ANALYZING AND IMPROVING GENERAL CHEMISTRY TEACHING AND ASSESSMENT USING ITEM RESPONSE THEORY

by

KIMBERLY DAWN SCHURMEIER

(Under the Direction of Charles H. Atwood)

ABSTRACT

General Chemistry is a demanding course that challenges its participants who comprise largely freshman science majors. We have found that many of these students cite this class as the toughest they have to take in their first year, and they struggle to successfully complete the course. In this study, we employ Item Response Theory (IRT) to analyze previous computer-administered examination data and elucidate those areas of chemistry that are problematic for students. We investigate the potential for specific questions to discriminate between students' abilities and show the types of questions that will separate A and B students, B from C, etc. Additionally, it is shown that a range of these topics must be present on an examination to accurately and fairly ascribe a grade to students. To further identify difficult topics that represent a barrier for learning, we find some common misconceptions that students have about certain key concepts; without correction, these can lead to misinterpretations of theories. Specific topics are analyzed to determine for whom, and why, these topics are difficult. Armed with this analysis, the instructors modified their approach to teaching these key concepts in the 2006 and 2007 academic years. Improvement in student understanding of some of these problematic areas is accomplished.

INDEX WORDS: Chemical Education, Item Response Theory, General Chemistry Misconceptions

ANALYZING AND IMPROVING GENERAL CHEMISTRY TEACHING AND ASSESSMENT USING ITEM RESPONSE THEORY

by

KIMBERLY DAWN SCHURMEIER

B.A., Franklin College, 2004

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2009

ANALYZING AND IMPROVING GENERAL CHEMISTRY TEACHING AND ASSESSMENT USING ITEM RESPONSE THEORY

by

KIMBERLY DAWN SCHURMEIER

Major Professor:	Charles H. Atwood
Committee:	Richard W. Morrison Charles Kutal

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2009

DEDICATION

I dedicate this thesis to my parents who have always been supportive; thank you so much for everything.

ACKNOWLEDGEMENTS

I would have been unable to complete this degree without all of the help I received along the way. I would like to first thank my family for keeping me in touch with what really matters. Thank you Melissa and Eli for your wonderful visits to Athens, which gave me much-needed vacations. Thank you Mark for always reminding me that I should have fun once in a while, and that my little brother, the one who dislikes school so much, found a job before me. Kelsey you were so much younger when I moved here and you are still 12-years-old in my mind; knowing that you are now seventeen makes me feel old. I'm sorry I wasn't around as much as I would have liked; you are the best little sis I could have asked for. And thanks to Mom and Dad for always offering to help without me even asking. Throughout these five years you've always reminded me that you were proud of everything that I did— I am happy to be your favorite middle daughter.

On the school front, I would like to thank my committee, Dr. Atwood, Dr. Morrison and Dr. Kutal, for the wonderful guidance you have provided throughout my five years at UGA. My research would not be as thorough if it wasn't for all three of you pushing me and giving me support. I would also like to thank Gary Lautenschlager for always answering my questions about IRT no matter how silly they might have been. Carrie, Angie and Sonja— oh, how you made the chemistry department a fun place. Thank you for all of the wonderful support you gave me with research, classes and throughout my job search. Thank you Carrie for helping me with this research through our lengthy discussions and for being a great friend to lean on when things went a little crazy. And last but not least, Andrew, I could not have made it through this program

without you— I probably would have starved to death. Thank you for everything that you have done for me from helping me study for exams, to listening to me practice presentations and editing this lengthy dissertation.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

LIST OF TABLES x

LIST OF FIGURES xii

CHAPTER

1 INTRODUCTION 1

Importance of Chemical Education Research in General Chemistry 1

Item Response Theory 2

JExam Testing System 3

Dissertation Overview 6

Chemistry Misconceptions and Difficult Topics 6

2 ITEM RESPONSE THEORY 9

Item Response Theory - Parameters 9

Using the Item Response Theory Program – BILOG-MG 3 14

3 ANALYZING OLD EXAMINATIONS ON JEXAM USING ITEM RESPONSE THEORY 21

Examinations Given from Fall 2001 – Spring 2003 21

Examinations Given from Fall 2003 – Spring 2005 33

4 WRITING EXAMINATIONS USING IRT RESULTS 40

JExam 40

Writing a Multiple Choice Exam 47

5 ANALYSIS OF FALL 2005 – SPRING 2006 ACADEMIC YEAR EXAMINATIONS 54

IRT Analysis 54

Students' Understanding of Ions 57

Students' Understanding of Molecular Polarity 59

Students' Understanding of Quantum Numbers 61

Students' Understanding of the Terms "Strong, Weak, Concentrated and Dilute" 63

Students' Understanding of Molecular Image Problems 66

Students' Understanding of Inorganic Nomenclature 70

Students' Understanding of Mole Concepts 71

Students' Understanding of Solution Calorimetry 71

Topics that are Easy for Students 72

Importance of Wording 72

6 ANALYSIS OF FALL 2006 – SPRING 2007 ACADEMIC YEAR EXAMINATIONS 74

Comparison of Students' Abilities and IRT Analysis 74

Students' Understanding of Ions 77

Students' Understanding of Molecular Polarity 78

Students' Understanding of Quantum Numbers 83

Students' Understanding of Molecular Image Problems 84

7 ANALYSIS OF FALL 2007 – SPRING 2008 ACADEMIC YEAR EXAMINATIONS 89

IRT Analysis 89

Students' Understanding of Ions 91

Students' Understanding of Molecular Polarity 92

Students' Understanding of Quantum Numbers 94

Students' Understanding of Molecular Image Problems 95

Students' Understanding of Mole Concepts 99

Analysis of Tries on JExam Examinations 100

8 COMPARISON OF STUDENTS' ABILITIES 105

9 CONCLUSIONS 115

REFERENCES 120

LIST OF TABLES

Page

Table 2.1: KR-21 reliability for each individual examination given from the fall of 2004 to the spring of 2008 17

Table 2.2: Reliability index of academic year examinations given from the fall 2004 to the spring 2008 calculated using BILOG-MG 3 18

Table 3.1: Equivalency of question 12 in fall 2001 first semester exam 27

Table 3.2: Equivalency of question 11 in fall 2002 first semester exam 28

Table 3.3: Equivalency of question 16 in spring 2002, question 12 in fall 2001 and question 18 in spring 2003 first semesters exams 29

Table 3.4: Equivalency of question 18 in fall 2001 first semester exam 30

Table 3.5: Item response theory analysis of question 19 in spring 2002 and question 17 in fall 2003 second semester exams 32

Table 3.6: Classical test analysis of question 12 in fall 2003 and question 6 in fall 2004 first semester exams 34

Table 3.7: Item response theory analysis of question 12 in fall 2003 and question 6 in fall 2004 first semester exams 35

Table 3.8: Classical test analysis of question 12 in fall 2003 and question 6 in fall 2004 first semester exams 36

Table 3.9: Comparison of item response theory analysis before and after the introduction of JExam homeworks 37

Table 3.10: Comparing item response theory analysis before and after the introduction of JExam homeworks.....38

Table 4.1: Ability needed for each letter grade for fall 2006 – spring 2007 academic year 42

Table 5.1: Ability needed for each letter grade for fall 2005 – spring 2006 academic year 55

Table 5.2: IRT parameters for items on the 2005 – 2006 academic year examinations.....56

Table 6.1: Ability needed for each letter grade for fall 2004 – spring 2005 academic year 75

Table 6.2: Comparing parameters of items given on both 2005 and 2006 academic years 76

Table 6.3: IRT parameters for items on the 2006 – 2007 academic year examinations.....77

Table 7.1: Ability needed for each letter grade for fall 2007 – spring 2008 academic year 89

Table 7.2: IRT parameters for items on the 2007 – 2008 academic year examinations.....91

Table 7.3: IRT analysis of items with different abilities for first, second and third tries on fall 2007 first examinations.....101

Table 7.4: IRT analysis of items with increased discrimination on first, second and third tries on fall 2007 first examinations 103

LIST OF FIGURES

Page

Figure 2.1: Ideal item characteristic curve 12

Figure 2.2: Poorly discriminating item characteristic curve 13

Figure 2.3: Excerpt of data file to analyze using BILOG-MG 3 19

Figure 4.1: Regression of Ability vs. Percent Correct generated from IRT analysis of fall 2006 – spring 2007 academic year. 41

Figure 4.2: Item characteristic curve of item number 11206 44

Figure 4.3: Item information curve for item number 11206 45

Figure 4.4: Total information curve generated from the fall 2006 - spring 2007 academic year 46

Figure 4.5: Item characteristic curve of item number 2016 given in the spring of 2007 49

Figure 4.6: Item characteristic curve of item number 1045 given in the fall of 2007 51

Figure 4.7: Item characteristic curve of item number 2038 given in the spring of 2008 52

Figure 8.1: Number of students at each ability for the fall 2004 – spring 2005 academic year 105

Figure 8.2: Number of students at each ability for the fall 2005 – spring 2006 academic year 106

Figure 8.3: Number of students at each ability for the fall 2006 – spring 2007 academic year 107

Figure 8.4: Number of students at each ability for the fall 2007 – spring 2008 academic year 108

Figure 8.5: Number of students at each ability before modification occurred in the classroom 111

Figure 8.6: Number of students at each ability after modification occurred in the classroom 111

Figure 8.7: Percentage of students at each ability before and after modification occurred in the classroom 113

CHAPTER 1

INTRODUCTION

Importance of Chemical Education Research in General Chemistry

For many students, general chemistry is a difficult, career determining course. Those that excel move on to take other chemistry courses while those that do poorly frequently choose nonscientific careers. In the 2004-2005 academic year, the first general chemistry hour exam at the University of Georgia (UGA) was administered to 1430 students; whereas, the second semester final exam was administered to 882 students. 38.3% of the students dropped out or chose not to complete both semesters of general chemistry. (This calculation also assumes incorrectly that all students enrolled in the second semester course were also enrolled in the first semester course.) With these statistics, it is necessary to change the way the students are being taught so that more students succeed in the class without lowering the course standards. Some students are encountering the majority of the material for the first time, while other students are reviewing what they were taught in high school. By finding the topics that cause difficulty to students, changes can be made to the way those topics are taught in class. Hopefully with these changes, more students will succeed in general chemistry classes.

The goal of this research is to develop accurate procedures for assigning grades while simultaneously identifying chemistry topics that are difficult for specific groups of students. Item Response Theory (IRT) is used to analyze examinations at UGA. The results of the analysis will show which topics were difficult for students. The analysis can also determine if

grades at UGA are assigned accurately and consistently for the examinations given during the 2005, 2006 and 2007 academic years. By using IRT, any areas of weakness in teaching and learning in the UGA General Chemistry program can be found.

Item Response Theory

Item Response Theory (IRT) has been used in various ways to determine the intrinsic ability of students. The theory was originally designed to analyze psychological and educational abilities; however, in the last fifteen years, it has been employed to analyze students' abilities in a wide range of subjects (*1-7*). Prior to this research, IRT had not been used to analyze any questions in chemistry. By using IRT, questions can be analyzed in more detail. For example, by looking at patterns in the assigned difficulty of questions, difficult topics for students can be identified. Since most chemical educators are familiar with Classical Test Theory (CTT), the major differences between Item Response Theory (IRT) and CTT are highlighted below to help provide a better understanding of the importance of IRT. Both methods provide insight into certain aspects of assessment. CTT is usable for small sample sizes and simpler to perform. IRT is preferable for a larger sample size (preferably 200 or more) and involves choosing an IRT model, adjusting the model parameters, followed by computer iteration until the model converges on the data.

The mean, median, and Gaussian probability distribution for the test can be calculated using CTT. For each test item, CTT can also give an item discrimination factor, which is a comparison of the performance of the top quartile of students versus the bottom quartile. However, CTT has a greater dependence than IRT upon the subject group whose exams are being analyzed as well as upon the nature of the examination. For example, if the same group of

individuals were given two different assessments on the same subject using different assessment items, CTT analysis would yield different results for the two assessments. IRT analysis is independent of the individuals assessed and the assessment items used. Rather than assigning each student a percentage correct on the exam, IRT assigns each student an "ability level" that is based upon his or her responses to the assessment items. Furthermore, each item in the examination is also assigned an ability that is based upon those students with a given ability level and higher who have a high probability of correctly answering the question. In effect, after the IRT analysis is performed each test item's ability indicates how that item discriminates between students within the entire ability range.

Reliability of the examinations can be analyzed using both CTT and IRT. With CTT the reliability of the examination is analyzed by calculating Kuder-Richardson 21 (KR-21) values, which is determined relative to the test mean. With IRT, reliability for an examination can be found by producing a total information curve, which informs us how accurate the assignment of student abilities are across the entire ability scale. The reliability of the examinations discussed in the following chapters will be analyzed by calculating both the KR-21 values and total information curves (*8, 9*). A detailed explanation of the IRT models used in this analysis is discussed in Chapter 2.

JExam Testing System

At UGA, there are around 1400 students that take general chemistry each year. Because of the sheer number of students that go through the program, examinations have to be given in a manner that both correctly assesses the students accurately and affords grading in a timely manner. Written examinations are clearly not practical because of the exorbitant time required to

grade them and the potential for errors in the grading and/or data entry stages. One possible solution to this problem is to offer examinations that are administered using pencil and paper, but can be graded automatically using the Scantron system. While the Scantron system is currently used for the end-of-semester examinations, it is severely limited because only multiple-choice questions can be used.

To address these problems, the JExam computerized examination system was written for examinations at UGA (*10-12*). Written in the Java scripting language, JExam comprises a database of questions, which are delivered by a server that can be accessed remotely via an internet connection. With this design, students can answer homework questions at their convenience, and examination rooms can be easily established using standard desktop PCs for the purpose of delivering the examinations. All that is required to access the JExam system is an internet connection and a small piece of software that is freely distributed by the UGA chemistry department. Furthermore, many types of question can be delivered with this software, making it much more flexible than the Scantron system. As mentioned previously, JExam contains a database of questions that are entered by the instructors; this currently contains over 12,000 questions.

On a written examination, partial credit is commonly awarded to students who use the correct methodology but make a simple mistake that leads to an incorrect answer. With an automated grading system, this is difficult to implement, so the JExam system offers multiple attempts to students who do not answer questions correctly on the first try. For homework questions, students are given three attempts to answer a question correctly and students will receive full credit for answering it correctly regardless of the number of attempts needed. On the examinations, students receive full credit for a question if they answer it correctly on the first try,

50% credit for a question if they answer it correctly on the second try and 25% credit if they answer it correctly on the third try. Therefore, if a student understands the underlying principles of a given question but makes an innocent error in the calculation, they will be told that their answer is incorrect and will be able to re-work it to receive 50% credit. It should be noted that questions requiring a numerical answer are constructed with a tolerance level to account for round-off errors in students' responses.

Around 40 students currently take the examinations at a given time and each exam comprises 20-30 questions. To prevent cheating, variations of each question (referred to as items) are written. The JExam program randomly chooses which item a given student will receive to make the exams as unique as is reasonably possible. This randomization would be difficult to implement for written examinations where there is always a danger of one of the early exam participants removing an exam from the testing center and circulating it.

Regarding this study, the most important feature of JExam is the grading mechanism. The examinations are graded real-time so that students can see which responses were incorrect and remedy them if needed. The students' responses for each try are logged in a database and can be used to verify a student's grade in case of discrepancies. Furthermore, if a mistake in a question is discovered after the conclusion of the examination, re-grading of the examination or awarding of credit can be performed retroactively, as appropriate. The database of student responses is crucial for the IRT analysis discussed herein. The data output by JExam is easily reformatted to make them compatible with the BILOG program, which performs the IRT analyses.

Dissertation Overview

Using IRT, students' responses to computerized examinations given with JExam are analyzed, and questions that separate students based upon their varying levels of chemical understanding are elucidated. Chapter 3 discusses the IRT analysis of the examinations given on JExam from the fall of 2001 to the spring of 2005. The questions on these examinations were assessed to determine equivalency so that non-equivalent questions could be removed from future examinations (*5, 13-15*).

The IRT results from these previous examinations facilitated the exam writing ability of both new and experienced instructors in large lecture classes. Questions were analyzed by not just looking at what percent of the students answered the questions correctly, but if the questions are "good" questions. This is possible because IRT can easily show if a question does not fit the IRT model. The most common reason for a question poorly fitting the IRT model is poor question wording (*13*). Once good questions are selected these questions are used to help write examinations that accurately assign student grades. Chapter 4 discusses how to use the initial IRT analysis to write examinations that thoroughly determine the student's knowledge of chemistry. Chapter 4 contains a detailed method for writing examinations both for JExam, a computer program utilized to give examinations, and multiple choice format exams.

Chemistry Misconceptions and Difficult Topics

Chemistry misconceptions are held by undergraduate and graduate students as well as by some chemistry teachers (*16-20*). Many of these misconceptions have been identified, and it has been shown that it is difficult for a student holding the misconception to overcome it (*21, 22*). Some of the common misconceptions identified previously include: bonding, compounds in

solution, equilibrium and energy (*16, 17, 23-26*). Other topics with which it has been found that students struggle include: quantum numbers, geometries, chemistry on a molecular level and nomenclature (*17, 27-31*). The difference between a difficult topic and a misconception is that a difficult topic is something that is hard for students to learn and understand thoroughly, whereas a misconception occurs when a student believes they understand a concept, but what they think is correct is actually false.

It takes years of experience, along with many tests, for teachers to start seeing patterns in the topics that are difficult for students to learn. One of our research goals was to determine some areas of a typical general chemistry curriculum that confuse a majority of students, and thus, require greater classroom emphasis. In pre-existing research, most of the above mentioned difficult topics and misconceptions were found by looking at a small sample, usually around 100 students or less (*17, 18, 23, 31-41*). Our research uses a much larger sample, around 1200 students each academic year, thus giving us more information about topics that are difficult. In addition to determining topics that are difficult for all chemistry students at UGA, we use IRT to determine which portions of our student population found these topics to be difficult. Chapters 5, 6 and 7 discuss the misconceptions and difficult topics that UGA students have. These topics are found by looking at specific questions on examinations and then determining which students, whose ability levels have been calculated, answered the question correctly.

This research also looked at what point in a given academic year students stop understanding chemistry and start struggling with specific topics. The basics of general chemistry are taught very early in the semester, and if they are understood well, students can build upon that knowledge throughout the year. It is important that those difficult topics that

appear early in the first semester of general chemistry be thoroughly addressed, so that the understanding can serve as a foundation later in the course.

CHAPTER 2

ITEM RESPONSE THEORY

Item Response Theory – Parameters

Initially, we describe the various model possibilities that are associated with Item Response Theory (IRT). In each case, the basic IRT equation is employed but the number of parameters used is increased. IRT analysis comes in three dichotomous varieties, namely the one, two and three-parameter models. A dichotomous model differs from a more general model, as the students receive all- or no-credit for each item; students' are not given partial credit when the dichotomous model is used for analyses. The one-parameter model, known as the Rasch model, employs only the difficultly parameter, *b*, which describes how difficult (or hard) a question (or item) is. As the value of *b* increases, the more difficult the item is. The two-parameter IRT model includes the difficulty parameter, *b*, and the discrimination parameter, *a*, which describes how discriminating a question is between students with different amounts of knowledge. If a question does not discriminate at all, its IRT analysis will yield an *a* parameter of zero, indicating that every student can answer the item correctly. The most complex IRT model, the three-parameter model, incorporates the aforementioned parameters, augmented with a guessing parameter, *c*, designed to indicate the probability with which a student can "guess" the answer correctly. Theoretically, a multiple choice question with four possible answers has a *c* value ≈ 0.25. For non multiple-choice questions, the guessing parameter will approach zero.

