of ability in which an examinee's position is represented by a vector of relationships among these traits, items and his/her test performance (Thissen & Steinberg, 1984). Among them, there is a dominant factor that will play a major role in influencing an examinee's test performance. For example, an examinee's test performance can be determined by various factors, such as his/her academic achievement, extent of test preparation and the test site conditions, such as lighting and physical comfort. However, it can be reasonably assumed that the major factor of his/her performance depends on his/her academic achievement which, all by itself, sufficiently explains his/her test performance.

The relationship between the observable test performance and the unobservable latent traits can be described by various mathematical functions which are based upon different assumptions about the test data, such as the extent of examinee guessing and the heterogeneity of item difficulties. When certain assumptions are met and a mathematical form of this relationship is specified, an IRT model results. The appropriateness of this mathematical model with any set of test data can be established by conducting a suitable goodness of fit investigation. With a reasonable fit between the chosen model and the data
set, IRT can provide both consistent item statistics and ability estimates to account for test performance.

At the core of IRT are a series of mathematical models that aim to assess such relationships between examinees' abilities and their test performance. The two models to be employed in this dissertation research are the three-parameter IRT model for dichotomously scored MC items and the graded response model for essay scores.

Although different in forms, all IRT models share one common mathematical starting point -- to assess the probabilities of correct responses to test items as functions of examinees' abilities. Through the estimation process, items and persons are placed on the ability scale in such a way that there is as close a relationship as possible between the expected examinee probability parameters and the actual probabilities of performance for examinees positioned at each ability level. Item parameter estimates and examinee ability estimates are approximated continually until the maximum agreement possible is obtained between predictions based on the ability and the item parameter estimates and the actual test data. With the ultimate estimates of the model parameters, IRT models can predict the probability that a respondent at a particular score level will answer a given item correctly. Normally, examinees with higher scores on the traits have higher expected probabilities of answering the item correctly than examinees with lower scores on the traits.

**Item Response Theory vs. Classical Measurement Theory**

Although its roots can be traced back to the beginning of this century in the work of Thurstone (1925), Thorndike (1927) and others, it was not until a little more than a decade
ago when IRT came into wide application in the fields of educational and psychological measurement. It is gaining acceptance because it provides more adaptable and effective methods of test construction, analysis, and scoring than those derived from classical measurement theory (CMT). In sharp contrast to CMT as elaborated in detail by Gulliksen (1950), IRT offers three advantages (Thissen, 1990; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991):

1. Invariance of item parameter and examinee ability estimations;
2. High levels of precision and comparability of students' abilities;
3. High levels of accuracy of item parameter estimation and flexibility of test construction.

Each of the above features is explicated below.

**Invariance of Item Parameter and Examinee Ability Estimation:**

As the cornerstone of IRT, the invariance property of IRT reflects a two-fold fact, when a chosen IRT model fits the data. First, unlike the classical measurement theory (CMT) whose estimates of examinees' abilities depend on how difficult or easy a test happens to be, the IRT estimates that characterize examinee abilities do not depend on any particular set of items being used to estimate the abilities, as long as the items all measure the same trait. In other words, examinee ability estimates are independent of the particular sample of test items. The second component of the invariance property is that unlike the CMT whose values of item difficulty and item discrimination vary with the strength of a particular sample of examinees, the IRT parameters that identify the functions of items do not hinge on the ability distribution of the examinees, when a chosen IRT model fits the data. In other words, IRT item difficulty and discrimination indices do not change significantly for examinees who
are either high or low in their abilities. These two components of IRT invariance property are commonly referred to as "item-free" and "examinee-free" properties of IRT, respectively.

The reason for such an invariance is that the probability of success for a given ability level, \( P(\Theta) \), is the average of all items responses in the subpopulation of examinees with a specified ability level \( \Theta \). A fit IRT model specifies an exact relationship between this probability of success and the examinees at a particular ability level. In other words, the observed probability of success of ALL the examinees at any \( \Theta \) level is exactly equal to \( P(\Theta) \). Such relationships can be valid for actual tests of any length. However, it should be pointed out that the invariance is a property of population which is defined here as all examinees at one specific ability level of \( \Theta \). Invariance property does NOT hold with samples from a particular \( \Theta \) level. For example, if a sample is obtained from the subpopulation of examinees with the specified ability value \( \Theta \), it is extremely unlikely that the average observed probability of correct responses will be exactly equal to that of the whole group. It is clear that the invariance property of IRT offers unprecedented flexibility that has revolutionized the enterprise of test construction. For example, because item statistics are no longer examinee-dependent, a test developer can choose the items that meet some relatively absolute standards and the items will perform as desired no matter who take them. Furthermore, because examinee ability is no longer item-dependent, it is now possible to develop short, but highly efficient tests that are tailored to measure the specific levels of examinees. The latter is called "adaptive testing", a significant innovation in test construction.

When applying IRT, it is essential to demonstrate whether invariance holds for the data, since every IRT application, such as equating, item banking, item bias and adaptive
testing, capitalizes on this property. This dissertation research is no exception. It has been demonstrated in the previous chapter on instrument design that the item parameters for the 20 MC items selected for the "AP Chemistry Survey and Exam Kit" hold relative stable across various random samples. In addition, it has also been shown that the correlation of the ability estimates of 1,000 random examinees between the original 75 and selected 20 MC items is .94, indicating high levels of consistency.

Advantages of Item Response Theory over Classical Measurement Theory:

Because of its use of well defined mathematical models to fit examinee responses to items, IRT offers high levels of precision and efficiency in ability assessment. First of all, while CMT assumes uniform variance of error of measurement for all examinees, IRT examines the standard error of measurement at every ability level in light of the maximum information an item can offer. This naturally results in increased accuracy of ability estimates. It is common knowledge that some people, especially those of high ability, perform a task more consistently than others, especially those of low abilities. The CMT assumption of uniform error of measurement is clearly inaccurate. As a matter of fact, IRT provides information about the error of measurement for every score estimation for every examinee independently.

Second, because of its robust estimation of errors of measurement, CMT is most reliable for measuring middle-ability students but very relaxed about the ability estimates for either high- or low-ability examinees. IRT does not have the disadvantage. Because of its stable and accurate parameters that describe the properties of items, such as item difficulty,
tests can be tailored to desired proficiency levels exactly on target with a relatively small number of items. Increased measurement accuracy can be obtained when the test difficulty matches the approximate ability level of each examinee (Lord, 1980a; Weiss, 1983).

Third, it is well-known that the concept of reliability of CMT largely lies in the concept of parallelism of either items or forms of a test. Yet, it is common that a moderate level of parallelism is difficult to achieve, even if one is to administer the same test twice over time. Yet, IRT is completely free of such a dilemma. Due to invariance properties, IRT offers enhanced levels of flexibility of test construction and comparability of students abilities. With test items that have been fitted to an IRT model with known parameters, it is possible to estimate an examinee’s ability on the same ability scale from any subset of items in the domain of items that have been fitted to the model. Regardless of the number of items administered as long as the number is not too small, the ability estimate for each examinee will be an asymptotically unbiased estimate of true ability. Any variation in ability estimates obtained from different sets of test items is due to measurement error only.

Fourth, because of its lack of theoretical and mathematical model, CMT provides no basis for predicting how an examinee might perform when confronted with a test item. IRT, on the other hand, can project the probability that an examinee with known ability will answer a particular question correctly. This is because when a chosen model fits the data set, the same item characteristic curves (ICC) are obtained regardless of the distribution of ability in the sample of examinees. The probability that an examinee with a known ability estimate can answer the item correctly can be determined from this ICC.

Finally, IRT offers high levels of comparability. Respondents may be comparably scored on more than one forms of a single test with great ease, as long as every two forms
of tests share some items in common. Because IRT employs one universal standard to carry out item and test analyses, short, long, easy, hard, parallel or other alternate forms of tests are all treated in the same way.

Three-Parameter and Nominal IRT Models:

Two IRT models will be used in the dissertation: the three-parameter dichotomous model for MC items and Bock's 1972 nominal model with monotonicity constraints on some of its parameters for graded responses. The following is a brief sketch of each of them.

The mathematical expression for the three-parameter logistic model is:

\[ P_i(\theta) = c_i + (1 - c_i) \left[ \frac{\exp^{D \alpha_i (\theta - b_i)}}{1 + \exp^{D \alpha_i (\theta - b_i)}} \right] \]

where

- \( P_i(\theta) \) = the probability that a randomly selected examinee with ability \( \theta \) answers item \( i \) correctly.
- \( \alpha_i \) = discrimination parameter;
- \( b_i \) = item difficulty parameter;
- \( c_i \) = pseudo-chance level or guessing parameter;
- \( D \) = scaling factor introduced to make the logistic function as close as possible to normal ogive function. It is normally equal to 1.7.

The parameters of the MC items and the true chemistry ability estimates of the participants have all been estimated through this model. BILOG 3 (Mislevy & Bock, 1990) will be used to estimate the item parameters of the 75 MC items in this dissertation, including the 20 MC items in the survey.
The mathematical expression of the nominal IRT model (Bock, 1970) is:

\[
T_{j|x}(\theta) = \frac{\exp[\alpha_{jx}\theta+c_{jx}]}{\sum_{m} \exp[\alpha_{jm}\theta+c_{jm}]}
\]

where the \{\alpha_k, c_k\}, k=0,1,..., m are the item category parameters that characterize the shape of the individual response trace lines that characterize the probabilities that examinees can obtain each of the graded responses. The model is used to estimate the parameters of essay problems, examinees' abilities and expected scores on the essay problems. MULTILOG 6 (Thissen, 1990) computer package will be used to estimate the parameters of the five essay items in this dissertation.

**Test Score Equating**

As explicated earlier, one of the major objectives of this dissertation is to investigate a sound method to adjust examinees' scores on the items that they have chosen, but have differential difficulty levels. The following is a brief review of the definitions and assumptions of a few commonly used equating methodologies. More attention will be devoted to the equating methodology to be used in this dissertation -- IRT test equating or linking.
Definition and Assumption of Test Score Equating:

What is test score equating? Generally speaking, test score equating is a statistical adjustment process to convert the scores on Test X to the metric of Test Y to compensate for the differences in their relative item difficulties (Hambleton, Swaminathan & Rogers, 1991; Braun & Holland, 1982). Its basic assumption is that the test items on the two forms to be equated measure the same or a similar psychological trait. They are homogeneous in form and content (Levin, 1955) and parallel in structure including timing and item types. They differ, however, only in their relative difficulties. It is only the form-to-form variation in test difficulty that makes test equating necessary (Braun & Holland, 1982).

In simple terms, the points earned on a harder test are "worth more" than those earned on an easier form of the test (Angoff, 1971). Four specific equating conditions have been outlined by Angoff (1971) and Lord (1977, 1980):

1. **Equity** -- Only tests measuring same or similar traits with similar reliabilities can be equated. Fallible scores on tests X and Y cannot be equated.

2. **Invariance across groups** -- Independent of students' abilities, equating should yield similar results no matter which sample group is used.

3. **Symmetry** -- Unlike regressive analysis whose parameter estimates vary depending on which variable is used as the dependent variable, equating should not depend on which test is used as the reference test.

4. **Unidimensionality** -- The current equating methodologies can be used only with unidimensional tests.
In summary, the general rationale of any equating process should be a matter of indifference to the examinees at every given ability level whether they take test X or test Y.

There are two major theoretical camps of equating methodologies: the classical method of the observed-score test equating and the IRT method of true score test linking.

Classical Equating Methods:


Equipercentile equating considers the scores on tests X and Y to be equivalent if their respective percentile ranks in any given group are equal. Strictly speaking, in order to equate scores on two tests, the test must be given to the same groups of examinees. In practice, this process typically is carried out by giving the tests to randomly equivalent groups of examinees.

Linear equating is accomplished if score x on Test X and the score y on Test Y satisfy the following linear z-score regression equation:

$$y = \frac{\sigma_y}{\sigma_x} (x-\mu_x) + \mu_y$$

where \(\mu_x, \mu_y, \sigma_x, \sigma_y\) are means and standard deviations of the scores on tests X and Y, respectively. This relationship can also be expressed in terms of standard scores, namely, by equating the standard score on test X to the standard score on test Y.
The basic assumption of linear equating is that if two test score distributions can be equated, they differ only with respect to their means and standard deviations. Angoff (1971, 1982) has explicated detailed procedures to take into account many different procedures, for example, outliers and the unreliability of the test scores.

However, classical equating methods have several shortcomings (Hambleton & Swaminathan, 1985). The most salient one is that the conditions of equity can hardly be met when using classical methods. For example, it can be seen that both equipercentile and linear methods are highly group dependent, because both percentiles and regression are group dependent by nature. Furthermore, although it ensures that the transformed score distributions are identical, equipercentile method results in a nonlinear relation not only between the raw scores but also between true scores.

The second shortcoming is that it is possible that linear equating method which employs z-score regression does not warrant the assumption of symmetry. It is known that regressing on x from y will yield a different set of parameters from regressing y from x, unless the two groups of examinees correlate equally with a criterion and the two test forms measure the criterion with equal amount of measurement error. Clearly, it is extremely difficult, if not impossible, to achieve these two characteristics simultaneously with the test instrument and examinees, respectively.
Item Response Theory Linking Methods:

Items response theory methods obviate the need for equating because of its inherent property of invariance of item and ability parameters (Lord, 1981). Theoretically, if an IRT model fits the data, direct comparison of the ability parameters of two examinees who take different tests is made possible by this invariance property. Such invariance still holds even when examinees share only subsets of items. In short, apart from measurement error, ability estimates will be invariant across either whole tests or subsets of items. As a result, test score equating is often referred to as test score linking in the IRT framework. What must be ensured, however, is that item and ability parameter values based on two tests are on a common scale through an appropriate linking process. The common scale can be easily achieved using any of the four IRT linking designs: (1) single-group, (2) equivalent-group, (3) common-anchor, (4) common-person.

The single-group design is one that allows one group of examinees to take the two test forms to be equated. The use of a common group of people across two tests results in a common scale to reveal the relative difficulties of the two tests, and consequently, makes it possible to adjust for them. Although straightforward, this design is often not practical because it usually causes fatigue and boredom on the part of examinees.

The second design is the equivalent-group design which administers the two tests to two groups of examinees of equivalent ability distributions. Because of the equivalence of the ability distributions of the two groups, the difference in the test performance between the two groups must be due to the differences of the test item difficulties. Although theoretically plausible, it is hardly feasible to accurately match examinees in their abilities.
The third design is the common-person design to give the two tests to be equated to two groups of examinees, with a common group of examinees from both groups taking both tests. The two tests are brought down to the same scale through the performance of the common group of examinees taking both tests. It can be seen that this problem is a revision of the single-group design mentioned above. Although it shares the same shortcomings of the single-group design, this design is mostly used to equate two tests after they have been given to a large number of students.

The final design is called the common-anchor design which administers the two tests to be equated to two different groups of examinees with a common set of items embedded in both tests. The two tests are brought down to the same scale through the bridging of the common anchor items. It can be seen that this design is the least intrusive but most efficient. Because of such advantages, this design will be used in this dissertation to equate the scores on the chosen items.

Methods of Unidimensional and Multidimensional Scaling

In addition to IRT estimations of item parameters and student abilities and IRT equating methodology, scaling methodologies will be employed to discern the psychological processes that underlie examinees' item choices. As presented in the last chapter, both SIMILARITY and PREFERENCE data have been collected through the AP Chemistry Survey Kit. First, the pair-wise similarity data from 'How similar are these two questions in content?' will be analyzed through AVELOMAT (Dunn-Rankin, 1983) to obtain a lower triangular matrix of averages. The matrix is, then, fed into KYST II (Dunn-Rankin, 1983)
to discern the content dimensionality of both the MC items in the Survey Kit and the five essays. It is hypothesized that items of closer dimensionality are more likely to be chosen together.

The pair-wise comparison data from "Which item seems easier for you?" is analyzed through RANKO (Dunn-Rankin, 1983) to obtain the rank order of all the MC items and essay problems. Critical difference ranges are calculated at .05 significance level. With the critical ranges from RANKO, it is possible to evaluate which MC items or essays are perceived significantly different from one another, as well as the impact of such perceived difficulty on students' choices.

**Chapter Summary**

This chapter has reviewed the definitions, assumptions and the rationale of the three research methodologies of IRT, test equating and unidimensional and multidimensional scaling with an emphasis on their applications in this dissertation research. It has been pointed out that the major reasons to employ IRT methodology and IRT equating are to obtain relative high degrees of accuracy in the estimation of item parameters and student abilities, to aid instrument design, and to assess the impact of leaving choice scores unadjusted for their differential difficulties. Furthermore, the reason to use scaling methodology is to uncover the psychological processes underlying students' choices. With the methodology in place, the next step is to apply them to the data that are already in hand.
CHAPTER VI

OVERALL ESSAY CHOICE TENDENCIES AND PERFORMANCE

How do examinees choose essays? What are the tendencies of their overall choices? What is the relationship between their general tendencies of choices and their mean performance? These are the questions to be researched in this chapter. It is the purpose of this chapter to ascertain:

1. the relative difficulty levels of the five essays in terms of their expected scores;
2. the overall examinee choice tendencies;
3. the overall mean performance on both individual essays and essay combinations.

Relative Difficulty Levels of the Five Essays

Unlike a multiple-choice question whose difficulty level can be unequivocally described by the parameter called "item facility or difficulty" based on proportions of correct responses, an essay question does not possess such specially-designated parameters, because the probability to each of its graded responses is estimated separately. For example, all the five essays are measured on a 8-point scale, and students' responses can be scored from 0 through 8. Examinees of different ability levels have different probabilities to obtain each of the 9 points in the scale.
Two alternative measures can be used as indicators of relative essay difficulties -- the IRT expected scores and the observed raw scores for each of the five essays conditioned on abilities. The IRT expected score is the preferred choice because IRT can treat the five different essays in a similar way through a common-item anchorage. It is known that all the ten essay combinations share, at least, one essay in a chained fashion. For example, Essay Combination 5, 6 & 7 is connected with Essay Combination 7, 8 & 9 through Essay 7, called the "anchor item". Through such anchoring, the probability curves of answering each of the graded responses is calculated through

\[ T_{jk}(\theta) = \frac{\exp[\alpha_{jk}\theta + c_{jk}]}{\sum_{k=0}^{m} \exp[\alpha_{jk}\theta + c_{jk}]} \]

as a function of abilities. Then the expected scores can be computed by aggregating across the trace lines through

\[ E(\text{Score} | \theta) = \Sigma x_j T_{jk}(\theta) \]

However, it should be pointed out at this point that the accuracy of either of the two indices is open to debate for two reasons. First, the IRT expected scores are obtained under the violation of one of the key assumptions of IRT missing-at-randomness, because examinees chose the essay items they attempted. The effect of such a violation is one of the subjects of investigation in this dissertation. Second, the conditional observed mean scores might not be an unbiased choice either, because they are based on the examinees who chose their own essay items and the difficulty levels of the essay items are not adjusted.
Essay Difficulty as Reflected by IRT Expected Scores:

Let's first look at what IRT expected scores can hint to us in terms of the difficulty of the five essays. Figure 6.1 displays the IRT expected scores of the five essays on the basis of 18,462 examinees in the 1989 AP Chemistry Exam.

![Figure 6.1: IRT Expected Scores of the Five Essays](image)

It can be observed from Figure 6.1 that there is substantial amount of interaction between the expected scores of the five essays and ability distributions. In general, Essays 5 and 8 are highly similar to each other in their expected scores or difficulty levels across the entire ability distribution.
Essay 9 is the second most difficult for the very low ability range but becomes the easiest from middle ability onward. Essay 6 seems to be the most difficult of all the essays all the way from low ability to upper-middle ability range and switches to be the second or third easiest for high ability students. Essay 7 parallels Essays 5 and 8 in its difficulty within the low to upper-middle ability range and then zooms to be the easiest for the extremely high ability students. Table 6.1 below summarizes the difficulty ranks of the five essays across three ability levels. The higher the rank is, the more difficult the essay is.

**Table 6.1**

Summary of Difficulty Rank of Five Essays Across Three Ability Ranges

<table>
<thead>
<tr>
<th></th>
<th>Essay 5</th>
<th>Essay 6</th>
<th>Essay 7</th>
<th>Essay 8</th>
<th>Essay 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Ability</td>
<td>1</td>
<td>5</td>
<td>2.5</td>
<td>2.5</td>
<td>4</td>
</tr>
<tr>
<td>Middle Ability</td>
<td>2</td>
<td>5</td>
<td>3.5</td>
<td>3.5</td>
<td>1</td>
</tr>
<tr>
<td>High Ability</td>
<td>4.5</td>
<td>3</td>
<td>2</td>
<td>4.5</td>
<td>1</td>
</tr>
<tr>
<td>Average</td>
<td>2.5</td>
<td>4.3</td>
<td>2.7</td>
<td>3.2</td>
<td>2</td>
</tr>
</tbody>
</table>

The correlation between the average ranks in Table 6.1 above and the mean scores on the five essays shown in Table 6.3 is .95, significant at .05 level.

**Essay Difficulties as Reflected by Conditional Mean Scores:**

Figure 6.2 below shows the mean scores of the five essays for ten ability levels. It can be seen that the mean score curves of the five essays parallel those of the expected scores.
of the five essays in Figure 6.1, except the conditional mean score curves indicate little interaction. Although Essay 9 remains the easiest, Essay 6 becomes the hardest essay instead of Essay 5 as shown earlier. Essays 7 & 8 are virtually identical in terms of their difficulties throughout the entire range. Although Essay 5 is close to Essays 7 & 8 for the lower half of the ability distribution, it becomes slightly easier for the upper half of the ability distribution. In their entirety, the difficulty rank of the five essays is Essay 9 as the easiest, followed by Essays 5, 8 & 7 in between, and tailed by Essay 6 as the hardest.

![Figure 6.2: Essay Difficulties As Reflected by Conditional Mean Scores](image)

Figure 6.2: Essay Difficulties As Reflected by Conditional Mean Scores
Overall Essay Choice Tendencies

Allowing examinees to choose three out of the five essays yields 10 essay combinations ($\binom{5}{3}$): Essays 5,6,7; Essays 5,6,8; Essays 5,6,9; Essays 5,7,8; Essays 5,7,9; Essays 5,8,9; Essays 6,7,8; Essays 6,7,9; Essays 6,8,9 and Essays 7,8,9. Given the fact the five essays are not equally difficult, the first question to be asked is "Did the students choose these ten combinations randomly, or differentially according to different essay difficulties?"

A look at Figure 6.3 would tell that students did NOT choose the essay combinations randomly. Essay Combination 5,6,8 is the most popular one, while Essay Combination 6,7,9, the least popular combination. The rank of the choice frequencies is summarized in Table 6.2 on the next page.

It should be pointed out that 9425 out of the total of 18462 examinees (55% of the total population) chose the two essay combinations: 5,6,8 and 5,7,8, while less than 1% of people chose Essay Combination 6, 7 & 9. Reasons for such disparate differences in examinee essay preferences are presented in the next chapter.
Figure 6.3: Frequencies of Essay Choice Combinations Based on the Total Population of 18,435 Examinees

Table 6.2

Rank of Choices of Essay Combinations and Actual Frequencies

<table>
<thead>
<tr>
<th>Rank</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essay Choice</td>
<td>5,6,8</td>
<td>5,7,8</td>
<td>5,6,7</td>
<td>5,8,9</td>
<td>6,7,8</td>
<td>7,8,9</td>
<td>5,7,9</td>
<td>5,6,9</td>
<td>6,8,9</td>
<td>6,7,9</td>
<td></td>
</tr>
<tr>
<td>Choice Freq.</td>
<td>5,227</td>
<td>4,198</td>
<td>2,555</td>
<td>1,707</td>
<td>1,392</td>
<td>898</td>
<td>753</td>
<td>457</td>
<td>407</td>
<td>121</td>
<td>18,435</td>
</tr>
<tr>
<td>Percent</td>
<td>28.35%</td>
<td>26.68%</td>
<td>14.86%</td>
<td>9.26%</td>
<td>7.55%</td>
<td>4.87%</td>
<td>4.08%</td>
<td>2.48%</td>
<td>2.21%</td>
<td>.66%</td>
<td></td>
</tr>
</tbody>
</table>
The popularity rank of the five individual essays is obtained by summing the number of times that each of the five essays were chosen and is displayed in Figure 6.4. It can be seen that Essay 5 on valence and electronic configuration is the most popular, followed by Essay 6. Essay 9 on nuclear chemistry is the least chosen.

![Figure 6.4: Popularity Rank of the Five Essays](image)

**Figure 6.4:** Popularity Rank of the Five Essays

**An Overview of the Mean Performance on the Five Essays and Ten Essay Combinations**

Table 6.3 summarizes the mean scores on the five essays. Note that not all examinees answered every essay problem, since only three essays were required of the examinees.
Table 6.3
Mean Scores on Five Essays

<table>
<thead>
<tr>
<th>Essays</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. Score</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Mean Score</td>
<td>2.71</td>
<td>1.61</td>
<td>2.50</td>
<td>2.65</td>
<td>3.41</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>1.96</td>
<td>2.02</td>
<td>1.88</td>
<td>2.01</td>
<td>2.32</td>
</tr>
<tr>
<td>Mean as % of Max.</td>
<td>34</td>
<td>20</td>
<td>31</td>
<td>33</td>
<td>43</td>
</tr>
</tbody>
</table>

The order of the five essays in terms of mean scores is 9, 5, 8, 7, and 6, with Essay 9 having the highest mean of 3.41 and Essay 6 having the lowest mean of 1.61 points. The means for Essays 7 and 8 are very similar. Compared with the maximum score of 8 points possible on each essay, the mean scores ranging from 1.61 to 3.41 points, are fairly low, at only 20% to 40% of the maximum scores. Such averages imply that these essays were not only substantially difficult for the students as a whole, but also considerably different in terms of difficulty. The difference between the highest and lowest means is 1.8 points, occupying 23% of the maximum score. This difference indicates the necessity for equating the scores for these five essays.

What is the relationship between the mean scores on the five essays and their popularity rank? In other words, do more popular essays have higher means than the less popular ones? The answer is negative. The correlation between the popularity rank of the five essays and their corresponding means is -.60, although not statistically significant. This negative correlation is somewhat disturbing, because logically, people expect higher scores for the more popular essay combinations, since students are expected to choose the items that they feel confident in. This paradox will be dealt with later in Chapter X.
What is the mean performance like if the mean scores on the five essays are examined conditionally on students' ability levels? This question is important because it provides information on the mean performances relative to students of different ability levels. Figure 6.5 below shows that these five essays are differentially difficult to students of similar abilities.