The one-parameter Rasch model was not used in our analysis because the inherent lack of flexibility proved too restrictive (*42-44*). All the examinations discussed in this dissertation were analyzed using the two and three-parameter item response theory logistic models. As noted previously, a multiple choice question will have an associated guessing factor due to the limited number of responses, while there are, in principle, an infinite number of responses for a free-response question. For this reason, the two-parameter model is optimal for free-response question analysis, while the three-parameter model is most suitable for the analysis of multiple choice questions. For reasons to be discussed later, certain free response questions did not fit the two-parameter model well. In these cases, the three-parameter model was employed in our analysis. This approach of adjusting the analysis to suit the problem affords a more rigorous analysis than the approach taken in previous studies, in which a single model was assumed and questions not fitting that model were excluded from the analysis (*2, 4, 5, 7*). Even with the more complex three-parameter model being used for many questions, there were still items that did not fit the model properly. The reasons for their possible poor fit are discussed later in this chapter.

Although many detailed treatises describing the fundamentals of IRT can be found in the literature (*8, 45, 46*), we will provide the basic mathematical details here for clarity. The IRT analysis operates by constructing an item characteristic curve (ICC) using the three-parameters described above, which determines the probability, $P(\theta)$, that a student with an ability, θ, will correctly answer the question being analyzed. This is accomplished by fitting the student response data to the a, b and c parameters in the equation

$$P(\theta) = c + (1-c)\frac{1}{1+e^{-a(\theta-b)}}, \qquad (2.1)$$

where *b* is the difficulty parameter, *a* is the discrimination parameter and *c* is the guessing parameter (*46*). When the two-parameter model is used, the guessing parameter, *c*, in Equation (2.1) is constrained to be zero.

An example ICC is shown in Figure 2.1. Ideally these sigmoidal curves should have a steep slope, *a*, (greater than 1) with a midpoint, *b*, anywhere on the ability scale. In the IRT program used in this analysis, BILOG-MG 3, ability levels are normalized to fall in the range of -4 to +4. An ICC with a zero slope indicates there is an equal probability of every student answering the question correctly regardless of that student's ability (chemical knowledge). A question having an ICC with a zero slope does not provide any measure of a student's ability. In the limit of an infinite slope, the ICC is a step function centered at an ability level *b*. In this case, a student with an ability level greater than *b* will always get the correct answer, while a student below this level will only answer the question correctly by guessing. This is an ideal scenario for separating students by ability level. An example of a highly discriminating ICC is shown in Figure 2.1.

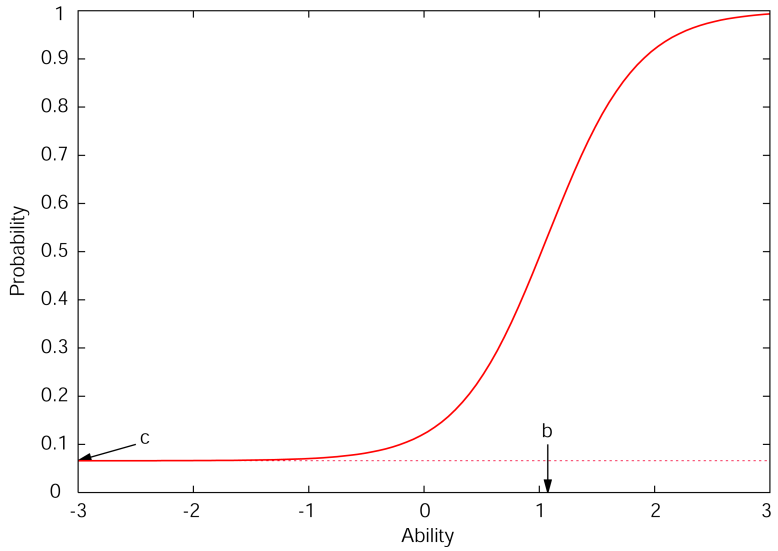


Figure 2.1: Ideal item characteristic curve. The slope of this item characteristic curve (*a*) is 2.566. This question has an ability of 1.074 (*b*) and lies on the 0.066 (*c*) asymptote.

The slope, *a*, of the curve in Figure 2.1 is 2.566 indicating that this is a highly discriminating question. The ability or difficulty, *b*, extracted from midpoint of the slope on the item's ICC is 1.074. This indicates that students with an ability of θ=1.074 have a probability of 0.533, calculated using Equation (2.1), of correctly answering the question (*46*). Students with ability less than 1.074 have a decreasing probability of correctly answering the question, and students with abilities greater than 1.074 have an increasing probability of correctly answering the question. Also notice that the guessing parameter, *c*, for this question (indicated by the lower asymptote) is 0.066. This indicates that all students, irrespective of their ability level, have at least a 6.6% chance of correctly answering the question. In its present format, IRT assigns a probability of "guessing" the correct answer independent of the students' ability; this is a pitfall of IRT. One would expect an increasing probability of a student "guessing" the question correctly with an increasing ability, and students with lower abilities should be less likely to "guess" the question correctly.

Figure 2.2 shows an example of a poorly discriminating ICC. The ability, *b*, extracted from this ICC is 0.603 indicating that it discriminates between students with a lower ability than the question shown in Figure 2.1. Even though the ability level for this ICC is slightly lower than that of Figure 2.1, this is not a concern. It is the other parameters that indicate this question was not well constructed. The small slope, *a*, of 0.399 indicates that it poorly discriminates students of this ability. Furthermore, this ICC has a guessing parameter, *c*, of 0.500. This indicates that all students have a fifty percent chance of guessing the item correctly. Theoretically, guessing parameters can range from $0.0 \leq c \leq 1.0$. However, *c* values greater than 0.35 are undesirable due to the high likelihood of students successfully answering the question regardless of their knowledge (*46*).

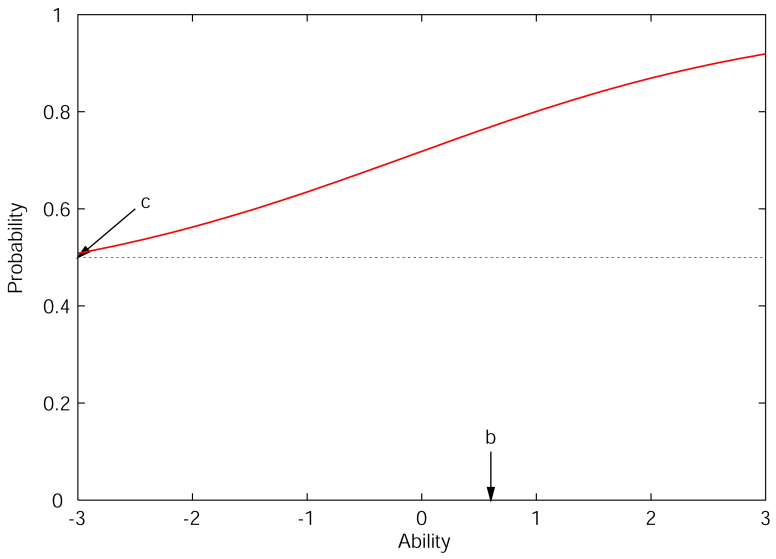


Figure 2.2: Poorly discriminating item characteristic curve. The slope of this item characteristic curve is 0.399. This question has an ability of 0.603 (*b*) and lies on the 0.500 (*c*) asymptote.

Using the Item Response Theory Program – BILOG-MG 3

In the interests of reproducibility, we will now outline some more practical aspects of the analyses performed in this work. While probably not of interest to the general reader, the remainder of this chapter will focus on the more pragmatic issues of IRT germane to the studies presented in this dissertation; hopefully this will aid future work along similar lines.

For the IRT model used (either the one, two, or three parameter) to fit the student response data, it is assumed that the test items exhibit a quality called unidimensionality. Unidimensionality indicates that the students' responses to the test items are a function of the students' abilities and that an underlying trait (in this context general chemistry) is associated with their responses. Unidimensionality of the response data permits us to place students on the ability scale (-4 to +4) (*8*). To show unidimensionality is present in our data, a correlation matrix was generated using SAS 9.1.3. Each item's students' responses on the examinations were correlated with every other item on the exam. If an item is correlated with itself, it has a correlation of 1.0. A positive manifold between questions resulted from this analysis, which demonstrates a positive correlation between questions (*47*). This is sufficient to prove that the same trait is being measured in these questions, which in this case is general chemistry. Thus, unidimensionality between questions is present.

IRT analysis was performed using the BILOG-MG 3 program (*9, 48*). The marginal maximum likelihood estimate (MMLE) was used to estimate item parameters. When using MMLE the students' locations (their abilities) are calculated independent of each other indicating that a student's location depends solely on which items that particular student answered correctly or incorrectly. MMLE assumes that the calibration sample is a random sample of the population of students. Using this assumption, estimating the item parameters for each question is

independent of the estimation of the student locations. However, an assumption of a normal distribution of student location is made with MMLE; this in turn makes the item parameters possibly dependent on this "normal" population distribution (*8*). Such a distribution is not an unrealistic assumption when using IRT, since a large sample of students is being analyzed.

After MMLE was used to estimate the item parameters, the student parameters were estimated using the Bayesian *expected a posteriori* (EAP) procedure. EAP is analyzed similarly to MMLE, where a normal population distribution is assumed at the start of the calculation (*8*). The primary advantage of EAP is that it enables the assignment of abilities to students who provide all answers correct, or to those who incorrectly answer all questions, unlike other IRT estimation procedures (*8, 48*). The students are initially assigned a mean location of 0.0, which is used to start the iterative procedure to determine the students' locations. Biweights can be used to determine the amount of influence a specific student's ability has when calculating students' locations on the ability scale. If a student did not fit the IRT model well, (the students answered and missed a variety of easy and difficult questions) their location on the ability scale would have a reduced weighting during the EAP iterations (*9, 48*). Biweights were not used in this analysis, since most students fit the model well.

A question's fit to the model is determined by the chi-square value statistics and the degrees of freedom (DF). The probability of each question fitting the chosen model was also calculated. If the aforementioned diagnostics reveal that a question does not fit the model well, then the question either does not discriminate equally among students with the same ability (knowledge of chemistry) or the wrong model was chosen. BILOG-MG 3, like other IRT programs, generates item characteristic curves, which were discussed above, by independently fitting the data for each test item (Figures 2.1 and 2.2) (*9, 49-51*).

Occasionally, an item does not fit the IRT model; these poor item fits can stem from a variety of causes. The question itself may be intrinsically poor, *e.g.* if its wording confuses students. Questions that poorly fit the model must be identified and rectified by rewording, or changing the answer options for multiple-choice items. One of the benefits of the IRT analysis is that these inferior questions are identified by the poor fits to the models used for the analysis. Another possible reason for question misfit occurs when brighter students think about the question in a more advanced way, due to something else they learned in lecture, and answer the question incorrectly (*13*). In this case it is possible that students might arrive at the correct answer with less chemistry knowledge. Although the questions themselves can cause inadequate data fits, an inappropriate choice of the specific IRT model used to analyze the data can also be problematic.

Many of the original items that were analyzed using IRT had been asked for four or more years. In 2004, the JExam test bank contained more than 9,000 questions and currently contains over 12,000 questions. The internal reliability of the examinations given for the year prior to the initiation of IRT analysis along with tests from 2005 academic year and later were analyzed using Kuder-Richardson formula 21 (KR-21) and the standard error of measurement (SEM) values (Table 2.1). The KR-21 analysis shows that all the examinations are reliable but the examinations given using JExam are more reliable than paper multiple choice examinations. IRT also can calculate reliability of the items that are analyzed. Since examinations for each academic year were analyzed together with IRT, only one internal reliability index was calculated for each academic year (Table 2.2).

Table 2.1: KR-21 reliability for each individual examination given from the fall of 2004 to the spring of 2008.

	KR-21	SEM
Fall 2004		
Exam 1	0.976	1.88
Exam 2	0.981	2.44
Exam 3	0.966	2.60
Final Exam	0.847	3.72
Spring 2005		
Exam 1	0.966	2.02
Exam 2	0.978	1.87
Exam 3	0.978	1.78
Final Exam	0.841	3.84
Fall 2005		
Exam 1	0.975	2.00
Exam 2	0.978	2.28
Exam 3	0.968	2.14
Final Exam	0.810	3.85
Spring 2006		
Exam 1	0.970	2.10
Exam 2	0.983	1.62
Exam 3	0.979	1.59
Final Exam	0.811	3.79
Fall 2006		
Exam 1	0.973	2.23
Exam 2	0.981	2.13
Exam 3	0.969	2.09
Final Exam	0.823	4.00
Spring 2007		
Exam 1	0.974	2.23
Exam 2	0.971	1.85
Exam 3	0.982	1.68
Final Exam	0.773	3.55
Fall 2007		
Exam 1	0.979	2.05
Exam 2	0.982	2.12
Exam 3	0.973	2.19
Final Exam	0.830	3.99
Spring 2008		
Exam 1	0.982	1.93
Exam 2	0.984	2.07
Exam 3	0.984	1.84
Final Exam	0.838	3.63

Table 2.2: Internal reliability index of academic year examinations from the fall 2004 to the spring 2008 calculated using BILOG-MG 3.

Academic Year	Reliability Index
fall 2004 – spring 2005	0.979
fall 2005 – spring 2006	0.990
fall 2006 – spring 2007	0.990
fall 2007 – spring 2008	0.989

Data input for the BILOG-MG 3 program should contain the following student response information. The first set of numbers is the student identification number. Each data column, after the student identification number represents one question asked of the students. If the student did not receive the question on their version of the examination, a 9 is assigned for this question. If a student answers an item correctly, a 1 is assigned, and if a student did not answer the item correctly a 0 is present. There are only 0s, 1s and 9s in the input file besides the student's ID numbers. Figure 2.3 is an excerpt of a data file, where the first row contains only 9s to indicate that a 9 represents a student not receiving the item on the examination. One limitation of this IRT program is that partial credit for questions cannot be analyzed. If a question has more than one part, the student will receive a 1 if they answered all parts correctly and a 0 if they answered one or more parts incorrectly.

```
                  99999999999999999999999999999999999
96461796 90999099199919991999919999199919991999099
71523873 91999099199919991999919999199919991999199
71589529 90999099199919991999919999199919991999199
79204159 90999099199919991999909999099919991999099
70927820 99099909919990999199990999099099199919
61138873 99099919909990999199991999919991999199919
70908479 99099909909991999099990999909990999099909
71548085 99099909909990999199990999919991999099909
70724565 99099909909991999199990999919991999199909
15315917 99099909909991999199991999919991999199919
78602667 99199919919991999199991999919991999199919
08276494 99099909909991999199991999909991999199919
70048371 99099909919991999199991999909991999199919
96982504 99099909909991999199990999909990999199909
70083376 99900999990999099919999099990999919990
17890423 99900999990999199919999199990999919990
04037268 99910999990999199919999099990999919990
70387650 99900999990999099909999199990999919990
06691777 99900999990999199919999199990999099990
70907090 99900999990999199919999199991999199919990
```

Figure 2.3: Excerpt of data file to analyze using BILOG-MG 3.

While it is possible to use Excel to format the data for use in BILOG-MG 3, Excel has a major limitation for large data files. Excel only allows a maximum of 256 columns to be present in a data file. This is much less than the requisite 594 columns necessary to analyze the 2005 – 2006 academic year data. Data can also be inputted into BILOG-MG in a .dat file format. If the data needs to be edited, such as a question needs to be removed, perl or another scripting program should be used.

To analyze items using the three-parameter model in BILOG-MG 3, in the presence of questions analyzed using a two-parameter model, constraints must be made on the *c* parameter for those items analyzed using the two-parameter model. Alpha and Beta commands under the

"Priors" command make this constraint possible. If no constraints are necessary, Alpha is set to 5.00 and Beta set to 17.00, which permits the guessing parameter to be freely estimated. If the guessing parameter must be constrained to zero (a two-parameter model analysis) the Alpha prior should be set to 2.00 and the Beta prior to 1000.00. Based upon these constraints, the items ICC's guessing factor is constrained to 0.0 and the ICC of the question will have a guessing factor, *c*, of 0.001 with a standard error of 0.001.

Multiple examinations must be analyzed simultaneously in order for the items to correspond to the same ability scale. One examination, containing questions that fit the IRT model well, should first be analyzed separately. In the multiple examination analysis, the parameters for the items on the well fitting exam are fixed as the parameters of items from other examinations are calculated. For this to work in BILOG-MG 3, a command of Prname must be in the command file under Global commands. It is important to note that when checking for syntax for errors in BILOG-MG 3, the program will inform the user that the command Prname is invalid, which is not true. The command Prname should be set equal to the name of the file containing the discrimination, difficulty and guessing parameters for each item that needs to remain fixed. The Fix command also needs to be used under the test command to inform the program which items the Prname file is constraining.

The ability scale is partitioned equally into a specified number of quadrature points. When using BILOG-MG 3, quadrature points are used to estimate student locations at the start of the MMLE and EAP procedures. Twenty-five to thirty-one quadrature points were used to estimate student locations in the analysis that is presented here. The number of quadrature points used in the analysis should be at minimum two times the square root of the number of items students answered (*8, 48*).

CHAPTER 3

ANALYZING OLD EXAMINATIONS ON JEXAM USING ITEM RESPONSE THEORY

Examinations Given from Fall 2001 – Spring 2003

At the University of Georgia, the computerized testing system JExam is used to administer examinations during the semester (*10-12*). For each hour-long examination, students taking either first or second semester general chemistry self select one of approximately 40 test sessions involving 38 students per session (reflecting the number of computers in the testing center). Course instructors prepare the examinations by inputting questions into JExam. For clarification, let us make the following definitions. As described in Chapter 1, unique examinations are constructed for individual students by selecting each question from a pool of equally difficult variants of the same question. Hereafter we will use the word "question" when referring to the topic being examined and the word "item" to denote to the specific variant of that question. Once a specific number of questions for the examination are selected, each individual question in the examination consists of a subset of, hopefully equivalent, items. JExam then assembles a unique test for each of the 38 students by choosing one test item from each question subset. For example, a student might have 25 questions on their examination; their 25 items are different but equivalent to another student's 25 items. The rationale of this is to help prevent cheating on examinations and students discussing specific questions that they received on their exam. Prior to the start of IRT analysis in 2005, items in each question (sometimes as many as twenty-five) were assumed to be equivalent by the instructor making the examination.

For IRT to be statistically valid, each item to be analyzed must have responses from around 200 students. In the fall semester of 2001, approximately 1000 students took examinations where some of the questions contained 25 items. Consequently, there were significantly fewer than 200 students answering each item. To analyze these old examinations using IRT, each question must be analyzed instead of each item. This is possible by assuming that each question contains only equivalent items. As we shall see, that assumption was not strictly correct.

Initially, exam 1 from fall 2001 was analyzed using the three-parameter logistic model. The three-parameter logistic model allows the discrimination, difficulty and guessing parameter for each question to converge and then produces an item characteristic curve (ICC) for each of the thirty questions. Only the first try on the examination was analyzed. Many of the questions were discriminating but did not fit the IRT model well, exhibiting poor chi squared values. Four of the 15 items from a question that had a poor fit to the IRT model are discussed in detail below. Item index numbers corresponding to the item numbers within the JExam question database are used for reference purposes.

One question that did not fit the IRT model well was, "What is the formula of the ionic compound formed when the *x* ion reacts with the *y* ion?" Listed below are the ions used for *x* and *y* on the examination. (item numbers 7121-7135, respectively)

- Sodium/hydroxide
- Magnesium/hydroxide
- Aluminum/hydroxide
- Sodium/nitrate
- Magnesium/nitrate
- Aluminum/nitrate
- Sodium/sulfate
- Magnesium/sulfate
- Aluminum/sulfate
- Sodium/carbonate

- Magnesium/carbonate
- Aluminum/carbonate
- Sodium/phosphate
- Magnesium/phosphate
- Aluminum/phosphate

While the IRT analysis was performed, the assumption was made that the fifteen items were equivalent. Clearly, there are a few major differences in these items. There are different charged cations (+1, +2, +3), and five different anions are used (OH^-, NO_3^-, SO_4^{2-}, CO_3^{2-} and PO_4^{3-}). These differences affect the difficulty of the items, yielding items that are not equivalent.

A second nomenclature question also did not fit the IRT model. This free response question contained two parts. 1) "What is the correct name of this ionic compound? *x* 2) How many ions are present in one formula unit of the compound shown above?" Listed below are the ionic compounds used in the examination (item numbers 7136-7148 respectively).

- $Mg(OH)_2$
- $Al(OH)_3$
- $NaNO_3$
- $Mg(NO_3)_2$
- $Al(NO_3)_3$
- Na_2SO_4
- $MgSO_4$
- $As_2(SO_4)_3$
- Na_2CO_3
- $MgCO_3$
- $Al_2(CO_3)_3$
- Na_3PO_4
- $AlPO_4$

Because this question contains many different cations and anions as well as differing numbers of ions, the questions' difficulties were different which resulted in nonequivalent items and a poor fit to the IRT model.

A third example of a question that poorly fit the IRT model was a series of multiple choice items on empirical formulas of compounds:

- "What is the empirical formula for the substance with this analysis: Na 54.0% B 8.50% O 37.5%. The atomic molar masses are B = 10.8 g/mol, Na = 23.0 g/mol, and O = 16.0 g/mol." (item 2648)
- "A compound is found to consist of 34% sodium, 16.4% boron and 48.6% oxygen. Its simplest formula is _____. (The atomic molar masses are B = 10.8 g/mol, Na = 23.0 g/mol, and O = 16.0 g/mol" (item 2649)
- "A compound of sodium, sulfur, and oxygen contains: 2.08% Na, 40.56% S, and 30.36% O. Which formula is correct? (The atomic molar masses are: Na = 23.0 g/mol, S + 32.1 g/mol, and O = 16.0 g/mol)" (item 2652)
- "A gaseous compound contained 90% carbon and 10% hydrogen by mass. What is the simplest formula for the gas? The atomic molar mass of C is 12.0 g/mol and for H is 1.0 g/mol" (item 2655)
- "The empirical (simplest) formula of a compound containing 54.3% carbon, 9.15% hydrogen, and 36.32% oxygen is" (item 2658)
- "A compound of hydrogen, chlorine and oxygen contains 1.18% H and 42.0% Cl. What is the simplest formula of this compound?" (item 2661)
- "A compound contains 46.7% nitrogen and 53.3% oxygen by mass. What is the empirical formula of this compound?" (item 2939)
- "An oxide of nitrogen contains 25.9% nitrogen and 74.1% oxygen by mass. What is the empirical formula of this compound?" (item 2940)
- "A compound was analyzed and found to contain 36.9% nitrogen and 63.1% oxygen. What is the empirical formula for this compound?" (item 2941)

There are several significant differences in the items for this question. Item 2648 has 4 multiple choice answers while the remainder have 5. As a result, the probability that a student will guess the answer correctly for this item is different than the others. Some of these items have compounds with two elements present while others have three. In item 2661 three elements were present, but only two elements were given percentages requiring that the students calculate the third. In some items, molar masses were given while in others they were not. With this many subtle differences, these test items probably were not equivalent causing the question not to fit the IRT model used.

The final question that did not fit the IRT model was a free response solution dilution question. Below are the five items that were assumed to be equivalent (items 1589-1593 respectively).

- "How many mL of 18.4 *M* H_2SO_4 are needed to prepare 600.0 mL of 0.10 *M* H_2SO_4?"
- "A laboratory stock solution is 1.50 *M* NaOH. Calculate the volume, in mL, of this stock solution that would be needed to prepare 300.0 mL of 0.200 *M* NaOH.
- "Calculate the molarity of the resulting solution if 25.0 mL of 2.40 *M* HCl solution is diluted to 3.00 x 10^2 mL."
- "Calculate the molarity of the resulting solution if enough water is added to 50.0 mL of 4.20 *M* NaCl solution to make a solution with a volume of 2.80 L."
- "Calculate the resulting molarity of a solution prepared by mixing 25.0 mL of 0.160 *M* NaBr and 55.0 mL of 0.0320 *M* NaBr."

These items exhibit subtler differences than the questions discussed previously. Items 1589 and 1590 both ask for a final volume in mL given initial and final molarities, and the initial volume in mL. Items 1591 and 1592 ask the student to calculate the final molarity given the initial and final volumes and the initial molarity. In item 1591, the final volume is given both in mL and scientific notation whereas in item 1592 the second volume is given in liters. Item 1593 is not similar to any of the previous items because it asks students to mix two solutions and calculate the molarity of the resulting solution. These differences cause the items not to be equivalent and, as a consequence, the question is unable to fit the IRT model.

Because it is possible that the individual items in a question are not equivalent, it proved impossible to analyze the data in IRT assuming item equivalency. Thus we had to find a different analysis method. Our next attempt was to analyze each item individually and then compare the IRT parameters for each item to other items in a question. Analyzing the questions in this fashion will show that the items are not equivalent in a question, or that the questions analyzed earlier contain items that do not fit the three-parameter logistic model. Because of the aforementioned lack of students receiving each item, data from exam one fall 2001 was pooled

together with exam one fall 2002 in an attempt to have a statistically meaningful sample for each item. This is one clear advantage that IRT has over CTT. Because CTT analysis is more dependent upon the students taking the examination, pooling items from two examinations is not possible. However, with IRT, the results are independent of the students taking the examination because an ability scale is formed, where each item related directly to the other items on the examination. This makes it possible to analyze pooled exam data from more than one academic year.

Even after the data from the fall 2001 and fall 2002 exams were pooled, many of the test items still had too few student responses for analysis. For example, out of the 351 items from the fall 2001 exam one, only 77 items could be analyzed. From the fall 2002 exam, 105 of the 220 items could be analyzed. To increase the number of items analyzed, the data from exam one in spring semesters of 2002 and 2003 was pooled with the fall 2001 and 2002 data. While the spring semester exams were given to fewer students, their addition to the pool should push the number of students that answered each item to over 200.

Further pooling of data from this time frame was deemed unacceptable because a significant change to JExam occurred in the fall of 2003 with the addition of homework problems for the students. Due to concern that the increased familiarity from the homework feature with the JExam questioning process might change the difficulty of many of the questions, data from fall 2003 and fall 2004 were not combined with the exams from fall 2001 and 2002.

After four semesters of data were pooled to increase the number of student responses above 200, many more items could be analyzed. For example, our initial poor fit question, "What is the formula of the ionic compound formed when the *x* ion reacts with the *y* ion?"

(where *x* is a cation and *y* is a polyatomic anion) could now be analyzed. Shown below in Table 3.1 is the IRT data for 13 of the 15 items initially believed to be equivalent.

Table 3.1: Equivalency of question 12 in fall 2001 first semester exam. Data was generated using pooled data from exam 1 for fall 2001, 2002, spring 2002 and 2003.

Item Number	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D.F.	Number of Students
7122	1.464 (0.246)	-0.424 (0.148)	0.074 (0.050)	9.0	7.0	361
7123	1.778 (0.320)	-0.176 (0.125)	0.079 (0.049)	10.9	7.0	400
7124	1.048 (0.222)	0.327 (0.217)	0.118 (0.068)	8.8	8.0	441
7125	1.152 (0.231)	0.091 (0,181)	0.076 (0.050)	11.7	7.0	269
7126	1.217 (0.290)	0.466 (0.182)	0.088 (0.055)	11.5	6.0	307
7127	1.550 (0.359)	-0.520 (0.173)	0.075 (0.051)	1.6	5.0	239
7128	1.367 (0.260)	-0.480 (0.181)	0.100 (0.064)	1.0	6.0	354
7129	1.366 (0.288)	-0.167 (0.165)	0.074 (0.050)	10.5	7.0	248
7130	1.339 (0.274)	-0.078 (0.174)	0.081 (0.054)	4.0	8.0	236
7131	1.503 (0.278)	-0.196 (0.130)	0.066 (0.044)	3.5	5.0	339
7132	1.595 (0.331)	0.003 (0.137)	0.065 (0.044)	3.4	6.0	269
7133	1.767 (0.468)	-0.073 (0.158)	0.112 (0.065)	5.1	5.0	265
7134	2.130 (0.568)	-0.072 (0.121)	0.063 (0.043)	8.9	5.0	221

Out of these 13 items, with sufficient data pooling many of the items proved to be equivalent. As can be seen from Table 3.1, item 7122 is equivalent with items 7127 and 7128, and items 7130, 7133 and 7134 are also equivalent. However, not all items in the question are equivalent. The most difficult item was item number 7126 (ability, *b*, of 0.466), which asked the students to determine the formula of the ionic compound formed when the aluminum ion reacts with the nitrate ion. The least difficult item, number 7127 with $b = -0.520$, asked the students for the formula of the ionic compound formed when the sodium ion reacts with the sulfate ion. These thirteen items have ability levels that differ by greater than one and therefore cannot as a whole

be considered equivalent (*14, 15, 52*). Smaller sets of these items can be used on future examinations since it has been shown that they are equivalent.

IRT analysis of items 9011-9015 which were used as a question on the first examination in fall 2002 shows that these items are not equivalent. Each of these items instructs the students to "Convert *a x* to *y*" where *a* is a variable amount and *x* and *y* are the different linear units listed below (item numbers 9011-9015, respectively). Results of the IRT analysis are presented in Table 3.2.

- miles/km
- miles/Mm
- yards/km
- yards/Mm
- feet/Mm

Table 3.2: Equivalency of question 11 in fall 2002 first semester exam. Data was generated using pooled data from exam 1 for fall 2001, 2002, spring 2002 and 2003.

Item Number	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D.F.	Number of Students
9011	0.991 (0.214)	-1.378 (0.324)	0.105 (0.070)	4.2	8.0	266
9012	1.282 (0.264)	0.090 (0.185)	0.095 (0.061)	5.8	6.0	286
9013	0.731 (0.164)	-1.326 (0.375)	0.099 (0.066)	5.6	9.0	264
9014	1.219 (0.258)	0.252 (0.168)	0.075 (0.049)	8.3	6.0	277
9015	1.320 (0.281)	0.320 (0.163)	0.070 (0.046)	6.1	7.0	244

From Table 3.2 we can see that item numbers 9011 and 9013 have similar *c* values (guessing factors) of around 0.1 and similar ability values of -1.3, indicating that these items discriminate between students with relatively low abilities. While items 9011 and 9013 are mutually equivalent, they are not equivalent to item numbers 9012, 9014 and 9015. This proves that

asking students to convert from miles to kilometers or yards to kilometers is easier than having them convert from miles, yards or feet to mega meters.

IRT analysis of the fall 2002 2nd and 3rd first semester examinations found that many of the sets of items were equivalent, as intended, but some were surprisingly not. For example, items 6400 to 6403 asked students "Which of the following molecules will exhibit resonance?" using a multiple answer problem with each item having one set of the 5 possibilities listed below (items 6400-6403 respectively).

- SO_2, NO_3^{1-}, SO_3^{2-}, ClO_3^{1-}, PO_4^{3-}
- SO_2, NO_2^{1-}, SO_4^{2-}, ClO_2^{1-}, PO_4^{3-}
- SO_2, CO_3^{2-}, SO_3^{2-}, ClO^{1-}, PO_4^{3-}
- SO_3, NO_3^{1-}, SO_4^{2-}, ClO_4^{1-}, PO_4^{3-}

Since the phosphate ion appears in all four items, the other four possible answers must have caused the discrepancy in the item difficulties. Item number 6402, containing ClO^{1-} and CO_3^{2-} as two possible answers, which was not present in the other three items, is the most difficult having an ability of 1.173 (See Table 3.3). All of these items have subtle differences resulting in a large range of abilities for the question. As a result, these items cannot be used within the same question on future examinations.

Table 3.3: Equivalency of question 16 in spring 2002, question 12 in fall 2001 and question 18 in spring 2003 first semester exams. Data was generated using pooled data from Exam 3 for Fall 2001, 2002, Spring 2002 and 2003.

Item Number	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D.F.	Number of Students
6400	1.402 (0.385)	0.816 (0.187)	0.053 (0.036)	3.8	6.0	219
6401	1.071 (0.253)	0.463 (0.199)	0.066 (0.044)	9.0	6.0	235
6402	0.946 (0.278)	1.173 (0.324)	0.079 (0.051)	7.3	7.0	216
6403	0.811 (0.213)	0.073 (0.291)	0.099 (0.065)	6.4	7.0	212

Some item subsets, 5658-5663 for example, contained poorly discriminating items that did not separate students of different ability levels. This item set was a multiple choice question, asking the students to "Choose the compound that contains the strongest ionic bond." Each item contained four possible answers, which are given below (Items 5658-5663 respectively).

- Al_2O_3, $MgCl_2$, NaBr SrO
- Ga_2O_3, $CaCl_2$, KBr, SrS
- Al_2S_3, $BaBr_2$, RbI, BaI_2
- AlN, CaI_2, KBr, Cs_2Se
- GaN, SrI_2, BaS, Rb_2Se
- GaP, $SrBr_2$, BaSe, Cs_2Te

Each of these items fit the three parameter IRT model, but discriminate poorly between student abilities as seen by their small slope values ($a < 1.0$) Table 3.4. Due to the small slopes of their ICCs, the midpoint and, hence, abilities of these items are ill-defined, exhibiting great variation. These items have a wide variety of abilities ranging from -3.235 to 0.483 with very small slopes (the largest was 0.875). Interestingly, these items all have lower than expected guessing factors. For a four-option multiple choice question, the guessing factor if based off of random guessing (the asymptote of the ICC) should be around 0.25. Item 5659 has an asymptote of just 0.104 while the others are even smaller.

Table 3.4: Equivalency of question 18 in fall 2001 first semester exam. Data was generated using pooled data from exam 1 for fall 2001, 2002, spring 2002 and 2003.

Item Number	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D.F.	Number of Students
5658	0.452 (0.128)	-3.235 (0.985)	0.103 (0.069)	10.1	4.0	230
5659	0.630 (0.169)	-0.935 (0.405)	0.104 (0.069)	9.4	7.0	236
5660	0.776 (0.187)	-1.261 (0.369)	0.097 (0.065)	6.5	7.0	230
5661	0.875 (0.223)	0.351 (0.270)	0.102 (0.064)	2.4	9.0	258
5662	0.618 (0.155)	0.483 (0.337)	0.095 (0.063)	10.4	8.0	303
5663	0.584 (0.146)	0.203 (0.358)	0.097 (0.064)	3.3	8.0	273

IRT analysis on the second semester general chemistry examinations proved to be more difficult than for the first semester exams. Because there were so few responses to any individual item, items were pooled with other semesters to increase the number of student responses. Data was pooled from the in-sequence semesters of spring 2002 and spring 2003, as well as the off-sequence fall 2003 and fall 2004 semesters. In the pooled data set, there were 391 items in total for the second exam. A total of 2017 students in the pooled data set took the second examination. Only 41 items out of the 391 had over 200 students answering them permitting their analysis with IRT. Similarly, on the third examination, only 45 items could be analyzed using IRT. Unfortunately, even though these 41 items from exam two and 45 items from exam three could be analyzed using IRT, the results are not dependable. On average, each student out of the 2017 only answered roughly 5 of the 41 items included in the analysis. Many of the 2017 students answered only one of the 41 items. The basis of IRT is that a difficulty is assigned to an item based upon how many students answered that item correctly or incorrectly along with performance of that same student on other items examined. When trying to analyze exam 2 and exam 3 data with very few students answering more than 12 items, the data does not fit the IRT model. It is necessary for students to answer 20 items in order for the chi square and the probability of the data to fit the model to be reliable (*8*). The data from the 3 examinations from the spring 2002, spring 2003, fall 2003 and fall 2004 semesters will have to all be pooled together to see if IRT can be performed in a more reliable manner.

For the second semester general chemistry course, all three examinations were pooled together for four semesters. Afterwards, the items that had fewer than 200 responses were omitted; only leaving 177 items to be analyzed. Even though so few items could be analyzed using IRT, the analysis is now reliable because students answered on average 22 items out of the

177. Now, with sufficient data available, the items fit the model in contrast to the situation in which data from just four semesters was used for each exam. Similar to the first semester analysis, there were a few questions that did not contain equivalent items. One of the most surprising non-equivalent items was contained in a question about buffer solutions. The question was multiple choice with five options and asked the students about the pH of a buffer solution. Below is a list of the non-equivalent items 8686 – 8689 and Table 3.5 shows the IRT analysis of these items.

- "Calculate the pH for a buffer solution prepared by mixing 100.0 mL of 0.100 *M* HF and 200.0 mL of 0.100 *M* KF." (item 8686)
- "If 400.0 mL of 0.100 *M* CH_3COOH and 200.0 mL of 0.100 *M* $NaCH_3COO$ solutions are mixed, what is the pH of the resulting solution?" (item 8687)
- "A buffer solution is prepared by mixing 250.0 mL of 1.00 *M* CH_3COOH with 500 mL of 0.500 *M* calcium acetate, $Ca(CH_3COO)_2$. Calculate the pH. The K_a of acetic acid is $1.8x10^{-5}$" (item 8688)
- "A buffer solution is prepared by mixing 250.0 mL of 1.00 *M* HNO_2 with 500.0 mL of 0.500 *M* calcium nitrite, $Ca(NO_2)_2$. Calculate the pH." (item 8689)

Table 3.5: Item response theory analysis of question 19 in spring 2002 and question 17 in fall 2003 second semester exams. Data was pooled from exam 1, 2 and 3 for spring 2002, 2003, fall 2002 and 2003.

Item Number	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D.F.	Number of Students
8686	0.957 (0.250)	0.491 (0.269)	0.108 (0.068)	7.7	8.0	240
8687	1.315 (0.427)	0.518 (0.237)	0.148 (0.075)	12.3	7.0	225
8688	0.845 (0.274)	1.689 (0.432)	0.121 (0.064)	6.1	8.0	274
8689	0.884 (0.247)	0.877 (0.319)	0.125 (0.070)	4.4	9.0	263

All four items involve a weak acid and a corresponding basic salt, solution amounts in mL with their corresponding concentrations in molarity. However, item 8688 is much more difficult than the other three items having an ability, *b*, of 1.689. In this item the students are given the K_a, but

for the other 3 items in this group students are expected to look up K_a values from the resources. Because of this, it would be expected that item 8688 would be easier than the other three items, but this is not the case. One significant difference is that in items 8688 and the next most difficult item, 8689, the basic salt has a one to two mole ratio, but in items 8686 and 8687 the mole ratio is one to one. It is probable that this mole ratio caused items 8688 and 8689 to be more difficult than items 8686 and 8687.

Examinations Given from Fall 2003 – Spring 2005

As stated previously, at the beginning of the fall 2003 semester, the JExam computer program was now being used for homework as well as examinations. It was our belief that increased student familiarity with JExam would impact their performance on the examinations. Many of the same difficulties arose when trying to analyze the data from these semesters, most notably the problematic small sample sizes (< 200). Data from the on-sequence semesters, fall 2003 and fall 2004 were pooled with the off-sequence semesters, spring 2004 and spring 2005, data. Just as the analysis for the first semester general chemistry in the fall 2001, exam one data was pooled for four semesters to see if the items in a question were equivalent.

New questions were written and used for the fall 2003 – spring 2005 exams that were not present on the fall 2001 – spring 2003 examinations. Initial IRT analysis of data is performed using CTT. Results of the CTT analysis are then used to initiate the IRT analysis. Based upon the CTT analysis, if an item is too difficult or too easy, the IRT analysis cannot converge due to the extreme differences in item difficulties. For the first exam in fall 2003 – spring 2005, many of the items on the exam were too easy. Consequently, the range between the difficult and easy problems was too large for the BILOG-MG 3 program to converge to the specified criterion. To

resolve this problem, items that were too easy can be either given a constrained set of parameters or removed from the analysis of the examinations. Because the vast majority answered the items correctly, we know that the items must be equivalent. Therefore they can safely be removed from the analysis. One question that proved to be too easy for the students was "The melting point of *x* is *y* °C. What is the melting point of *x* in °F?" Where *x* is an ionic compound and *y* is its melting point. Since questions of this type do not pose any conceptual difficulty for the students, they correctly determine the strategy needed to make this simple conversion. Moreover, students were given in the exam resources the conversion factor between °C and °F so no memorization skills were required. See Table 3.6.

Table 3.6: Classical test analysis of question 12 in fall 2003 and question 6 in fall 2004 first semester exams. Data was generated using pooled data from Exam 1 for Fall 2003, 2004, Spring 2004 and 2005.