Two general trends can be observed. First, as expected, the higher the ability levels, the better the respondents perform across all five essays. The mean difference between level-1 and level-10 students is 3.88.
Second, the effect of choice has a bigger impact on the students of middle ability levels than those of extreme ability levels. For example, the maximum differences resulting from choosing among the five essays are 0.89 points (1.15 - 0.26) for level-1 students as compared to 1.18 points (5.56 - 4.38) for level-10 students. Yet, the differences from choosing among these five essays range from 1.72 points (3.03 - 1.31) for level-5 students to 1.83 points (3.76 - 1.93) for level-7 students. Although small in its magnitude, a difference of 1.83 points takes up 22% of the maximum score of 8 points.

The above findings seem to suggest that the scores of middle-level students are affected more than the scores of either extremely high- or low-ability students. A little thought will find such an phenomenon justifiable. If some examinee is of either extremely high or low ability, it is highly likely that he/she would either get an item right or wrong, respectively. Yet, if someone is of middle ability, the choice of an item that is just "right" for his/her level would make an essential difference in how he/she would perform. If the chosen item is above his/her level, he/she might perform poorly. Please note that these findings concur with the expected scores of the five essays in Figure 6.2 which shows wider gaps for middle-ability students.

Now, let's examine how big the effect of choosing would be after students choose three out of five essays. Two trends can be observed from Figure 6.6. First, Essay Combination 5, 8 & 9 has the highest mean score of 9.89 points, while Essay Combination 5, 6 & 7, the lowest mean of 4.57. The difference between these two combinations is 5.32, which is 22% of a total score of 24 points. The difference between Essay 9 and Essay 6 takes up 23% of the maximum score on an individual essay. It seems to suggest that the
mean score differences of the essay combinations mirror the mean score differences among individual essays proportionally. This is shown to be true by the second trend.

Figure 6.6: Mean Scores on Ten Essay Combinations

The second trend is that the rank of the mean scores of the various essay combinations can be predicted, in most cases, by referring to the rank of mean scores of the individual essays. For example, if one is to predict which of the two essay combinations has a higher mean, Essay Combination 5, 6, 8 or 5, 7, 9? The answer is Essay 5, 7 & 9, because Essays 7 and 9 in the latter combination have higher means than Essays 6 and 8 in the former combination. This quick comparison is possible only because of the lack of interaction between essay means and student ability levels.
What is the relationship between the popularity of essay combinations and group mean levels of performance? More specifically, "Do more popular essay combinations imply higher mean scores?" Table 6.4 summarizes the mean scores for each of the ten essay combinations.

Table 6.4

<table>
<thead>
<tr>
<th>Rank</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essay Choice</td>
<td>5,6,8</td>
<td>5,7,8</td>
<td>5,6,7</td>
<td>5,8,9</td>
<td>6,7,8</td>
<td>7,8,9</td>
<td>5,7,9</td>
<td>5,6,9</td>
<td>6,8,9</td>
<td>6,7,9</td>
</tr>
<tr>
<td>Choice Freq.</td>
<td>5,227</td>
<td>4,198</td>
<td>2,555</td>
<td>1,707</td>
<td>1,392</td>
<td>898</td>
<td>753</td>
<td>457</td>
<td>407</td>
<td>121</td>
</tr>
<tr>
<td>Mean Scores</td>
<td>7.10</td>
<td>8.14</td>
<td>4.57</td>
<td>9.89</td>
<td>7.18</td>
<td>9.72</td>
<td>8.37</td>
<td>7.85</td>
<td>8.75</td>
<td>7.12</td>
</tr>
</tbody>
</table>

Similar to the earlier correlation of -.60 between the popularity rank of individual essays and their mean scores, the correlation between the rank of essay combination preferences and their mean scores is -0.22. This finding will also be explained in the next chapter.

What do the mean scores of the ten three-essay combinations look like at different ability levels? Figure 6.7 displays similar patterns of mean scores as those observed earlier. Again, there is less variation in the mean scores at either very high or very low ability levels, while the mean scores of middle-ability students are affected a great deal more.
levels. Again, the biggest difference of 5 points occurs at level 7, equivalent to about 20% of the total score of 24 points.

Figure 6.7: Mean Scores on Three Chosen Essays Conditioned on Ability Levels

Chapter Summary

This chapter has reported four areas of investigation regarding the nature and choices of the five essays. The first area of investigation was on the difficulty level of the five essays. Because of the lack of a specific parameter to describe the difficulty of an essay, both IRT expected scores and mean performances conditional on student ability levels have been examined. It has been found that the two measures agree with each other in terms of
the rank order of the essay difficulties, although the former seems to present a more accurate picture because IRT analyzes the five essays on a common scale through common-anchor linking.

The second area of investigation was the general essay choice patterns. It has been found that drastic differences exist in the frequencies with which the five essays were chosen by the 18,462 examinees. For example, Essay Combination 5, 6 & 8 was chosen by nearly 27% of the total population, while Essay Combination 6, 7 & 9 was chosen by less than 1% of the total examinees.

The third area of investigation focused on the consequences of choosing among essays of differential difficulty levels without properly adjusting for their difficulties. Substantial differences have been found in the mean scores of the examinee groups of similar abilities. The differences can be as big as 20% of the maximum scores or over 50% of the mean scores.

The final area of investigation is the relationship between the popularity of essays and essay combinations and the mean performance on them. Contrary to the common belief that examinees would choose the essays they feel confident in, it has been found that the mean performance on the essays and essay combinations is a reversal of the essay popularity.

What are the reasons behind such differential patterns of essay choices? What is common among those essays or essay combinations that were chosen the most or least? Who chose the most vs. the least popular essays or essay combinations? Who chose the essays or essay combinations that observed the highest or lowest mean scores? Who made deliberate choices and who did not? How can a psychometrician adjust for the score differences as the result of choosing more or less difficult essays or essay combinations?
These important questions are investigated in detail in the remainder of this dissertation. As
the dissertation unfolds, the answers to these questions will gradually unveil themselves.

Turn to the next chapter to see what the chemistry teachers have to say about the findings
in this chapter.
CHAPTER VII

COMPREHENSIVE ANALYSES OF THE RELATIONSHIPS AMONG STUDENTS’ ESSAY CHOICES, ABILITY PROFILES, AND PERFORMANCES

With the revelations of the disparate essay choice patterns and performances, the next step is to account for them. This chapter consists of three components. The first component summarizes teachers’ judgment of essay difficulties and the curricular reasons behind the differential essay choices. The second component explores the relationship between students’ ability levels and their choices. The third component investigates the performance differences between those who chose deliberately and those who chose casually. By the end of this chapter, all the three research questions regarding the relationship among students’ choices, abilities and performances will be answered.

Understanding Students’ Differential Choices of Essays and Essay Combinations through Teachers’ Perspectives

Two findings from the last chapter are specially striking. Why did so many students choose Essay Combinations 5, 6 & 8 and 5, 7 & 8, while so few chose Essay Combination 6, 7 & 9? Intrigued by such a mystery, the author conducted a follow-up survey, called "Chemistry Teacher Expert Judgment", to consult with the twelve local chemistry teachers on their interpretations of the surprising and seemingly paradoxical findings. Eight teachers
replied in time. The following is the summary of such a consultation on the basis of the eight teachers’ comments.

Chemistry Teacher Expert Judgment Survey:

"Chemistry Teacher Expert Judgment Survey" consists of three components (see Appendix C):

1. Rank of essay difficulty;
2. Textbook coverage of the chemistry topics represented by the five essays;
3. Elaboration of the common characteristics shared by both the most and least frequently chosen essay combinations.

There are three purposes to these areas of investigation. First, it was informative to see how differently or similarly teachers ranked the five essays in terms of their difficulty levels. It has been shown that the popularity of essay choices correlated negatively with mean performance. If teachers’ rankings agree with students’ choice preferences, it signifies that students’ choices have been significantly influenced by teaching. If the teachers’ ranking does not agree with students’ choice preferences, it implies that there are essential differences between the perceptions of teachers and students.

The second purpose of this specific inquiry was to determine the effects of textbook coverage on teachers’ ranking and students’ choice preferences.

The third purpose of the teacher survey was to ask the teachers to offer their expert judgment on what is shared by the most and the least frequently chosen essay combinations.
From such an inquiry, it is possible to infer on the content familiarity as reasons for students’ choices.

It is through these three areas of investigation that the mysteries of differential choice patterns and the negative correlation between choice preference and mean performance have been enlightened. Because it has been shown that the 618 students in Hawaii chose the essay combinations in the same way as the examinees in the 1989 AP Chemistry Exam, and because Hawaii AP chemistry students performed at the same level as the national norm, the information from the eight local AP chemistry teachers could be generalizable to the national data.

Teachers’ Rankings of Essay Difficulties and Textbook Coverage:

As the first area of consultation, the eight chemistry teachers were asked to carry out two tasks -- first to rank order the five essays in terms of their relative difficulties from the easiest to the hardest, and then rate each of the five essays in terms of the coverage of the commonly used AP chemistry textbooks. Although the essay difficulty levels have been empirically estimated through their IRT expected scores and conditional mean scores, it is believed that asking the teachers to independently rank them will provide, at least, two extra areas of information -- to see the differences or similarities between the perceptions of teachers and students, and to yield information on the relationship between teachers’ perceptions and textbook coverage. Please note that the teachers were asked to perform the first task before they were shown the drastically different choice patterns. Table 7.1 below summarizes the eight teachers’ ratings.
Three tendencies can be observed from Table 7.1. First, there is an order of essay difficulty across the five essays, except for Essays 7 & 8 which rank equally with each other. It is more surprising than expected to observe that the teachers’ ranking of the essay difficulties correspond almost perfectly with the popularity of the five essays -- the correlation is .92. In other words, if the teachers were to choose three essays, they would probably choose them in a similar fashion as the students did.

Table 7.1

Summary of Teacher Ratings of Essay Difficulty Levels

<table>
<thead>
<tr>
<th>Teachers</th>
<th>Difficulty Ranking</th>
<th>Extent of Textbook Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5 6 7 8 9</td>
<td>5 6 7 8 9</td>
</tr>
<tr>
<td>1</td>
<td>2 5 1 4 3</td>
<td>5 5 4 3 5</td>
</tr>
<tr>
<td>2</td>
<td>1 4 3 4 5</td>
<td>2 2 2 4 1</td>
</tr>
<tr>
<td>3</td>
<td>1 3 2 4 5</td>
<td>5 4 3 2 1</td>
</tr>
<tr>
<td>4</td>
<td>1 2 4 3 5</td>
<td>3 4 2 4 3</td>
</tr>
<tr>
<td>5</td>
<td>2 1 5 3 4</td>
<td>3 2 1 3 2</td>
</tr>
<tr>
<td>6</td>
<td>1 3 5 2 4</td>
<td>5 5 3 5 4</td>
</tr>
<tr>
<td>7</td>
<td>1 . . .</td>
<td>4 3 3 4 1</td>
</tr>
<tr>
<td>8</td>
<td>. 1 . .</td>
<td>3 3 3 3 3</td>
</tr>
<tr>
<td>Avg. Rank</td>
<td>1.4 3 3.3 3.3 4.3</td>
<td>3.75 3.5 2.6 3.7 2.3</td>
</tr>
</tbody>
</table>

* Note: The "." in the above table represents a missing entry which is excluded from calculating the average ranks. The last two teachers only indicated the essays that they viewed as the easiest.

Based on Question 70 of *Teacher Evaluation Sheet* (see Appendix B), it has been found that out of eleven chemistry teachers, three chose Essay Combination 5, 6 & 8, the
most popular essay combination; another three chose Essay Combination 5, 7 & 8, the
second most popular essay combination; another three chose Essay Combination 5, 6 & 7,
the third most popular essay combination; two chose Essay Combination 5, 8 & 9, the essay
combination with the highest mean score.

The only discrepancy between the teachers’ average rating and the observed mean
scores is the positioning of Essay 7 and Essay 8. According to three of the eight teachers,
Essay 7 turns out to be a little harder basically because it requires substantially more lab
experience to solve it.

Second, most teachers agreed that the textbooks seem to have provided sufficient
coverage on Essays 5, 6 & 8, while substantially less coverage on Essays 7 & 9. Please note
that Essay Combination 5, 6 & 8 has been shown to be the most chosen combination. Essay
7 requires additional lab experience, which is not always equally available to many students.
In addition, Essay 9 on nuclear chemistry is shown to be too specialized a topic for teachers
to teach extensively.

Finally, from the correlation between the teachers’ ranking of essay difficulty and
ratings of textbook coverage (r = -.73), it can be inferred that textbook coverage of the five
essays is related to how the teachers ranked the essay difficulty levels, and probably to how
students chose, too. The next section summarizes teachers’ explanations of the observed
essay choice patterns.
Teachers' Perspectives on the Popularity of the Five Essays or Essay Combinations

What caused both the teachers and students to view the difficulty levels of the five essays in a virtually identical way?

The teachers indicated that the four topics represented by Essays 5, 6, 7 & 8 belong to what is considered as the "core" of general chemistry. Essays 5, 6 & 8 ask about the basic theories and principles, while Essay 7 is oriented towards laboratory performance and implementation of chemistry principles. Essay 9 is on a highly specialized area of nuclear chemistry. From this description, it is not hard to infer on the reasons behind the differential choices for the ten essay combinations.

Explanations for Essays 5, 6 & 8, the Most Popular Essay Combination:

Why did 29% of the 1989 total examinee population choose Essay Combination 5, 6 & 8? There seem to be two reasons for choosing this combination. First, content-wise, these three questions address the chemistry theories and principles that begin to be covered in regular high school chemistry, and re-emphasized in AP chemistry textbooks. Second, task-wise, they involve just accounting for observations and describing concepts, which seems to be less complicated than actually solving a problem. Students should feel "safer" with problems that address theoretical issues. It is much easier to talk about something than actually implementing something.

Students might have felt that they could, at least, obtain partial credit by theorizing the chemistry phenomena in question. For low-ability students, they must have felt that they
could "fake", as one teacher put it, their answers more easily with this combination. In addition, this is the best place to avoid lab experiences and the more advanced subject of nuclear chemistry.

Why does this most popular combination have the second lowest mean score? Although familiar to students, the sub-questions or parts of the three essays, especially Essays 6 & 8, are very challenging, because they require deep understanding of the system within which these questions are framed. For example, being the most difficult, Essay 6 involves deep theoretical understanding of "the expected trends in the melting points of the four compounds LiF, NaCl, KBr, and CsI" through bonding principles. According to two teachers, the knowledge involved is so subtle that it "often escapes most students the first time around."

Explanations for Essays 5, 7 & 8 -- the Second Most Popular Combination:

As the second most popular essay combination, Essays 5, 7 & 8 were chosen by 27% of the total 1989 examinee population. In this combination, Essay 7, a lab-experience-oriented question, replaced Essay 6, the question that requires deep theoretical knowledge. Essay 7 is more difficult because it involves not only theoretical chemistry knowledge, but also the ability to perform chemistry experiments. Three teachers suggested that it is much harder to solve a chemistry problem by providing and specifying actual data and elements in compounds than to offer an explanation or description. As a result, students are less confident in getting this problem correct. Furthermore, laboratory experiments are not equally accessible to all students. Essay 7 asks heavily about reactions, focusing on the
applications of theory. According to the teachers, students normally "fear" the writing of reactions the most. This might be why this combination was not as preferable as Essay Combination 5, 6 & 8. However, probably because of its relatively heavy emphasis in the AP chemistry course, students were attracted to this question by its familiarity.

Although seemingly more difficult, Essay 5, 7 & 8 ranks the sixth in terms of its mean score. The reason for such an increase in the mean score is probably that whoever chose this combination, he/she must have had a fairly solid proficiency in not only chemistry knowledge but also performance.

Explanations for Essays 5, 6 & 7 -- the Combination with the Lowest Mean Score:

Essay Combination 5, 6 & 7 is unique in that it includes not only the theoretically challenging problem of Essay 6, but also the performance-demanding problem of Essay 7, although Essay 5 is relatively easy. According to one teacher, this combination could be a "nightmare" for students who wished to maximize their points. Three reasons may explain why this combination had the third highest frequency. First, the familiarity of these three essays must have attracted many students in the first place, who might not have realized how "tricky" they were. According to teachers' expect judgment, the content of these three essays was most extensively covered by AP chemistry textbooks.

The second reason is "lethargy" as three teachers suggested. Because these were the last three questions in the last part of the whole test, some students might have chosen this combination to "get over" with the test. It will be shown later in this chapter that those students who chose Essays 5, 6 & 7 in the sequential order scored significantly lower (two-
points on the average) than those students who answered these three essays in a selective order, say, first Essay 5, then Essay 7, and finally, Essay 6.

The third reason is that according to the teachers, some students probably did not realize they could choose three out of the five essays. This could be true of those students who were "rushing" through this last part of the exam, because they spent too much time in the previous essay parts.

From the three reasons above, it is not surprising to see that this combination has the lowest mean score of all ten essay combinations.

**Explanations for Essays 6, 7 & 9 -- the Least Popular Combination:**

Essay Combination 6, 7 & 9 has the lowest choice frequency, which is not surprising from our understanding of Essays 6 & 7 from the above. How is Essay 9? According to Figure 6.5, Essay 9 has the lowest choice frequency. The reason that Essay 9 is the least favored essay might have been because of relatively less instructional coverage according to Table 7.1. The teachers offered three reasons to this fact. First, because of its highly specialized usage, nuclear chemistry, as represented by Essay 9, is often located towards the end of most AP textbooks, even though it is covered adequately, if not extensively. Two examples of AP chemistry textbooks are *Chemical Principles* by Masterton, Slowinski and Stanitski (1985) and *Chemistry* by R. Chang (1988). They are two of the most widely adopted textbooks for AP Chemistry. In these textbooks, nuclear chemistry is one of the terminal chapters, for example, as the 23rd chapter or later. As a result, the topic is often
either not reached or glossed over by many chemistry teachers. It is highly likely that this insufficient instructional coverage has led to the much lower choice frequencies of Essay 9.

Second, nuclear chemistry is something that teachers only talk about, because it does not lend itself to experimentation in any high school chemistry laboratory. Some knowledge required by Essay 9 goes beyond a normal AP chemistry curriculum. Therefore, many students may feel unprepared.

Third, in addition to being less taught, the content of Essay 9 is peculiar in that it does not seem to address a theory or application of a theory. Rather, this question appears to be testing only the recall of particular facts on the decay of a particular isotope. In general, what is taught in nuclear chemistry is the common decay process of types of isotopes. Students are not expected to memorize the decay process of a particular isotope. Two teachers indicated that they even had to look up in textbooks for the specific facts in order to solve this problem.

Is Essay 9 hard? The answer is both Yes and No. Compared with the other essays, Essay 9 is undisputedly the least complicated one, if studied, because it does not require extensive interconnected chemistry knowledge as the other essays do. It is mostly an either "know" or "don't know" situation. If you have studied it before, you have a high chance of solving it. But if you have not studied this particular problem, it is virtually impossible to score any point. As a result, those who did not know anything about it would not attempt it. Those who did know about it would score high. This explains why this essay has the highest mean score in light of the lowest choice frequency, as shown by Table 6.3 and Figure 6.5 in the last chapter.
Explanations for Essays 5, 8 & 9 -- the Combination with the Highest Score:

Why did Essay Combination 5, 8 & 9 have the highest mean score? With the help of the previous description, it is not hard to narrow down on three reasons. First, all these three essays share one distinct feature --- they require students to explicate on three areas of chemistry that are detail and fact oriented. In order to successfully complete the theoretical and conceptual explication, the students for this combination had to have an exceptionally thorough knowledge of chemistry.

Second, as explained earlier, Essay 9 on nuclear chemistry is often glossed over because of its specialized application and relatively terminal location in AP textbooks. It can be deduced that once a student reached and studied nuclear chemistry to a certain extent, he/she should be already extensively educated in chemistry. It will be shown later in this chapter that more higher-ability than lower-ability students chose this essay combination.

Finally, the non-inclusion of Essays 6 & 7, which required deep theorization and extensive lab experience, respectively, must have slightly eased the overall difficulty level for this combination. That is probably why Essay Combination 7, 8 & 9 and Essay Combination 6, 8 & 9 exhibit slightly lower mean scores, respectively.

With the above answers to the nature of the essay combinations, the next question is "Who can accurately identify the relative difficulties of the essays and choose the easier ones to their advantage?" This question will be addressed in the next section of this chapter.
Exploring the Relationship between Students' Ability Profiles and Choices

It is known from Chapter VI that the different preferences for the five essays are largely due to the fact that some essays are more presented earlier in the textbook and are more familiar to students than the others. Light has been shed on the negative correlation between students’ preferences of essays or essay combinations and the mean scores on them -- the more familiar general chemistry essay problems tapped more deeply into theories or concepts, rather than the specific facts demanded by the highly specialized nuclear chemistry problem.

Given the knowledge of the properties or nature of the ten essay combinations, the next question is "Who chose each of the ten combinations?" More specifically, what is the ability profile of the students who chose each of the ten essay combinations? The main purpose of this investigation is to examine the relationship among student ability profiles, essay choices and performance. Five areas of inquiry constitute this investigation. The first area of inquiry is "What are the mean group abilities behind the ten essay combinations when all the ten groups of students are measured on the same scale through polytomous IRT?" Although the mean scores for the ten essay combinations are already known from Figure 6.7, it is important to point out that those means are slightly biased because they were based on the students who had already chosen the ten combinations without adjusting for the differential essay difficulties. The IRT measure to be shown will have adjusted for such difficulty differentials through a common-anchor linking described in Chapter IV.
The second area of inquiry is "What do the student ability distributions look like for the ten essay combinations?" The purpose of this inquiry is to identify "who chose what" and who can correctly choose the essay combination that is most appropriate for their ability.

The third area of inquiry is "How differently do students of different abilities levels choose?" This question is a complement to the previous inquiry to look at the relationship between ability and choice from a different point of view.

The fourth area of inquiry is "Who chose Essay Combination 5, 8 & 9 that had the highest mean score, although it was chosen by only 9.2% of the 1989 AP Chemistry examinees?" The identification of the student ability composition of this group is of special importance in that it can further illuminate on the relationship between performance and ability.

The fifth area of inquiry is "Did every examinee choose? If not, who did not or could not choose at all? What is the consequence of not choosing deliberately the appropriate items?" This question is important because the basic assumption of allowing choices is that every examinee is able to choose to his/her advantage. To the extent that this assumption is found incorrect, we can see the necessity to equate scores.

**Mean Group Abilities behind the Ten Essay Combinations as Estimated through Polytomous IRT:**

Although the mean scores of the ten essay combinations are known, we still do not know the mean abilities of the ten groups of students, because the mean scores are not on a comparable scale due to the fact that the essay difficulties are not adjusted to be equal.
Yet, this adjustment is automatically fulfilled if the expected values of the probabilities of the students attempting each of the ten combinations are estimated through IRT through a common anchor design. The mean ability estimates for the ten groups of students are summarized in Figure 7.1:

Figure 7.1: Group Mean Abilities for Ten Essay Combinations (In IRT Standard Scale)

Note that all the mean estimates in Figure 7.1 are expressed in terms of IRT standard scale, normally ranging from -3.5 to 3.5. The mean estimate of the students attempting Essay Combination 5, 6 & 8 is set to be zero against which the other group mean abilities are measured. It can be seen that Essay Combination 5, 8 & 9 has the highest group mean
ability, while Essay 5, 6 & 7, the lowest group mean ability. The second and third highest group means belong to the students choosing Essay Combinations 7, 8 & 9 and 6, 8 & 9.

The rank of the group mean abilities for the ten combinations correspond perfectly with their mean scores shown in Chapter VI. As a matter of fact, the correlation between the mean scores and the mean group abilities for the ten essay combinations is .96.

However, it should be pointed out that despite the almost perfect correlation, these two indices are fundamentally different in their perspectives. While the mean scores are based on the outcomes of performance, the IRT group means originates from estimating a student’s probability of answering an item right. An analogy of cause and effect may clarify their distinction a little more. The mean scores reflect the outcome, while IRT mean indices start from the cause and measure the effect.

Another advantage of using IRT mean group ability estimates is that it shows a clearer sense of reference because of the use of a standard scale. For example, the group mean ability of -1.02 for Essay Combination 5, 6 & 7 indicates that the whole group is one standard deviation below the population mean, while the group mean of .47 for Essay Combination 5, 8 & 9 is about half standard deviation above the mean. Yet, by looking at the mean scores for these two groups, one would not have such a clear sense on how really different these two groups are. How different are the two groups of Essay Combination 5, 6 & 7 and Essay Combination 5, 8 & 9 on the average? In probability terms, they are 43% different!
Student Ability Profile Behind the Ten Essay Combinations:

Although the information on the mean group abilities behind the ten essay combinations offers a general picture of the relative strengths of the ten groups of students, more insight can be obtained with a look into the detailed student ability profiles that compose the mean abilities --- in other words, to investigate "who chose what." Figure 7.2 consists of ten plates, each of which depicts the numbers of the ten levels of students choosing each of the ten essay combinations.

Figure 7.2: Essay Choice Frequencies vs. Ability Levels
Figure 7.2: (Continued) Student Ability Profiles behind Each Choice Pattern
Three tendencies can be observed from the ten plates in Figure 7.2. First, eight of the ten essay combinations were attempted by approximately equal numbers of students of ten ability levels. For example, approximately about 25-27% of the students across levels 2-9 attempted Essay Combination 5, 7 & 8, while approximately 4% students from all ten ability levels attempted Essay Combination 5, 7 & 9. However, this is not true of Essay Combination 5, 6 & 7 and Essay Combination 5, 8 & 9. Remember it has been shown in Chapter VI that these two essay combinations have the lowest (4.57) and the highest (9.89) mean scores, respectively.

The second tendency is that Essay Combinations 5, 6 & 8 and 5, 7 & 8 attract the most students. Note that these two essay combinations do not involve Essay 9 on nuclear chemistry which is often "glossed over" if not skipped, according to the teachers' perspectives. Furthermore, these two essay combinations do not include Essays 6 and 7 simultaneously, which are described as "the most involved" by the teachers.

The third tendency is that the combination that contains Essay 9 or Essays 6 & 7 simultaneously was shunned by most students who made definite choices. This is why the latter six combinations have greatly reduced numbers of students across all the ten ability levels.