Item	Ionic compound (*x*)/ melting point (*y*) °C	Number of times attempted	Number correct	Percent correct
7060	NaF / 993	470	426	90.6
7061	NaCl/801	460	427	92.8
7062	NaBr/747	521	495	95.0
7063	KCl/770	446	421	94.4
7064	CaF_2/1423	468	443	94.7
7065	Na_2S/1180	495	476	96.2
7066	K_2S/840	475	445	93.7
7067	MgO/2800	486	471	96.9
7068	CaO/2580	440	418	95.0
7069	BaO/1923	486	461	94.9

On the basis of the data presented in Table 3.6, it was decided to remove items 7062, 7065, 7067 and 7068 from the data then perform the IRT analysis using the three-parameter logistic IRT

model. Analyzed items that remained on the exam had the ability levels displayed in Table 3.7. These questions are equivalent, but are incredibly easy for the students, with the most difficult having an ability of -2.405.

Table 3.7: Item response theory analysis of question 12 in fall 2003 and question 6 in fall 2004 first semester exams. Data was generated using pooled data from exam one for fall 2003, 2004, spring 2004 and 2005.

Item Number	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D.F.	Number of Students
7060	1.066 (0.221)	-2.405 (0.411)	0.100 (0.067)	3.9	6.0	470
7061	1.162 (0.253)	-2.443 (0.427)	0.100 (0.067)	3.9	6.0	460
7063	0.972 (0.223)	-3.109 (0.621)	0.099 (0.067)	4.7	7.0	446
7064	1.551 (0.340)	-2.400 (0.342)	0.094 (0.064)	1.5	5.0	468
7066	1.157 (0.242)	-2.600 (0.452)	0.096 (0.065)	6.2	6.0	475
7069	1.318 (0.287)	-2.676 (0.416)	0.098 (0.066)	3.9	6.0	486

Just as some of the items were too easy for the program to use in its analysis, there were also some very difficult items that had to be removed from the data set for the analysis to converge. For example, the third examination in the first semester of general chemistry contained item numbers 9299-9313. Of these items 9303-9309, 9311 and 9312 proved to be too difficult for convergence in BILOG-MG 3 to the assigned criterion. Items 9299-9302 and 9313 were not analyzed because there were fewer than 200 students answering each item. Table 3.8 shows the CTT analysis of items 9303-9309, 9311 and 9312. These items all have two parts: "Enter the correct formula of this compound. *x*," where *x* is a ternary acid salt along with "How many ions are present in one formula unit of the compound written above? For a covalent species enter zero." Once these difficult items were removed from the analysis of exam three,

given in the fall 2003, fall 2004, spring 2003 and spring 2005 semesters, the IRT analysis ran smoothly.

Table 3.8: Classical test analysis of question 12 in fall 2003 and question 6 in fall 2004 first semester exams. Data was generated using pooled data from exam 1 for fall 2003, 2004, spring 2004 and 2005.

Item	Ternary Acid Salt (*x*)	Number of times attempted	Number correct	Percent correct
9303	lithium dihydrogen borate	215	14	5.9
9304	lithium dihydrogen borite	228	14	6.1
9305	lithium hydrogen borate	226	19	8.4
9306	lithium hydrogen borite	228	16	7.0
9307	rubidium hydrogen carbonate	221	31	14.0
9308	rubidium dihydrogen arsenate	221	13	5.9
9309	rubidium dihydrogen arsenite	218	13	6.0
9311	rubidium hydrogen arsenite	211	11	5.2
9312	sodium hydrogen selenate	215	14	6.5

To assess the effect of JExam homework on item parameters, the three parameters for each item (discrimination, difficulty and asymptote) were compared before and after homework was assigned. For most items, the three parameters did not change, *e.g.* items 9016, 9022 and 5663, shown in Table 3.9. An asterisk next to the item number indicates that IRT analysis was generated after homework was administered using JExam while no asterisk indicates that the IRT analysis was generated before homework was assigned. For items where the parameters changed slightly, usually the discrimination parameter, *a*, was altered. The item's difficulty, *b*, and asymptote, *c*, parameters stayed within their respective standard errors before and after homework was assigned. Interestingly, the discrimination parameter typically became larger

indicating that the items are more discriminating after JExam homework was instituted. For example, see items 7073, 5303, 9158 and 127 in Table 3.9.

Table 3.9: Comparison of item response theory analysis before and after the introduction of JExam homeworks.

Item Number	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D.F.	Number of Students
9016	1.071 (0.197)	-0.503 (0.206)	0.081 (0.055)	6.1	7.0	313
9016*	1.085 (0.245)	-0.493 (0.234)	0.094 (0.061)	2.6	7.0	202
9022	1.005 (0.215)	-1.329 (0.321)	0.089 (0.060)	4.1	7.0	231
9022*	1.036 (0.203)	-1.114 (0.244)	0.095 (0.063)	5.3	8.0	340
5663	0.584 (0.146)	0.203 (0.358)	0.097 (0.064)	3.3	8.0	273
5663*	0.654 (0.173)	0.230 (0.345)	0.110 (0.071)	4.5	8.0	280
7073	1.010 (0.211)	-1.508 (0.325)	0.092 (0.062)	2.9	7.0	315
7073*	1.999 (0.519)	-1.620 (0.236)	0.091 (0.062)	1.5	5.0	291
5303	0.757 (0.158)	-0.596 (0.275)	0.102 (0.068)	4.3	8.0	412
5303*	1.475 (0.234)	-0.708 (0.138)	0.068 (0.046)	6.0	7.0	438
9158	0.787 (0.214)	-0.876 (0.354)	0.102 (0.068)	1.0	7.0	201
9158*	1.322 (0.182)	-0.833 (0.155)	0.077 (0.052)	9.9	7.0	501
127	0.791 (0.162)	-0.545 (0.268)	0.097 (0.064)	8.9	8.0	331
127*	1.570 (0.393)	-0.475 (0.162)	0.079 (0.053)	10.9	4.0	266

Data was generated using pooled data from Exam 1, 2 or 3 for Fall 2001, 2002, Spring 2002 and 2003.
*Data was generated using pooled data from Exam 1, 2 or 3 for Fall 2003, 2004, Spring 2004 and 2005.

We believe that the difference in the slope was caused by brighter students answering questions incorrectly simply because they were unfamiliar with JExam prior to the assignment of homework using the program. The students' lack of familiarity with the program is unrelated to their intrinsic chemical knowledge. After the assignment of homework with JExam, the students became more familiar with the program making the test questions a better predictor of their chemical knowledge rather than how comfortable they were with the program. The lack of familiarity with the program would be exacerbated by the lower ability students ignoring the

homework assignments, while the more academic students participate and gain familiarity with JExam. This would have the effect of increasing the discriminating quality of the question for the aforementioned reason.

Interestingly enough, the analogous pre- and post JExam homework analysis for the second semester of general chemistry yielded no increase in the discrimination parameter *i.e.* the items were just as discriminating before and after homework was assigned. Table 3.10 contains examples of items in which the three item's parameters did not change from pre- and post JExam homework.

Table 3.10: Comparing item response theory analysis before and after the introduction of JExam homeworks.

Item Number	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D.F.	Number of Students
203	0.709 (0.189)	1.101 (0.386)	0.096 (0.061)	13.9	8.0	272
203*	0.643 (0.133)	0.940 (0.301)	0.101 (0.063)	10.8	9.0	751
8006	1.052 (0.198)	-1.294 (0.267)	0.094 (0.063)	7.3	8.0	401
8006*	1.110 (0.268)	-1.247 (0.296)	0.101 (0.067)	4.7	8.0	236
8072	1.199 (0.241)	-0.473 (0.210)	0.107 (0.068)	10.5	8.0	340
8072*	1.298 (0.315)	-0.655 (0.225)	0.119 (0.075)	13.3	7.0	259
8675	1.017 (0.270)	1.302 (0.280)	0.075 (0.047)	5.1	7.0	293
8675*	1.089 (0.366)	1.148 (0.306)	0.146 (0.072	7.6	8.0	254

Data was generated using pooled data from Exam 1, 2 or 3 for Spring 2002, 2003, Fall 2002 and 2003.
*Data was generated using pooled data from Exam 1, 2 or 3 for Spring 2004, 2005, Fall 2004 and 2005.

The similarity of the parameters before and after homework was assigned was most likely caused by the students' familiarity with the program for the second semester of the course. This familiarity removes anomalous false responses caused by their unfamiliarity with JExam, which adds credence to our proposition that program familiarity was a cause of the changes in discrimination before and after homework was instituted.

Results of the analysis of these old examinations afforded identification of items with the best discrimination values and the smallest guessing parameters. Furthermore, the analysis resulted in an accurate determination of which database questions on a specific topic were equivalent in difficulty and ability level. Items were sorted by ability levels, and poorly discriminating items were removed from examinations starting in the fall 2005 academic year (*13*). Nonequivalent items on a specific topic were also eliminated from use in future examinations.

CHAPTER 4

WRITING EXAMINATIONS USING IRT RESULTS

JExam

Starting in the fall 2005 academic year, tests were written using high quality questions that were discriminating with low guessing factors. In general, tests should consist of questions with large ICC slopes, whose midpoints occur at a variety of ability levels, as this affords categorization of students across the entire ability spectrum. A specific number of items that discriminate between each grade level should be on the examination. Initially, determination of which items separate A students from B, C, D, and F students was based upon the IRT analysis performed on the old examinations.

Once examination items have been analyzed using IRT, the students are assigned abilities based on which items they answered correctly. As useful as IRT is for accurately analyzing individual items, students' examination grades in general chemistry at UGA are still assigned based upon the number of items they missed and not their calculated IRT abilities. Final course grades are assigned to the students not based upon their IRT abilities but calculated from their examination, homework and pop quiz scores. Figure 4.1 shows the correlation of the students' assigned IRT abilities versus their percent correct to compare the students' ability with the grade they would have received on their examinations if they were only given one try.

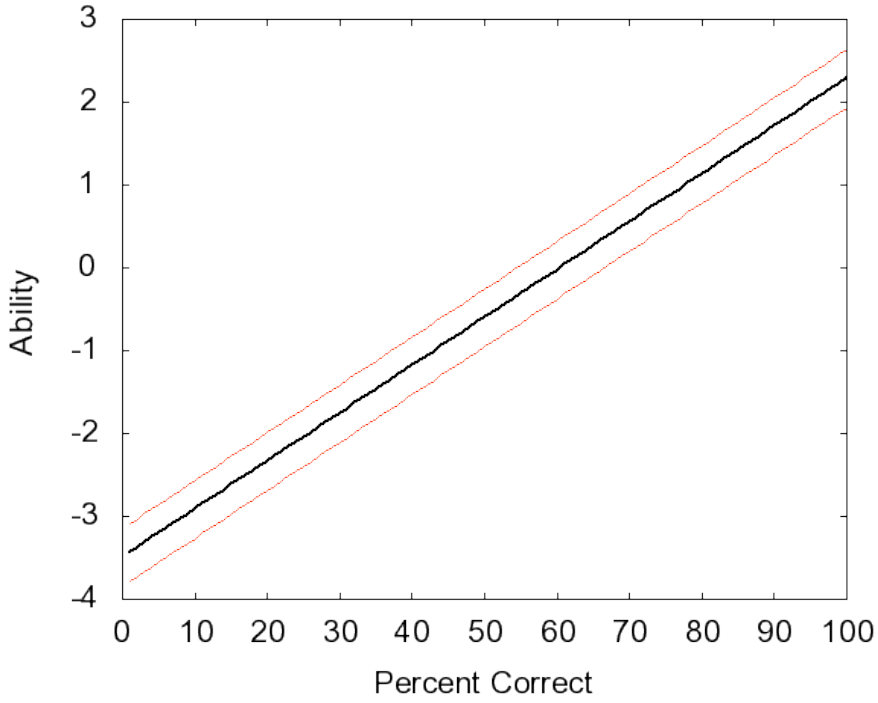


Figure 4.1: Regression of Ability vs. Percent Correct generated from IRT analysis of fall 2006-spring 2007 academic year. The black line is the student's ability versus the percent correct; the red lines are 95% confidence intervals.

Figure 4.1 indicates that for a student to receive an A on the examination (test score ≥ 90%), they had an IRT ability of 1.70493 or higher. For a student to receive a B (assuming 90%≥ B≥ 80%), the student had an ability between 1.12743 and 1.70493 and so forth. See Table 4.1 for the full analysis for each letter grade, assuming the following grade assignments: A ≥ 90%, B 80%-90%, C 70%-80%, D 60%-70% and F < 60%. From this analysis, each test item can be identified as an A, B, C, D or F discriminating item. These designations help determine item difficulty levels and are useful when composing examinations and determining for whom topics are difficult, which will be discussed in Chapters 5, 6 and 7.

Table 4.1: Ability needed for each letter grade for fall 2006 – spring 2007 academic year.

A	B	C	D	F
1.70493	1.12743	0.549928	-0.0275721	< -0.0275721

The ability that a student is assigned derives from their performance on the examinations. The students are separated into ability groups based upon their performance on a given topic. These ability groups are most accurately described as students with very high, high, medium, low and very low abilities; however, for this analysis A, B, C, D and F grades are used as descriptors of these different abilities for brevity. Our assignment of a particular ability level to a student does not preclude them enhancing their knowledge and improving future performance on a given topic. This does not imply this is the grade that a student earned for the class, but it is based upon the number of correct responses for the first try on the examinations, coupled with the subject matter in which these correct responses came.

Once IRT analysis has been performed on previous exam items, the parameters of these items can be used to help build appropriately discriminating future examinations for the student population. It is required that enough questions of each discrimination level are placed on the test to adequately assess the students. The number for each level depends upon the total number of test questions on the examination along with the grading scale used. For example on a twenty-five question exam using the previously mentioned grading scale, there should be three questions that discriminate A students (ability $\geq$ 1.70493) from B, C, D, and F students. If a student misses two of these three more difficult questions, they could still answer the remaining simpler questions correctly; they would receive a 92 on the exam correlating to a low A on the previous grading scale. As we construct the examination, we assume that a student capable of

answering A level difficulty items correct, ability of 1.70493 or higher, would also answer correctly B, C, D and F discriminating questions. Likewise, students that miss three questions on the exam should miss the most difficult questions— the three A discriminating questions. If a student missed all three A discriminating questions, the highest grade that student could receive is an 88, or a B, which is appropriate if they did not answer a single A/B discriminating question correctly. Using similar logic to ensure that if a student completely misses all questions of a specific grade ability, there should be two questions that separate A and B students from C, D and F students. There should also be three questions that discriminate A, B and C students from D and F students, two questions that discriminate A, B, C and D students from F students, with the remainder of the questions being F discriminating questions. These F discriminating questions will still have large slopes, *a* values, but A, B, C and D students should be capable of answering these questions correctly.

It is important to choose questions within a grade level that have unique ability levels. For example, for the A discriminating questions, one question should separate the high A students from the middle A students. A second question should separate the middle A students from the lower A student and the third would separate the lower A students from the higher B students. By using this technique for all of the questions on the exam, the students will accurately be assigned the correct grade based upon the chemistry knowledge that the student possesses. If a grading scale other than A ≥ 90%, B 80%-90%, C 70%-80%, D 60%-70% and F < 60% is used, the quantity of each ability level question must be adjusted accordingly.

Once the examination has been given, it can be tested for reliability using both IRT and the CTT reliability index. The CTT inner reliability index, KR-21, was calculated for each examination and is shown in Table 2.1. With IRT, each item characteristic curve has a

corresponding item information curve that indicates which group of students are best assessed by the question. Figure 4.2 is an example of an item characteristic curve for item 11206 having an ability, *b*, of 0.723, discrimination factor, *a*, of 5.835 and guessing factor, *c*, of 0.130. Figure 4.3 is the corresponding item information curve for the same item.

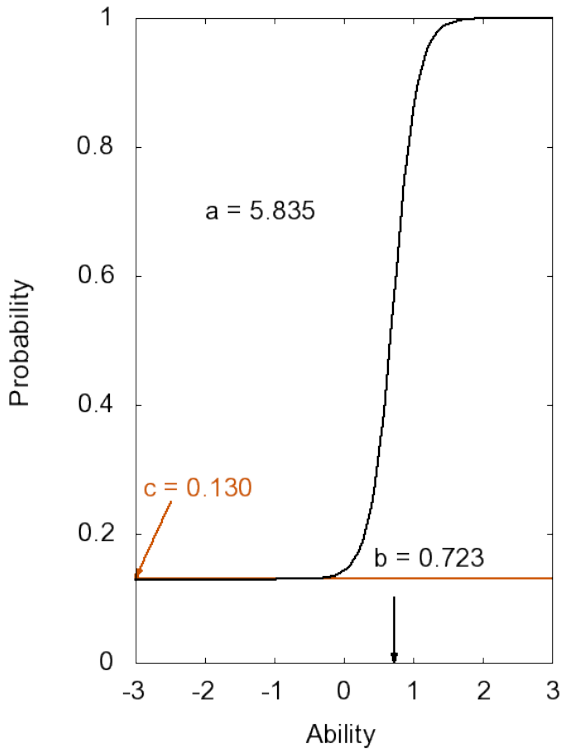


Figure 4.2: Item characteristic curve of item number 11206. Figure generated from data gathered on exam 1 fall 2003, 2004, spring 2004 and 2005. Item characteristic curve indicates that this item has an ability, *b*, of 0.723, asymptote, *c*, of 0.130 and discrimination parameter, *a*, of 5.835.

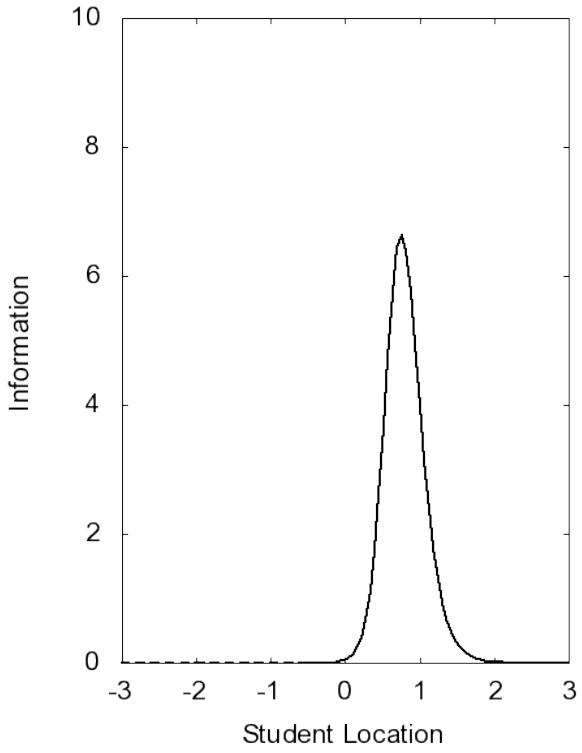


Figure 4.3: Item information curve for item number 11206. Figure generated from exam 1 fall 2003, 2004, spring 2004 and 2005 data

The peak of the item information curve is centered on the ability, *b*, which was calculated for that item. Item information curves for every item are symmetrically distributed around the item's ability. Just as a steep slope for an item characteristic curve is indicative of a highly discriminating item, a tall, thin peak on an item information curve is indicative of the same. A shorter, wider item information curve indicates greater uncertainty in the distinguished ability levels, which corresponds to a shallow slope in the item characteristic curve having an ill-defined midpoint. The curve shown in Figure 4.3 indicates that this item informs us a lot about a student with an ability near 0.723, but gives us no information about a student with an ability greater than two, since they all answer it correctly, or less than zero, since they all answer it incorrectly.

Total information curves (TIC) are generated by integrating all of the item information curves for an examination into a single graph. The TIC indicates the collective amount of information gained from the examination as a function of student ability. The TIC takes into account the amount of information each item contributes to minimize the uncertainty of the student abilities (*8*). Figure 4.4 is an example of a good TIC. An exam written with a suitable range of item abilities having highly discriminating items generates a total information curve that has a large centered peak with small standard errors on both ends. For an exam written to determine if a student should pass or not pass a class, the total information curve should be tall, slender and centered at the ability necessary to pass the class.

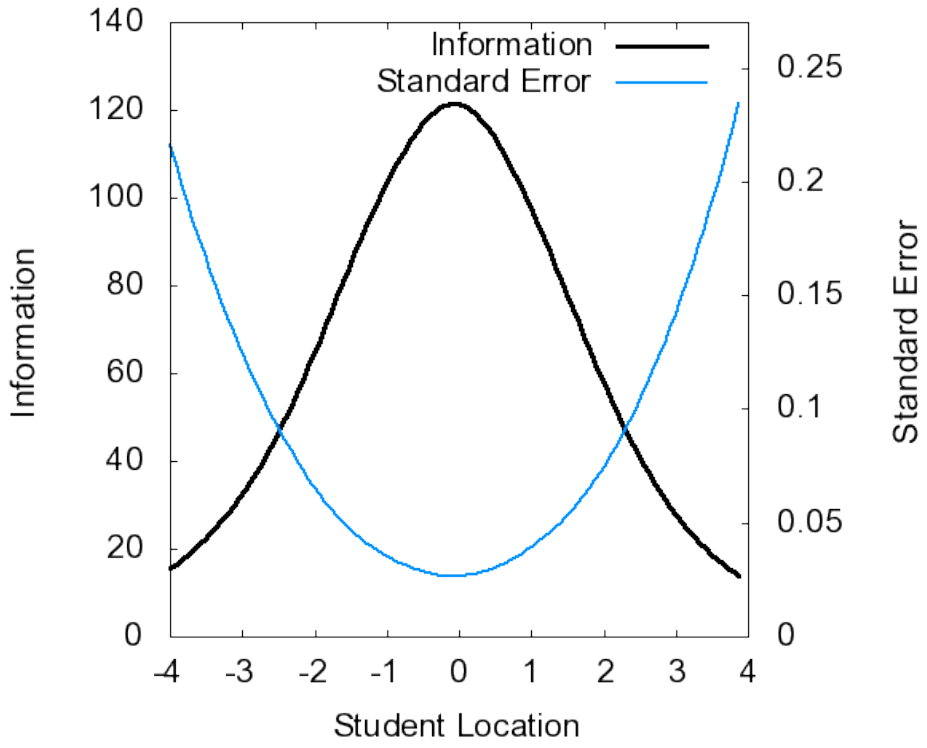


Figure 4.4: Total information curve generated from the fall 2006 –spring 2007 academic year

The TIC shown in Figure 4.4 indicates that the students' abilities are assessed with a small amount of error. As expected, the error in a student's ability (amount of chemistry knowledge) is very small, toward the center of the graph. The error is greater on both of the edges of the student location, below -2.5 and above 2.6. A student with an ability of 3.0 has a standard error of 0.129. Contrastingly, the standard error for a student with an ability of 0.0 is 0.0276. By referring to Table 4.1, it is seen that a student with an ability of 3.0 would be considered an A student. Including the standard error of 0.129, a student with an ability of 2.871, would still lie in the A ability range. A student with an ability of 0.0 is considered a D student. Even with the error bar surrounding 0.0, the student's location would be from -0.0276 to + 0.0276. An ability of 0.0276 would still be considered in the D ability range, whereas an ability of -0.0276 would be considered an F ability by 0.0001. The standard error on the total information curve is extremely small when looking at student performance and knowledge.

Writing a Multiple Choice Exam

Writing good multiple choice questions is more difficult than writing good discriminating open-ended questions. On multiple choice exams each question has 4 or 5 choices. These incorrect options should all be well designed distractors. Flawed reasoning or an incorrect calculation should lead the students to one of these incorrect choices. Using IRT, questions can be analyzed to determine the effectiveness of their distractors. The three-parameter logistic IRT model is used to calculate the guessing factor.

With purely random guessing, the guessing factor for every item with 5 options would be 0.20, where students would have a 20% chance of answering the item correctly by randomly guessing. Guessing factors that are calculated using IRT are obtained by monitoring the

performance of the students with lower abilities that should have answered the question incorrectly, but were able to answer it correctly. Guessing factors can deviate from the value expected by the naïve assumption that every answer has an equal probability of beeing chosen. These guessing factors also take into account the idea that the options to a multiple choice question "vary in their degree of attractiveness" (*8*). Specific keywords in an option might cause students to veer towards that answer, whereas answers that seem out of place might cause students to avoid choosing them (*53*).

Examples of multiple choice examination items with guessing parameters significantly different from the random guessing expected values are listed below. In the first example, it is seen that poor distractors cause the guessing factor to be larger than what would be expected from random guessing. This indicates that students are able to "guess" the answer correctly by the process of elimination of one or more of the item's distractors. For example, on the second semester final exam given in the spring of 2007, on item 2016, students were asked to "Consider the equilibrium system shown below. Select the response below that includes all of the true statements, and none that are false.

$$2\ CO_2(g) + 4\ H_2O(\ell) \rightleftharpoons 2\ CH_3OH(\ell) + 3\ O_2(g) \qquad \Delta H = 1453\ kJ/mol$$

I. Removal of CH_3OH will increase the relative concentration of oxygen gas.
II. Increasing the concentration of carbon dioxide gas will decrease the relative concentration of CH_3OH.
III. Lowering the reaction temperature will increase the relative concentration of CH_3OH.
IV. Quadrupling the volume of the reaction vessel will decrease the relative concentrations of CO_2.
V. Decreasing the partial pressure of oxygen will increase the relative concentration of CH_3OH."

Possible answers were

- II, III, and IV
- I, II, and IV
- I, IV, and V
- III and V
- I and V

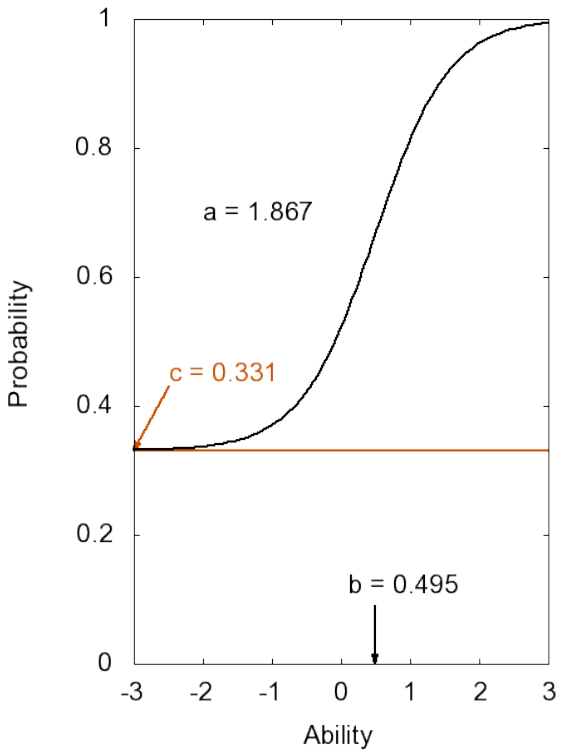


Figure 4.5: Item characteristic curve of item number 2016 given in the spring of 2007. This is a multiple choice item with a larger than expected guessing factor

The guessing factor of the ICC for item 2016 (Figure 4.5) is 0.331 indicating that students had a 33.1% chance of guessing this question correctly, even if their chemical knowledge was minimal. Out of the 867 students that answered this item, 441 answered it correctly. The fifth possible answer (I and V) proved to be the largest distractor with 230 students choosing this answer. Because so few students chose the first and fourth options (II, III and IV or III and V) as their answers, it appears many students eliminated one or both of these "distractors" before they "guessed" their answer.

Smaller than expected guessing factors imply that at least one of the distractors is working well, garnering a selection from lower ability students not based on guessing. Item 1045, given on the first semester fall 2007 final exam, is an example of an item with a smaller than expected random guessing factor. The item asked students to "Compare the two images below. Image one best represents which of the following aqueous solutions?"

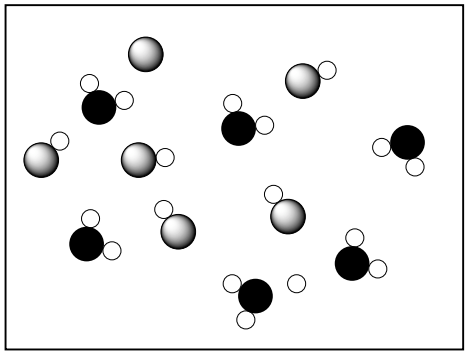


Image One

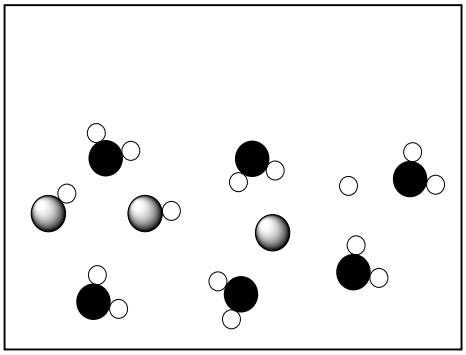


Image Two

The five possible answers were:

- a concentrated solution of HCl
- a dilute solution of HCl
- a concentrated solution of HF
- a dilute solution of HF
- none of these

Examination of the ICC for item 1045 (Figure 4.6) indicates that it has an item difficulty of 1.047, discrimination factor of 1.299, and a guessing factor of 0.098. The correct answer, a concentrated solution of HF, was the most common answer. The next most common answer was the last option "none of these" followed by "a concentrated solution of HCl." Low guessing resulted from high responses to an "attractive" distractor. Out of the 1200 students who answered this item, only 258 students answered either "a dilute solution of HCl" or "a concentrated solution of HF." Lower ability students did not try to guess on this question because they were "distracted" by one of the more "attractive" answers.

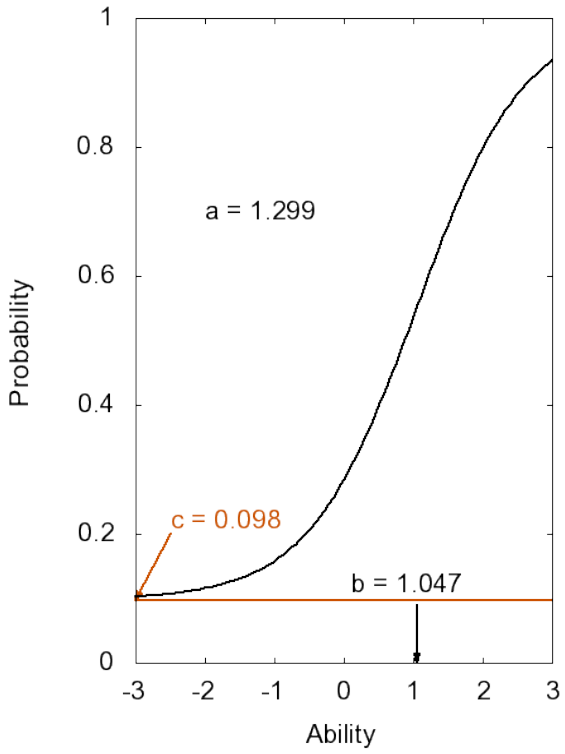


Figure 4.6: Item characteristic curve for item number 1045 given in the fall of 2007. This is a multiple choice item with a smaller than expected guessing factor.

Item 2038, appeared on the second semester general chemistry final in the spring of 2008. This item is an example of a five option multiple choice question with a guessing factor close to the expected random guessing factor of 0.200. The item asked the students "Which of the following compounds has the weakest intermolecular forces?" The five options were "RbCl, I_2, CH_3Cl, CH_3NH_2 and CH_3COOH." As demonstrated by the item 2038's ICC (Figure 4.7), this item has an ability of 0.147, a slope of 1.475 and a guessing factor of 0.194. Students have a 19.4% chance of guessing the item correctly; this is close to the 20% chance of randomly guessing the item correctly.

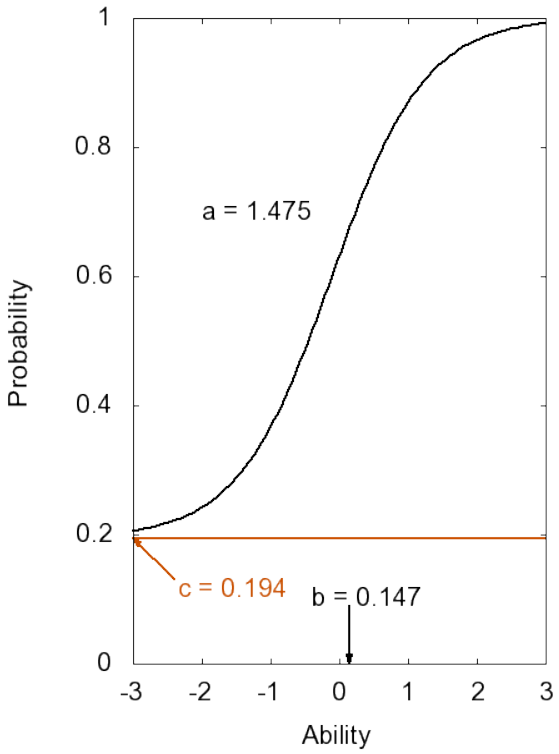


Figure 4.7: Item characteristic curve of item number 2038 given in the spring of 2008. This is a multiple choice item with a guessing factor similar to the random guessing factor.

Each of the ICCs shown above has a corresponding item information curve. The item information curve contains a smaller amount of information about a person's location when a three-parameter model is used. For this reason comparison between two and three-parameter item information curves should be performed cautiously. Since the total information curve is the sum of the item information curves, when using the three-parameter model, the total information curve will contain less information than if the analysis was performed with the two-parameter model.

Similar to other question types, multiple choice questions have a variety of abilities. Examinations using a multiple choice format should be constructed similarly to the computerized examinations, using the procedure outlined above. Since we want to correctly assign student abilities across the continuum, the examination items should have a wide span of difficulties in which all items are discriminating. The number of each type of discriminating question depends on the number of questions on the examination and the grading scale used for the course.

CHAPTER 5

ANALYSIS OF FALL 2005 – SPRING 2006 ACADEMIC YEAR EXAMINATIONS

IRT Analysis

Once the examinations from the 2001 to 2004 academic years were analyzed with IRT, their analysis was used to guide the writing process for the fall 2005 – spring 2006 academic year examinations. Then the examinations given during the 2005 academic year were analyzed to find those topics on which the students struggled. For an accurate topic comparison, all eight examinations given during the academic year must be analyzed simultaneously to be on an equivalent scale. A student did not have to complete both semesters of the course to be included in the analysis; a student needed to take at minimum one examination in the academic year. This gives the most accurate comparison of topics across examinations and semesters. To accomplish scale equivalency, IRT was first applied to the first semester exam two data using both the two and three-parameter logistic models. The item being analyzed determined the model that was used; if a student could guess on an item and answer it correctly, the three-parameter model was used. For example, test items that were multiple choice or multiple answer were analyzed using the three-parameter model. If a student could not guess an item correctly, *i.e.* the students freely responded to the question, the guessing parameter was constrained to 0.0 effectively limiting the model to two parameters. These parameters include the discrimination and difficulty parameters.

Once the parameters for all the items on the second exam were calculated, that data was combined with data from the other seven exams given during the 2005-2006 academic year. The

resulting second examination item parameters (*a*, *b* and *c* values) were then fixed while the parameters for the other exam items were freely estimated (with constraints on the guessing parameter for free response items). IRT analysis was repeated with the full set of data. This approach yielded item and student ability values on the same scale for all 2005-2006 examinations.

Table 5.1 indicates the minimum ability level needed to achieve each letter grade for the 2005-2006 academic year. The results were determined by the overall percentage of correct answers needed to obtain a certain letter grade, compared to the students' ability levels calculated in the IRT analysis. Combining this knowledge with the known ability of a given question, the potential of that question to discriminate between A, B, C, D and F students was determined. Specific general chemistry topics and their associated discrimination ability levels are discussed below. Table 5.2 shows the IRT analysis for each item discussed in the 2005-2006 analysis.

Table 5.1: Ability needed for each letter grade for fall 2005-2006 academic year.

A	B	C	D	F
1.64177	1.04417	0.446575	-0.151024	< -0.151024

Table 5.2: IRT parameters tem for items on the 2005 - 2006 academic year examinations.

Item Number	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D. F.	Number of Students
231	1.456 (0.350)	-1.623 (0.324)	0.178 (0.081)	3.0	5.0	312
1009	2.564 (0.589)	-2.172 (0.279)	0.194 (0.084)	1.2	3.0	1053
1011	1.949 (0.271)	1.145 (0.074)	0.120 (0.025)	13.3	9.0	1053
1032	0.812 (0.231)	2.026 (0.313)	0.212 (0.057)	9.8	9.0	1053
1034	1.451 (0.203)	-2.149 (0.281)	0.200 (0.088)	7.0	6.0	1053
1055	1.898 (0.210)	-0.485 (0.120)	0.192 (0.065)	7.1	8.0	1053
1056	1.426 (0.353)	1.552 (0.149)	0.217 (0.038)	6.8	9.0	1053
1408	1.522 (0.258)	-0.415 (0.148)	0.001 (0.001)	2.4	6.0	239
1724	1.580 (0.400)	-1.580 (0.400)	0.189 (0.085)	7.7	7.0	206
1726	1.124 (0.236)	-1.849 (0.374)	0.190 (0.086)	5.4	7.0	232
2009	1.270 (0.202)	-1.208 (0.276)	0.185 (0.082)	8.5	8.0	713
2013	1.038 (0.450)	2.828 (0.612)	0.293 (0.046)	11.5	9.0	713
2026	1.200 (0.240)	0.736 (0.212)	0.233 (0.077)	4.6	9.0	713
2062	0.998 (0.163)	0.734 (0.185)	0.139 (0.061)	15.9	8.0	713
2066	0.794 (0.277)	2.416 (0.479)	0.255 (0.066)	2.7	8.0	713
2942	1.073 (0.265)	-2.364 (0.492)	0.198 (0.089)	2.9	6.0	316
5683	0.636 (0.134)	0.501 (0.368)	0.191 (0.077)	10.2	9.0	692
7140	0.857 (0.101)	0.108 (0.097)	0.001 (0.001)	11.1	9.0	677
7144	1.230 (0.117)	0.290 (0.075)	0.001 (0.001)	9.9	9.0	692
7147	0.843 (0.102)	0.149 (0.099)	0.001 (0.001)	6.1	9.0	677
7238	2.229 (0.706)	-1.370 (0.230)	0.189 (0.085)	2.4	4.0	260
7398	0.920 (0.192)	0.431 (0.150)	0.001 (0.001)	2.6	8.0	250
7416	1.195 (0.247)	1.666 (0.242)	0.001 (0.001)	3.8	6.0	286
7438	1.037 (0.250)	-1.028 (0.353)	0.191 (0.086)	2.3	7.0	237
7480	1.306 (0.242)	-0.740 (0.229)	0.168 (0.076)	1.7	7.0	370
7532	1.295 (0.238)	-0.420 (0.144)	0.001 (0.001)	5.7	7.0	250
8078	0.989 (0.401)	-3.154 (1.247)	0.203 (0.090)	0.8	2.0	239
8178	0.902 (0.209)	1.240 (0.215)	0.001 (0.001)	2.2	6.0	245
8316	1.149 (0.313)	-1.306 (0.465)	0.201 (0.089)	5.9	5.0	245
8342	0.907 (0.285)	-1.967 (0.747)	0.201 (0.090)	1.0	6.0	245
8346	1.648 (0.577)	-1.935 (0.603)	0.191 (0.086)	0.5	2.0	239
8411	1.665 (0.425)	-1.208 (0.357)	0.176 (0.081)	1.5	3.0	202
8435	1.149 (0.329)	-1.222 (0.441)	0.225 (0.096)	5.7	5.0	202
8824	1.519 (0.511)	1.869 (0.322)	0.073 (0.032)	1.8	5.0	200
9018	1.101 (0.108)	-0.411 (0.088)	0.001 (0.001)	11.1	9.0	677
9062	1.151 (0.213)	-1.198 (0.211)	0.001 (0.001)	3.0	8.0	260
9194	1.847 (0.656)	1.419 (0.217)	0.102 (0.035)	1.8	7.0	253
9204	1.660 (0.308)	1.180 (0.155)	0.001 (0.001)	2.8	6.0	316
9262	1.325 (0.233)	1.212 (0.173)	0.001 (0.001)	5.5	5.0	336
9264	1.518 (0.271)	1.107 (0.139)	0.001 (0.001)	3.7	5.0	316
9271	1.211 (0.233)	-0.931 (0.208)	0.001 (0.001)	5.2	8.0	225
9329	1.520 (0.449)	0.387 (0.238)	0.250 (0.086)	3.4	7.0	200