The conclusion that can be made from the above three tendencies is that in eight of the ten combinations, both the "popular" and "unpopular" essay combinations are embraced and shunned equally by both high and low ability students. In other words, both low and high ability students tended to both perceive and choose similarly in these eight combinations. The next step is to find out how differently students of different ability levels tended to choose.
Essay Choice Patterns Across Ten Ability Levels:

Complimentary to the previous investigation of "Who chose what" is another look into the essay choice patterns of the students across the ten ability levels. Do students of different abilities have similar or different essay choice patterns? This information is revealed in Figure 7.3. Again, Figure 7.3 consists of ten plates, each of which represents the choice patterns of each of the ten levels of examinees.

Figure 7.3: Essay Choice Patterns of Students of Ten Ability Levels
Figure 7.3: (Continued) Essay Choice Patterns of Students of Ten Ability Levels

Group Figure 7.3 presents an overview of how students of each of the ten ability-levels chose the ten essay combinations. Each of the ten smaller pictures displays the essay...
choice patterns for each of the ten ability levels. It can be seen that all the smaller pictures, except for combination 5, 6 & 7 for level-1 students, are almost identical. In other words, students of Level 2 through Level 10 tended to choose the ten essay combinations similarly. The extent of the choice pattern similarity is reflected by the significant intercorrelations among the choice frequencies of the ten ability levels, as summarized in Table 7.2:

Table 7.2
Intercorrelations Among Choice Preferences of Students of Ten Ability Levels

<table>
<thead>
<tr>
<th>Levels</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.88</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.79</td>
<td>0.98</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.73</td>
<td>0.97</td>
<td>0.99</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.71</td>
<td>0.96</td>
<td>0.98</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.65</td>
<td>0.92</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.58</td>
<td>0.90</td>
<td>0.95</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.61</td>
<td>0.91</td>
<td>0.95</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.57</td>
<td>0.88</td>
<td>0.93</td>
<td>0.96</td>
<td>0.98</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.49</td>
<td>0.81</td>
<td>0.86</td>
<td>0.90</td>
<td>0.93</td>
<td>0.94</td>
<td>0.96</td>
<td>0.96</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>All</td>
<td>0.74</td>
<td>0.96</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
<td>0.92</td>
</tr>
</tbody>
</table>

(Note: Correlations below 0.65 is not statistically significant at 0.05 level.)

Two trends can be observed from Group Figure 7.3 and Table 7.2. First, the choice pattern of level-1 students are significantly different from those of level-7 through level-10 students. Second, the choice patterns of students of similar ability levels are similar. The higher the ability, the more clear this tendency becomes. Specifically, the choice preferences
of students at and above Level 2 are highly similar, while the choice preferences of students above Level 5 are virtually identical.

One conclusion can be made from the above two tendencies. There is a substantial amount of agreement regarding the choice patterns of the students of all ability levels, except for level-1 students. The biggest variation occurred with level-1 students.

This finding is contrary to what has been hypothesized. A lot more variations in the choice patterns are expected across the ten levels of students, due to the common belief that high-ability students would choose differently from low-ability students. Equal numbers of students from the ten ability levels balked at answering Essay 9 on nuclear chemistry. Again, equal numbers of students from the ten ability levels consciously stayed away from confronting Essays 6 and 7 simultaneously. Such a finding seems to suggest that the majority of the students across the ten ability levels could determine the nature and properties of all the five essays.

Given the regularity in the student ability profiles behind the ten essay combinations, the two questions are "Who were really 'bold' enough to confront the 'risky' combination of 5, 8 & 9? Who were really 'ambitious' enough to take the 'heavy-duty' combinations of 5, 6 & 7? What are the consequences of being both 'bold' and 'ambitious'?"? These are the questions to be answered next.
"Going Against the Odds":

Although it ranks the fourth in terms of its popularity, Essay Combination 5, 8 & 9 has the highest mean score (9.89), and the most highest-ability students attempting it. What is the student ability profile behind this combination? A look at Figure 7.4 will provide a clear answer.

![Student Ability Composition in Essay Combination 5, 8 & 9](Image)

Figure 7.4: Student Ability Composition in Essay Combination 5, 8 & 9

Figure 7.4 shows that this essay combination has more higher ability students than low-level students. It is interesting that there are increasing numbers of higher-ability students who chose this combination. The higher the abilities, the more students chose this
combination. This kind of "step-up" ability distribution seems to suggest two interesting phenomena. First, if this essay combination represents a relatively difficult one because of the infrequently-taught Essay 9, the "step-up" ability distribution seems to imply that some of the higher ability students tended to choose a more challenging set of essays. On the other hand, if this essay combination represents a relatively straightforward one because these three essays have been described as the least "involved" set, then this "step-up" ability distribution may imply that as the ability gets higher, students become more capable of accurately identifying the real nature of the five essays.

The next interesting question to be asked is "Is it true that this group of students are really at a statistically significant higher level of chemistry competence to start with than the rest of the population?" Although this question cannot be answered directly, it is possible to infer an answer by comparing the mean scores on the 75 MC items of the ten levels of students who chose this combination with the ten levels of students who answered any other essay combination. For this investigation, three groups of students are chosen:

1. the group for Essay Combination 7, 8 & 9, because it has the second highest mean score;
2. the group for Essay Combination 5, 7 & 8, because it attracted the biggest number of respondents
3. the group for Essay Combination 5, 6 & 7, because it has the lowest mean score;

If there exist significant differences, it can be concluded that the students who chose Essay Combination 5, 8 & 9 are significantly different from the rest of the population who
took the 1989 AP Chemistry Exam. Can such a conclusion be made? A look at Figure 7.5 negates this idea.

![Bar chart showing mean scores across ability levels and essay combinations](chart.png)

Figure 7.5: Mean Score Comparison of the Ten Levels of Students Across Four Essay Combinations

Figure 7.5 shows that the ten levels of students who answered these four essay combinations were virtually identical to start with in terms of their chemistry abilities on the basis of the 75 MC items. Please note that this is still true even with the ten levels of students who chose Essay Combination 5, 6 & 7 which is known to have the lowest group mean among the ten groups.

The above finding strongly suggests that whatever differences there are across the ten essay combinations, they are due to the relative difficulties of the essays that consist of the
ten essay combinations. This is another strong piece of evidence for the importance of adjusting students’ scores after their choices.

However, it should be pointed out that this finding does not negate the possible fact that the students choosing Essay Combination 5, 6 & 9 might be different from the other students in terms of their curricular experiences. It is possible that this group of students might have studied nuclear chemistry more extensively than the rest of the population, which made it possible for them to identify the relative straightforwardness of Essay 9.

"Facing the Challenge":

In order to answer "who chose Essay Combination 5, 8 & 9", which has the highest group mean, it is now time to check out "who chose Essay Combination 5, 6 & 7" that has the lowest mean score. As described earlier, Essays 6 and 7 are the most involved question. Who would choose them? The ability composition of the students attempting this combination is revealed in Figure 7.6.
Figure 7.6:  Student Ability Composition in Essay Combination 5, 6 & 7

Figure 7.6 portrays a completely opposite picture of student ability profiles as compared to Figure 7.4. With the lowest mean score, Essay Combination 5, 6 & 7 has the highest number of level-1 students and the smallest number of level-10 students. As the ability level increases, the numbers of students steadily decreases. Such a "step-down" choice pattern seems to suggest that low-ability students did not seem to recognize the relatively high difficulty levels of this essay combination.

Why did so many level-1 students choose this combination, given that Essays 6 and 7 are so complicated? The fact is that a disproportionate 35% of level-1 students vs. only 6% of level-10 students chose this pattern. Given the fact that Essays 6 & 7 are so blatantly
difficult, one can't help wondering how many of these level-1 students really made their conscious choices. Is it possible that a certain number of the level-1 students might have just attempted the three essays in the "first-come-first-tackle" basis, since Essays 5, 6 & 7 happen to be the first three essays? If so, why didn't they choose? How big is the difference between the level-1 students who did choose and their counterparts that did not. If they did not choose, then to what an extent did their scores suffer? Before investigating these question, it is necessary to distinguish *sequential* vs. *selective* choosing.

**Exploring The Effect of Sequential vs. Selective Choosing on Performance**

Although no one can be perfectly sure who deliberately chose and who did not, it seems reasonable to assume, given the structure of Part D, that most of the non-deliberately-choosing examinees must have been those that answered Essay 5, 6 and 7 in the sequential order of these three essays. Those examinees who did not choose Essays 5, 6 and 7 in such sequential order, as well as those who chose the other essay combinations can be classified as choosing examinees.

Two points are worth emphasizing here. First, the order that Essays 5, 6 and 7 are chosen is essential here. If one student first worked on Essay 5, then Essay 7, and finally Essay 6, he/she is classified as a choosing examinee, because he/she made some conscious choice in deciding which essay problem should be tackled first, even though his/her essay choice combination is Essays 5, 6 and 7. Second, not all the examinees who chose to answer Essays 5, 6 and 7 in the sequential order are necessarily non-choosing students. It is admitted here that some choosing students might have deliberately chosen to answer the three
Essays in the sequential order. However, there is no information in the data that can help to identify such examinees, whose number is expected to be small, given the complexity of Essays 6 and 7.

The Effect of Sequential vs. Selective Choosing on Level-1 Student Performance:

What is the difference between those who chose to answer the three essays sequentially and those who chose to answer these three essays selectively? Let's first limit our discussion to the level-1 students, as shown in Figure 7.7.

Two trends are obvious from Figure 7.7. First, in general, there is a significant difference between the sequentially-choosing and selectively-choosing level-1 students. The former has a mean of 0.60 out of 24 points, while the latter, a mean of 2.7. A t-test of the difference yields a t-value of 18.23, highly statistically significant at .0001 probability level.

Second, although there were almost equal numbers of sequentially-choosing and selectively-choosing students (330 vs. 344) at ability level 1, seven times more sequentially-choosing students received zero scores than selectively-choosing students (210 vs. 30).

Why did the sequentially-choosing students perform so poorly compared with selectively-choosing students on Essay Combination 5, 6 & 7? Is it because the sequential group is much less competent than the selective group? The answer is maybe. Because these two groups differ by 1.45 in terms of their mean scores on Section I of the 75 MC questions (The sequential group has a mean of 4.25, while the selective group, 5.70.) However, would anyone attribute the dramatic performance difference observed in Figure 7.7 to this minute mean difference? There must have been other factors that attributed to this. It is the belief
of both this author and some teachers that it must have been a combination of factors that attributed to the drastic differences between sequential choosing and selective choosing. Among the main factors are physical fatigue, mental stress and boredom etc. However, the most significant factor must have been personal motivation or lethargy. It is highly possible that the sequentially-choosing students must have lost their motivation to achieve their best scores. This hypothesis can be further strengthened by the following investigation of the overall performance difference between the two groups across the ten ability levels.

![Performance Comparison Between Sequential vs. Selective Choosing of Level-1 Students Only](image)

Figure 7.7: Performance Comparison Between Sequential vs. Selective Choosing of Level-1 Students Only
The Overall Effect of Sequential vs. Selective Choosing on the Performance:

Is it true that the effect of sequential and selective choosing only occurred to level-1 students? Before answering this question directly, let's first see how many students chose sequentially vs. selectively across the ten ability levels in Figure 7.8:

![Graph](image)

**Figure 7.8:** Numbers of Sequentially vs. Selectively Choosing Students Across Ten Ability Levels

Two tendencies can be clearly seen from Figure 7.8. First, across the ten ability levels, more students selectively chose. Specifically, out of the 2,566 examinees who responded to Essay Combination 5, 6 & 7, at least, 1,856 examinees deliberately chose to
answer these three essays, while approximately 710 examinees seemed to have answered these three essays because they were presented first. Second, as the ability increases, the number of sequentially-choosing students drastically decreases. These two findings seem to suggest that student competence is one of the major factors in determining whether or not a student would choose sequentially. The higher the ability is, the more likely a student would choose selectively.

Is there a consistent performance difference between these two groups? The answer is yes, as shown in Figure 7.9. It can be observed that there is a consistent mean difference between the sequentially and selectively choosing students, ranging from the minimum of 2.06 points at level 6 to the maximum of 4.77 points at level 10. On the average, the selective group scored about 3.95 points better than the sequential group (5.66 for the selective group vs. 1.71 for the sequential group). This difference is shown to be statistically significant by a two-way ANOVA analysis, as shown in Table 7.3.

Table 7.3
Summary of ANOVA Results Based on Randomly Selected 11 Students across the Sequential and Selective Groups and Ten Ability Levels

<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>Anova SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob &gt; F</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>19</td>
<td>2043.436</td>
<td>107.549</td>
<td>10.64</td>
<td>0.0001</td>
<td>45.42%</td>
</tr>
<tr>
<td>Group</td>
<td>1</td>
<td>371.800</td>
<td>371.800</td>
<td>36.77</td>
<td>0.0001</td>
<td>8.87%</td>
</tr>
<tr>
<td>Level</td>
<td>9</td>
<td>1607.800</td>
<td>178.644</td>
<td>17.67</td>
<td>0.0001</td>
<td>37.21%</td>
</tr>
<tr>
<td>Group*Level</td>
<td>9</td>
<td>63.836</td>
<td>7.093</td>
<td>0.70</td>
<td>0.7020</td>
<td>Negligible</td>
</tr>
<tr>
<td>Error</td>
<td>200</td>
<td>2022.364</td>
<td>10.11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>219</td>
<td>4065.800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-Square</td>
<td></td>
<td>0.503</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 7.3 clearly shows a highly significant main effects between choosing and non-choosing students, and across the ten ability levels, while there is virtually no interaction between the dimensions of groups and ability levels. More specifically, choosing students performed significantly better than non-choosing students. This performance advantage is consistent across the ten ability levels. As expected, the higher the ability levels, the better the performance. There is virtually no interaction between the way examinees answered the essay questions and their levels of abilities.
Please note that the above two-way ANOVA analysis is conducted on the basis of the equal sample size of 11 randomly selected students in each cell. As shown in Figure 7.8, there are drastically different numbers of examinees in each group -- the biggest group size is 344, while the smallest group size is only 11. Since extremely different N sizes may ordinarily render ANOVA analyses invalid (Keppel, 1982), the author employs one of the commonly-used solutions to rectify the problem -- random sampling of 11 cases from each of the 20 groups to make the sample sizes equal. Please note that the reduction of sample sizes makes the ANOVA analysis much more conservative -- making it more difficult to obtain significant results. Compared with Figure 7.9. Figure 7.10 verifies that the random reduction of group sizes has essentially maintained all original performance characteristics of both the sequentially and selectively choosing students.
It should be pointed out here that the observed consistent difference is somewhat surprising. Although differences between the two groups are expected at lower levels, they should gradually taper off as ability levels increase, because high ability students solve problems more consistently. In both Figures 8.9 and 8.10, however, the difference between the two groups of students is the biggest at level 10.

As hypothesized earlier in this chapter, this difference must have been largely due to the factor of motivation. Remember that Part D is the last part of a three-hour "grueling" examination. It is no doubt that a substantial amount of mental vigilance and effort is
required to think about what it takes to solve an essay problem and to determine which essays are the most appropriate. Yet, it is highly likely that a considerable number of the sequentially choosing students must have said to themselves, "I have been working on the test too long and I am too tired to read through all of the five essays. Let me just get it done with and go home."

![Image showing a graph comparing mean performance between sequential and selective choosing on Essay Combination 5, 6 & 8.]

Figure 7.11: Mean Performance Comparison Between Sequential and Selective Choosing on Essay Combination 5, 6 & 8

Does such a difference exist for any other essay combination? The answer is No, since none of the other nine essay combinations involves the first three essays in Part D of the 1989 AP Chemistry Exam. Let's take Essay Combination 5, 6 & 8 as an example. Essay 8 is the fourth essay and its choice involves a deliberate selection process on the part
of the student. As Figure 7.10 shows, there is virtually no difference between the sequentially and the selectively choosing examinees for this essay combination.

Chapter Summary

Through three areas of investigation, there are nine major findings. As for the curricular reasons for the differential essay choices, three factors seem to be able to account for them as well as the negative correlations between students' mean performances and the popularity of essays and essay combinations. The first factor is the familiarity with essay content (Finding 1). The majority of students tend to choose essays that are covered the most by textbooks, like Essays 5, 6, 8 and 7. Textbook coverage transforms into familiarity. Upon choice, familiarity seems to be equated with difficulty. To a large extent, the actual difficulty levels of the essays are not realized until way after they have started solving the problems. Furthermore, his factor of familiarity seems to apply to teachers' choices, too, due to the fact their average difficulty rank corresponds with students' rank of choice preferences.

The second factor that comes into play is lab-experience (Finding 2). Students tend to avoid a lab-related problem if possible, even though it is very familiar to them. This is understandable because it is much harder to actually implement a solution to a problem than to describe a phenomenon.

The third factor is the order that textbooks present different topics (Finding 3). It seems to dictate how teachers teach, and consequently, how students learn. Essay 9 on nuclear chemistry is a unique example. Whichever essay combination that involves Essay 9 is shunned by the majority of students. Altogether, only 24% of the total 1989 AP
examinee population chose the six essay combinations that involve Essay 9.

As for the relationship between students' ability levels and essay choices, four major findings have been obtained. First, on the basis of standard polytomous IRT scale, the mean abilities of the ten groups of examinees for the ten essay combinations are found to be drastically different from one another, ranging from as far as more than one standard deviation below the mean to nearly half a standard deviation above the mean (Finding 4). Note that these mean abilities measures are unique in that all the ten combinations were measured on the same scale.

Second, in spite of the different group mean abilities, it has been found that eight of the ten essay combinations were chosen by approximately equal numbers of students from the ten ability levels (Finding 5). This is an indication that most of the ten essay combinations were equally attractive to students of the ten ability levels.

Third, as a result of being equally attracted, the essay choice patterns for the ten ability groups are substantially similar, with five ability levels having virtually identical essay patterns (Finding 6).

In spite of the predominant similarity in choices, there are some unique dissimilarities, as well. Not surprisingly, these unique dissimilarities occurred in the two essay combinations that have the highest and lowest mean scores, respectively. The fourth finding of this chapter is that two factors must have attributed the highest mean score of Essay 5, 8 & 9 (Finding 7). The first factor is that Essay 9, although less likely to be covered in AP chemistry, is relatively straightforward in its solution. The second factor is that the group of students choosing this combination must have completed an extensive and solid AP chemistry curriculum to have included such unique topics as nuclear chemistry. As
a result, they were able to identify the relative straightforwardness of Essay 9 and score high points.

Finally, two important findings have been obtained regarding the performance differences between deliberate and casual choosing. First, a disproportionate number of level-1 examinees chose the first three essays in Part D, Essay 5, 6 & 7 which has the lowest mean score. It has also been found that those level-1 students who chose the three essays selectively scored significantly higher than those level-1 students who chose the three essays sequentially (Finding 8). Again, two major factors can be attributed to these phenomena. First, it is possible that those level-1 students who just answered the three essays in their sequential order might be too poor in their ability to choose. However, it is argued that it is more likely that it is the factor of motivation that has attributed to the significant performance difference, because these two level-1 groups were found not to be significantly different on Section I of the 75 MC items.

In addition to the significant performance difference for level-1 students, significant performance differences between the sequential and selective choosing were also found cross the ten levels of abilities (Finding 9). What is especially surprising is that the biggest mean difference (4.8 points) occurred at level 10. Such a finding provides a direct measure of the effect of motivation. The more capable a student is, the bigger loss he/she is to suffer if he/she loses his/her motivation to try.

So far, substantial amount of information has been uncovered from the 1989 national AP Chemistry Data on the relationship among essay choices, ability levels and performance. It is time now to look at the same relationship in light of the underlying psychological processes which are delineated in detail in the next chapter.
CHAPTER VIII

EXPLORING THE RELATIONSHIP AMONG

PSYCHOLOGICAL PROCESSES UNDERLYING CHOICES, STUDENT ABILITIES

AND PERFORMANCE

The relationship among students' essay choices, ability and performance has been investigated in the previous chapter on the basis of the national data of the 1989 AP Chemistry Exam. This chapter looks into a different dimension of the research --- the relationship among the psychological processes underlying choices, student abilities and performance. There are three major investigations in this chapter -- 1) analyses of psychological processes underlying essay choices; 2) accounting for the discrepancy between essay choice popularity and mean performance through multiple choice questions; 3) examining the relationship among student ability, extent of study, item choices and performance through multiple choice questions. By the end of this chapter, all the three research questions surrounding the second research issue of this dissertation will be answered.

Analyses of Psychological Processes Underlying Essay Choices

Because the national AP Chemistry examination data do not contain information regarding the reasons behind students' choices, an instrument called the "Advanced Placement Chemistry Survey and Test Kit" (The Kit) was administered to a wide spectrum
of chemistry students in Hawaii to collect information regarding the psychological processes behind the choices. Of the wide range of information collected, three major aspects were essential -- the extent of study on a certain item, perception of the similarity of the paired items, and a hypothetical choice. The purpose of this first investigation is to:

1. describe the performance of Hawaii students on the AP Chemistry Test Kit in order to bridge and validate their performance with that of the 1989 national examinees.

2. to report the general essay choice patterns of the Hawaii students.

3. to assess the dimensionality of the five essays through multidimensional scaling and to show the relationship between students' choices and essay dimensionality.

4. to demonstrate the relationship between the students' perceptions of essay difficulty levels and choices.

General Performance of the Hawaii Subjects on AP Chemistry Test Kit:

It has been described in Chapter IV that 20 MC items were selected from the original 75 MC items of Section I of the 1989 AP Chemistry Examination and built into the two parts of the Kit -- the Mini AP Chemistry Test and AP Chemistry Item Comparison and Performance. Altogether, 618 students responded to the 20 items. Table 8.1 below summarizes their performance on the 20 items in reference to the national norm.
Table 8.1
Descriptive Statistics of the Performance of Survey Subjects

<table>
<thead>
<tr>
<th>Student Category</th>
<th>Student Number</th>
<th>Min. Score</th>
<th>Max. Score</th>
<th>Mean</th>
<th>Std</th>
<th>KR-20</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>HI AP Chemistry</td>
<td>237</td>
<td>1</td>
<td>20</td>
<td>8.72</td>
<td>4.41</td>
<td>.83</td>
<td>2.04</td>
</tr>
<tr>
<td>Non-AP Chemistry</td>
<td>333</td>
<td>0</td>
<td>16</td>
<td>4.96</td>
<td>2.11</td>
<td>.26</td>
<td>1.92</td>
</tr>
<tr>
<td>UH Upper Division</td>
<td>15</td>
<td>6</td>
<td>20</td>
<td>12.36</td>
<td>4.09</td>
<td>.83</td>
<td>2.04</td>
</tr>
<tr>
<td>UH Chem Students</td>
<td>33</td>
<td>2</td>
<td>15</td>
<td>5.73</td>
<td>1.98</td>
<td>.52</td>
<td>1.98</td>
</tr>
<tr>
<td>Total Hawaii Sample</td>
<td>618</td>
<td>0</td>
<td>20</td>
<td>7.94</td>
<td>3.78</td>
<td>.76</td>
<td>1.97</td>
</tr>
<tr>
<td>National AP Sample</td>
<td>1000</td>
<td>0</td>
<td>20</td>
<td>8.42</td>
<td>3.66</td>
<td>.76</td>
<td>2.10</td>
</tr>
</tbody>
</table>

It can be seen that the AP chemistry students in Hawaii scored about the same as the national norm. Please note that the 237 AP students includes every student registered in the AP classes in the 12 high schools surveyed. Not all of these 237 AP students decided to take the official 1992 AP Chemistry Exam. As expected, since most of the non-AP students had completed only about one year of general chemistry study by the time they responded to this instrument, their mean performance on this instrument is only half of that of their AP counterparts.

The mean scores for the UH upper division class and the other UH chemistry classes were 12.36 and 5.73, respectively. Three logistic conditions attributed to these varied results. First, the UH students were told to respond to the Kit on a completely anonymous basis, because they did not have any obligation to fulfil this task. Second, they were not as motivated as their high school counterparts to perform well on the 20 MC items because AP
chemistry was considered much less relevant to their college study than it was to the high school students. Third, most of the UH students had barely had enough time to work on the survey, because most of the UH students had to go to another class once the allotted time was up, while their high school counterparts had more ample time. (My instruction for this survey was to take as much time as possible to work on this survey.)

As the result of these three logistic factors, the observed mean scores may not be accurate reflections of the chemistry abilities of the college students. As a matter of fact, there were significantly more unreached and omitted items in this group than in any other group. Fifty of the sixty deleted cases belong to this group.

The mean score for the total sample of 618 respondents was calculated to be 7.94, slightly below the national norm. Figure 8.1 summarizes the score distributions of the national, Hawaii Total, Hawaii AP and non-AP students.
Essay Choice Preference of Students in Hawaii:

Given that the students in Hawaii performed closely to the national norm on the average, it is interesting to ask "Did they choose the five essays in a similar way as the 1989 AP Chemistry Students did?" Note that the students in Hawaii were asked only to indicate in Part D of the Kit which of the ten essay combinations they would like to answer. Figure 8.2 reveals strikingly familiar essay choice patterns between the Hawaii and national students.
It can be seen from Figure 8.2 that based on 554 Hawaii students who responded to the essay choice question, their overall choice pattern of the ten essay combinations substantially mirrors that of the 18,462 examinees of the 1989 AP Chemistry Exam. The correlation between the two patterns is .87, significant at .05 level.

Figure 8.2: Comparison of the Overall Essay Choice Pattern of 554 Hawaii and 18,462 National Students

Three features are worth pointing out. First, like the 1989 AP Chemistry Exam population, the Hawaii students also liked Essay Combination 5, 6 & 8 the best. Essay Combination 5, 7 & 8 still remains as the second most popular combination, although the absolute number of examinees has considerably decreased.
Second, almost the same proportion of Hawaii students (13%) as the 1989 AP examinees (14%) chose Essay Combination 5, 6 & 7. Note that Essay 7 was presented as the last essay in the Kit because of the change of position in order to test the effect of positioning on Essay 9. There seems to be something inherent about Essay Combination 5, 6 & 7 that attracts some examinees.