9340	0.854 (0.233)	-1.329 (0.478)	0.194 (0.087)	5.3	8.0	217
9528	0.889 (0.237)	1.377 (0.295)	0.132 (0.059)	7.5	7.0	295
9545	0.812 (0.206)	-2.679 (0.722)	0.001 (0.001)	3.9	6.0	295
10592	1.006 (0.180)	-1.150 (0.224)	0.001 (0.001)	10.0	7.0	312
10836	1.052 (0.131)	0.051 (0.087)	0.001 (0.001)	8.4	8.0	746
11192	1.052 (0.158)	-2.315 (0.336)	0.196 (0.087)	2.8	7.0	692
11213	0.653 (0.137)	0.817 (0.321)	0.161 (0.067)	9.5	9.0	692
11232	0.977 (0.234)	-0.053 (0.309)	0.213 (0.086)	9.0	9.0	316
11233	0.967 (0.247)	0.017 (0.344)	0.256 (0.091)	6.7	8.0	336
11234	0.703 (0.199)	1.086 (0.484)	0.237 (0.086)	14.9	9.0	312
11235	0.953 (0.251)	0.977 (0.305)	0.186 (0.071)	2.1	9.0	316
11237	0.622 (0.207)	2.572 (0.627)	0.243 (0.061)	5.3	9.0	692
11337	0.959 (0.164)	1.150 (0.180)	0.001 (0.001)	7.9	7.0	295
11449	1.681 (0.645)	2.486 (0.441)	0.041 (0.016)	2.7	6.0	428
11507	1.005 (0.250)	2.854 (0.575)	0.001 (0.001)	0.8	5.0	201
11511	1.025 (0.275)	1.650 (0.276)	0.148 (0.055)	1.8	8.0	370
11587	1.291 (0.414)	0.854 (0.278)	0.283 (0.086)	9.1	8.0	271
11632	1.037 (0.332)	1.465 (0.336)	0.189 (0.076)	4.4	7.0	232

Students' Understanding of Ions

Many students have difficulty grasping the idea that ions have different structures than atoms and covalent compounds, particularly for polyatomic ions (*17, 30, 34, 54-56*). In fact, our analysis shows that lower ability students cannot distinguish the difference between an ion, atom or molecule. For example, when asked, "How many ions are present in one formula unit of the compound shown above?" The compound given contained a polyatomic ion, such as Na_2CO_3, (item 7144) UGA students that make a D or F on their first exam either count the number of atoms or indicate that there is only one ion (Table 5.2) Analysis of other items indicates that their confusion is independent of the polyatomic ion (items 7147 and 7140). On this same examination students were asked, in a multiple answer format, to "Choose all of the *x* compounds from the list below," where *x* is either the word "ionic" or the word "covalent". For example, item 11213's compounds were $Ca(OH)_2$, Li_3N, Sr_3N_2, CO_2, NI_3 and CBr_4. This item

has an ability of 0.817, indicating that only A, B and higher C students could answer this item correctly. This item's ICC has a small slope of 0.653, thus it did not discriminate well between students. The shallow slope indicates that some of the lower end C students answered the item correctly whereas some of the higher C students did not.

Another example of this confusion of atoms, ions and molecules is item 9262 that asked, "How many mL of *x M* $Sr(OH)_2$ are required to make *y* mL of a *z M* $Sr(OH)_2$ solution? What is the molar concentration of the Sr^{2+} ions in the *z M* $Sr(OH)_2$ solution? What is the molar concentration of the OH^- ions in the *z M* $Sr(OH)_2$ solution?". The first part of the item was not difficult for the students, as shown by another solution dilution item asked on the same examination (item 10592), which had an ability of -1.150. Such a low ability demonstrates that the item is extremely easy; therefore, it is an F discriminating item which A, B, C, D and many F students were able to answer the item correctly. It was, however, the second and third parts of item 9262 that made the problem difficult for students; this more difficult ion question discriminated effectively between high B and low B students. Only A and high B students could correctly calculate the ionic solution concentrations of Sr^{2+} and OH^-. These results are independent of the polyatomic ionic compound given, as shown in item 9294. On this same examination, students were also assessed on another item requiring the understanding of ions. The item read, "There are 20.0 drops of solution in 1.00 mL of solution. How many bromide, Br^{1-}, ions are present in *x* drops of *y M* $MgBr_2$?" This item, 9204, also discriminated between high B and low B students. Our analysis shows that unit conversion problems have always been easy for our students (item 9018), so the difficulty of this question was not due to the unit conversions that arise when working this problem. We therefore conclude that the difficulty arises with the ion concept.

On our first-semester final exam we asked the students "What mass of $FeCl_3$ would contain the same **total** number of ions as 16.8 g of $Al_2(SO_4)_3$." This item, 1011, with an ability of 1.145, effectively discriminates between our B and C students. The students can easily convert from grams to atoms and grams to moles (items 9062 and 1009, respectively) showing that the general understanding is there for most of the problem in item 1011. The real difficulty for this problem is the result of either the concept of ions or the large number of steps in the problem. This is just conjecture at this point requiring further research to verify one or the other; the independent steps can be examined individually to elucidate the cause of the students' problems.

Students' Understanding of Molecular Polarity

Students struggle with molecular polarity because they have specific misconceptions including the notion that individual atoms have polarity (*17, 20*). Many students do not combine the concept of electronegativity with that of polarity believing that electronegativity has nothing to do with polarity (*17, 57*). Other students have trouble visualizing the molecular shape of molecules (*33, 57*). During the first semester, students were asked on item 1055, "Which of the following molecules are nonpolar? CCl_4, CH_2Cl_2, CH_3Cl, $CHCl_3$, SiH_2Cl_2." This item discriminates well between high F and low F students. However, if the students are asked to take two seemingly related concepts and put them together, the question discriminates between high B and B students at UGA. For example, students were asked to choose, from a list of molecules, the one that is nonpolar but contains polar covalent bonds (item 1056). The molecules given for the five options were "NH_3, H_2Te, $SOCl_2$ (S is the central atom), $BeBr_2$, and HF." Only A and

very high B students could answer this question correctly. Correctly responding to this item is contingent mainly on determining the shape of the molecule.

An understanding of polarity is required to determine the intermolecular forces present in the liquid and solid states of a molecule. Thus, it comes as no surprise that students at UGA also struggle with intermolecular forces (*17, 34, 35, 37, 58*). Students were asked on their first examination in the second semester of general chemistry to "Choose the dominant (strongest) intermolecular attraction in the liquid state for each of the substances listed below" (item 9528). The substances to choose from were:

- strontium sulfide – Sr_2S
- dimethyl ether – $H_3C\text{-}O\text{-}CH_3$
- methyl amine $–CH_3\text{-}NH_2$
- carbon tetrachloride $–CCl_4$

Each of these compounds had a corresponding choose box with the following four options: "ion-ion attraction, London dispersion forces, dipole-dipole attraction, hydrogen bonding." This item was difficult for C, D and F students; lower F students are unable to determine if a molecule is polar or nonpolar, making it nearly impossible for them to answer item 9528. Given that the intermolecular forces are contingent on the polarity of the interacting molecules, it is extremely difficult for lower ability students to determine the intermolecular force present. Many students believe that intermolecular and intramolecular attractions are the same (*i.e.* they believe that covalent bonds and intermolecular attractions are one and the same) (*35, 37, 59*). Inappropriate wording of questions can compound this confusion. For example: "What intermolecular force exists in molecule X?" should be worded as "What intermolecular force exists between X molecules in the liquid state?" The initial example question could lead to significant confusion between intramolecular and intermolecular forces.

Another intermolecular force item that proved difficult was item 2013, which asked students "Which response correctly identifies all the interactions that might affect the properties of BrI?" Listed below are the options that were given to the students.

- Dispersion force, ion-ion interaction
- Hydrogen bonding force, dispersion force
- Permanent dipole force
- Permanent dipole force, dispersion force
- Dispersion force

Only high A students at UGA could successfully determine all of the intermolecular interactions present in BrI. Two conclusions can be drawn from the IRT analysis of polarity and intermolecular forces: 1) more time needs to be spent in the classroom discussing these topics; 2) wording used while discussing intermolecular forces or while asking questions must be thought out very carefully to avoid reinforcing the students' misconceptions.

Students' Understanding of Quantum Numbers

Another topic that warrants more attention in class is quantum numbers. This can be easily accomplished because atomic electron configurations and quantum numbers are usually taught simultaneously. However, understanding the meaning of quantum numbers is a very difficult topic for most students, while atomic electron configurations are quite easy. Less time should be spent in class covering electron configuration so more time can be spent on understanding quantum numbers. It has been found that many students can remember the rules for quantum numbers, i.e. n, ℓ, m_ℓ and m_s; however, the physical meaning associated with these numbers tends to elude them (*17, 31, 60-62*). This is symptomatic of the recurring theme throughout chemistry that students place stronger emphasis on memorization than garnering physical understanding. On an hourly examination, students were asked to "Choose the set of

quantum numbers which would not be correct for any of the electrons in the ground state configuration of the element S." The students were given the following five options:

- $n = 3$, $\ell = 2$, $m_\ell = 0$, $m_s = +1/2$
- $n = 2$, $\ell = 1$, $m_\ell = -1$, $m_s = -1/2$
- $n = 3$, $\ell = 0$, $m_\ell = 0$, $m_s = +1/2$
- $n = 2$, $\ell = 0$, $m_\ell = 0$, $m_s = -1/2$
- $n = 1$, $\ell = 0$, $m_\ell = 0$, $m_s = -1/2$

This item asked the students to use their knowledge of quantum numbers, as well as their understanding of the aufbau principle and Pauli exclusion; only the A, B and C students answered item 9329 correctly. The addition of the conceptual part of the question made the item much more difficult than asking the students to simply assign quantum numbers to a specific orbital such as in item 7480.

Application of a physical understanding to quantum numbers also provides great confusion for students. On the second exam of the first semester, students were asked in item 11511, "Which quantum number describes the specific orbital within a subshell that an electron occupies?" Only A students could select, from the four choices, the magnetic quantum number as their answer. A similar question was asked on the first semester final exam and, again, only the A students answered it correctly (item 1032).

Out of all of the electron configuration items on the examinations throughout the academic year, only one question proved problematic for B and lower students. For item 8824 students were asked four free response questions:

- "How many d electrons are there in the ground state electron configuration for the element Fe?"
- "What is the value of the *n* quantum number for the d electrons in Fe?"
- "How many of the d electrons in Fe have +1/2 spins?"
- "How many of the d electrons in Fe have -1/2 spins?"

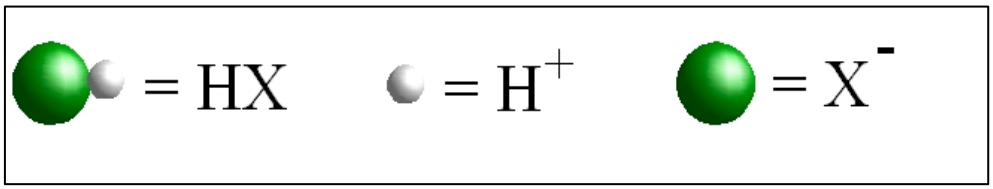
Even though these were free response questions, there are only so many logical answers a student should use; therefore, this question was analyzed using the three rather than the two-parameter model. This item's ICC has an ability of 1.869, which is indicative of a high degree of difficulty for students, and only the A students answered it correctly. This item's ICC also has a large slope of 1.519 indicating that the item discriminates well between A and B students; B, C, D and F students were unable to answer it successfully. Furthermore, the guessing parameter for this item is 0.073, which shows that students, with any amount of chemical knowledge, had a 7.3% chance of guessing the item correctly. Even though there were a limited number of logical responses, students were unable to "guess" the item correctly.

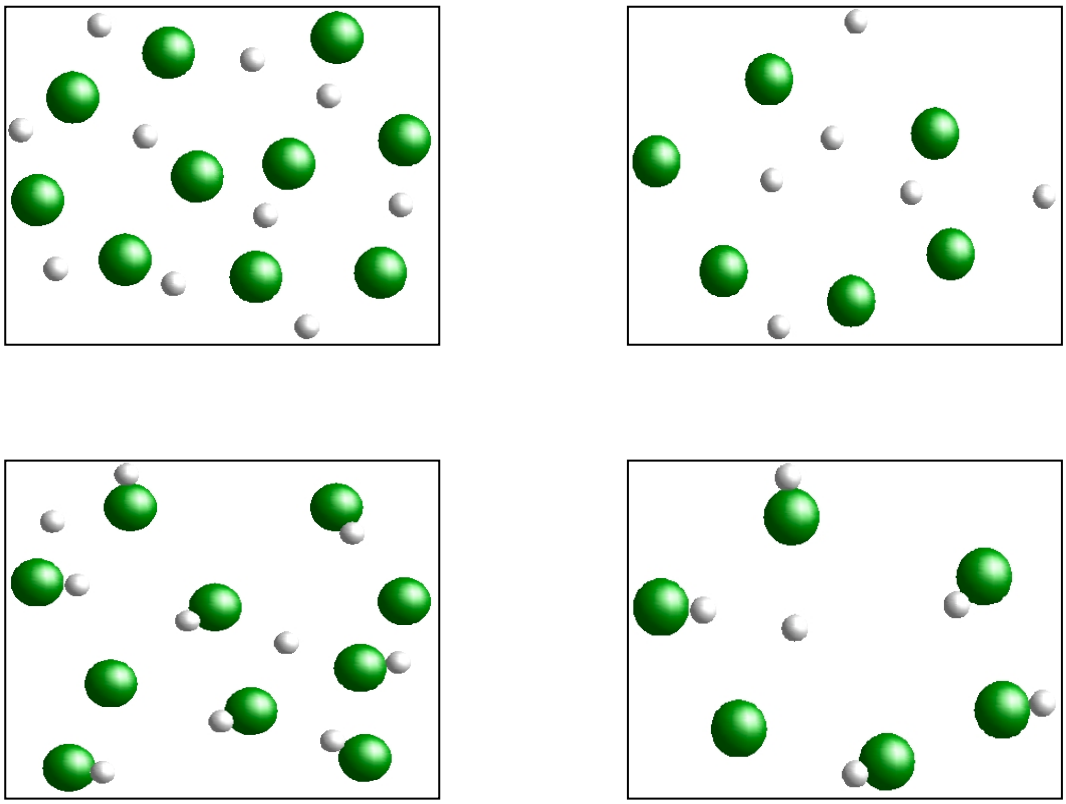
Students' Understanding of the Terms "Strong, Weak, Concentrated and Dilute"

Confusion of the terms strong, concentrated, weak and dilute is a pervasive problem in general chemistry courses (*30*). Several students harbor the misconception that strong and concentrated are equivalent terms and that weak and dilute are as well. This confusion might stem from everyday use of the words in an interchangeable fashion, such as written in Gabel's paper, we say "the coffee is strong", rather than "the coffee is concentrated" (*63*). Exam questions on this subject were designed to force the students to recognize strong electrolytes, weak electrolytes and nonelectrolyes as well as apply their understanding of concentrated and dilute. For example, almost all students could correctly pick a weak or strong acid from a list of molecules, items 1726 (86.6% students correctly responded) and 1724 (84.5% students correctly responded). Both items strongly discriminated between low and lower ability F students, as evidenced by their large ICC slopes and item abilities

Even though students do not have difficulty determining if an acid is strong or weak, they do have difficulty deciding if it can form a concentrated or dilute solution. On the same exam, the students were given a list of water-soluble compounds, (HCl, HF and CH_3OH) and were asked to choose which could form dilute solutions (item 11449). This item had an ability of 2.486 indicating that only the high A students understand that all three compounds can form dilute solutions. The second part of item 11449 asked which of the three compounds could form concentrated solutions; once again, only the high A students answered this item correctly. Most other students selected HCl to form a concentrated solution and HF to form a dilute solution, further confirming the students' misconception that concentrated = strong and weak = dilute.

It has been shown in previous studies that students have difficulty with molecular level understanding (*24, 30, 39, 54, 55, 64, 65*). We found that at UGA that students struggle with molecular level understanding of the terms concentrated, dilute, strong and weak. The previous item, 11449, was on the same examination as a molecular image problem that also examined the students' understanding of the terms concentrated, dilute, strong and weak. Students were presented four molecular images of solutions (shown below) and asked to choose the one that best fits one of the four following solution descriptions - concentrated/strong, concentrated/weak, dilute/weak or dilute/strong (items 11232-11235, respectively). The students were also given a key for the colored spheres; the key and the four images the students were given are shown.





On the surface these four, seemingly related, items appear very similar; however, they discriminate between ability levels very differently. The most difficult combinations are the dilute/weak acid and dilute/strong combinations, having similar abilities of 1.086 and 0.977 that discriminate between B and C students. Contrastingly, A, B, C and high D students could correctly choose the image of a concentrated/weak acid or concentrated/strong acid. This certainly indicates that students have difficulty with the term dilute, since both of the items that asked students to choose the dilute solution were more difficult than the ones asking the students

to choose the concentrated solutions. On the second-semester final examination, students again were asked, "which of the following images best represents a concentrated/weak solution". They were again given four images along with the option, "none of these." As in their first encounter with this problem, it was still difficult for students. The item had an ability of 0.736; low C, D and F students were unable to answer this item correctly (item 2026). The item actually became more difficult when the fifth option, "none of these," was added and when the item was not asked directly after it was covered in the class. The students' understanding of this concept was not corrected in the intervening timeframe.

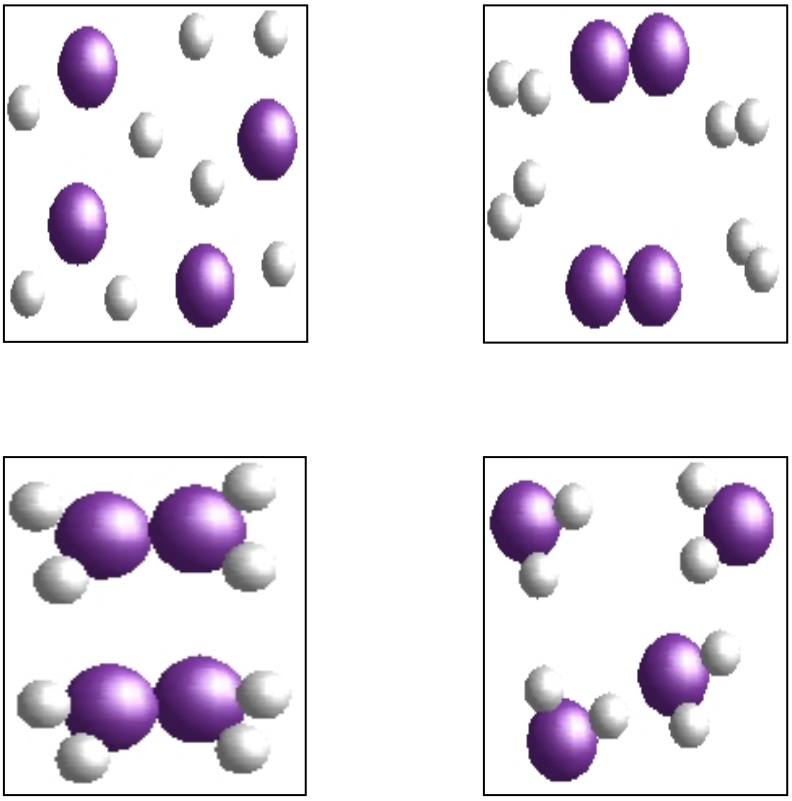
Students' Understanding of Molecular Image Problems

In similarity to other students, students at UGA struggled greatly with molecular level understanding, not just when asked about the terms concentrated, dilute, strong and weak (*28-30, 54, 65, 66*). Many of these items warrant elaboration and will be discussed below. One example of UGA students having difficulty with another type of molecular level questions is an item about the concept of chemical and physical changes. On the very first examination that students were given, they were asked on a macroscopic level, "Which answers are chemical changes and not physical changes?" Below is their list of optional answers (item 5683).

- "freezing of water"
- "rusting of iron"
- "dropping a piece of iron into hydrochloric acid (hydrogen gas is produced)"
- "burning a piece of wood"
- "emission of light by a kerosene oil lamp"

The difficulty for this item, found from its ICC, is 0.501, which indicates that A, B and C students were able to answer the item correctly, and D and F students could not. More importantly with this item, the students were asked about the "freezing of water," a phase

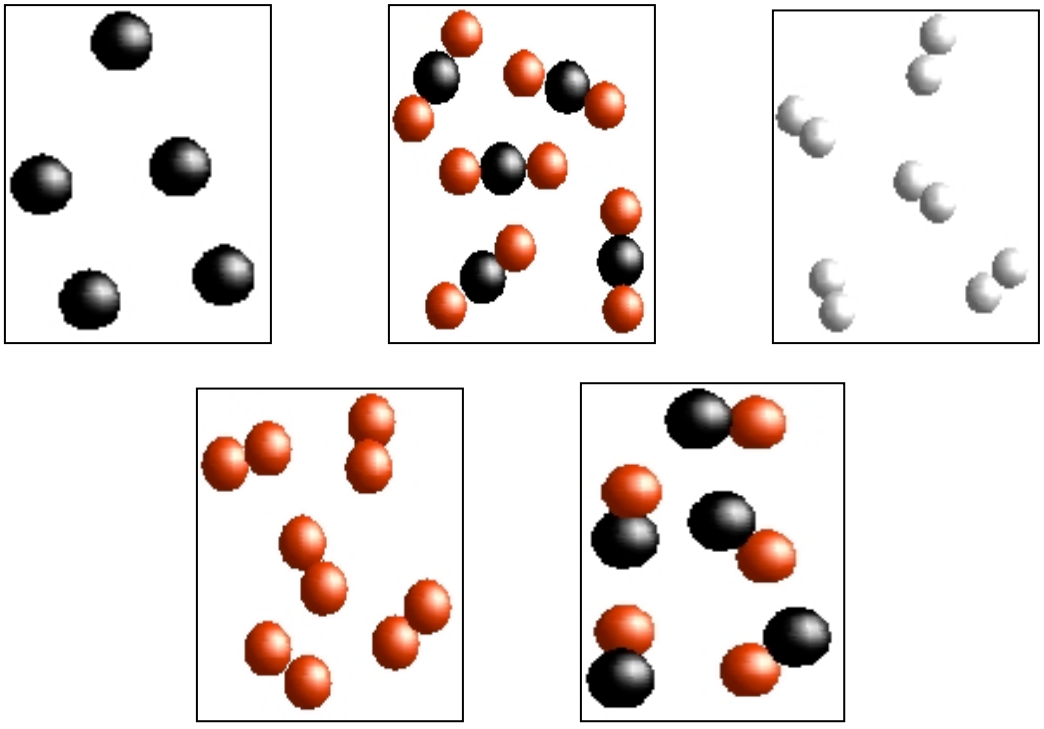
change. Students that answer this item correctly should understand the concept that phase changes are physical changes not chemical changes. On the same examination, the students were also asked a molecular level image item about boiling. Item 11237 reads, "The compound H_4P_2 boils at 57.5°C. Choose the picture that best represents what molecular H_4P_2 looks like after boiling has occurred. A purple ball represents a phosphorous atom, and a white ball represents a hydrogen atom." The four options were:



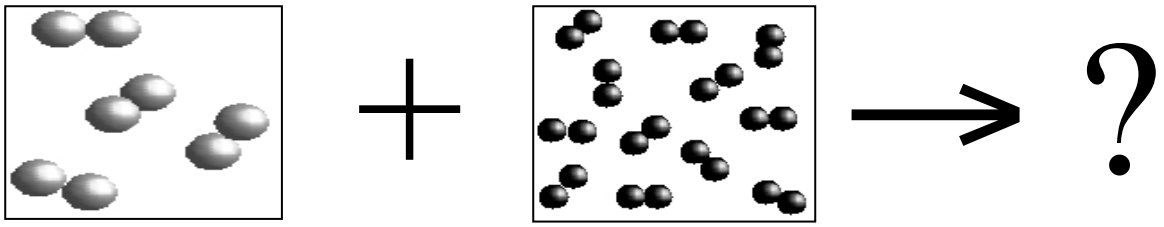
This item had an ability level of 2.572 indicating that only the highest A students were capable of answering it correctly. Asking the students the same concept on a molecular level question rather than a regular question, many A, B, C and D students who could successfully identify that

freezing is a physical change, were unable to correctly choose out of four options the image that represents a physical change.

Another example of a difficult molecular level image problem is an item that was given in second semester general chemistry. The students were asked to look at five images shown below and "rank the following atoms and molecules based on which would effuse the fastest." The students were told that, "A red ball represents an oxygen atom, a black ball represents a carbon atom, and a white ball represents a hydrogen atom." They were also told to "Assume all are gases at the same temperature" (item 11587). A, B and higher ability C students answered this item correctly. Lower ability C students along with D and F students could not answer the item correctly; they did not understand the concept of effusion or the molecular images.

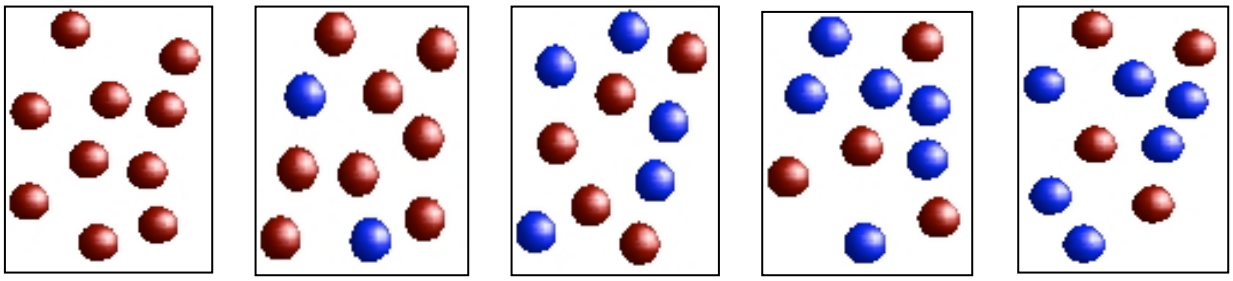


On the second semester final examination, students were given a molecular image problem, item 2066, that incorporated a balanced equation and Avogadro's law. This item was extremely difficult, and only the highest A students answered it correctly. The students were told, "The first image represents 4 L of gaseous nitrogen molecules. The second image represents hydrogen molecules. How many liters of NH_3 will be formed at constant temperature and pressure?" They were given the following equation:



Since this item was asked using molecular level images rather than a numerical question, the problem is much more difficult. The added conceptual understanding that is needed to answer an item like this makes the item too difficult for most students taking the class.

During the second semester general chemistry class, students were asked to solve a molecular image problem about equilibrium. The questions read, "The reaction occurring in these images is R ⇌ B, where R represents a red ball and B represents a blue ball." They were then told to use the following images to determine at what time equilibrium was first established (item 11632).



t = 0.0 s t = 5.0 s t = 10.0 s t = 15.0 s t = 20.0 s

Once again only A and high B ability students could answer this item correctly. Through our research we found that even with having molecular level images on every examination, students still struggle with them. These items are conceptually difficult, and only the brightest students can answer them correctly every time.

Students' Understanding of Inorganic Nomenclature

Our IRT analysis identified several nomenclature subcategories that highly discriminate between students' ability levels. As the nomenclature rules for a given compound become more complicated, not surprisingly, the questions become discriminating between higher-level students. For example, exam results reveal that some F students could correctly answer questions about ionic compounds with transition metal cations (item 7532), making this a poor candidate for separating grade levels. However, D and F students failed to correctly name binary covalent compounds (item 7398). All A students could correctly name ternary acids such as HIO_2 (item 7416), whereas only high A students correctly named ternary acid salts such as RbH_2AsO_4 (item 11507).

Students' Understanding of Mole Concepts

As found in earlier studies, (*67-70*) students at UGA also struggle with the concept of the mole. Through a conceptual item asked on an examination, it was seen that only the A and high B students conceptually understand the difference between a mole of a monatomic element and a mole of a diatomic molecule. Students were asked to compare a mole of Mg to a mole of H_2 (item 9194). The item read, "The density of Mg is 1.74 g/mL. The density of H_2 is 8.9×10^{-5} g/ml. In the following responses compare 1.0 mole of Mg to 1.0 mole of H_2." There were four different comparisons; they were:

- The mass of 1.0 mole of Mg is >, < or = the mass of 1.0 mole of H_2.
- The volume of 1.0 mole of Mg is >, < or = the volume of 1.0 mole of H_2.
- The number of atoms in 1.0 mole of Mg is >, < or = the number of **molecules** in 1.0 mole of H_2.
- The number of atoms in 1.0 mole of Mg is >, < or = the number of atoms in 1.0 mole of H_2.

This item's characteristic curve has a very large slope of 1.847 indicating that it discriminates well between students. This indicates that students with an ability above 1.419 completely understand the mole concept and can answer the item correctly. Students with abilities less than 1.419, struggle with the mole concept having only a 10.2% chance of guessing the item correctly.

Students' Understanding of Solution Calorimetry

Students also struggle with solution calorimetry in general chemistry (*71*). On the second semester, second examination, students were asked to calculate a calorimeter constant (item 10836), which was difficult for D and F students. On the same examination, students were asked to calculate the amount of heat liberated or absorbed when a reaction occurred between a specific amount of $PbBr_2$ and NaCl (item 8178). The students were given the balanced equation, initial and final temperatures and the heat capacity for the calorimeter. Only A and B students were

able to answer this item correctly. A similar problem was given on the semester final, but now the students were asked about NaOH and HCl (item 2026). Even though many solution calorimetry experiments are performed in the laboratory, this item on the final was still difficult for lower end C, D and F ability students. Solution calorimetry is taught near the same time as calculating ΔH_{rxn} from ΔH_f and calculating ΔH_{rxn} using Hess's Law (items 8316 and 8342 respectively). Both of these topics are easy for students so less time should be spent covering them leaving more time to be spent focusing on calorimetry.

Topics that are Easy for Students

Some topics in general chemistry are easier for students, and discriminate only between low student ability levels. These topics have sufficiently low abilities that virtually all students can answer questions regarding them. These topics include: unit conversions (9018), significant figures (11192), balancing equations (231), oxidation states (9271), stoichiometry (without excess reagent) (7238), electron configuration (1034), empirical formula (2942), numbers of protons, neutrons and electrons (7438), atomic radii (9340), dilutions (10592), the ideal gas law (9545), phase diagrams (8078), osmotic pressure (8346), freezing point depression (1408) ΔH_{rxn} from ΔH_f (8316), Hess's Law (8342) and rate law calculations (8411, 8435). These topics are relatively easy for the students to comprehend and should require less in-class instructional time.

Importance of Question Wording

It was found when analyzing the items given on the examinations throughout the academic year that wording plays an important role in determining the difficulty of a question

(*24, 69, 72-74*). This is illustrated in the comparison of two items regarding gas laws. During the first exam of second semester general chemistry, students were asked the following free response question: "A sample of dry air of total mass 1.000 g consists of 0.700 g of nitrogen and 0.300 g of oxygen. Calculate the total pressure (in atm) when this sample is in a flask of *x* L at *y* °C" (item 11337). Only A and B students were able to answer this question correctly. On the final examination for the course, students were asked "What is the pressure exerted by a mixture of 14.0 grams of N_2, 71.0 grams of Cl_2 and 16.0 grams of He in a 50.0-liter container at 0°C." This time 96.9% of students were able to answer the question correctly (item 2009).

There are three differences in these questions. One is that in the final exam question the sample consisted of three gases rather than two. The second difference is that the more difficult item was a free response item, and the easier item was given as a 5-option multiple choice. Furthermore, the first question used the words nitrogen and oxygen whereas the second question used the chemical symbols N_2, Cl_2 and He. The use of three gases would seemingly make the question more difficult since it is longer and presents more opportunities for errors. However, our data shows that it had a lower ability level. We therefore hypothesize that students forgot that the gases are diatomic when presented only with the name of the diatomic molecule. This led to use of the wrong molar masses and an incorrect calculation of the total pressure. Specific items should be designed for future examinations to test this hypothesis. Question wording is vital to the composition of a good examination, as for some topics it can determine the difficulty of the question. Indeed, even changes as small as the substitution of a single word can cause an easy question to be quite difficult or vice versa.

CHAPTER 6

ANALYSIS OF FALL 2006 – SPRING 2007 ACADEMIC YEAR EXAMINATIONS

Comparison of Students' Abilities and IRT Analysis

Upon completion of the 2005-2006 academic year and the IRT analysis of that year's data, all general chemistry instructors were informed of the IRT findings. Topics that were difficult for students such as the concept of ions, polarity, intermolecular forces, quantum numbers, the mole concept, and solution calorimetry were discussed with the instructors. Armed with this knowledge, the class instructors went into the classroom knowing what students found to be both difficult and easy. While it is impossible to know what actually occurred in the classroom on a daily basis, reminders were given to all instructors when a difficult topic was about to be taught in the classroom. The instructors were asked to spend more time on these tougher topics or to consider teaching these topics in a different, more effective way.

Upon conclusion of the 2006-2007 academic year IRT analysis was performed on the examinations. Similar to the previous academic year, the second examination given in the first semester of general chemistry was analyzed using the three-parameter IRT model with free response items constrained to a guessing factor, *c*, of 0.00. Once the second examination analysis was completed, the resulting parameters for these items were held constant while items from the other 7 examinations were allowed to freely converge to the two or three-parameter model dependent on the type of item. (Free response items were forced to have a guessing factor of 0.00 while the other items were able to converge to the guessing factor that best fit the ICC.)

Results of the IRT analysis demonstrated that students needed almost the same ability to achieve the same letter grade from the 2005-2006 and the 2006-2007 academic years (Tables 5.1 and 4.1). This indicates that even though many exam questions differ between the two academic years, the same amount of chemical knowledge is expected of students from one year to the next. This shows a consistency in the general chemistry program, *i.e.* the class is not easier one year in comparison to the other year.

Based upon this analysis it was of interest to see if students in the fall 2004 – spring 2005 academic year needed the same ability to receive an A, B, C, D and F in the class as in the latter years. We reran the 2004 – 2005 academic year examinations with IRT similarly to the 2005 and 2006 academic years analyses. Therefore, the analysis was performed by using a two-stage process with optimized parameters for examination two appearing in a constrained fashion in the remaining seven examinations' analyses. Due to the large numbers of each item in each subset, we could only analyze each question rather than the specific items, but from the overall analysis of the questions it is still possible to determine the ability needed to receive a specific grade in the class. In the 2004-2005 academic year, a slightly smaller ability (less chemistry knowledge) was needed to receive the same grade in the class than was needed for the latter two academic years (Table 6.1).

Table 6.1: Ability needed for each letter grade for fall 2004 – spring 2005 academic year.

A	B	C	D	F
1.48652	0.829556	0.172595	-0.484366	< -0.484366

Many of the items on the 2005-2006 academic year examinations were also used on the 2006-2007 examinations. These items proved to still discriminate between students of similar abilities comparable to the data from the previous year. See Table 6.2 for a small comparison of item parameters from items given during both the 2005 and 2006 academic years. Data from the 2006 academic year analysis is indicated with an asterisk next to the item number.

Table 6.2: Comparing parameters of items given on both 2005 and 2006 academic years.

Item Number	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D.F.	Number of Students
4950	2.352 (0.575)	-0.886 (0.232)	0.211 (0.087)	1.8	5.0	262
4950*	1.746 (0.532)	-1.092 (0.369)	0.191 (0.086)	0.9	5.0	283
7400	1.360 (0.234)	0.406 (0.109)	0.001 (0.001)	3.2	6.0	264
7400*	1.054 (0.144)	0.567 (0.110)	0.001 (0.001)	2.6	8.0	437
9468	1.189 (0.222)	1.109 (0.156)	0.001 (0.001)	2.7	6.0	271
9468*	1.248 (0.251)	0.921 (0.144)	0.001 (0.001)	1.2	6.0	207
9490	1.023 (0.166)	-1.231 (0.244)	0.001 (0.001)	6.2	9.0	426
9490*	0.953 (0.150)	-1.268 (0.243)	0.001 (0.001)	9.3	9.0	507
9743	0.948 (0.139)	-0.973 (0.175)	0.001 (0.001)	6.5	9.0	544
9743*	1.073 (0.101)	-1.211 (0.129)	0.001 (0.001)	7.3	9.0	1244

An asterisk next to the item number indicates that the item's parameters were from the 2006-2007 academic year IRT analysis.

In addition to the existent items in JExam used on previous years' examinations, new items were programmed into JExam and asked on the 2006-2007 examinations. Some of these new items were written to have more conceptual items on the examinations; others were written to better assess the topics, which we found were most problematic for UGA students. Item parameters for these new items are found in Table 6.3.

Table 6.3: IRT parameters for items on the 2006 – 2007 academic year examinations.

Item Number	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D.F.	Number of Students
1033	1.057 (0.197)	1.142 (0.157)	0.160 (0.051)	4.8	9.0	1231
1042	0.975 (0.318)	2.462 (0.370)	0.184 (0.040)	7.6	8.0	1231
1057	1.086 (0.184)	0.665 (0.174)	0.201 (0.060)	10.0	9.0	1231
2014	0.640 (0.131)	0.057 (0.389)	0.211 (0.089)	11.0	9.0	867
2024	0.839 (0.138)	-1.605 (0.383)	0.182 (0.082)	13.5	9.0	867
2030	1.734 (0.423)	1.176 (0.137)	0.319 (0.051)	9.0	9.0	867
2033	1.857 (0.324)	1.207 (0.091)	0.149 (0.037)	7.6	8.0	867
11604	0.926 (0.277)	1.450 (0.306)	0.197 (0.064)	3.9	9.0	509
11683	1.074 (0.195)	1.104 (0.208)	0.001 (0.001)	1.8	8.0	230
11777	1.249 (0.221)	1.530 (0.22)	0.001 (0.001)	2.5	6.0	303
11786	0.828 (0.206)	2.157 (0.468)	0.001 (0.001)	6.5	7.0	250
11910	1.149 (0.531)	3.035 (0.900)	0.125 (0.041)	0.8	6.0	190
11911	1.183 (0.400)	1.934 (0.393)	0.114 (0.050)	4.3	5.0	207
11912	0.678 (0.178)	-1.712 (0.580)	0.004 (0.002)	11.0	8.0	215
11913	0.678 (0.189)	-1.774 (0.574)	0.004 (0.002)	5.8	7.0	181
11914	0.706 (0.200)	-2.667 (0.816)	0.004 (0.002)	0.9	5.0	193

Students' Understanding of Ions

The previous year IRT analysis showed that students struggled with determining the number of ions in a compound. Items from the previous year asked the students about both nomenclature and the number of ions. It was decided to ask the students an item that only pertains to the number of ions in a compound. On the first examination given in the fall of 2006, students were asked, "How many potassium, oxygen, carbon and carbonate ions are present in one formula unit of this compound? K_2CO_3." The students were given 4 different free response boxes to answer the number of ions for each ion mentioned (item 11683). This item was difficult for students having ability levels below 1.104. Only A and B students answered the item correctly. Since the item has a slope of 1.074, it discriminated well between B and C students. From these results, we conclude that C, D and F students are unable to understand the difference

between an ion and an atom. These students also do not understand that the polyatomic carbonate ion, CO_3^{2-}, does not break up into oxygen and carbon ions.