Third, probably due to the changed position of Essay 9, there seems to be an increase in the choice frequency of Essay 9 by Hawaii students. For example, only about 3% of the 1989 examinees chose Essay Combination 5, 6 & 9, while about 11% of the Hawaii students chose it. There are also more Hawaii students choosing Essay Combinations 5, 7 & 9 and 7, 8 & 9. It may be argued that this is probably because the chemistry teachers in Hawaii have emphasized nuclear chemistry more. However, since seven of the eight teachers indicated in their responses to the "Chemistry Teacher Expert Judgment Survey" that nuclear chemistry is not their priority, the above argument can hardly be substantiated.

What can be concluded from the above findings is that there seems to be some universal regularity in the way students choose. It is highly probable that the cause for the regularity is due to the commonality of the AP chemistry textbooks. Although there is no standardized national AP Chemistry curriculum, it is the textbook commonality that dictates what, when and how will be taught, all of which transcends into the student’s varying familiarity with various subjects and eventually how they choose.

127
Dimensionality of the Five Essays:

In light of the high levels of similarity in the choice tendencies involving the five essays, one might wish to discern the cognitive dimensionality underlying the five essays, which is the purpose of the following investigation.

These five essay questions are known to involve the following areas of chemistry knowledge:

Essay 5: valence, electronic configuration, covalent bonding, molecular geometry.
Essay 6: periodic trends, stability, ionization, energy, properties of halogens, properties of alkali metals.
Essay 7: properties of metals, writing and balancing chemical equations, conservation of mass, double displacement reactions.
Essay 8: rates of reaction, physical behavior of gases, energy changes.
Essay 9: nuclear chemistry.

The above description clearly indicates that these five essay questions are different from one another, with some questions much further away from others in terms of content, such as Essay 9 on nuclear chemistry.

Because not all examinees responded to these five essays, traditional factor analysis based on students' scores cannot be applied here to ascertain the dimensionality of the five essays. In order to offset this information deficit, pair-wise similarity comparisons ("How similar are Problem I and II?") have been incorporated into Part D of the AP Chemistry Survey and Test Kit. Based on the similarity data from the Hawaii students, the dimensionality in terms of the content similarity of the five essays is revealed through multidimensional scaling in Figure 8.3.

As reflected by the content descriptions, these five essays are depicted as substantially different from one another, since the five essays are located in the four quadrants of the two
dimensional space. However, under the seemingly large differences lies a certain extent of regularity. First of all, if we examine the two halves divided by the vertical horizon, we can see that the five essays are divided into two parts -- Essays 5, 6 & 8 in the left half, Essays 7 and 9 in the right half. What characteristics do Essays 5, 6 & 8 share in common? They reflect the most frequently taught topics and constitute the common core of general chemistry.

On the other hand, requiring extensive lab experience, Essay 7 is usually more conveniently demonstrated through hands-on lab sessions than through classroom lecturing. Since not all students have equal access to laboratory conditions, Essay 7 cannot be dealt with as readily by average students. Furthermore, in Chapter VII, Essay 9 on nuclear chemistry has been shown to be the least taught because of its highly specialized application. It can be seen that Essay 9 is further away from the rest of the four essays.

Therefore, it is appropriate to designate the horizontal dimension as "Extent of Textbook Coverage". The left half contains the "Core Chemistry" topics, while the right half, the "Non-Core Chemistry" topics.

How can the vertical dimension be interpreted? From earlier analyses in Chapter VII, it is known that both Essays 6 and 7 tap into the deeper and more complex chemistry theories and structures, while Essays 5, 8 & 9 are more specific and straightforward in terms of their answers. Therefore, it is reasonable to denote the vertical dimension as "Complexity of Problems", The upper half contains the relatively "Straightforward" questions, while the lower half, the relatively "Complicated" questions.
Figure 8.3: Dimensionality of the Five Essay Questions
Relationship between Essay Dimensionality and Essay Choices:

With the dimensionality of the five essays revealed, the next question is "What is the relationship between the essay dimensionality and student choices?" More specifically, do students tend to choose essays of the same or similar dimension? The answer is Yes.

A review of the findings in Chapter IV on the general essay choice tendencies as shown in Figure 6.4 will substantiate the above affirmative answer. For example, Essay Combination 5, 6 & 8 has been found to be the most frequently chosen combination. These three essays happen to form the "Core Chemistry" half of the horizontal dimension. It is also known that any essay combination with Essays 7 & 9 is usually avoided by examinees. For example, Essays 6, 7 & 9 has been shown to be the least favored combination, basically because Essays 7 & 9 form the "Non-core Chemistry" half of the horizontal dimension.

Moreover, as shown in Figure 6.7, we know that Essay Combination 5, 8 & 9 has the highest mean score. This is because these three essays are from the "Straightforward" half of the vertical dimension. It is also known whichever combination including Essays 6 & 7 has a lower mean score. For example, Essay Combination 5, 6 & 7 has the lowest mean score of all the ten essay combinations, basically because Essays 6 & 7 were from the "Complicated" dimension of the essays.

What can be concluded from the above findings is that not only are students' choices influenced by item dimensions, but also their scores. Item dimensions can be attributed to various factors. In the case of the five essays, it has been found that the two dimensions are attributed to the order of textbook presentation and tasks involved to solve the problems. It is argued here that with most subject tests like AP tests, the predominant mode of textbook
presentation has a long-lasting effect on how students are taught, which influences how they will choose essay questions.

After confirming the relationship between the item dimensions and student choices, the next step is to look into the relationship among item difficulty, student ability and item choices. "Do students tend to choose items that they perceive as easier?" Now let's turn to the relationship between students' perceptions of item difficulty and their choices.

Relationship between Students' Perceptions of Essay Difficulty and Their Choices:

In Part D of the Kit, Hawaii students were asked to conduct pairwise comparisons of the relative difficulty levels of the five essays through the question "Which one (essay) seems easier for you?" Such pairwise comparison data were analyzed through Ranko (Dunn-Rankin, 1983), which carries out variance stable rank scaling analysis. The linear plot with scale scores from RANKO is reproduced in Figure 8.4 below:

![Figure 8.4: Perceived Order of Essay Difficulties](image)

Hawaii students ranked Essay 5 as the easiest, followed by Essays 8, 6, 7 and 9 in that order. It is both expected and surprising that this rank explains and predicts the popularity of the ten essay combinations as shown in Figure 6.4. For example, we know that the first three easiest essays, Essays 5, 6 & 8 in Figure 8.4 above, form the most popular essay combination seen in Figure 6.4, while Essays 5, 8 & 9 and 5, 6 & 7 form the second
and third most popular essay combinations. The three essays that form the least popular Essay Combination 6, 7 & 9 in Figure 6.4, turn out to be the last three essays in the above scale.

Table 8.2 further shows that all the five essays were perceived significantly different from one another even at .01 level. This finding reinforces the dimensions of the five essay - these five essays not only address different content areas but also require different levels of competence to solve them.

<table>
<thead>
<tr>
<th>Table 8.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank Differences</td>
</tr>
<tr>
<td>Essays</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>9</td>
</tr>
</tbody>
</table>

Note: The critical differences are 137 at .05 level and 163 at .01 level.

**Accounting for the Discrepancy between Choice Popularity and Mean Performance in Light of the Psychological Processes Underlying Multiple Choice Questions**

On the top of the list of the research questions regarding the relationship among student ability, item difficulty and choices is "Do students choose to answer the items which they consider as easier and closer to each other in dimensionality, and perform more favorably on such choices?" The answer to the first half of the question is Yes since it has
been found that students did tend to choose more frequently the essays they perceived as easier. However, the answer to the second half of the question seems to be No due to the negative correlation between the popularity of the essay combinations and the mean scores on them. Is it possible, however, that the majority of the 1989 examinees were not able to identify between what is 'really' easy or hard?"

It is suspected here that the negative correlation between essay popularity and essay performance might have been caused by some artificial reason — Essays 5, 6, 7 & 8 might have been constructed to be much more difficult and complicated than the students had expected than Essay 9. Unlike MC items which are relatively equal in terms of the tasks involved and can be scored on a relatively absolute criterion of being either correct or wrong, essay items may have varying degrees of scoring stringency. For example, some essay topics, such as women’s rights to abortion, may not have a universal criterion for scoring. It is hypothesized here that it might have been a combination of factors that have caused the observed negative correlation between the popularity of essay choices and student mean performance.

How can the relationship between students’ ability to discriminate items of different difficulties and their performance be assessed in a less confounded way? The answer may be found by utilizing responses to MC items. Comparison IV in Part C of the Kit (See Appendix A) has been specially designed for this purpose. Comparison IV has four MC items — Items IV1, IV2, IV3 and IV4. The topics represented by these four items range from molarity for Item IV1, electrochemistry for Item IV2, acid-base equilibrium for Item IV2, and energy and phase changes for Item IV4. In addition to being different in content,
these four items are also very different in difficulty ranging from being very easy (-1.48 for Item IV1) to being very hard (2.56 for Item IV2).

The reason for the four MC items to have four levels of difficulty and to cover four different topics is to mirror the five essays in Part D of the 1989 AP Chemistry Exam which have been shown to possess very different difficulty levels and content areas. This second investigation of this chapter is intended to further confirm whether or not students would perform better on the items they choose, after the possible confounding of different operating entities is excluded by the use of MC items. There are four parts in this investigation:

1. investigation of the dimensionality of the four MC items;
2. assessment of the student perception of item difficulty;
3. comparison of the item difficulties as reflected by IRT, student perception and mean scores;
4. evaluation of the relationship between item choice combination, extent of study, student ability and mean performance.

**Dimensionality of the Four MC Items:**

Like the dimensionality investigation of the five essays, the dimensionality of the four MC items was analyzed through the similarity data using KYST and is displayed in Figure 8.5 (stress = 0.0006) (Dunn-Rankin, 1988). As expected, none of the four items cluster closely with each other. However Item IV1 on molarity and Item IV3 on acid and base share the same quadrant because they both involve stoichiometry, the computational aspect of chemistry. Opposite to Items IV1 and IV3 is Item IV4 on energy and phase change, which mainly involves the identification of the physical states of matter. Therefore, the horizontal dimension can be labelled as "More Chemistry vs. More Physics". It is interesting to note
that Item IV2 on electrochemistry is located at the midpoint of this axis. It is known that electrolysis, in many ways, reflects the unique combination of — physics and chemistry. As for the vertical dimension of the graph, since Item IV2 on electrolysis involves deduction of the effects of the principles in both chemistry and physics, the lower half of the vertical dimension can be called "Deduction", while the upper half, "Non-deduction", which combines both computation and identification.

Figure 8.5
Dimensionality of the Four MC Items in Comparison IV
Perceptions of MC Item Difficulty Levels:

Again, based on the pair-wise comparison data, the difficulty levels of MC items as perceived by Hawaii students are revealed in a linear mode by RANKO (Dunn-Rankin, 1988) in Figure 8.6:

Very Hard

<table>
<thead>
<tr>
<th>Item</th>
<th>Perceived Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV1</td>
<td>Easy</td>
</tr>
<tr>
<td>IV2</td>
<td>Medium</td>
</tr>
<tr>
<td>IV3</td>
<td>Medium</td>
</tr>
<tr>
<td>IV4</td>
<td>Medium</td>
</tr>
<tr>
<td>IV1</td>
<td>Very Easy</td>
</tr>
</tbody>
</table>

0 ... 30-----*-----40-----*-----50-----*-----60-----*-----70-----*-----80-----100

Figure 8.6: Perceived RANKO of MC Item Difficulties

Item IV1 on molarity was perceived the easiest, followed by Items IV4 on physical states of matter, and IV3 on acid and base. Item IV2 on electrolysis was perceived as the most difficult. All items except Items IV2 and IV3 are significantly different from one another.

Comparisons of Item Difficulties as Reflected by IRT, Student Perception and Mean Scores:

Does the above order of perceived item difficulties agree with the actual difficulty estimates as measured by IRT? Did the students more frequently choose the items they perceived easier? Did the students perform better on the items that they considered easier and chose more frequently? A look at Table 8.3 will show that the answers to all the three questions are Yes. Three conclusions may be in order.
First, there is a perfect rank correlation between the item difficulty estimates as measured by IRT and perceived by students. This is the second time to confirm this finding after it was found true with essay questions in this chapter. This verification signifies that student perceptions are largely accurate in terms of judging item difficulties, at least, as long as the items are fairly different from one another.

Table 8.3
Comparisons of IRT Item Difficulty Estimates, RANKO Scale, Choice Frequencies and Mean Scores

<table>
<thead>
<tr>
<th>Item</th>
<th>IV1</th>
<th>IV2</th>
<th>IV3</th>
<th>IV4</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRT Estimates</td>
<td>-1.48 (1)</td>
<td>2.56 (4)</td>
<td>1.09 (3)</td>
<td>0.36 (2)</td>
</tr>
<tr>
<td>RANK Scale</td>
<td>70 (1)</td>
<td>36 (4)</td>
<td>39 (3)</td>
<td>55 (2)</td>
</tr>
<tr>
<td>Choice Frequencies</td>
<td>549 (1)</td>
<td>338 (4)</td>
<td>443 (3)</td>
<td>468 (2)</td>
</tr>
<tr>
<td>Mean Scores</td>
<td>0.56 (1)</td>
<td>0.19 (4)</td>
<td>0.31 (3)</td>
<td>0.20 (2)</td>
</tr>
</tbody>
</table>

Note: the numbers in the parentheses indicate ranks within each category.

Second, what is more important is that the IRT item difficulties and student perceptions are also perfectly correlated with item choice frequencies. This reinforces the earlier finding that students do tend to choose more frequently the items that they consider easier.

Third, what is most important is the IRT item difficulties, students perceptions and choices are also perfectly correlated with the mean scores on the four items. This finding has clarified the mystery originated from the previous chapters concerning the negative
relationship between essay choice popularity and mean performance. What is manifested is that students do choose for a reason. It reinforces the hypothesis that the negative correlation might have been caused by some extraneous factor, most likely the artificial efforts to make the essay questions on the "core" subjects much harder than the essays on the "uncore" subjects.

Someone may argue that there is a reason to such artificial manipulation, since there is less textbook coverage on the "Non-core" subjects, like nuclear chemistry. This author has no problem with such an argument. However, what this author does have problems with is the practice of letting examinees choose among the drastically different items without adjustment for their difficulties. It is the lack of adjustment that will systematically advantage some groups of examinees over others, and consequently, make the test scores unfair. In the case of the 1989 AP Chemistry Exam, more than half of the 18,435 examinees might have been affected, since more than 50% percent of the total population chose the essay combinations that are made up of the "core" chemistry subjects.

What can be concluded from the findings so far is that the student-perceived relative difficulty order of the five essays is not the "real" difficulty rank in terms of how difficult it is to obtain a certain score, but a reflection of their relative familiarity with each of the five essay topics. Such familiarity, in turn, results in the popularity rank of both the five essays and the ten essay combinations. If one recalls, these two popularity ranks correlate negatively with students' mean performance on them. The reason for the negative correlation is largely due to the fact that the familiar "core" topics of chemistry in the 1989 AP Chemistry Exam were more demanding or "involved" than the less familiar "Non-core" topics, according to the teachers' evaluation. Therefore, there is an artificial element to the
nature of the five essays. The only way to find out if students’ perceptions of item
difficulties truly match the "real" difficulty levels of items is through the MC items. This
is the main research objective of the next investigation.

Accounting for the Relationships among Student Ability, Extent of Study, Item Choices
and Performance through Multiple Choice Questions

Students’ perceptions of item difficulties and dimensionality through uni- and multi-
dimensional scaling have been shown to be considerably consistent with IRT estimates of
item difficulties in terms of their rank, and highly revealing in conveying the various content
areas and instructional properties of both essay and MC items. It has also been shown that
students tended to choose the items they perceived as easier and performed better on them.
The next question is "How fine-tuned are these perceptions when items with very similar
difficulties are evaluated?" Remember all the essay and MC items used so far are
substantially different from one another. In order to generalize this methodology, it is
necessary to investigate the extreme possibilities -- "To what extent can students distinguish
items of very close item difficulties? To what extent do students’ choices of items affect
their performance?"

In addition, some further research questions have to be investigated regarding the
effect of choices on performance. The second question to be asked is "To what extent do
choices enhance scores?" Since choices have been discussed so extensively up to now in this
dissertation, the reader may feel that choice determines everything. The author feels that it
is important to elucidate the role of choice in reference to item difficulty, student ability and performance.

Third, what is the relationship among students' extent of study, choice, performance and item difficulty? Do students tend to choose items that they study more? The effects of extent of study on choice and performance is an non-ignorable issue because of its relationship to instruction.

The third investigation of this chapter will investigate these three research questions. The data that provide relevant information are the four comparisons in Part C of the Kit.

As described in Chapter IV, Comparisons I, II and III have two paired MC items each. The paired MC items in Comparisons I, II & III were specially selected with differing difficulty levels. Specifically, Comparison I has one very easy and one very hard item. Comparison II has two middle-difficulty items; and, Comparison III has two very hard items. The reason for such arrangements is to assess the extent to which students of different ability levels can accurately distinguish between the paired items in these comparisons.

In addition to the special arrangements mentioned above, three additional survey questions were composed to help solicit more information. The first question, "To what extent have you studied a problem like ...?", was asked to see if students would choose the items that they have studied more extensively. The second question, "Which of the above two questions seems easier for you?", was to look into the extent that students of various ability levels can accurately identify the easier of the two paired items. Third, "If you were allowed to choose between these two questions, which one would you choose?" was to confirm whether or not students would indeed choose to answer the items they identified as easier. In order to appraise the stability of students' decisions about their choices, a final
question was asked, "If you would like to choose between these two questions again after you have tried to answer them, which one would you choose now?" It is hoped that the relationship among student ability, extent of study, choices, performance and item difficulties can be sorted out in a controlled manner.

The Mean Ability of the Four Groups of Hawaii Students

As reported in Table 9.1, the mean scores for UH upper Division Class, UH chemistry students, Hawaii AP and Non-AP students on the twenty MC items in the Kit were 12.36, 5.73, 8.72, and 4.96 points, respectively. Due to the relative small numbers of the UH upper and lower division students (15 vs. 33), these two groups are combined into one group called "College Students" for the analyses to follow. Table 8.4 below summarizes the IRT mean ability estimates of Hawaii AP, College and non-AP students.

Table 8.4
Mean Group Ability Estimates of Hawaii AP, College and Non-AP Chemistry Students

<table>
<thead>
<tr>
<th>Student Category</th>
<th>IRT Mean Ability Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP Chemistry</td>
<td>.88</td>
</tr>
<tr>
<td>College</td>
<td>.51</td>
</tr>
<tr>
<td>Non-AP Chemistry</td>
<td>.00</td>
</tr>
</tbody>
</table>

The purpose of reporting the above IRT mean ability estimates is to provide a reference framework for the later analyses to compare the mean group ability with item
difficulties. IRT ability estimates are in the same scale as IRT item difficulty estimates. With a common scale, it is possible to directly assess the likelihood of a group of students to correctly answer an item of a certain difficulty. For example, an item with a difficulty estimate of -2.00 should be considered easy for most students of the three groups of Hawaii students, because it is way below the mean group ability of even the non-AP students. Another note is worthwhile here. The reason that the mean ability estimate of the non-AP students is 0 is because this group was used as the baseline reference group, based on which the mean abilities of the AP and College students were calculated.

Student Judgment, Performance, Choice and Choice Change in Comparison I:

Comparison I has two MC items. The first item represents one of the most basic topics in chemistry -- oxidation and reduction, while the second item, a somewhat more advanced topic, solution and molarity. Based on the 1989 national AP Chemistry data, these two items are drastically different in their difficulty. The first has a difficulty parameter of -2.63, way below the mean group ability of non-AP students, while the second item, fairly hard, with a difficulty parameter of 1.59, way above the mean group ability of AP students. Table 8.5 below summarizes students' responses to the survey questions regarding the two items.

Four tendencies can be observed from the responses of the three groups of Hawaii AP, College and non-AP chemistry students in Table 8.5. First, they all had studied Item 11 more than Item 12, because Item 11 is a more basic topic. Second, probably because of more study on Item 11, the majority of the three groups considered Item 11 as easier. Third,
probably because it was studied more and easier, the majority of them chose Item I1 as their preferred choice item. Fourth, the majority of the three groups of students (ranging from 93% for AP students to 73% of non-AP students) answered the first item correctly. Please note, less than half of the students got the second item right, because it is much harder. The final finding is the most important one. Probably because of their relative success in performing on the first item, 5% to 10% of students switched to Item I1 as their preferred item after they had attempted these two items.

Table 8.5

Relationship among Extent of Study, Judgment of Item Difficulty, Choices and Choice Change

(Comparison I: Item Difficulty Estimates -- Item I1 = -2.63; Item I2 = 1.59)

<table>
<thead>
<tr>
<th>Student Type</th>
<th>N</th>
<th>Extent of Study</th>
<th>Which is Easier? (%)</th>
<th>First Choice(%)</th>
<th>% of Getting Item Right</th>
<th>Second Choice Change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I1</td>
<td>I2</td>
<td>I1</td>
<td>I2</td>
<td>I1</td>
<td>I2</td>
</tr>
<tr>
<td>AP</td>
<td>237</td>
<td>2.85*</td>
<td>2.367</td>
<td>81*</td>
<td>81*</td>
<td>79*</td>
</tr>
<tr>
<td>College</td>
<td>48</td>
<td>2.88*</td>
<td>2.208</td>
<td>73*</td>
<td>73*</td>
<td>67*</td>
</tr>
<tr>
<td>Non-AP</td>
<td>333</td>
<td>2.07*</td>
<td>1.712</td>
<td>72*</td>
<td>72*</td>
<td>59*</td>
</tr>
</tbody>
</table>

Note that the "**" marks the bigger number in each pair.

What can be concluded from the above findings is that students do choose the items they study more and consider easier. This is especially true when two items are drastically different, with one being very easy and the other, very hard. In this case, the majority of students of all ability levels can and will make fairly accurate judgment and consistent
choices. Varying percentages of students of various ability levels would adjust their choices if they make an inappropriate one in the first place.

**Student Judgment, Performance, Choice and Choice Change in Comparison II:**

The two items in Comparison II are on solution chemistry and physical behavior of gases (Charles’ Law). Their difficulty estimates based on the 1989 national AP chemistry data indicate that they are of medium difficulty, -0.07 for Item III and .089 for item II2. Item III is slightly easier than Item II2. Note that these two items are very close to each other in difficulty, which makes it hard for students to distinguish between them. Although the difficulty levels of these two items are below the mean ability levels of AP and College students, they are close to the mean ability level of non-AP students. How did the Hawaii students respond to these two items?

Table 8.6

<table>
<thead>
<tr>
<th>Student N</th>
<th>Extent of Study</th>
<th>Which is Easier (%)</th>
<th>First Choice (%)</th>
<th>% of Getting Item Right</th>
<th>Second Choice</th>
<th>Choice Change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Item III</td>
<td>Item II2</td>
<td>Item III</td>
<td>Item II2</td>
<td>Item III</td>
<td>Item II2</td>
</tr>
<tr>
<td>AP 237</td>
<td>1.84</td>
<td>2.60*</td>
<td>38</td>
<td>62*</td>
<td>38</td>
<td>62*</td>
</tr>
<tr>
<td>College 48</td>
<td>1.75</td>
<td>2.31*</td>
<td>38</td>
<td>62*</td>
<td>40</td>
<td>60*</td>
</tr>
<tr>
<td>Non-AP 333</td>
<td>1.40</td>
<td>1.77*</td>
<td>59*</td>
<td>41</td>
<td>39</td>
<td>61*</td>
</tr>
</tbody>
</table>

Note the "*" marks the bigger number in each pair.
Four tendencies can be observed from Table 8.6 about AP and college students. First, AP and college students had studied Item II2 substantially more than Item III, even though Item II is measured as easier on the national data. Sample fluctuation is expected, especially when the two items are so close to each other. Second, because of more study on Item II2, 62% of both the AP and college students indicated Item II2 as their easier item. Third, as the result of their more study, slightly more students got Item II2 right. Remember that these two items are similar in difficulty. Finally, 5% AP vs. 14% College students changed their choices in favor of Item III, an indication that they might have realized that these two items were not as different as they originally judged.

How did the non-AP students do? Because the difficulty levels of the two items are within the vicinity of the mean ability of this group, their responses were fairly inconsistent. First, although they indicated that they had studied Item III a little more than Item II2, 59% of the non-AP students judged Item III as the easier one. This is the first inconsistency. Second, instead of judging Item III as the easier one, 61% of the non-AP chose Item II2 as their first choice. This is the second inconsistency. Third, in spite of the "flip-flops" of their decisions, 34% of the non-AP students answered Item III correctly vs. only 21% for Item II2. Finally, 30% of this group changed their mind to switch to Item III, the easier item, as their preferred choice.

Three conclusions can be made from the above observations. First, due to the fact the two items are close to each other in their difficulty on the national data, the number of students in Hawaii who answered the two items right at each ability level were almost equal, although more of the more advanced students answered them right.
Second, because the mean ability of the non-AP students was similar to the difficulty levels of the two items (i.e., average), substantial amount of inconsistency has occurred in their decisions on their judgment and choices.

Third, and more importantly, 30% of the non-AP group, in spite of the "flip-flops" in their decisions, could realize through actually working on the problems that they had made inappropriate choices in the first place, and changed their choices, after they attempted the two items. Remember that 30% of the 333 non-AP students is equal to a substantial number of 109 students. This finding can suggest that there might have been a substantial number of students who had wished that they could switch essays during the 1989 AP Chemistry Exam. However, switching on an essay item is far more time-consuming and much less practical than switching on a MC item. This finding further strengthens the importance of making a correct choice in the first place on an essay test, or students could be penalized by inappropriate choices.

**Student Judgment, Performance, Choice and Choice Change in Comparison III:**

Comparison III has two very hard items, way above the mean group ability of Hawaii AP chemistry students. Item III1 has a difficulty level of 2.28, while Item III2 has a difficulty level of 2.09. In terms of content, Item III1 is on water solution, while Item III2, on ions in water solution. These two items were purposefully chosen for their content similarity. In face of the two extremely difficult but similar items, how did the Hawaii students respond? As expected, there is increased inconsistency across the three groups.
Three observations can be observed. First, although all the three groups of students had studied the first item substantially more than the second item, only AP students predominantly considered the first item as easier (67%), while half of the other groups (51%) considered the second item easier. This is the first choice inconsistency in the AP group. Second, 64% of the AP students chose the first item as their preferred choice, while still about half of both the college and non-AP students chose the second. Third, more students of all the three groups answered the second item right than the first item. Remember that the second item is less studied than the first one, but was measured as easier on the national AP data. Finally, after their relative success on Item III2, 9% of AP students adjusted their choice for the easier question, while virtually no student in the college and non-AP groups changed.

Table 8.7

Relationship among Extent of Study, Judgment of Item Difficulty, Choices and Choice Change

(Comparison III Item Difficulty Estimates -- Item III1 = 2.28; Item III2 = 2.09)

<table>
<thead>
<tr>
<th>Student Type</th>
<th>Extent of Study</th>
<th>Which is Easier (%)</th>
<th>First Choice (%)</th>
<th>% of Getting Item Right</th>
<th>Second Choice Change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Item III1</td>
<td>Item III2</td>
<td>Item III1</td>
<td>Item III2</td>
<td>Item III1</td>
</tr>
<tr>
<td>AP</td>
<td>2.00*</td>
<td>1.40</td>
<td>67*</td>
<td>33</td>
<td>64*</td>
</tr>
<tr>
<td>College</td>
<td>2.11*</td>
<td>1.81</td>
<td>49</td>
<td>51*</td>
<td>49</td>
</tr>
<tr>
<td>Non-AP</td>
<td>1.44*</td>
<td>1.07</td>
<td>49</td>
<td>51*</td>
<td>48</td>
</tr>
</tbody>
</table>

Note that the "*" marks the bigger number in each pair.
Two conclusions can be made from Table 8.7 about Comparison III. First, when items are very close to each other in their difficulty, students tend to hesitate about which one to choose and decision inconsistency is bound to happen. The eventual outcome is that all the groups tended to be equally divided about which one to choose. This conforms to the finding from Comparison II.

Second, when the mean ability of a group is close or slightly below the difficulty of the items, this group tends to have more decision inconsistencies. The AP students in Comparison III and the non-AP students in Comparison II are two examples. Two reasons can be offered for this phenomenon. On one hand, these students are less certain, because they have about 50% chance to get the items right. On the other hand, it may be because they tend to experiment more and would like to explore the possibilities, because they still have some ability to try. On the contrary, there is very little movement or inconsistency in the groups whose ability level is either way above or below the item difficulty level. AP group in Comparison II and both the college and non-AP students are examples of this phenomenon.

What can be concluded in general from the above three comparisons is that the extent of study seems to play a major role in determining which item is to be considered as easier and which is to be chosen eventually. Students tend to choose items which they feel they have studied more, although such extent does not agree with the real item difficulty parameters. Choices tend to agree more with the extent of study rather than the item difficulty parameter. However, the item difficulties estimated on the 1989 AP Chemistry Exam have proven to be correct in terms of the final outcome percentages of students who
would answer them correctly. No matter how students choose in the first place, the items with lower difficulty indices always are always answered correctly by more students.

Summarizing the Relationship among Student Performance, Ability, Extent of Study, and Item Choice Through Logistic Regression Models

In face of the complex relationship of students' extent of study, choices, item difficulty, and students' performance, one may wish to have a statistical model that can predict and explain the effects of choice and extent of study on performance. It is the purpose of this section to find and explain such a model.

Through CATMOD procedure of SAS6.03 (SAS, 1988), six logistic regresional analyses have been carried out to assess the relative contributions of students' extent of study, item choice and their chemistry ability to their actual performance outcome on each of the six items. Note that the dependent variable, SCORE, is the dichotomous vector of the students' correct (1) or incorrect (0) responses to each item, while the independent variables are the student ABILITY, EXTENT of study and item CHOICE. The variable of student ABILITY is defined as students' scores on the ten MC questions in Part B of the survey Kit. Table 8.8 on the next page summarizes the regression results.

Three conclusions can be drawn from Table 8.8. First, the most important variable to predict students' success on each of the six items is student ABILITY, because it is the only variable that has significant $X^2$ values for all the six items. This is expected because it takes substantial chemistry competence to answer the items correctly, especially when they are very hard.

150
Second, the variable of CHOICE is a significant predictor only for the very easy Item II. There are two important implications from this finding. On one hand, it suggests that anyone who cannot even identify an extremely easy item from a hard one, it is hardly worthwhile for him/her to attempt the item, let alone succeed in getting it right. On the other hand, it warns that if a student fails to select this easy item, he/she is likely to pay a price. This is especially true when students are supposed to choose among essays which are often worth multiple points. Of course, for the items that are substantially hard and close to each other in difficulty, it takes more than just students’ choosing to answer them right.

Third, the variable of EXTENT of study is only useful in predicting up to middle-level items (up to II2). What is signified by this finding is that when confronted with extremely difficult items, the extent of study of most students may not be sufficient enough to conquer them. What counts is the true competence or "innate or internalized" ability of students that has been accumulated through years of chemistry study.

No one should be confused in considering the above findings of the relatively small effects of choice on performance as evidence to negate the impact of choices on test scores. This is particularly true when students are allowed to choose items of drastically different difficulties, like the essays in Part D in the 1989 AP Chemistry Exam. Note that CHOICE was shown to have an insignificant role in predicting the performance ONLY on items of SIMILAR difficulties. Remember that CHOICE was shown to be a highly SIGNIFICANT predictor of the performance on items of different difficulties in Comparison I. It is known already that the five essays in Part D do have different difficulties. Therefore, the findings in this section should serve as validation of what has been found so far in this dissertation.
Table 8.8
Summary of Logistic Regressional Models

<table>
<thead>
<tr>
<th>Paired Items</th>
<th>Diff</th>
<th>Effects</th>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>-2.63</td>
<td>INTERCEPT</td>
<td>1</td>
<td>-1.0756</td>
<td>0.2353</td>
<td>20.89</td>
<td>0.0000*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ABILITY</td>
<td>2</td>
<td>0.1515</td>
<td>0.0486</td>
<td>9.72</td>
<td>0.0018*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Very</td>
<td>3</td>
<td>0.4622</td>
<td>0.2200</td>
<td>4.41</td>
<td>0.0357*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Easy</td>
<td>4</td>
<td>0.4235</td>
<td>0.0803</td>
<td>27.80</td>
<td>0.0000*</td>
</tr>
<tr>
<td>I2</td>
<td>1.59</td>
<td>INTERCEPT</td>
<td>1</td>
<td>-3.3223</td>
<td>0.3398</td>
<td>95.60</td>
<td>0.0000*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ABILITY</td>
<td>2</td>
<td>0.3037</td>
<td>0.0502</td>
<td>36.68</td>
<td>0.0000*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hard</td>
<td>3</td>
<td>0.2786</td>
<td>0.2787</td>
<td>1.00</td>
<td>0.3174</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EXTENT</td>
<td>4</td>
<td>0.1332</td>
<td>0.1127</td>
<td>1.40</td>
<td>0.2373</td>
</tr>
<tr>
<td>II1</td>
<td>-0.07</td>
<td>INTERCEPT</td>
<td>1</td>
<td>-1.5028</td>
<td>0.2289</td>
<td>43.11</td>
<td>0.0000*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ABILITY</td>
<td>2</td>
<td>0.2385</td>
<td>0.0429</td>
<td>30.91</td>
<td>0.0000*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medium</td>
<td>3</td>
<td>-0.2075</td>
<td>0.1867</td>
<td>1.23</td>
<td>0.2665</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EXTENT</td>
<td>4</td>
<td>0.2870</td>
<td>0.0942</td>
<td>9.29</td>
<td>0.0023*</td>
</tr>
<tr>
<td>II2</td>
<td>0.089</td>
<td>INTERCEPT</td>
<td>1</td>
<td>-2.0317</td>
<td>0.2298</td>
<td>78.19</td>
<td>0.0000*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ABILITY</td>
<td>2</td>
<td>0.2855</td>
<td>0.0465</td>
<td>37.74</td>
<td>0.0000*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medium</td>
<td>3</td>
<td>0.3050</td>
<td>0.1993</td>
<td>2.34</td>
<td>0.1259</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EXTENT</td>
<td>4</td>
<td>0.1737</td>
<td>0.0846</td>
<td>4.22</td>
<td>0.0400*</td>
</tr>
<tr>
<td>III1</td>
<td>2.28</td>
<td>INTERCEPT</td>
<td>1</td>
<td>-2.2733</td>
<td>0.2782</td>
<td>66.75</td>
<td>0.0000*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ABILITY</td>
<td>2</td>
<td>0.1870</td>
<td>0.0474</td>
<td>15.57</td>
<td>0.0001*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Very</td>
<td>3</td>
<td>-0.2882</td>
<td>0.2522</td>
<td>1.31</td>
<td>0.2532</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hard</td>
<td>4</td>
<td>0.0653</td>
<td>0.1064</td>
<td>0.38</td>
<td>0.5392</td>
</tr>
<tr>
<td>III2</td>
<td>2.09</td>
<td>INTERCEPT</td>
<td>1</td>
<td>-2.2524</td>
<td>0.2567</td>
<td>77.00</td>
<td>0.0000*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ABILITY</td>
<td>2</td>
<td>0.2188</td>
<td>0.0442</td>
<td>24.50</td>
<td>0.0000*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Very</td>
<td>3</td>
<td>-0.0340</td>
<td>0.2362</td>
<td>0.02</td>
<td>0.8855</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hard</td>
<td>4</td>
<td>0.0759</td>
<td>0.1029</td>
<td>0.54</td>
<td>0.4605</td>
</tr>
</tbody>
</table>

Note that the "*" marks the significant predictors

With the findings on how students choose between two items and what accounts for the success in performing on two items, a curious reader may ask, "How do students choose among more than two items and perform on them? Do students tend to choose items that are..."
close to each other in content dimensionality? Do students tend to choose clusters of items that are relatively easier in difficulty and studied more? Do they perform better on the items they choose?" These are the questions to be answered in the next section.

Choosing among More Than Two Items:

In order to investigate how students choose among more than two items, it is necessary to turn to the four MC items in Comparison IV again, but to focus on different aspects of information from what was used previously. Since students were asked to choose three out of four items, there exist four combinations \( \binom{4}{3} \): 1, 2 & 3; 1, 2 & 4; 1, 3 & 4; 2, 3 & 4. Table 8.9 summarizes the item difficulties (in parentheses), extent of study on the items, frequencies of the item combination choices and performance across the three groups of AP, college and non-AP students.

In terms of extent of study, it is no surprise that all groups agreed that Item IV1, the easiest item as measured by the national data, had been studied the most. Although there are some differences across the three groups, Item IV2, IV3 and IV4 are equal to one another in their extent of study in terms of the total Hawaii sample.

Did students choose the items whose content they had studied the most? The answer is Yes. An examination of the second portion of Table 8.9 reveals that the most frequent item combination for each group of students consists of the items whose content had been studied the most by that group. For example, the most frequent combination for the AP group is Items IV1, IV2 & IV3. The content of these three items was the most studied ones among the four items. Choosing items with most studied content is also true of the other
groups of the college and non-AP groups. What is more striking is that very few students changed their mind at their second chance to choose. This is probably because these items are very different from one another in their difficulty, and therefore, easy to distinguish.

Did the students perform the best on the combination that they chose the most? The answer is Yes for most students, as shown in the lower portion of Table 8.9. Except for non-AP students, all the other students performed the best on Item Combination 1, 3 & 4, the most frequently chosen combination. Remember that these three items are the easiest items as measured on the national data. The reason that the non-AP students do not have the highest mean on Item Combination 1, 3 & 4 is probably due to the fact that their mean ability is too low for them to perform consistently on this combination.
Table 8.9
Summary Statistics for Comparison IV

### Average Extent of Study

<table>
<thead>
<tr>
<th>ITEMS</th>
<th>Item IV1</th>
<th>Item IV2</th>
<th>Item IV3</th>
<th>Item IV4</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIFFICULTY ($b=-1.48$)</td>
<td>($b=2.56$)</td>
<td>($b=0.36$)</td>
<td>($b=1.09$)</td>
<td></td>
</tr>
<tr>
<td>Hawaii Total</td>
<td>2.33*</td>
<td>1.71</td>
<td>1.71</td>
<td>1.69</td>
</tr>
<tr>
<td>AP</td>
<td>2.86*</td>
<td>2.06</td>
<td>2.17</td>
<td>1.79</td>
</tr>
<tr>
<td>College</td>
<td>2.76*</td>
<td>1.84</td>
<td>2.18</td>
<td>1.93</td>
</tr>
<tr>
<td>Non-AP</td>
<td>1.89*</td>
<td>1.44</td>
<td>1.31</td>
<td>1.59</td>
</tr>
</tbody>
</table>

### Item Combinations & Choice Frequencies

<table>
<thead>
<tr>
<th>First Choice</th>
<th>Second Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3</td>
<td>1 2 4</td>
</tr>
<tr>
<td>Hawaii Total</td>
<td>24%</td>
</tr>
<tr>
<td>AP</td>
<td>34%*</td>
</tr>
<tr>
<td>College</td>
<td>23%</td>
</tr>
<tr>
<td>Non-AP</td>
<td>20%</td>
</tr>
</tbody>
</table>

### Mean Performance

<table>
<thead>
<tr>
<th>Item Combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3</td>
</tr>
<tr>
<td>Hawaii Total</td>
</tr>
<tr>
<td>AP</td>
</tr>
<tr>
<td>College</td>
</tr>
<tr>
<td>Non-AP</td>
</tr>
</tbody>
</table>

Note: 1. The numbers in parentheses are IRT item difficulty estimates.
2. The "*" marks the biggest number in the group.
Chapter Summary

This chapter has confirmed some of the most important questions surrounding the relationships among students' psychological processes underlying choices, ability, performance, extent of study and item difficulties in a relative controlled fashion. Altogether, six conclusions can be made from the three areas of investigations in this chapter. The first important finding is that there seems to be a high level of generality in the essay choice tendencies, since the Hawaii students almost replicated the essay choice patterns of the 1989 AP chemistry students (Finding 1). Such a high level of generality is attributed to the highly similar mode of content presentation predominant in major AP chemistry textbooks.

Second, the two dimensions of the essays revealed through multidimensional scaling are highly meaningful, reflecting a close match between the textbook coverage and the choice preference of students (Finding 2). It has been found that students did tend to choose items from the same or similar dimension.

The third important finding is that students' perceptions of essay difficulty levels, when transformed into a rank of difficulty, are able to accurately account for and predict the popularity of choices of essay combinations (Finding 3). Furthermore, students' perceptions of item difficulties are sufficiently compatible with psychometrically assessed item difficulties, as long as the items are not too close to each other in their difficulties.

Based on MC questions, this chapter has also confirmed that there is a perfect correlation among item difficulties as assessed by IRT methodology, student perceptions, choice frequencies and their mean performance (Finding 4). Such a finding is important
because it points a finger at the unreasonable negative correlation between the essay popularity and performance. It has been suggested that the negative correlation might have been caused by the deliberate efforts on the part of test maker or item writer to make the "core" subject items unusually hard. Normally, this practice would not have been a problem if all the essay items were to be taken by all examinees. However, in the case of AP Chemistry exams, it causes a tremendous equity problem since the essays of unequal difficulties are to be chosen and treated as though they were equally difficult. The raw effects of such a practice have already been demonstrated in Chapters VI and VII. For example, it has been shown in Chapter VII that as many as five points of difference could result from choosing one essay combination over another.

As for the relative roles of ability, choice and extent of study on performance, it has been found that it is the student true ability that significantly and consistently determines students' success on items (Finding 5). This explains why the most popular IRT models are still unidimensionally based on students' innate ability in accounting for test performance.

As for the role of extent of study, it seems to largely dictate students' judgment of item difficulty and their choices (Finding 6). It is highly correlated with students' claimed familiarity with items.

Furthermore, no matter how students choose in the first place, substantial numbers of students are able to adjust their choices after working through the items (Finding 7). However, the amount of choice inconsistency increases with the item difficulties. Most inconsistency occurs in the group whose mean ability is closest to the item difficulty (Finding 8).
CHAPTER IX

EQUATING SCORES ON STUDENT SELECTED ITEMS

With the understanding of the tendencies, reasons and consequences of students' choices, an investigation of how students' scores on chosen items can be best adjusted for their varying item difficulties is in order. Before going to details, it is necessary to clarify two differences between the conventional equating practice and equating on the scores on the items student purposefully choose. Normally, test items are selected by test developers according to certain criteria that are independent of students' knowledge. Test score equating, in such situations, is a statistical adjustment process used to convert the scores on one test to the metric of another in order to equalize the differences in item difficulties. For example, Test of English As A Foreign Language (TOEFL) is administered, at least, four times a year. It is a routine practice not only to equate the four TOEFL versions within each year, but also over the previous years in order to maintain a relative comparability of scores across all TOEFL administrations.

However, as described in Chapter I, equating the scores on the subsets of items students purposefully choose is different in, at least, two respects. First, it has been confirmed in this dissertation that students tend to select items which they feel advantageous to them, The basic assumption of missing-at-randomness for equating has been violated. It is not known to what extent this violation can affect the accuracy of equating if a traditional equating method, like IRT equating, is used. It is the primary purpose of this chapter to assess the effect of such a violation.
The second difference in equating choice scores is to compensate for the differences in the difficulties of the items not shared by the chosen subsets of items. It is known from this dissertation, Essay Combination 5, 6 & 7 is different from Essay Combination 5, 6 & 8 because Essays 7 is more difficult than Essay 8. The purpose of equating the scores on the two combinations is to "expand" the scores on the Essay Combination 5, 6 & 7 towards the scores on Essay Combination 5, 6 & 8, while "scaling" down the scores on Essay Combination 5, 6 & 8 towards the scores on Essay Combination 5, 6 & 7. Therefore, since most students select items to their best advantage, their equated scores can be considered the maximum possible equated scores. These have been called "the upper bound of equating accuracy" by Wainer, Wang and Thissen (1991). It is the second mission of this chapter to investigate to what extent and under what conditions this such equating occurs.

How is it possible to assess the effects of the violation of missing-at-randomness in IRT equating? In order to obtain a clear picture of what equating does to choice scores, it seems mandatory to simultaneously possess two batches of information -- (1) students' complete responses to an entire test and (2) students' partial responses to their chosen items. It is through the differences in the ability estimates from these two kinds of information that the accuracy of IRT equating on choice items can be reasonably measured. In the case of this dissertation, such unique information is especially solicited through Part C of the Survey Kit.

For each of the first three comparisons in Part C, there is a survey question, "If you were allowed to choose between these two questions, which one would you choose?" This question identifies the chosen items for each student. Furthermore, because all
the students are instructed to perform on both items, their complete answers to the entire test are known.

With such two-sided information, three methods can be employed to probe into the equating accuracy on the chosen items --- (1) the correlation between the two sets of ability estimates based on the complete and chosen items, respectively; (2) a t-test of the differences of the two ability estimates; and (3) the differences in the students' ranking between the two estimates.

The analyses in this chapter are based on the 237 Hawaii AP students. The reason to employ this group of students only is that this is the only group of students that have sufficient correct responses to the six items in Comparisons I, II and III in Part C of the Kit. Considerably fewer non-AP and college students in Hawaii answered these items right. It is suspected that the lack of sufficient correct responses would contribute more noise than information to the desired investigation of the accuracy of IRT equating the scores on chosen items. The reason that the UH upper-division class is not included either is that the number of this group is relative small, and its inclusion might introduce estimation bias since this group is at a much higher level of chemistry study.

Six sets of related scores are used in this chapter. The first two related sets of scores are scores on the three chosen vs. unchosen items. These two sets of scores are called "CHOSEN3" and "UNCHOSEN3". The second set of two related scores are two raw scores earned on the 20 vs. 17 items in the Kit. These two scores are called "RAW20" vs. "RAW17". Corresponding to RAW20 and RAW17 are the third set of two related IRT scores, IRT20 and IRT17, which are equated IRT scores. It is through the relationship
among these six sets of scores that the effects of equating vs. not equating the scores of chosen items will be analyzed.

**Investigating the Effects of Choices Through MC Items**

Did the AP students perform better on their chosen items than on the unchosen items? The answer is *Yes*. Table 9.1 summarizes the descriptive statistics of the performance of the AP students on the 3 chosen and unchosen items. On the average, the AP students performed better on the chosen items than on the unchosen items (1.44 vs. 1.08). The mean difference of .36 point is shown to be statistically significant.

**Table 9.1**

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Min.</th>
<th>Max.</th>
<th>Mean</th>
<th>StD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHOSEN3</td>
<td>237</td>
<td>0</td>
<td>3.00</td>
<td>1.44</td>
<td>0.89</td>
</tr>
<tr>
<td>UNCHOSEN3</td>
<td>237</td>
<td>0</td>
<td>3.00</td>
<td>1.08</td>
<td>0.87</td>
</tr>
</tbody>
</table>

**T-Test of the Difference Between the Scores of the 3 Chosen and the 3 Un-chosen Items**

| Variable                                               | N Obs | Mean | Std Error | T    | Prob > |T| |
|--------------------------------------------------------|-------|------|-----------|------|--------| |
| Difference Between Chosen and Un-chosen Scores         | 237   | 0.36 | 0.07      | 5.25 | 0.0001 |

161
How many students scored higher, the same or lower between the chosen and un-chosen items? The purpose of this question is to check into the proportions of students who made appropriate or inappropriate choices. Table 9.2 indicates that 45% of AP students scored higher on the chosen items, while about 22%, scored lower. Allowing choices does not make any difference to about 33% of the students. However, 22% is a substantial number of subjects who did not choose effectively. This finding, again, challenges the common assumption that the majority, if not all, students would score better on the chosen items.

Table 9.2

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same Score</td>
<td>79</td>
<td>33.3</td>
<td>79</td>
<td>33.3</td>
</tr>
<tr>
<td>Lower Score</td>
<td>51</td>
<td>21.5</td>
<td>130</td>
<td>54.9</td>
</tr>
<tr>
<td>Higher Score</td>
<td>107</td>
<td>45.1</td>
<td>237</td>
<td>100.0</td>
</tr>
</tbody>
</table>

What are the ability distributions that scored the same, higher and lower on the chosen items? According to Figure 9.1, the majority of the middle-level students who scored 7 to 12 points on the 20 MC items benefitted the most from making choices. Three times as many as middle-level students scored higher on the chosen items than on the unchosen items. For example, for those students who scored 7 points, as many as 18 students scored better on the chosen items, while only 2 and 3 students scored the same or lower on the chosen items. For the students of the lower-third ability level, approximately equal numbers...
of students scored the same, higher or lower. For the students of upper-third ability level, there are substantially more students who scored higher or the same than those who scored lower.

![Graph showing ability distributions of AP Chemistry students who scored the same, better, and lower on the basis of chosen items.]

**Figure 9.1: Ability Distributions of AP Chemistry Students Who Scored the Same, Better, and Lower on the Basis of Chosen Items**

The results indicate that allowing choices gives ample leeway for middle-level students to avoid the exposure of their academic deficiencies by avoiding the items they do not know, because their scores on the basis of the chosen items are higher than their scores on the unchosen items. In the case of this dissertation experiment where students were asked
to choose three out of six items, the middle-level students could avoid as much as half of their deficiencies.

Investigating the Relationship between the Raw Scores and IRT Equated Scores

What is the difference between using raw scores vs. the IRT equated scores on the 20 MC items? The difference lies in that raw scores treat items of different difficulties uniformly, while IRT weighs items of various difficulties differentially through internal automatic linking as described in Chapter V. The effect of such a difference on ranking examinees can be investigated through two measures. The first measure is the correlation (both rank and Pearson correlations) between the raw scores and IRT equated scores. This correlation shows to what extent the two sets of scores agree with each other in terms of ranking students. The second measure is the range of students’ rank changes between these two sets of scores. The bigger the changes are, the less compatible these two sets of scores are with each other. The equating conducted on the 20 MC items is traditional equating and does not involve the violation of missing-at-randomness, because every item was attempted by every student.

Table 9.4 shows that the correlation between the raw and the IRT equated scores on the 20 MC items is .90 and .93 as measured by Kendall’s tau and Pearson correlation coefficients, respectively. Figure 9.2 indicates that it is the lower-level students whose scores seem to have reduced the correlation, since they levelled off from the diagonal trend of the main stream of the scatterplots. The reason for such levelling off is that the score of 1 is as low as the raw score reporting can go, while IRT can go down much further.
The fact is that IRT makes a more detailed estimation of students' abilities than classical measurement theory (CMT), especially with those examinees who have the same raw scores. While students with a similar raw score are assigned the same ranking by CMT, IRT may rank them quite differently, depending on the item difficulty levels of the items they answer correctly. Suppose that two students have the same raw score of 1 point. If the item scored by Student A is harder than the item scored by Student B, Student A is assessed by IRT to be more able than Student B.

Table 9.3

Descriptive Statistics of Raw and IRT Equated Scores on the 20 MC Items

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std</th>
<th>Median</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRT20</td>
<td>224</td>
<td>0.04</td>
<td>1.13</td>
<td>0.20</td>
<td>-3.02</td>
<td>2.43</td>
</tr>
<tr>
<td>Raw20</td>
<td>224</td>
<td>8.95</td>
<td>4.24</td>
<td>8.00</td>
<td>2.00</td>
<td>19.00</td>
</tr>
</tbody>
</table>

Note: Thirteen students are deleted from this table because of non-convergence of IRT ability estimation.

Table 9.4

Kendall's Tau and Pearson Correlations between Raw and IRT Scores on the 20 MC Items.

<table>
<thead>
<tr>
<th></th>
<th>IRT Score20</th>
<th>Raw Score20</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRT20</td>
<td>Kendall Tau</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Pearson R</td>
<td>1.00</td>
</tr>
<tr>
<td>Raw20</td>
<td>Kendall’s Tau</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>Pearson R</td>
<td>0.93</td>
</tr>
</tbody>
</table>
To what extent have students' rankings changed between these two scores? In spite of the high correlations, Table 9.5 shows that 98% of the AP students changed their rankings from using raw scores to using IRT equated scores. More specifically, the rankings of 94 students (42%) moved down, while the rankings of 126 students (56%) moved up from raw scores to IRT equated scores. Only 4 students (1.8%) maintained their rankings. The largest downward rank change is 54. The largest upward rank change is 30.

Table 9.5

<table>
<thead>
<tr>
<th>Rank Change</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Downward</td>
<td>94</td>
<td>42.0</td>
<td>94</td>
<td>42.0</td>
</tr>
<tr>
<td>Same</td>
<td>4</td>
<td>1.8</td>
<td>98</td>
<td>43.7</td>
</tr>
<tr>
<td>Upward</td>
<td>126</td>
<td>56.2</td>
<td>224</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Note: Thirteen students are deleted from this table because of non-convergence of IRT ability estimation.
Frequency Distributions of Rank Changes

<table>
<thead>
<tr>
<th>Rank Change</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>-55 ~ -45</td>
<td>1</td>
<td>0.4</td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>-44 ~ -35</td>
<td>8</td>
<td>3.6</td>
<td>9</td>
<td>4.0</td>
</tr>
<tr>
<td>-34 ~ -25</td>
<td>5</td>
<td>2.2</td>
<td>14</td>
<td>6.3</td>
</tr>
<tr>
<td>-24 ~ -15</td>
<td>15</td>
<td>6.7</td>
<td>29</td>
<td>12.9</td>
</tr>
<tr>
<td>-14 ~ -5</td>
<td>36</td>
<td>16.1</td>
<td>65</td>
<td>29.0</td>
</tr>
<tr>
<td>-4 ~ 5</td>
<td>82</td>
<td>36.6</td>
<td>147</td>
<td>65.6</td>
</tr>
<tr>
<td>6 ~ 15</td>
<td>54</td>
<td>24.1</td>
<td>201</td>
<td>89.7</td>
</tr>
<tr>
<td>16 ~ 25</td>
<td>15</td>
<td>6.7</td>
<td>216</td>
<td>96.4</td>
</tr>
<tr>
<td>26 ~ 35</td>
<td>8</td>
<td>3.6</td>
<td>224</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Which measure is more accurate, the raw scores or IRT equated scores in terms of ranking students? Table 9.7 summarizes the differences between using the raw and IRT scores in terms of ranking 43 students who scored 3, 8 and 18 points. It can be seen that all the students with the same raw score share the same rank. For example, the first 12 students with the raw score of 3 points share the same rank of 12, while the students with the raw score of 18 points are allocated the rank of 221. This is because traditional scoring techniques treat the items with differential difficulties equally.

However, Table 9.7 reveals that the IRT scores for the students with the similar raw score points vary substantially. For example, the IRT scores for the students with the raw score of 3 points change from -2.99 symbolizing very low ability, to -.99 representing medium ability. Furthermore, students’ rankings also change correspondingly, namely from 4 to 41. All of this is because IRT differentially weighs items of varying difficulties.
Table 9.7
Comparison of Students' Rankings as Reflected by Raw Scores and IRT Scores

<table>
<thead>
<tr>
<th></th>
<th>Raw Score</th>
<th>IRT Score</th>
<th>Standard Error</th>
<th>IRT Score Rank</th>
<th>Raw Score Rank</th>
<th>Rank Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>-2.99</td>
<td>3.68</td>
<td>4</td>
<td>12</td>
<td>-9</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>-2.99</td>
<td>3.68</td>
<td>4</td>
<td>12</td>
<td>-9</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>-2.50</td>
<td>2.95</td>
<td>7</td>
<td>12</td>
<td>-5</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>-2.50</td>
<td>2.95</td>
<td>7</td>
<td>12</td>
<td>-5</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>-2.50</td>
<td>2.95</td>
<td>7</td>
<td>12</td>
<td>-5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>-1.74</td>
<td>1.67</td>
<td>15</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>-1.71</td>
<td>1.61</td>
<td>16</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>-1.69</td>
<td>1.58</td>
<td>19</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>-1.08</td>
<td>0.79</td>
<td>37</td>
<td>12</td>
<td>25</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>-1.05</td>
<td>0.77</td>
<td>39</td>
<td>12</td>
<td>27</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>-1.05</td>
<td>0.77</td>
<td>39</td>
<td>12</td>
<td>27</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>-0.97</td>
<td>0.72</td>
<td>41</td>
<td>12</td>
<td>29</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
<td>-0.25</td>
<td>0.47</td>
<td>79</td>
<td>107</td>
<td>-28</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>-0.03</td>
<td>0.44</td>
<td>89</td>
<td>107</td>
<td>-18</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>-0.06</td>
<td>0.43</td>
<td>92</td>
<td>107</td>
<td>-15</td>
</tr>
<tr>
<td>16</td>
<td>8</td>
<td>0.02</td>
<td>0.42</td>
<td>98</td>
<td>107</td>
<td>-9</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
<td>0.05</td>
<td>0.41</td>
<td>101</td>
<td>107</td>
<td>-6</td>
</tr>
<tr>
<td>18</td>
<td>8</td>
<td>0.05</td>
<td>0.41</td>
<td>102</td>
<td>107</td>
<td>-6</td>
</tr>
<tr>
<td>19</td>
<td>8</td>
<td>0.10</td>
<td>0.40</td>
<td>104</td>
<td>107</td>
<td>-3</td>
</tr>
<tr>
<td>20</td>
<td>8</td>
<td>0.12</td>
<td>0.40</td>
<td>109</td>
<td>107</td>
<td>-2</td>
</tr>
<tr>
<td>21</td>
<td>8</td>
<td>0.12</td>
<td>0.40</td>
<td>109</td>
<td>107</td>
<td>-2</td>
</tr>
<tr>
<td>22</td>
<td>8</td>
<td>0.17</td>
<td>0.39</td>
<td>110</td>
<td>107</td>
<td>-4</td>
</tr>
<tr>
<td>23</td>
<td>8</td>
<td>0.18</td>
<td>0.39</td>
<td>111</td>
<td>107</td>
<td>-5</td>
</tr>
<tr>
<td>24</td>
<td>8</td>
<td>0.20</td>
<td>0.39</td>
<td>112</td>
<td>107</td>
<td>-6</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>0.20</td>
<td>0.38</td>
<td>114</td>
<td>107</td>
<td>-7</td>
</tr>
<tr>
<td>26</td>
<td>8</td>
<td>0.20</td>
<td>0.38</td>
<td>114</td>
<td>107</td>
<td>-7</td>
</tr>
<tr>
<td>27</td>
<td>8</td>
<td>0.21</td>
<td>0.38</td>
<td>117</td>
<td>107</td>
<td>11</td>
</tr>
<tr>
<td>28</td>
<td>8</td>
<td>0.22</td>
<td>0.38</td>
<td>118</td>
<td>107</td>
<td>12</td>
</tr>
<tr>
<td>29</td>
<td>8</td>
<td>0.23</td>
<td>0.38</td>
<td>120</td>
<td>107</td>
<td>14</td>
</tr>
<tr>
<td>30</td>
<td>8</td>
<td>0.24</td>
<td>0.38</td>
<td>121</td>
<td>107</td>
<td>15</td>
</tr>
<tr>
<td>31</td>
<td>18</td>
<td>1.68</td>
<td>0.26</td>
<td>215</td>
<td>221</td>
<td>-6</td>
</tr>
<tr>
<td>32</td>
<td>18</td>
<td>1.94</td>
<td>0.32</td>
<td>220</td>
<td>221</td>
<td>-1</td>
</tr>
<tr>
<td>33</td>
<td>18</td>
<td>2.12</td>
<td>0.41</td>
<td>221</td>
<td>221</td>
<td>0</td>
</tr>
<tr>
<td>34</td>
<td>18</td>
<td>2.16</td>
<td>0.43</td>
<td>222</td>
<td>221</td>
<td>1</td>
</tr>
</tbody>
</table>
How does IRT weigh items of varying difficulties? As described in Chapter V on IRT equating, IRT gives more weight to difficult items than easy items. More specifically, a correct answer on a hard item is worth more than a correct answer on an easy one. In Table 9.8, the response vectors of Students 8 and 9 are compared, one can see that although they both scored 3 points, their third correct responses are different. Student 8 answered Item 19 correctly, while Student 9 got Item 13 right. It has already been shown in Table 4.4 that Item 19 is more difficult than Item 13 -- the former has a difficulty of .36, while the latter has the difficulty of -0.07.

Table 9.8
Illustrations of the Enhanced Accuracy of IRT Scores over Raw Scores.

<table>
<thead>
<tr>
<th>Student Number</th>
<th>Response Vector</th>
<th>Raw Score</th>
<th>IRT Score</th>
<th>Std Error</th>
<th>IRT Rank</th>
<th>Raw Rank</th>
<th>Rank Diff (IRT - Raw)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0100000000010000010000001000</td>
<td>3</td>
<td>-1.05</td>
<td>0.77</td>
<td>39</td>
<td>12</td>
<td>27</td>
</tr>
<tr>
<td>11</td>
<td>0100000000010000010000001000</td>
<td>3</td>
<td>-1.05</td>
<td>0.77</td>
<td>39</td>
<td>12</td>
<td>27</td>
</tr>
<tr>
<td>34</td>
<td>01101100010010001010</td>
<td>8</td>
<td>0.20</td>
<td>0.38</td>
<td>114</td>
<td>107</td>
<td>7</td>
</tr>
<tr>
<td>35</td>
<td>01101100010010001010</td>
<td>8</td>
<td>0.20</td>
<td>0.38</td>
<td>114</td>
<td>107</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>0010000000010000001000000010</td>
<td>3</td>
<td>-1.69</td>
<td>1.58</td>
<td>19</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>00100000000100000000000000</td>
<td>3</td>
<td>-1.08</td>
<td>0.80</td>
<td>37</td>
<td>12</td>
<td>25</td>
</tr>
<tr>
<td>40</td>
<td>11111111111101101111</td>
<td>18</td>
<td>1.68</td>
<td>0.26</td>
<td>215</td>
<td>221</td>
<td>-6</td>
</tr>
<tr>
<td>42</td>
<td>11111111111111111111111001</td>
<td>18</td>
<td>2.12</td>
<td>0.41</td>
<td>221</td>
<td>221</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>11111111111101111111111011</td>
<td>18</td>
<td>2.34</td>
<td>0.52</td>
<td>223</td>
<td>221</td>
<td>2</td>
</tr>
</tbody>
</table>

This situation is also true of Students 42 and 43 who both scored 18 points. They both missed Item 18. Student 42 missed Item 19 (difficulty = .36), while Student 43 missed
Item 11 (difficulty = -2.67). In other words, since Student 43 answered correctly the more difficult Item 19, he/she is deemed by IRT to be a more able student.

In what circumstances does IRT judge students to be of the same ability? IRT offers similar IRT estimates only when students have identical response vectors. This is exemplified by Students 10, 11, 34 and 35 in Table 9.8 who have exactly the same response vectors. Consequently, their rankings are the same, as well.

Based on the above examples and results, it can be concluded that IRT scores are much more accurate than raw scores in terms of determining students’ abilities and rankings. This is because IRT technology simultaneously takes into account not only the total number of scored points, but also the differential difficulties of the items students attempt.

So far, we have examined the advantages of IRT scores over raw scores in the circumstance when all the items are attempted by all students. The IRT equating or linking that is inherently implemented in such a case does not involve the violation of its missing-at-random assumption. What happens to its accuracy when it is applied to the scores on the items students deliberately choose? The students’ selection process involves non-ignorable non-responses. Ironically, it is just in this situation that the automatic and internal adjustment process of IRT is especially desired. This is because the different items chosen by students often vary substantially in terms of item difficulties. How accurate are the IRT scores as compared with traditional raw scores on the chosen items? What is the nature of the equated IRT scores based on the chosen items? What is the effect of the violation of its assumption of missing-at-randomness?
Investigating the Accuracy of IRT Equating on Choice Items

Due to choices, four sets of scores can be formed on the basis of the complete 20 MC items and the partial 17 MC excluding the unchosen items -- two of the four sets are in the raw score format, and the other two, in IRT raw score format. The two sets of raw scores are simply the numbers of students' correct answers to the 20 and 17 MC items, respectively. The two sets of IRT scores were obtained through two separate IRT analyses of students' abilities using MULTILOG (Thissen, 1991). The first analysis used students' complete responses to the entire 20 MC items, while the second analysis employed students' responses to the 17 MC items. These four sets of scores are referred to as RAW20, RAW17, IRT20 and IRT17, respectively.

The investigation of the IRT equating accuracy on the chosen items is to be evaluated in reference to the raw scores through two methods. The first method is to evaluate through two t-tests whether or not significant differences exist between IRT20 and IRT17, and RAW20 and RAW17, respectively. Nonsignificant t-tests symbolize negligible differences in students' ability estimates based on the full and partial data.

The second method is to cross-examine the intercorrelations among these four sets of scores of RAW20, RAW17, IRT20 and IRT17. If there are parallel patterns of correlations within and across the two sets of raw and IRT scores, reasonable confidence can be established on the accuracy of the IRT equating on chosen items.
T-Test of the Differences between Full and Partial IRT and Raw Scores:

Table 9.9 summarizes the descriptive statistics of the two sets of ability estimates as well as the t-statistics. It can be seen that the maximum, minimum, mean ability estimates as well as the standard deviations of both the full and equated ability estimates are virtually identical. The mean of the differences between these two IRT estimates is virtually zero. As a result, the t-test is nonsignificant. This is a good illustration of the advantage of IRT described in Chapter V, that IRT treats long and short tests equally, when the data are fit by a chosen IRT model.

Table 9.9

Summary Statistics and T-Statistics of IRT20 and IRT17 vs. RAW20 and RAW17

<table>
<thead>
<tr>
<th>Ability</th>
<th>N</th>
<th>Min.</th>
<th>Max.</th>
<th>Mean</th>
<th>Std</th>
<th>Min. Diff</th>
<th>Max. Diff</th>
<th>Mean Diff</th>
<th>Std Error</th>
<th>T</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRT20</td>
<td>224</td>
<td>-3.02</td>
<td>2.85</td>
<td>0.03</td>
<td>1.24</td>
<td>-1.76</td>
<td>0.54</td>
<td>0.01</td>
<td>0.02</td>
<td>0.68</td>
<td>0.49</td>
</tr>
<tr>
<td>IRT17</td>
<td>224</td>
<td>-3.07</td>
<td>2.70</td>
<td>0.02</td>
<td>1.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAW20</td>
<td>224</td>
<td>1.00</td>
<td>19.00</td>
<td>8.67</td>
<td>4.35</td>
<td>0</td>
<td>3.00</td>
<td>1.04</td>
<td>0.06</td>
<td>18.29</td>
<td>0.0001</td>
</tr>
<tr>
<td>RAW17</td>
<td>224</td>
<td>1.00</td>
<td>16.00</td>
<td>7.64</td>
<td>3.85</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: "IRT20 and RAW20" stands for the full ability estimates on the complete 20 items. "IRT17 and RAW17" stands for the partial ability estimate based on the 14 anchor and 3 chosen items. "Diff" stands for the difference between the full and partial ability estimates.

However, the situation with the raw scores is quite different. Because raw scores only count the numbers of correct responses, the effect of the unchosen items naturally shows
up in the scores. Consequently, the t-test of the difference between RAW20 and RAW17, namely, the effect of the unchosen items, is highly significant.

Intercorrelations Among the Full and Partial IRT and Raw Scores:

What are the inter-correlations among the four sets of ability estimates? As indicated in Table 9.10, the correlations between the scores of the same category are found to be identical. Specifically, the correlation between IRT20 and IRT17 is .98; so is the correlation between RAW20 and RAW17. Such identical correlations indicate high levels of parallelism and consistency in terms of ranking students within each score system.

Table 9.10
Intercorrelations among Four Sets of Scores (N=224)

<table>
<thead>
<tr>
<th></th>
<th>IRT20</th>
<th>IRT17</th>
<th>RAW20</th>
<th>RAW17</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRT20</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRT17</td>
<td>0.98</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAW20</td>
<td>0.93</td>
<td>0.91</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>RAW17</td>
<td>0.92</td>
<td>0.92</td>
<td>0.98</td>
<td>1.00</td>
</tr>
</tbody>
</table>

They are also indications of the success of equating. According to Morris (1982), equating matches the moments of the two distributions of test scores to be equated. At the minimum, it matches the means. At the next step, it should match the standard deviations and so on. The conclusion that can be drawn from the above investigation so far is that the violation of missing-at-randomness seems to have little effect on the accuracy of the
traditional equating methodology on the items that are not independently selected but are purposefully chosen by students.

Figures 12.3 and 12.4 display the scatterplots of the four sets of scores.

One clear difference between Figure 9.3 and Figure 9.4 is that there are more data points in Figure 9.4 than in Figure 9.3. This is because IRT makes more detailed estimations of students’ abilities than CMT.
Table 9.10 also shows that the correlations between the scores of different categories are slightly reduced. For example, the correlation between RAW20 and IRT20 is .93, while the correlation between RAW17 and IRT17 is .92. This reduction is expected since the scores of the two categories are computed differently. As shown earlier in this chapter, the essential difference between these two score systems is that the raw scores treat items of different difficulties equally, while the IRT scores weigh them differentially. Figures 12.5 and 12.6 display the scatterplots of the 224 AP students as reflected by the two different score categories of IRT and raw scores.

Figure 9.5: Scatterplot of the Ability Estimates of 224 Hawaii AP Chemistry Students As Measured by RAW20 and IRT20

Figure 9.6: Scatterplot of the Ability Estimates of 224 Hawaii AP Chemistry Students As Measured by RAW17 and IRT17
The common feature shared by Figures 12.5 and 12.6 is that the lower ends of the two scatterplots level off at the raw score of 1 point. This is because 1 point is as low as the raw scores can go, while IRT still has plenty room to "place" students. In Figure 9.6, for example, a 1-point raw score can mean as much as -1.2 to -3 on the IRT scale. This is exemplified by the three students in Table 9.11. It can be seen that students' IRT scores are directly dependent on the item difficulties of the items the students answered correctly. The more difficult the items, the higher IRT ability estimates.

Table 9.11
Examples of the Relationship between IRT Scores and Item Difficulties.

<table>
<thead>
<tr>
<th>Student Number</th>
<th>Raw Score</th>
<th>1st Correct Item #</th>
<th>Item Difficulty</th>
<th>2nd Correct Item #</th>
<th>Item Difficulty</th>
<th>Estimated Theta</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>1</td>
<td>-2.27</td>
<td>6</td>
<td>.89</td>
<td>-3.00</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>1</td>
<td>-2.27</td>
<td>8</td>
<td>1.20</td>
<td>-2.03</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>17</td>
<td>-1.49</td>
<td>18</td>
<td>2.56</td>
<td>-1.21</td>
</tr>
</tbody>
</table>

Comprehensive Analyses of the Effects of

EQUATING vs. NOT-EQUATING SCORES ON CHOSEN ITEMS

After equating the chosen items, the final two tasks of this dissertation are to investigate the effects of allowing choices on a test and the differences between equating vs. not equating scores on the chosen items. Such inquiries are important because the first task addresses the potentially profound differences between allowing vs. not allowing choices for items on a test. It is necessary to determine what allowing choices of items really does to
students. The second task concerns the prospective of alleviating differences due to choices. The ultimate goal of equating is to achieve equity and justice to as many examinees as possible.

**Comparison of Performance on the 20 vs. 17 MC Items Based on Observed Raw Scores:**

Let's first look at the three plates in Figure 9.7 which displays the relationship between RAW20 and RAW17 of the 223 AP students who are divided into three ability levels. The students are sorted in an ascending order with RAW17 and RAW20 as the primary and secondary sorting keys, respectively. The purpose of this display is two-fold. First, it is intended to show the relative performance of the students on both the 20 and 17 MC items. Second, it provides a foundation for later comparisons.

As expected, students’ scores on RAW20 are, at least, as high as those on RAW17. In other words, students’ scores on the basis of 17 items cannot surpass their counterparts on the basis of 20 items, regardless of how well a student performed on the items they chose. The lower the ability levels, the more RAW20 overlaps with RAW17. The overlap of the two scores means that the same number of items were scored on the two versions of the test.
Upper Third Level Students
- RAW17
++RAW20

Middle Level Students
- RAW17
++RAW20

Lower Third Level Students
- RAW17
++RAW20

Figure 9.7: Relationship between RAW20 and RAW17 without Equating
It should be noted that RAW20 and RAW17 are in different scales, because they have two different item totals, which results in two different means and standard deviations. In order to see how students performed relatively on each test, it is necessary to compare students in terms of standard z-scores. The relationship between RAW20 and RAW17 in standardized format is replicated in the three plates of Figure 9.8.

Comparison of Performance on the 20 vs. 17 MC Items Based on Standardized Scores:

Since z-scores of RAW20 and RAW17 reflect the relative locations of the students within the two distributions, the three plates in Figure 9.8 demonstrate dramatic changes of the students’ rankings from RAW20 to RAW17. It can be seen that the majority of the lower-third and middle-level students ranked lower on RAW20 than on RAW17. This is due to the fact that students deliberately chose the items they felt confident in and intentionally shunned the items they felt less sure of. In other words, for the majority of lower-third and middle-level students, performance was better on the 17-item test than on the 20-item test.

Yet, the performance of the upper-third-level students is the opposite. They maintained their superior rankings on the basis of RAW20. This reflects the fact that high-level students were able to perform consistently and allowing choices of MC items does not seem to affect their performance as much.

Yet, what is the relationship between students’ performances on the 20 and 17 MC items after the internal adjustment of the item difficulties? This relationship is illustrated by the three plates in Figure 9.9.
Figure 9.8: Relationship between RAW20 and RAW17 without Equating in Z-scores
Comparison of Performance on the 20 vs. 17 MC Items After IRT Equating:

The relationship between IRT20 and IRT17 is quite different as shown by the three plates in Figure 9.9. There are two general differences between the three plates in Figure 9.9 and the three plates in Figure 9.8. The foremost difference is that Figure 9.9 displays the ability estimates of the 223 students independently assessed from the 20 MC and 17 MC items separately. These independent estimation processes treat the item difficulties of the separate sets of chosen items differentially. Yet, it is known that the z-scores in Figure 9.8 treat every set of the chosen items equally regardless of the variations in item difficulties.

As the result of more fine-tuned estimations of student abilities, the second difference between Figure 9.9 and Figure 9.8 is that the IRT scores form considerably less ties of students' rankings than the z-scores. Such a substantial reduction of tied rankings ultimately transform into more accurate classifications of students.

What are the effects of IRT equating the scores on chosen items? Two general tendencies can be observed from the three plates in Figure 9.9. First, taking the total 223 students as a whole, the IRT equated scores on the 17 items tend to provide higher ability estimates than do the equated IRT estimates on the 20 MC items. This is expected because students were, in general, supposed to do better on the test that included the items that they chose to their advantage, as shown in Table 9.1. This supports the prediction of Wainer, Wang & Thissen (1991) that equating in such a circumstance provides a "plausible upper bound" of the ability estimates of the students. Such a phenomenon is in sharp contrast to the unequated z-scores in Figure 9.8 that showed the upper-third-level students predominantly ranked higher on the 20 items than on the 17 items, while the students below middle-level
abilities ranked the opposite way. Remember that z-scores specify relative standings of students in reference to the mean of the distribution, not students’ potential abilities on the basis of item information.

However, at the level of individual students, like regression, equating on the 17 items tends to predict higher scores for those students who actually scored lower on the full 20 MC items but scored well on the chosen items. On the other, it projects lower scores for those students who performed exceptionally well on the full 20 items, because 17 items impose a lower ceiling effect to cap their performance. However, it should be clarified here that the ceiling effect is not a problem of equating, but of allowing choices of a subset set of a full test. On the contrary, equating alleviates such a problem to a certain extent.

The second tendency in IRT equating is that there are substantially fewer discrepancies in the ability estimates of the upper-third-level and lower-third-level students than in the ability estimates of the students of the middle-ability level between IRT20 and IRT17. In other words, the ability estimates of the former two groups match more closely than those of the latter group and there are less adjustments necessary. This phenomenon is reasonable. For students of either high or low abilities, they tended to score the items either right or wrong. Except for a few students who really made exceptionally good or bad choices of items, the majority of the students in these two groups are expected to conform to this expectation.
Figure 9.9: Relationship between IRT20 and IRT17 After Equating
However, the situation with the middle-level students is quite different. Since they have certain ability to tackle some problems but not full ability to solve them all like high-level students. It is natural to observe ample ups and downs in their performance when they were allowed to choose, because some of them were bound to benefit more from choices than others. As a result, more adjustments of their scores were necessary.

Illustration of the Effects of Equating

In order to firmly grasp the above-described effects of equating, this study looks at some specific examples of the effects of equating on both the individual and group levels. First, the general effects of equating on chosen items on students' ability estimates and rankings are presented. How many students improved their ability estimates from IRT20 to IRT17? Figure 9.10 shows that 122 out of the 223 students (54%) boosted their ability estimates, 5 students (2%) maintained their ability estimates, while 95 students (43%) shifted their ability estimates downwards.
What happens to the rankings of the students from IRT20 to IRT17? It can be seen from Figure 9.11 that the distribution of rank changes are approximately symmetrically distributed. The maximum downward rank change is 45, while the maximum upward change is 35. Most students have very minor rank changes -4 and 5. Note that the "-" sign stands for downward rank change, while the "+" sign, upward rank change.
Now since we know who raised and lowered their rankings the most, let’s look at a few examples in order to understand what happened to them at the level of their response vectors.

Student 23 changed his/her ranking by 45 positions, which is the maximum change of ranking observed among the 223 students. What happened? The student scored 9 points in terms of raw score, which is slightly above the mean score of 8.72 of Hawaii AP students.
(shown in Table 9.1). Accordingly, his/her IRT20 ability estimate was .388. Now, as shown by his/her item response vector on the six comparison items, the student scored 3 items right. However, because he/she scored zero on the items he/she chose, the IRT17 ability estimate of this student dropped to -.20, slightly below the mean. As the result, his/her ranking decreased to 88. In short, Student 23 is the worst scenario of a student’s performance on choice items. The situation with Student 50 is similar to that of Student 23.

However, let’s look at Students 5 and 180 who moved up their rankings the most. Both of them made appropriate choices because the two of the three items they chose happened to be the two items they answered correctly out of the total six comparison items. The rankings of both of them have moved up. In this regard, these two students should be a good example of students making appropriate choices.
### Table 9.12
Illustration of the Effects of Choices and Choice Item Difficulties on IRT Ability Estimates and Rankings

<table>
<thead>
<tr>
<th>Ss #</th>
<th>Total Score</th>
<th>Full IRT $\Theta$</th>
<th>Full Rank</th>
<th>Response Vector on Six Items</th>
<th>Score on Six Items</th>
<th>Response Vector on 3 Items</th>
<th>Score on 3 Items</th>
<th>Choice $\Theta$</th>
<th>Choice Rank</th>
<th>Rank Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>9</td>
<td>0.388</td>
<td>133</td>
<td>011010</td>
<td>3</td>
<td>122</td>
<td>000</td>
<td>-0.20</td>
<td>88</td>
<td>-45</td>
</tr>
<tr>
<td>50</td>
<td>7</td>
<td>-0.086</td>
<td>94</td>
<td>111100</td>
<td>4</td>
<td>121</td>
<td>100</td>
<td>-0.71</td>
<td>52</td>
<td>-42</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>0.54</td>
<td>119</td>
<td>000101</td>
<td>2</td>
<td>222</td>
<td>011</td>
<td>0.64</td>
<td>150</td>
<td>31</td>
</tr>
<tr>
<td>180</td>
<td>9</td>
<td>0.13</td>
<td>110</td>
<td>100100</td>
<td>2</td>
<td>121</td>
<td>110</td>
<td>0.48</td>
<td>142</td>
<td>32</td>
</tr>
</tbody>
</table>

**Note:**
- Ss # = student's observation number.
- Total Score = total raw score on the 20 MC items.
- Full IRT $\Theta$ = IRT score estimated from 20 MC Items (IRT20).
- Full Rank = the student's ranking based on IRT20.
- Response Vector on Six Items = response vector on the six choice MC items, namely Item 11 through Item 16 in Comparison III of the Kit.
- Score on Six Items = student's score on the six choice items.
- Choice Pattern = the items that students chose in the three comparisons. Each of the three digits has two meanings: its position stands for the comparison number; its magnitude, the number of the two comparison items. For example, Choice Pattern 122 stands for the Item 1 in Comparison I, Item 2 in Comparison II, Item 2 in Comparison III.
- Response Vector on 3 Items = the response vector on the three items that a student chose.
- Score on 3 Items = the raw score observed on the three chosen items.
- Choice $\Theta$ = IRT score estimated from 17 MC items (IRT17).
- Choice Rank = a student's ranking based on his/her IRT17.
- Rank Change = change in a student's ranking from IRT20 to IRT17. The "-" sign symbolizes downward change, while the "+", upward change.
Although Student 180 answered one more item right than Student 5 (9 vs. 8), the IRT ability estimate of Student 180 is lower than that of Student 5 (.13 vs. .54). Why is this so? This is due to the fact that the set of items answered correctly by Student 5 is more difficult than those answered right by Student 180. Their response vectors to the six comparison items provide evidence. The difference between Student 5 and Student 8 is that the former answered correctly the first comparison item which is known to be very easy (-2.63), while Student 5 answered correctly the sixth comparison item which is known to be very hard (2.09). (Just a reminder that the item parameters are summarized in Table 4.4 in this dissertation.)

After examining the effects of equating on the chosen items on the ability estimates and rankings of individual students, it is time to look at the effects of such equating on the mean performance of different groups of students who chose the six MC comparison items differentially. The reader should be able to recall that in Chapters VI and VII, discussion was given to the relationship between mean essay difficulty levels and the mean performance for the ten groups of students with the ten essay combinations. One of the major findings was that the essays were differentially difficult for students of similar ability and students’ scores on such essays should have their scores adjusted. Since detailed illustrations have already been offered on the inner workings of equating chosen items, the following discussion is intended to shed light on what equating chosen items does to the mean performances of students with various MC item combinations.

It is known that for each of the three comparisons, the 223 students were asked to hypothetically choose one of the two paired MC items. This results in eight choice combinations (2*2*2): Item Combinations 111, 112, 121, 122, 211, 212, 221 and 222.
Remember that the value of each of the three numbers stands for the item and its position represents the comparison. For example, Item Combination 111 stands for the first MC item in each of the three comparisons. Naturally, these eight item combinations divide the 223 students into eight groups. Figure 9.11 summarizes their mean performance in terms of RAW20, RAW17 and RAW3, their mean ability estimates as measured by IRT20 and IRT17, as well as the mean item difficulty levels of the eight item combinations.

Figure 9.12: Comparison of Raw and IRT Equated Scores in Light of Mean Item Difficulties

Let's first examine the average item difficulty levels of the eight combinations. The average difficulty levels of Item Combinations 111, 112, 121 and 122 are fairly similar with slight variation. Although the average difficulty levels of Item Combinations 211, 212, 221
and 222 are also similar to one another, they are much harder than the previous four combinations.

Observe the curve representing RAW3, students' average performance on their three chosen items. It can be seen that this curve changes along with the curve presenting the average item combination difficulty levels. Note that the value of RAW3 for Item Combination 212 is zero, meaning no students answered any of these three questions right.

Observe the two curves representing RAW20 and RAW17. With the two curves running parallel, Curve RAW20 is above Curve 17, since it has a higher mean. The students that are shown to be the most and least able are those who chose Item Combinations 121 and 212, respectively. In contrast to the two curves representing RAW20 and RAW17, the curve for IRT20, for the most part, is below the curve for IRT17. This is additional evidence for the prediction of the upper bound of ability estimates due to equating on chosen items (Wainer, Wang & Thissen, 1991). However, Curve IRT17 does "touch down" on Curve IRT20 at Item Combination 212, because nobody in this group correctly answered any of the three items, as shown by Curve RAW3.

Another sharp contrast between Curves RAW20 and RAW17 vs. IRT20 and IRT17 is that the latter two curves appear smoother. Most importantly, they are dictated more by the average item difficulties of the item combinations. The most salient phenomenon is that although very few students got Item Combinations 212 and 222 right as shown by RAW20 and RAW17, the mean IRT ability estimates of the two groups substantially compensated for the increased item difficulties. This is why there are no dramatic "ups and downs" in Curves IRT20 and IRT17 as there are in Curves RAW20 and RAW17.
Chapter Summary

This chapter has investigated the issue on the effects of equating vs. not equating the scores on student-selected items from four perspectives: 1) the differences in students’ performances between the chosen and unchosen items; 2) the differences between the equated IRT and unequated raw scores on the 20 items; 3) the violation of missing-at-randomness on the accuracy of equating scores on chosen items; 4) the differences between equating and not equating scores on chosen items.

As for students’ performance on the chosen and unchosen items, it has been found that students performed significantly better on the items they chose than on the items they did not choose. It is the middle-ability students that benefitted the most from choices, since the majority of this group moved up their rankings substantially.

On the basis of the 20 MC items, IRT equated scores have been found to be more accurate than raw scores in reflecting students’ ability levels. This is because IRT ability estimation is more sensitive to the item difficulties of the items students answered correctly. Given two students who answered the same number of items correctly, the student who correctly answered more difficult items is deemed to have a higher ability by IRT. Yet, according to the unequated raw scores, they are considered to have equal ability levels.

The violation of missing-at-randomness seems to have a negligible effect on the accuracy of equating on chosen items. It has been found that the correlation between students’ ability estimates on the basis of the 20 complete and 17 partial items (including three chosen items) is .98. The t-test of the differences between these two sets of ability estimates is nonsignificant.
As for the differences between equating and not equating the scores on chosen items, it has been found that IRT equating offers not only more accurate assessment of students' abilities, but also more detailed distinctions of students' rankings. It has been shown that not equating the scores on chosen items tends to result in indiscriminant changes in the rankings of groups of students. However, IRT equating allows students to compensate their scores for the difficulties of the items they chose. The harder the items they answered correctly, the higher their ability estimates were.
CHAPTER X

DISCUSSION AND RECOMMENDATIONS

A review of the research findings in this dissertation suggests four directions to improve current practices of allowing choices for constructed response (CR) items.

1. reduce the disparities between students' choices of CR items.
2. control the difficulty levels of CR items.
3. improve students' general performance on essays.
4. encourage more students to choose deliberately and selectively instead of sequentially.

How to Reduce the Disparities of Students' Choices

It has been shown that the 1989 AP Chemistry Exam examinees predominantly chose the core-chemistry essays. Dramatic choice disparities reflect the extent of inappropriate item selection on the part of test developer. Poor item selection fails to measure the commonality of students' knowledge. In order for students to choose more evenly among all the choice items, there seems to be two plausible solutions. The first, a simplistic one, is to limit the choice items to only the core content of a subject, such as the "core" AP chemistry topics. The reason for such a solution is reflected in the fact that only a small number of students in the dissertation chose the non-core items. For example, since Essay 9 of nuclear
chemistry failed to attract the majority of students, why should it be included? The failure for an item to appeal to the most students would deem it an "odd" item.

The second solution to spreading students’ choices evenly is to offer more groups of CR items. Each of the groups would consist of items with similar content, and students would be instructed to choose one from each group. For example, three groups of essay questions could have been offered in Part D of the 1989 AP Chemistry Exam -- a pair of general chemistry problems; a pair of lab-oriented problems; and finally a pair of non-core chemistry problems including nuclear chemistry or organic molecules. The advantage of this solution is to ensure a common basis to compare students’ scores, and to provide reasonable flexibility for students’ choices.

Once a set of choice items have been selected by the examiner, a preliminary investigation of students’ potential choice preferences can be conducted by surveying a small group of students for their perceptions of item difficulties. Such perceptions can then be analyzed using the variance stable scaling method (RANKO) to yield a unidimensional item difficulty scale. It has been demonstrated in this dissertation that an item difficulty scale can accurately account for choice preferences.

**How to Control the Difficulty of CR Items**

It has been shown in this dissertation that two out of the five essays have substantially different item difficulties. Specifically, Figure 6.2 indicates that Essays 5, 7 & 8, have very similar item difficulties. Essay 6, a general chemistry problem, is unusually hard, while
Essay 9 on nuclear chemistry, a non-core chemistry subject, is fairly easy. Since ETS has a rigorous system to measure and classify MC items by item difficulty and discriminations, the unusual variations in item difficulty among the five essays in Part D of the 1992 AP Chemistry Exam is surprising. Three reasons are suspected. The first reason seems to be the lack of a widely accepted system to correctly measure and record the difficulties and discrimination of essay items. Unlike a MC item (which has specialized parameters of item difficulty, discrimination and guessing to describe the probability that an person with a certain ability can answer it right), an essay item has various graded responses, for example 1 to 9. The graded responses are described by different sets of parameters. However, the IRT expected score method employed in this dissertation was not operationally available until recently.

The author has noticed that ETS reports essay difficulties as the mean score of an essay. Although traditionally used, the mean scores of essays are known to be highly biased because they are considerably influenced by the ability levels of the students who choose to answer them. Such a bias is expected to be substantial when the number of examinees is relatively small.

Although unlikely, the second reason for variations in difficulty may be artificial manipulation. Since nuclear chemistry is taught less often, perhaps it was not made as difficult as the more often taught essay questions on core chemistry topics. In this case, it is highly unreasonable for an item like Essay 9, which requires specific recall of facts to be used on an equivalent basis with the other essays that require higher levels of processing in the cognitive taxonomy. If such selection process were true, the author is sympathetic with
such practice but strongly opposes it, because its direct consequence is to put students on an unequal basis of comparison.

Although equating chosen items helps to alleviate some differences caused by item difficulties, it cannot completely recover the loss in scores as a result of inappropriate choices. Equating fully awards students who perform well on chosen items. It has been shown in the dissertation that students' IRT ability estimates and rankings increased substantially when they scored high on their chosen items, but also suffered the most when they scored poorly on the items they chose. As a result, it is highly recommended that choice items be made as equal as possible in their difficulties so that the students who accidentally choose an inappropriate set of items will not be overly penalized.

If one does not have the access to MULTILOG or the expertise on polytomous IRT, what other method can be used to reflect essay choice difficulty? The other candidate to describe difficulty is the conditional mean. In this dissertation, IRT expected scores and conditional mean scores are highly comparable. This is because both of them share a common characteristic. That is, both take into account the ability levels of examinees. The advantage of IRT expected scores over the conditional means is that IRT technology automatically adjusts for item difficulties across the complete student ability spectrum, and therefore, is more accurate measure than the conditional means. The significance of demonstrating the high comparability of these two measures is to advocate, at least, the use of the conditional means when IRT measures are not feasible.

The need to accurately describe the difficulty levels of choice items becomes mandatory with the recent revolution in testing and measurement which is signified by the
increased use of constructed-response items on performance tests. Allowing students to choose among different performance tasks is common. Any measurement device, no matter how philosophically sound it is, ceases to serve its primary function, once it fails to provide fair and accurate assessment of student ability on an equal basis. This is especially true if the item difficulties of a test item are not described accurately in the first place.

How to Improve Students' General Performance on the Essay Parts

One dismal finding about the overall student performance on the essays is the unusually low mean scores. As shown in Table 6.3 and Figure 6.7, the mean scores on all the essays and essay combinations are only about 30% of the total scores. It is suspected that such low mean scores are probably caused by the resulting fatigue of the three-hour exam. The author talked to one local AP chemistry class after it took the complete 1989 AP Chemistry Exam for practice. The common reaction from the students was that they felt that their brains had been "fried". It is no doubt that such fatigue also greatly influences students’ ability to decide which items are most appropriate to choose. It is suspected that the lower the students abilities are, the bigger the effect of fatigue on choice accuracy.

It is suggested that the MC item section in the AP exams be switched with the essay section. In other words, the AP Chemistry Exam would start with the choice essay section and be followed by MC item section. The advantage of this switch is that it allows students to first tackle the complicated essay questions when mentally alert. This should allow students to more fully demonstrate their abilities. Because it takes much less energy and
cognitive processing to answer MC items than essay items, students should be well able to process MC items at the end of a test.

The author disagrees with the contention that starting with complicated essays contradicts the tradition of starting a test with easy items. Working on essay questions is different from working on MC items in that it involves more holistic and deeper cognitive processes. By the time students survey the essence of all the essay problems and make their choices, their mind should be in gear and ready to organize solutions. It is contended that since the AP Chemistry exams are an achievement test, it is the amount of knowledge and competence demonstrated on the test that really counts. If taking the essay section first improves scores, that is all that is desired. However, whether or not taking essay items first improves scores is still an empirical question. Future research is necessary in order to confirm or reject the proposition.

How to Encourage More Students to Choose Deliberately

A considerable number of low-ability students did not seem to choose selectively, but just picked the first three essays as they encountered them. The lower the ability level, the more students tended to behave this way. The performance of the students who chose sequentially has been estimated to be significantly lower than that of the students who chose selectively.

The solution to encouraging students to choose deliberately and to perform better is straightforward -- simply add an extra line of instruction such as "You are advised to choose
carefully among the following items in order to maximize your scores. This will remind students to take a moment to review each item’s content. This should result in more appropriate choices. Although most students know that they are supposed to choose, there may be some students who neither know about this possibility, as one AP chemistry teacher pointed out, nor fully realize its importance. Moving the essay section to the beginning of the exam may reduce sequential choosing, because students are more alert and energetic at the beginning of the test than at the end.

**How to Minimize the Differential Gains in the Scores of Students of Different Ability Levels**

It has been shown in Chapter IX that the middle-ability students benefitted the most from responding to self-selected CR items. This was followed by lower-level students, and thirdly, by the upper-level students. IRT equating seems to alleviate the problem of differential success by ability levels by compensating students choosing a harder set of items. The effect of such adjustment is marginal compared to the gains or losses of scores due to appropriate or inappropriate choices. However, it has been demonstrated that those students who scored well on their chosen items enhanced their IRT ability scores substantially, while those students who scored zero on their chosen items suffered dearly in their ability estimates.

In face of such differential benefits and losses due to choice questions, one would question the validity of allowing choices on a test. Is it still as fair and valid to compare
students using tests involving choices of items as it is using a test without choices of items? The answer is clearly No, because of two undesirable consequences.

The first consequence of allowing choices on a test is that it allows opportunities for lower- and middle-ability students to avoid exposing their deficiencies. The effect of such liberty to "sneak around" becomes bigger as the number of choice items gets bigger. For example, in this dissertation experiment where students chose 1 out of each of three pairs of items, lower- and middle-ability students could avoid confronting 50% of the six items presented. Imagine how much latitude there would be if five to ten pairs of items are open to choice. In Part D of the 1989 AP Chemistry Exam, choosing 3 out of 5 essays would give the examinees a chance to avoid as much as 40% of the content. Suppose that these five essays span the complete content range of a AP chemistry textbook, it means that hardly any difference can be detected between a middle-level student who has studied 60% of the textbook and another high-level student who has studied 100% of the textbook.

The second consequence of allowing choices is the reduced capability of a test to accurately identify students’ true abilities. First of all, the observed scores on chosen items are a mixture of varying degrees of reliability and accuracy. For example, at one extreme, the score of a student who happened to have chosen the set of items that he/she knew the best would be his/her best possible score. Note that his/her "best" score does not necessarily mean the best reflection of his/her holistic achievement, but only his/her performance on the test in connection with the set of items he/she has chosen. However, at the other extreme, the score of a student who ended up with the set of items that he/she knew the least, for whatever reason, would be his/her lowest possible score. The "worst" score may only stand
for his/her worst performance on a particular set of items. Between these two extremes is a gray area consisting of a "hodge-podge" of various combinations. Performance aberrance due to item variations will no doubt increase error variance.

The reduced accuracy of measurement may be highly magnified by the choices of big-unit test items, such as essays whose content and tasks often vary substantially. This is especially true when the freedom of choice is substantial. As a result, it is extremely difficult to tell the relative accuracy levels of the scores of various combinations of chosen items from one group of students to another.

The second reason for the reduced accuracy of ability estimation originates from reduced variability. Since students may avoid the items they do not know, their scores on the chosen items may vary less. It is a statistical fact that as the variability of scores reduces, the reliability decreases.

What can be done to minimize the differential gains in the scores of students of different ability levels and to maintain a relatively common basis to compare students? First, if students are supposed to perform on only one choice item, limit all choice items to the core content domain of a subject. There is no need to offer choice items of different content domains. Second, if varied content is required by a test, the solution is to pair different choice items of similar content together to prevent students from deliberately avoiding knowledge deficiency. Of course, all the items should be constructed with similar difficulty and complexity.
How to Equate Scores on the Choice items of Different Content

According to the content analyses by the eight AP chemistry teachers in Hawaii, the five essays in Part D of the 1992 AP Chemistry Exam address five different areas of chemistry -- three essays addressing what can be called "core" chemistry, such as molarity, stoichiometry and energy transfer, while the other two involve lab-experimentation and nuclear chemistry. In face of these essentially different essays, one would naturally ask "Is it legitimate to treat the scores on different topics as if they meant the same thing?"

There seem to be two kinds of attitudes towards this problem, depending on the assumptions one is willing to make. The most conservative attitude is No, since different topics reflect different expertise. Although they share some common knowledge and basic principles, these different essay items are more different than similar. For example, if nuclear chemistry were similar to general chemistry, there would not be such a special field of nuclear chemistry. Furthermore, these five chemistry topics are taught in a hierarchical fashion. General chemistry is the foundation for nuclear chemistry which is taught much later than general chemistry topics.

The second attitude seems to be a more liberal one. Different content dimensionality does not necessarily mean different psychometric dimensionality (Wilson, Wood & Gibbons, 1987). It is routine to equate dozens of MC items on a test that span a wide range of content domains. For example, as described in Chapter IV, the 75 multiple choice items in Section I of the 1989 AP Chemistry Exam spanned the complete range of AP chemistry curriculum and they are still shown to demonstrate unidimensionality. The difference between an essay
item and a MC item is that the content is magnified for an essay item because of its format. Whether or not essay test data are multidimensional should not be determined solely by their content, but estimated empirically.

Furthermore, as long as students have studied all the topics, the problem of topical differences is no longer a problem. Assuming that every student would choose items that they feel strong about, the differences in the scores of various chosen items would reflect relative by different academic strengths. As long as the differential difficulties of items are adjusted, the problem of equating scores becomes less serious.

The position of this author is a combination of the two attitudes. On one hand, it is prudent to examine how different choice items are in terms of content and task involvement, and on the other, it is necessary to empirically assess the psychometric dimensions of the choice items. When it has been agreed by experts that there is a continuation of a latent knowledge or competence running through all the choice items and when it has been empirically shown that all of the choice items form a uniform psychometric dimension, then equating can be carried out.
Appendix A

ADVANCED PLACEMENT
CHEMISTRY

SURVEY AND TEST KIT

General Instructions: This instrument contains BOTH survey and test questions. It is NOT an ordinary multiple choice test. It is IMPORTANT that you read the instructions for each section carefully in order to provide accurate answers. All the questions are numbered continuously. Please indicate your answer by completely filling in the corresponding oval on the supplied answer sheet. Give only one answer to each question.

Please note: Advanced Placement Chemistry is abbreviated as AP Chemistry throughout this kit.

Please do NOT make marks in this kit!

Please TURN IN this kit with your answer sheet!

THANK YOU VERY MUCH FOR YOUR COOPERATION!
Part A: General Information

Instructions: Please take out the answer sheet supplied to you and answer the following questions to the best of your knowledge. You may ignore the optional questions. Please note that the answers to the first five questions go on the left half of the answer sheet.

I. Your name (Optional)

II. Sex

(A) Yes  (B) No

III. Grade (Note: For college students, Freshman = 13, Sophomore = 14, Junior = 15, Senior = 16)

(A) Yes  (B) No

IV. Birthday (Optional)

(A) Yes  (B) No

V. ID Number (SS#)

(A) Yes  (B) No

VI. Special Code: With which of the following groups do you most closely identify?

(A) Caucasian  (B) Japanese  (C) Filipino  (D) Hawaiian & Part-Hawaiian  
(E) Chinese  (F) Samoan  (G) Portuguese/Spanish  (H) Black  (I) Other Asian  
(J) Others

(A) Yes  (B) No

If the answer to Question 3 is Yes, do you think it is a good test to reflect your chemistry ability?

(A) Yes  (B) No

Otherwise, mark Oval (E).

1. Have you ever taken an AP Chemistry course?

(A) Yes  (B) No

2. Do you consider an AP Chemistry course useful?

(A) Yes  (B) No

3. Have you taken an AP Chemistry test?

(A) Yes  (B) No

4. If the answer to Question 3 is Yes, do you think it is a good test to reflect your chemistry ability?

(A) Yes  (B) No

5. Do you plan to major in Chemistry in College?

(A) Yes  (B) No

6. Do you plan to obtain a Bachelor of Science degree in College?

(A) Yes  (B) No

7. Do you consider chemistry one of your favorite academic subjects?

(A) Yes  (B) No

8. Do you consider chemistry one of your strongest academic subjects?

(A) Yes  (B) No

9. How many high school chemistry courses have you taken?

(A) 1  (B) 2  (C) 3  (D) 4  (E) 5 or more

10. How many college chemistry courses have you taken?

(A) 0  (B) 1  (C) 2  (D) 3  (E) 4  (F) 5
Part B: Mini AP Chemistry Test

Instructions: For the following 11 multiple choice problems (11-21), please choose the most appropriate answer. For all problems involving solutions and/or chemical equations, assume that the system is in water and at room temperature unless otherwise stated. Feel free to use the attached tables whenever necessary.

11. Which is the most electronegative element?
   (A) O  (B) La  (C) Rb  (D) Mg  (E) N

12. Which is used to explain the fact that the four bonds in methane are equivalent
   (A) Hydrogen bonding  (B) Hybridization  (C) Ionic bonding  (D) Resonance  (E) van der Waals forces (London dispersion forces)

13. \[2 \text{K} + 2 \text{H}_2\text{O} \rightarrow 2 \text{K}^+ + 2 \text{OH}^- + \text{H}_2\]
    When 0.400 mole of potassium reacts with excess water at standard temperature and pressure as shown in the equation above, the volume of hydrogen gas produced is
    (A) 1.12 liters  (B) 2.24 liters  (C) 3.36 liters  (D) 4.48 liters  (E) 6.72 liters

14. The SbCl₅ molecule has a trigonal bipyramid structure. Therefore, the hybridization of Sb orbitals should be
    (A) \(sp^2\)  (B) \(sp^3\)  (C) \(dsp^2\)
    (D) \(dsp^3\)  (E) \(d^2sp^3\)

15. A white solid is observed to be insoluble in water, insoluble in excess ammonia solution, and soluble in dilute HCl. Which of the following compounds could the solid be?
    (A) \(\text{CaCO}_3\)  (B) \(\text{BaSO}_4\)
    (C) \(\text{Pb(NO}_3)_2\)  (D) \(\text{AgCl}\)
    (E) \(\text{Zn(OH)}_2\)

16. The solubility of CuI is \(2 \times 10^{-6}\) molar. What is the solubility product constant, \(K_{sp}\), for CuI?
    (A) \(1.4 \times 10^{-5}\)  (B) \(2 \times 10^{-6}\)
    (C) \(4 \times 10^{-12}\)  (D) \(2 \times 10^{-12}\)
    (E) \(8 \times 10^{-18}\)

17. A measured mass of an unreactive metal was dropped into a small graduated cylinder half filled with water. The following measurements were made.
    Mass of metal = 19.611 grams
    Volume of water before addition of metal = 12.4 milliliters
    Volume of water after addition of metal = 14.9 milliliters
    The density of the metal should be reported as
    (A) 7.8444 grams per mL  (B) 7.844 grams per mL  (C) 7.84 grams per mL
    (D) 7.8 grams per mL  (E) 8 grams per mL
18. $\text{H}_2\text{PO}_4^-$ + $\text{HBO}_3^{2-} \leftrightarrow \text{HPO}_4^{2-} + \text{H}_2\text{BO}_3^-$

The equilibrium constant for the reaction represented by the equation above is greater than 1.0. Which of the following gives the correct relative strengths of the acids and bases in the reaction?

**Acids** | **Bases**
--- | ---
(A) $\text{H}_2\text{PO}_4^-$ > $\text{H}_2\text{BO}_3^-$ and $\text{HBO}_3^{2-}$ > $\text{HPO}_4^{2-}$
(B) $\text{H}_2\text{PO}_4^-$ > $\text{H}_2\text{BO}_4^-$ and $\text{HBO}_3^{2-}$ > $\text{HPO}_4^{2-}$
(C) $\text{H}_2\text{PO}_4^+$ > $\text{H}_2\text{BO}_3^-$ and $\text{HBO}_3^{2-}$ > $\text{HPO}_4^{2-}$
(D) $\text{H}_2\text{PO}_4^+$ > $\text{H}_2\text{BO}_4^-$ and $\text{HBO}_3^{2-}$ > $\text{HPO}_4^{2-}$
(E) $\text{H}_2\text{PO}_4^+$ = $\text{H}_2\text{BO}_3^-$ and $\text{HBO}_3^{2-}$ = $\text{HPO}_4^{2-}$

19. (CH$_3$)$_3$CCl(aq) + OH$^-$ → (CH$_3$)$_3$COH(aq) + Cl$^-$

For the reaction represented above, the experimental rate law is given as follows.

$$\text{Rate} = k[(\text{CH}_3)_3\text{CCl}]$$

If some solid sodium hydroxide is added to a solution that is 0.010-molar in (CH$_3$)$_3$CCl and 0.10-molar in NaOH, which of the following is true? (Assume the temperature and volume remain constant.)

(A) Both the reaction rate and $k$ increase.
(B) Both the reaction rate and $k$ decrease.
(C) Both the reaction rate and $k$ remain the same.
(D) The reaction rate increases but $k$ remains the same.
(E) The reaction rate decreases but $k$ remains the same.

20. $\text{MnS(s)}$ + $2\text{H}^+$ $\leftrightarrow$ $\text{Mn}^{2+}$ + $\text{H}_2\text{S(g)}$

At 25°C the solubility product constant, $K_{sp}$, for MnS is $5 \times 10^{-15}$ and the acid dissociation constants $K_1$ and $K_2$, for H$_2$S are $1 \times 10^{-7}$ and $1 \times 10^{-13}$, respectively. What is the equilibrium constant for the reaction represented by the equation above at 25°C?

21. $\text{H}_2\text{O(s)}$ $\rightarrow$ $\text{H}_2\text{O(l)}$

When ice melts at its normal melting point, 273.16k and 1 atmosphere, which of the following is true for the process shown above?

(A) $\Delta H < 0$, $\Delta S > 0$, $\Delta V > 0$
(B) $\Delta H < 0$, $\Delta S < 0$, $\Delta V > 0$
(C) $\Delta H > 0$, $\Delta S < 0$, $\Delta V < 0$
(D) $\Delta H > 0$, $\Delta S > 0$, $\Delta V > 0$
(E) $\Delta H > 0$, $\Delta S > 0$, $\Delta V < 0$
Part C: AP Chemistry
Item Comparison and Performance

Instructions: In this section, there are four sets of comparisons of AP Chemistry problems and some questions about them. Please first read through the AP Chemistry problems and then answer the questions in the order they are presented.

Comparison I

Instructions: Please first examine Comparison Problems I and II below, and then answer questions 22-29 about them.

I. \( \text{Cr}_2O_7^{2-} \ldots e^- \ldots H^+ \rightarrow \ldots \text{Cr}^{3+} \ldots H_2O(l) \)

When the equation for the half reaction above is balanced with the lowest whole-number coefficients, the coefficient for H\(_2\)O is

(A) 2   (B) 4   (C) 6
(D) 7   (E) 14

II. A solution of toluene (molecular weight 92.1) in benzene (molecular weight 78.1) is prepared. The mole fraction of toluene in the solution is 0.100. What is the molality of the solution?

(A) 0.100 m   (B) 0.703 m
(C) 0.921 m   (D) 1.28 m
(E) 1.42 m

22. To what extent have you studied a problem like Problem I?

(Never) A B C D E (Extensively)

23. To what extent have you studied a problem like Problem II?

(Never) A B C D E (Extensively)

24. How similar are these two questions in content?

(A) Not at all similar
(B) Moderately similar
(C) Extremely similar

25. Which of the above two questions seems easier for you?

(A) Problem I   (B) Problem II

26. If you were allowed to choose between these two questions, which one would you choose?

(A) Problem I   (B) Problem II

27. Choose the correct answer to Problem I: (A) (B) (C) (D) (E)

28. Choose the correct answer to Problem II: (A) (B) (C) (D) (E)

29. If you would like to choose between these two questions again after you have tried to answer them, which one would you choose now?

(A) Problem I   (B) Problem II
Comparison II

Instructions: Please first examine Comparison Problems I and II below, and then answer questions 30-37 about them.

I. Which of the following is probably true for a solid solute with a highly endothermic heat of solution when dissolved in water?

(A) The solid has a low lattice energy.
(B) As the solute dissolves, the temperature of the solution increases.
(C) The resulting solution is ideal.
(D) The solid is more soluble at higher temperatures.
(E) The solid has a high energy of hydration.

II. A 2.00-liter sample of nitrogen gas at 27°C and 600. millimeters of mercury is heated until it occupies a volume of 5.00 liters. If the pressure remains unchanged, the final temperature of the gas is

(A) 68°C  (B) 120.°C
(C) 477°C  (D) 677°C
(E) 950.°C

30. To what extent have you studied a problem like Problem I?

(Never) A B C D E (Extensively)

31. To what extent have you studied a problem like Problem II?

(Never) A B C D E (Extensively)

32. How similar are these two questions in content?

(A) Not at all similar
(B) Moderately similar
(C) Extremely similar

33. Which of the above two questions seems easier for you?

(A) Problem I  (B) Problem II

34. If you were allowed to choose between these two questions, which one would you choose?

(A) Problem I  (B) Problem II

35. Choose the correct answer to Problem I: (A) (B) (C) (D) (E)

36. Choose the correct answer to Problem II: (A) (B) (C) (D) (E)

37. If you would like to choose between these two questions again after you have tried to answer them, which one would you choose now?

(A) Problem I  (B) Problem II
Comparison III

Instructions: Please first examine Comparison Problems I and II below, and then answer questions 38-45 about them.

I. Equal volumes of 0.10-molar \( \text{H}_3\text{PO}_4 \) and 0.20-molar KOH are mixed. After equilibrium is established, the type of ion in solution in largest concentration, other than the \( \text{K}^+ \) ion, is

\[
\begin{align*}
\text{(A) } & \text{H}_3\text{PO}_4^- \\
\text{(B) } & \text{HPO}_4^{2-} \\
\text{(C) } & \text{PO}_4^{3-} \\
\text{(D) } & \text{OH}^- \\
\text{(E) } & \text{HPO}_4^+ 
\end{align*}
\]

II. Adding water to some chemicals can be dangerous because large amounts of heat are liberated. Which of the following does NOT liberate heat when water is added to it?

\[
\begin{align*}
\text{(A) } & \text{KNO}_3 \\
\text{(B) } & \text{NaOH} \\
\text{(C) } & \text{CaO} \\
\text{(D) } & \text{H}_2\text{SO}_4 \\
\text{(E) } & \text{Na} 
\end{align*}
\]

38. To what extent have you studied a problem like Problem I?

(Never) A B C D E (Extensively)

39. To what extent have you studied a problem like Problem II?

(Never) A B C D E (Extensively)

40. How similar are these two questions in content?

(A) Not at all similar
(B) Moderately similar
(C) Extremely similar

41. Which of the above two questions seems easier for you?

(A) Problem I (B) Problem II

42. If you were allowed to choose between these two questions, which one would you choose?

(A) Problem I (B) Problem II

43. Choose the correct answer to Problem I: (A) (B) (C) (D) (E)

44. Choose the correct answer to Problem II: (A) (B) (C) (D) (E)

45. If you would like to choose between these two questions again after you have tried to answer them, which one would you choose now?

(A) Problem I (B) Problem II
Comparison IV

**Instructions:** Please first examine Comparison Problems I, II, III and IV below, and then answer questions 46-69 about them.

I. The weight of $\text{H}_2\text{SO}_4$ (molecular weight 98.1) in 50.0 milliliters of a 6.00-molar solution is

- (A) 3.10 grams
- (B) 12.0 grams
- (C) 29.4 grams
- (D) 294 grams
- (E) 300. grams

II. If a copper sample containing some zinc impurity is to be purified by electrolysis, the anode and the cathode must be which of the following?

- Anode
  - (A) Pure copper
  - (B) Pure zinc
  - (C) Pure copper
  - (D) Impure copper sample
  - (E) Impure copper sample

- Cathode
  - Pure zinc
  - Pure copper
  - Impure copper sample
  - Pure Copper

III. A 0.20-molar solution of a weak monoprotic acid, $\text{HA}$, has a pH of 3.00. The ionization constant of this acid is

- (A) $5.0 \times 10^{-7}$
- (B) $2.0 \times 10^{-7}$
- (C) $5.0 \times 10^{-6}$
- (D) $5.0 \times 10^{-3}$
- (E) $2.0 \times 10^{3}$

IV. For the substance represented in the diagram on the left, which of the phases is most dense and which is least dense at -15°C?

- Most Dense
  - (A) Solid
  - (B) Solid
  - (C) Liquid
  - (D) Liquid
  - (E) The diagram gives no information about densities.

- Least Dense
  - Gas
  - Liquid
  - Solid
  - Gas

212
46. To what extent have you studied a problem like Problem I?
   (Never) A B C D E (Extensively)

47. To what extent have you studied a problem like Problem II?
   (Never) A B C D E (Extensively)

48. To what extent have you studied a problem like Problem III?
   (Never) A B C D E (Extensively)

49. To what extent have you studied a problem like Problem IV?
   (Never) A B C D E (Extensively)

50. How similar are Problem I and II in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

51. Which one seems easier for you?
    (A) Problem I
    (B) Problem II

52. How similar are Problem I and III in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

53. Which one seems easier for you?
    (A) Problem I
    (B) Problem III

54. How similar are Problem I and IV in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

55. Which one seems easier for you?
    (A) Problem I
    (B) Problem IV

56. How similar are Problem II and III in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

57. Which one seems easier for you?
    (A) Problem II
    (B) Problem III

58. How similar are Problem II and IV in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

59. Which one seems easier for you?
    (A) Problem II
    (B) Problem IV
60. How similar are Problem III and IV in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

61. Which one seems easier for you?
   (A) Problem III
   (B) Problem IV

62. How similar are Problem III and IV in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

63. Which one seems easier for you?
   (A) Problem III
   (B) Problem IV

64. If you were allowed to choose three questions to answer, which three questions would you like to choose?
   (A) Problem I II III
   (B) Problem I II IV
   (C) Problem I III IV
   (D) Problem II III IV

65. Choose the correct answer to Problem I:
   (A) (B) (C) (D) (E)

66. Choose the correct answer to Problem II:
   (A) (B) (C) (D) (E)

67. Choose the correct answer to Problem III:
   (A) (B) (C) (D) (E)

68. Choose the correct answer to Problem IV:
   (A) (B) (C) (D) (E)

69. If you would like to choose three questions again after you have tried to answer the four questions, which three questions would you like to choose now?
   (A) Problem I II III
   (B) Problem I II IV
   (C) Problem I III IV
   (D) Problem II III IV
Part D: AP Chemistry
Essay Problem Comparisons

Instructions: In this section, there are five essay-type problems. Please first read through all five essays, and then answer the questions 70-95 about them.

Essay I: \[ \text{CF}_4, \text{XeF}_4, \text{ClF}_3 \]
(a) Draw a Lewis electron-dot structure for each of the molecules above and identify the shape of each.
(b) Use the valence shell electron-pair repulsion (VSEPR) model to explain the geometry of each of these molecules.

Essay II: The melting points of the alkali metals decrease from Li to Cs. In contrast, the melting points of the halogens increase from F\textsubscript{2} to I\textsubscript{2}.
(a) Using bonding principles, account for the decrease in the melting points of the alkali metals.
(b) Using bonding principles, account for the increase in the melting points of the halogens.
(c) What is the expected trend in the melting points of the compounds LiF, NaCl, KBr, and CsI? Explain this trend using bonding principles.

Essay III: \[ \text{C}_2\text{H}_4(g) + \text{H}_2(g) \rightarrow \text{C}_2\text{H}_6(g) \quad \Delta H^\circ = -137\text{kJ} \]

Account for the following observations regarding the exothermic reaction represented by the equation above.
(a) An increase in the pressure of the reactants causes an increase in the reaction rate.
(b) A small increase in temperature causes a large increase in the reaction rate.
(c) The presence of metallic nickel causes an increase in reaction rate.
(d) The presence of powdered nickel causes a larger increase in reaction rate than does the presence of a single piece of nickel of the same mass.
**Essay IV:** The carbon isotope of mass 12 is stable. The carbon isotopes of mass 11 and mass 14 are unstable. However, the type of radioactive decay is different for these two isotopes. Carbon-12 is not produced in either case.

(a) Identify a type of decay expected for carbon-11 and write the balanced nuclear reaction for that decay process.

(b) Identify the type of decay expected for carbon-14 and write the balanced nuclear reaction for that decay process.

(c) Gamma rays are observed during the radioactive decay of carbon-11. Why is it unnecessary to include the gamma rays in the radioactive decay equation of (a)?

(d) Explain how the amount of carbon-14 in a piece of wood can be used to determine when the tree died.

---

**Essay V:** Consider three unlabeled bottles, each containing small pieces of one of the following metals.

- Magnesium
- Sodium
- Silver

The following reagents are used for identifying the metals.

- Pure water
- A solution of 1.0-molar HCl
- A solution of concentrated HNO₃

(a) Which metal can be easily identified because it is much softer than the other two? Describe a chemical test that distinguishes this metal from the other two, using only one of the reagents above. Write a balanced chemical equation for the reaction that occurs.

(b) One of the other two metals reacts readily with the HCl solution. Identify the metal and write the balanced chemical equation for the reaction that occurs when this metal is added to the HCl solution. Use the attached table of standard reduction potentials to account for the fact that this metal reacts with HCl while the other does not.

(c) The one remaining metal reacts with the concentrated HNO₃ solution. Write a balanced chemical equation for the reaction that occurs.

(d) The solution obtained in (c) is diluted and a few drops of 1 M HCl is added. Describe what would be observed. Write a balanced chemical equation for the reaction that occurs.

---

216
70. To what extent have you studied a problem like Essay I?
   (Never) A B C D E (Extensively)

71. To what extent have you studied a problem like Essay II?
   (Never) A B C D E (Extensively)

72. To what extent have you studied a problem like Essay III?
   (Never) A B C D E (Extensively)

73. To what extent have you studied a problem like Essay IV?
   (Never) A B C D E (Extensively)

74. To what extent have you studied a problem like Essay V?
   (Never) A B C D E (Extensively)

75. How similar are Essay I and II in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

76. Which one seems easier for you?
   (A) Essay I (B) Essay II

77. How similar are Essay I and III in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

78. Which one seems easier for you?
   (A) Essay I (B) Essay III

79. How similar are Essay I and IV in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

80. Which one seems easier for you?
   (A) Essay I (B) Essay IV

81. How similar are Essay I and V in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

82. Which one seems easier for you?
   (A) Essay I (B) Essay V

83. How similar are Essay II and III in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar
84. Which one seems easier for you?
   (A) Essay II  (B) Essay III

85. How similar are Essay II and IV in content?
   (A) Not at all similar  (B) Moderately similar  (C) Extremely similar

86. Which one seems easier for you?
   (A) Essay II  (B) Essay IV

87. How similar are Essay II and V in content?
   (A) Not at all similar  (B) Moderately similar  (C) Extremely similar

88. Which one seems easier for you?
   (A) Essay II  (B) Essay V

89. How similar are Essay III and IV in content?
   (A) Not at all similar  (B) Moderately similar  (C) Extremely similar

90. Which one seems easier for you?
   (A) Essay III  (B) Essay IV

91. How similar are Essay III and V in content?
   (A) Not at all similar  (B) Moderately similar  (C) Extremely similar

92. Which one seems easier for you?
   (A) Essay III  (B) Essay V

93. How similar are Essay IV and V in content?
   (A) Not at all similar  (B) Moderately similar  (C) Extremely similar

94. Which one seems easier for you?
   (A) Essay IV  (B) Essay V

95. If you were allowed to choose three out of the five essays, which of the following combination would you choose?
   (A) = Essay I II III
   (B) = Essay I II IV
   (C) = Essay I II V
   (D) = Essay I III IV
   (E) = Essay I III V
   (F) = Essay I IV V
   (G) = Essay II III IV
   (H) = Essay II III V
   (I) = Essay II IV V
   (J) = Essay III IV V

   The End
Appendix B

TEACHER EVALUATION SHEET

Directions: The questions in this sheet are to ask the teacher to assess the extent to which he or she has taught the various chemistry subjects represented by the twenty one multiple choice questions and five essay problems in the Advanced Placement Chemistry Survey and Test Kit. The rating scale ranges from Never (A), through Somewhat (C), to Extensively (E).

**PART B: MINI AP CHEMISTRY TEST**

1. To what extent have you ever taught a problem like Problem 11?
   
   (Never) A B C D E (Extensively)

2. To what extent have you ever taught a problem like Problem 12?
   
   (Never) A B C D E (Extensively)

3. To what extent have you ever taught a problem like Problem 13?
   
   (Never) A B C D E (Extensively)

4. To what extent have you ever taught a problem like Problem 14?
   
   (Never) A B C D E (Extensively)

5. To what extent have you ever taught a problem like Problem 15?
   
   (Never) A B C D E (Extensively)

6. To what extent have you ever taught a problem like Problem 16?
   
   (Never) A B C D E (Extensively)

7. To what extent have you ever taught a problem like Problem 17?
   
   (Never) A B C D E (Extensively)

8. To what extent have you ever taught a problem like Problem 18?
   
   (Never) A B C D E (Extensively)

9. To what extent have you ever taught a problem like Problem 19?
   
   (Never) A B C D E (Extensively)

10. To what extent have you ever taught a problem like Problem 20?

    (Never) A B C D E (Extensively)

11. To what extent have you ever taught a problem like Problem 21?

    (Never) A B C D E (Extensively)
PART C: COMPARISON I

12. To what extent have you ever taught a problem like Problem I?

(Never) A B C D E (Extensively)

13. To what extent have you ever taught a problem like Problem II?

(Never) A B C D E (Extensively)

14. How similar are these two questions in content?

(A) Not at all similar
(B) Moderately similar
(C) Extremely similar

15. Which of the two questions seems easier for your students?

(A) Problem I  (B) Problem II

16. If your students were allowed to choose between these two questions, which one do you think that most of your students would probably choose?

(A) Problem I  (B) Problem II

PART C: COMPARISON II

17. To what extent have you ever taught a problem like Problem I?

(Never) A B C D E (Extensively)

18. To what extent have you ever taught a problem like Problem II?

(Never) A B C D E (Extensively)

19. How similar are these two questions in content?

(A) Not at all similar
(B) Moderately similar
(C) Extremely similar

20. Which of the two questions seems easier for your students?

(A) Problem I  (B) Problem II

21. If your students were allowed to choose between these two questions, which one do you think most of your students would probably choose?

(A) Problem I  (B) Problem II

PART C: COMPARISON III

22. To what extent have you ever taught a problem like Problem I?

(Never) A B C D E (Extensively)

23. To what extent have you ever taught a problem like Problem II?

(Never) A B C D E (Extensively)
24. How similar are these two questions in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

25. Which of the two questions seems easier for your students?
   (A) Problem I  (B) Problem II

26. If your students were allowed to choose between these two questions, which one do you think most of your students would probably choose?
   (A) Problem I  (B) Problem II

27. To what extent have you ever taught a problem like Problem I?
   (Never) A B C D E (Extensively)

28. To what extent have you ever taught a problem like Problem II?
   (Never) A B C D E (Extensively)

29. To what extent have you ever taught a problem like Problem III?
   (Never) A B C D E (Extensively)

30. To what extent have you ever taught a problem like Problem IV?
   (Never) A B C D E (Extensively)

31. How similar are Problem I and II in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

32. Which one seems easier for your students?
   (A) Problem I  (B) Problem II

33. How similar are Problem I and III in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

34. Which one seems easier for your students?
   (A) Problem I  (B) Problem III

35. How similar are Problem I and IV in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

36. Which one seems easier for your students?
   (A) Problem I  (B) Problem IV

37. How similar are Problem II and III in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar
38. Which one seems easier for your students?
   (A) Problem II
   (B) Problem III

39. How similar are Problem II and IV in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

40. Which one seems easier for your students?
   (A) Problem II
   (B) Problem IV

41. How similar are Problem III and IV in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

42. Which one seems easier for your students?
   (A) Problem III
   (B) Problem IV

43. How similar are Problem III and IV in content?
   (A) Not at all similar
   (B) Moderately similar
   (C) Extremely similar

44. Which one seems easier for your students?
   (A) Problem III
   (B) Problem IV

45. If your students were allowed to choose three questions to answer, which three questions do you think that they would like to choose?
   (A) = Problem I II III
   (B) = Problem I II IV
   (C) = Problem I III IV
   (D) = Problem II III IV

46. To what extent have you ever taught a problem like Essay I?
   (Never) A B C D E (Extensively)

47. To what extent have you ever taught a problem like Essay II?
   (Never) A B C D E (Extensively)

48. To what extent have you ever taught a problem like Essay III?
   (Never) A B C D E (Extensively)

49. To what extent have you ever taught a problem like Problem IV?
   (Never) A B C D E (Extensively)

50. To what extent have you ever taught a problem like Problem V?
   (Never) A B C D E (Extensively)
<table>
<thead>
<tr>
<th>Question</th>
<th>Content Comparison</th>
<th>Similarity</th>
<th>Easier for Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>51.</td>
<td>Essay I and II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>52.</td>
<td></td>
<td>(A) Not at all similar</td>
<td>Essay I</td>
</tr>
<tr>
<td>53.</td>
<td>Essay I and III</td>
<td></td>
<td>Essay III</td>
</tr>
<tr>
<td>54.</td>
<td></td>
<td>(A) Not at all similar</td>
<td>Essay II</td>
</tr>
<tr>
<td>55.</td>
<td>Essay I and IV</td>
<td></td>
<td>Essay IV</td>
</tr>
<tr>
<td>56.</td>
<td></td>
<td>(A) Not at all similar</td>
<td>Essay III</td>
</tr>
<tr>
<td>57.</td>
<td>Essay I and V</td>
<td></td>
<td>Essay V</td>
</tr>
</tbody>
</table>

223
65. How similar are Essay III and IV in content?

(A) Not at all similar
(B) Moderately similar
(C) Extremely similar

66. Which one seems easier for your students?

(A) Essay III  (B) Essay V

67. How similar are Essay III and V in content?

(A) Not at all similar
(B) Moderately similar
(C) Extremely similar

68. Which one seems easier for your students?

(A) Essay III  (B) Essay V

69. How similar are Essay IV and V in content?

(A) Not at all similar
(B) Moderately similar
(C) Extremely similar

70. Which one seems easier for your students?

(A) Essay IV  (B) Essay V

71. If you were allowed to choose three out of the five essays, which of the following combination would you choose?

(A) =Essay I II III
(B) =Essay I II IV
(C) =Essay I II V
(D) =Essay I III IV
(E) =Essay I III V
(F) =Essay I IV V
(G) =Essay II III IV
(H) =Essay II III V
(I) =Essay II IV V
(J) =Essay III IV V

Thank you so much for your support!
APPENDIX C

CHEMISTRY TEACHER EXPERT JUDGMENT SURVEY

TASK 1: Rank the Five Essays

Introduction: 1989 AP Chemistry Examination has five essays questions in Part D, Labelled as Essays 5, 6, 7, 8 & 9 (See Attached 1).

Instructions Please rank the five essays in terms of their relative easiness and your preferences to answer from (1) being the easiest essay to answer to (5) the most difficult essay to answer and rate Questions 2-6:

Q1: Your Rank: (1) (2) (3) (4) (5)

Q2: How likely do AP Chemistry textbooks address such a problem as reflected by Essay 5?

(Hardly) 1 2 3 4 5 (Extensively)

Q3: How likely do AP Chemistry textbooks address such a problem as reflected by Essay 6?

(Hardly) 1 2 3 4 5 (Extensively)

Q4: How likely do AP Chemistry textbooks address such a problem as reflected by Essay 7?

(Hardly) 1 2 3 4 5 (Extensively)

Q5: How likely do AP Chemistry textbooks address such a problem as reflected by Essay 8?

(Hardly) 1 2 3 4 5 (Extensively)

Q6: How likely do AP Chemistry textbooks address such a problem as reflected by Essay 9?

(Hardly) 1 2 3 4 5 (Extensively)

225
Task 2: Qualitative Descriptions of Common Characteristics:

Introduction: As for the five essays, students were asked to choose three out of the five essays to answer, which resulted in 10 possible combinations. Based on the 18,462 students who took the test in 1989, I found dramatic differences in students' choices, as shown in Attached Figure 1.

Instructions: Please write about your opinions on the reasons behind such dramatic choice distributions. Please attach more paper if necessary.

Question 7: Why so many students chose essay combination 5, 6 & 8? What characteristics do they share in common?

Afterthought for Question 7 (after you work through the survey once):

Question 8: Why so many students chose essay combination 5, 7 & 8? What characteristics do they share in common?

Afterthought for Question 8 (after you work through the survey once):

Question 9: Why so many students chose essay combination 5, 6 & 7? What characteristics do they share in common?

Afterthought for Question 9 (after you work through the survey once):

Question 10: Why so few students chose essay combination 6, 7 & 9? What characteristics do they share in common?

Afterthought for Question 10 (after you work through the survey once):

Question 11: Why so a moderate number of students chose essay combination 5, 8 & 9? What characteristics do they share in common?

Afterthought for Question 11 (after you work through the survey once):

Holistic Reflections or Additional Comments on the Overall Observed Choice Patterns:

Now you have completed your survey. If you wish, you may go over the survey again and add your afterthoughts to each of questions you have answered.
APPENDIX D

ESSENTIAL CHEMISTRY KNOWLEDGE COMPONENTS
REQUIRED BY THE MULTIPLE CHOICE ITEMS
IN PART B AND PART C OF
ADVANCED PLACEMENT CHEMISTRY SURVEY AND TEST KIT

Part B: Mini AP Chemistry Test

Item 11: Periodic table and trends.
Item 12: Covalent structure, electron configuration.
Item 13: Physical behavior of gases, stoichiometry.
Item 14: Molecular geometry.
Item 15: Solubility, acid-base chemistry, conic compounds.
Item 16: Solubility.
Item 17: Density, reporting numbers on physical characteristics of elements.
Item 18: Acid-base, equilibrium.
Item 19: Reaction rates.
Item 20: Equilibrium, solubility, acid-base.
Item 21: Thermodynamics.

Part C: AP Chemistry Item Comparison and Performance

Comparison I

Item I. Oxidation-reduction. Item II. Solution in molality.

Comparison II

Item I. Solution chemistry.
Item II. Physical behavior of gases (Charles’ law).

Comparison III

Item I. Ions in solution. Item II. Water solutions.

Comparison IV

Item I. Stoichiometry & molarity in solution chemistry.
Item II. Electrochemistry.
Item III. Acid-base Chemistry.
Item IV. Energy, phase, changes and triple point.
REFERENCES


Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. Psychometrika, 37, 29-51.


230