On the second semester final examination, students were asked for item 2033, "Which **aqueous** solution would have the **lowest** vapor pressure at 25°C?" The students were given the following options:

- 1 *m* NaCl
- 1 *m* Na_3PO_4
- 1 *m* $MgCl_2$
- 1 *m* $C_6H_{12}O_6$
- 1 *m* $C_{12}H_{22}O_{11}$

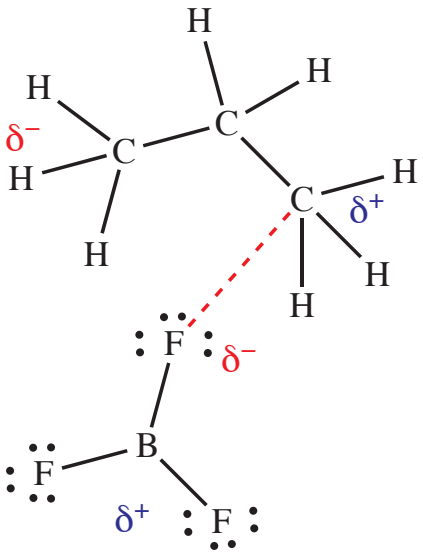
This item has an ability of 1.207 and a slope of 1.857; therefore, A and B students answered the item correctly whereas C, D and F students did not. The first option, NaCl, was the largest distractor having 237 students choose this option. Knowledge of ions is necessary but not sufficient to correctly answer the item. This is another example of how important it is for the students to learn the difference between ions, atoms and molecules at the beginning of the course.

Students' Understanding of Molecular Polarity

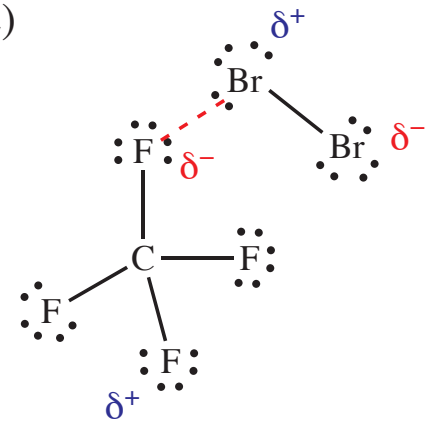
Another topic in the first semester general chemistry class that was found difficult for students was bond polarity and molecular polarity. This topic was difficult for students in both the 2005-2006 and 2006-2007 academic years, and affects their capability to fully understand intermolecular forces. On the first semester final examination, students were asked, "Which bond would have the smallest intrinsic bond dipole moment?" Their options were: H-Br, H-F, H-I, H-O in H_2O and H-Cl (item 1033). The ability of this item was 1.142 indicating that only A and B students could correctly answer this item.

On the first exam of the second semester, some new items on intermolecular forces were programmed into JExam. Each of these items showed the students Lewis structures of two molecules. The question read, "The intermolecular force depicted in this representation is: ion-ion attraction, dipole-dipole attraction, hydrogen bonding or London dispersion forces." The student received one of the following five images (items 11910-11914 respectively).

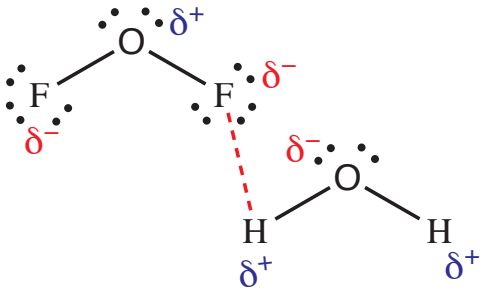
1)


H
H
H
δ−
C
C
H
H
δ+
H
H
F
δ−
B
F
δ+
F

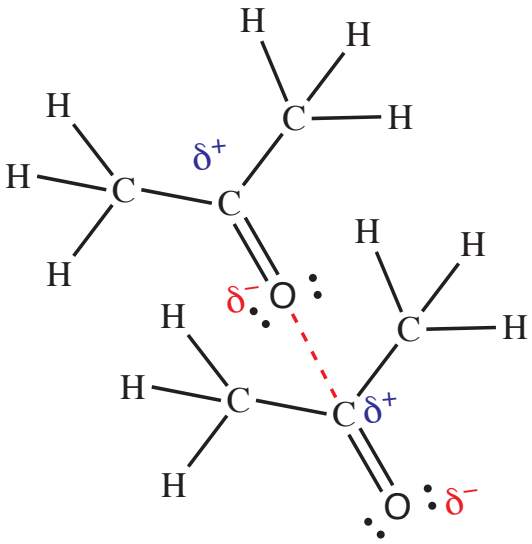
2)


δ+
Br
F
δ−
Br
δ−
C
F
F
F
δ+

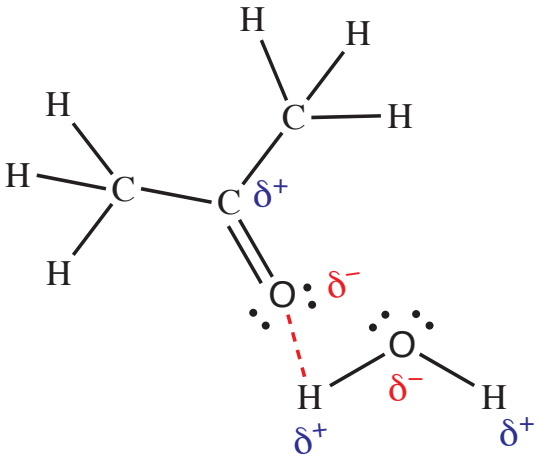
3)


O
δ+
F
δ−
F
δ−
δ−
O
H
H
δ+
δ+

4)


H
H
H
C
H
δ+
H
C
C
H
H
δ−
O
H
H
H
C
H
C
δ+
H
O
δ−

5)


H
H
H
C
H
H
C
C
δ+
H
O
δ−
O
H
δ−
H
H
δ+
δ+

Items 11910, 11913 and 11914 each had less than 200 students answering them. While it is suggested that 200 or more responses are preferred for each item to result in valid IRT analysis, this number is not a steadfast rule (*8*). Since more than 180 students responded to each of these items and the items fit the three-parameter model used, results from these items are valid. When writing the examination, it was believed that items 11910 – 11914 were equivalent; however, IRT analysis indicated that this was not the case. Item 11910 was the most difficult item of the group with an ability of 3.035. Item 11911 also discriminated between A students with an ability of 1.934. Both of these items contained images of nonpolar molecules with an induced dipole. It appears that students saw the indicated attraction and assumed it was a dipole-dipole intermolecular attraction without even looking at the molecules present. The other three items (11912 – 11914) were extremely easy for students to answer. Item 11913 contained an image of a dipole-dipole interaction and items 11912 and 11914 contained images of hydrogen bonding. Since these items are not equivalent, they should not be asked simultaneously on an examination again. If students are given the Lewis structure of the compound and the partially negative and positive charges are shown on the molecule, determining which intermolecular force present is really easy for students. However, if the students are only given the names and formulas of compounds and asked what intermolecular force is present; the item is much more difficult.

On the second semester final examination, students were asked, "Which of the following intermolecular forces is associated with ALL types of compounds?" Listed below are the options that were given to the students (item 2014).

- ion-ion interactions
- London Forces
- covalent bonding
- dipole-dipole interactions
- hydrogen bonding

Since they were not asked about a specific molecule, this item should have been extremely easy for all students. This item had an ability of 0.057, indicating that F ability students were unable to answer this item about intermolecular forces correctly. On the same final, students were asked, "Which of the following interactions is the **strongest** type of intermolecular force?" Below are the students' possible options.

- ion-ion interactions
- London Forces
- hydrogen bonding
- dipole-dipole interactions
- dispersion forces

This item, 2024, was much easier than the previous item that asked about the interaction that all molecules experience. 86.2% of students answered the item correctly, and the item has an ability of -1.605, which discriminates between F and low F students. Later on the same examination, the students were asked about miscibility, which is based on the polarity and intermolecular forces of the molecules. Item 2030 read, "Consider the following pairs of liquids. Which response contains all the pairs that are miscible and none that are immiscible?

I. benzene, C_6H_6, and hexane, C_6H_{14}
II. water and methanol, CH_3OH
III. water and benzene, C_6H_6"

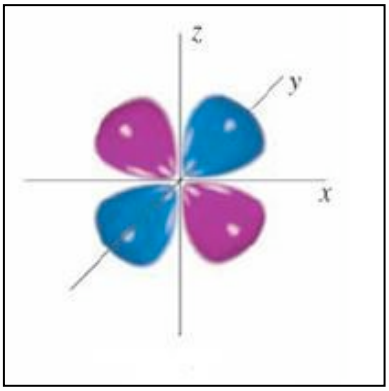
Possible answers were:

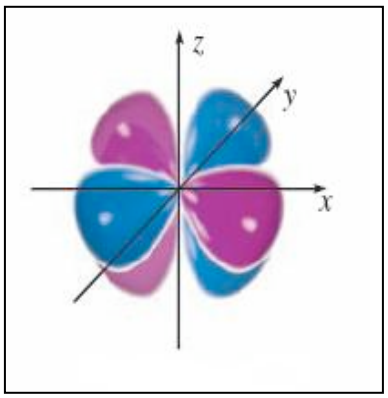
- I
- I and II
- II and III
- II
- I, II and III

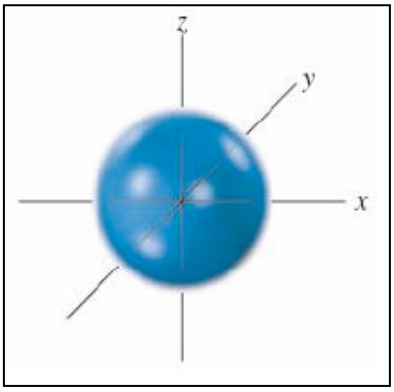
Determining if something is miscible is difficult for C, D and F students at UGA. The item had a difficulty of 1.176 indicating that A and B students answered the item correctly, and C, D and F students did not.

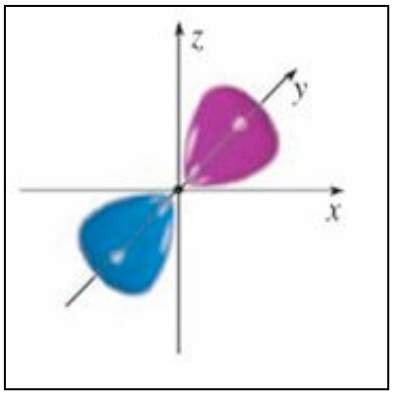
Students' Understanding of Quantum Numbers

In our previous year's IRT analysis we determined that quantum numbers caused difficulty for students. It was decided to try a different item in hopes to better ascertain the students' knowledge of quantum numbers. Instead of asking students a definition or the rules of quantum numbers, as in the previous year, the students were asked a two-part item (item 11786). The first part was free response asking students, "What is the maximum number of electrons in an atom that can be described by the following quantum numbers? $n = 4$, $\ell = 2$." The second part of the item was a multiple choice question asking, "Which image would best represent one of the orbitals the electrons described by $n = 4$ and $\ell = 2$ would be in?" The four images that were options are shown below; the fifth option was, "There is not enough information for the orbital to be determined."









This item's ICC had an ability of 2.157, indicating that it is extremely difficult for almost all students. Only the highest A students could answer this item correctly. This item was not expected to be this difficult; it was expected that the item would be difficult for C, D and F students, but was assumed that the A students could answer it correctly. There are very few steps to this problem, and only a limited understanding of quantum numbers is needed to answer it correctly.

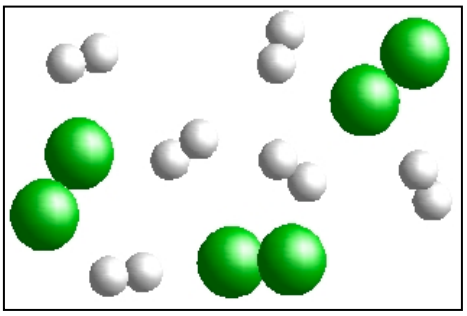
On the final examination, students were asked, "What is the maximum number of electrons in an atom that can be described by the following quantum numbers? n = 3, ℓ =1," which is very similar to the previous item. This item, 1042, was more difficult with a difficulty of 2.462, than item number 11786. The item was highly discriminating with a slope of 0.975, and once again only the high A students could answer this item correctly. Quantum numbers are consistently difficult for almost all of the students at UGA to understand. The difficulty with understanding quantum numbers for UGA students most likely stems from the students struggling to understand abstract concepts (*24, 63*).

Students' Understanding of Molecular Image Problems

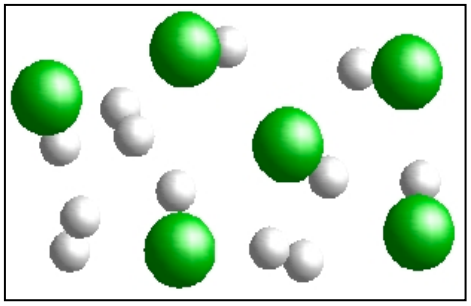
From the previous academic year, it was seen that students struggle with molecular image items on examinations. A new item (11604) was written to assess the students' understanding of a reaction on the molecular level is shown.

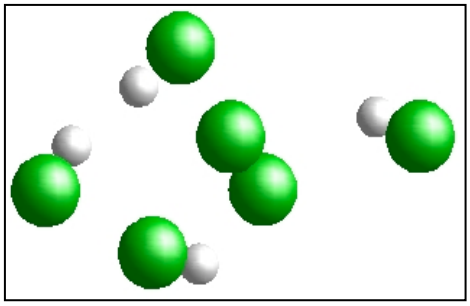
"Depicted below is a reaction vessel containing a mixture of H_2 and Cl_2 gas (before reaction begins). The green spheres represent chlorine atoms and the white spheres represent hydrogen atoms. Chlorine and hydrogen react to form hydrogen chloride, HCl.

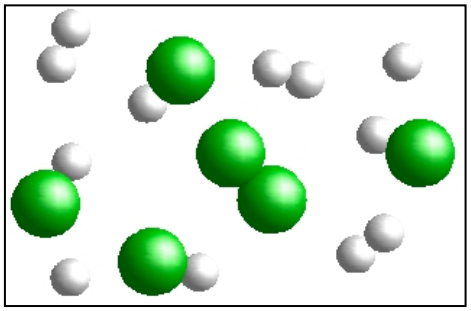
$Cl_2(g) + H_2(g) \rightarrow 2HCl(g)$

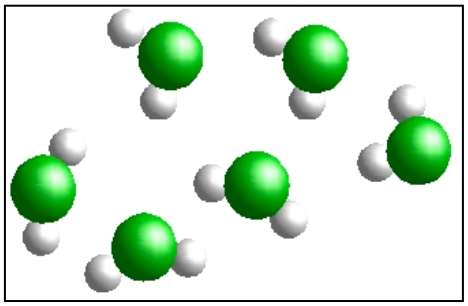


Choose the image below that best represents the reaction vessel if the reaction goes to completion."



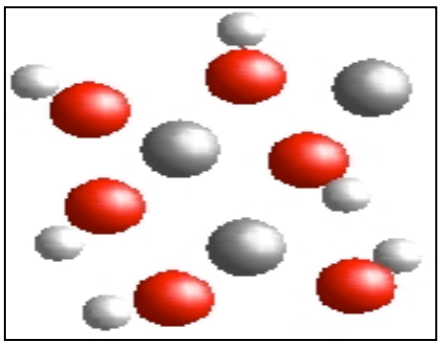


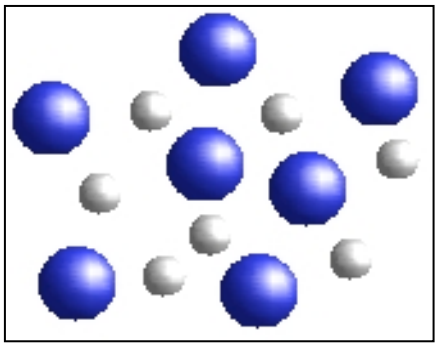




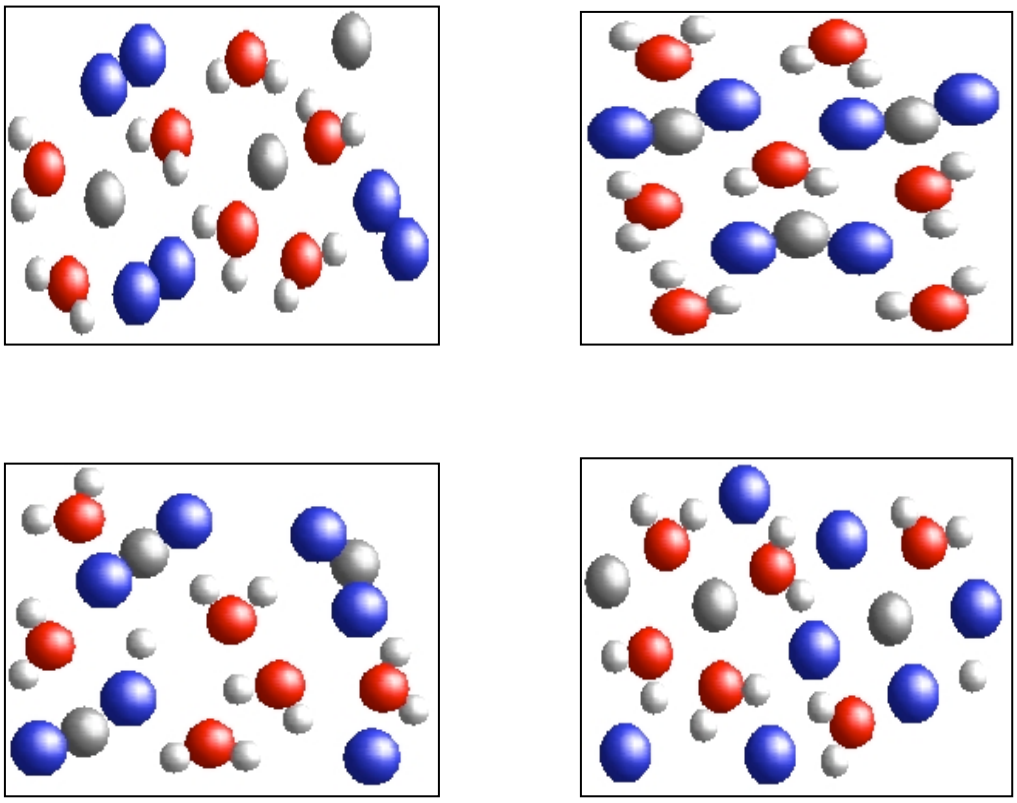
As can be seen in the item, the students were not told that this is a limiting reactant problem, which caused the item to be difficult. This item's ICC had a slope of 0.926 and an ability of 1.450, indicating that only A and high B students can answer the item correctly.

Another new molecular image problem is item number 11777, given on the third examination for the first semester of general chemistry. Here students were asked to once again look at the starting reactants and determine the products after the reaction goes to completion. The starting images were:





The students were informed that the image on the left contains $Ba(OH)_2$ and the image on the right contains HBr. The students were also informed that the reaction took place in water; however, they were not given a balanced reaction equation. The students were first asked to determine the products of the reaction and type them in the free response boxes provided. The students were then told to pick from the following images the "best representation of the products after the two vessels are mixed and the reaction goes to completion." The four options the students were given are:



To answer this item correctly the students must remember that $BaBr_2$ ionizes in water yielding a final image containing Ba^{2+} ions along with $2Br^{1-}$ ions, not $BaBr_2$. One of the item distractors contained the compound Br_2 as one of the products. This option would be chosen if students believed that $BaBr_2$ ionized to produce Ba^{2+} ions and Br_2 ions. Currently there is a limitation with JExam on exporting data. We are unable to analyze the number of students that chose each distractor, and therefore are unable to tell if this was a common choice for students. Item 11777 was analyzed using the two-parameter IRT model since the students had to type in the products that were formed. The item fit this model well indicating that students were unable to "guess" the item correctly. This molecular level understanding item was difficult for all but the A and high B students, yielding the same level of difficulty as the limiting reactant image problem, item 11604, on their previous examination.

On the first semester final examination, a limiting reactant molecular image problem was given. The students were given an image of a vessel before a reaction took place. They were told that the combustion of C_2H_4 in excess O_2 would take place. They were asked what image best represents the mixture after the reaction goes to completion (item 1057). This item was easier than the previously discussed molecular level image problems for students. The students were not given the reaction equation $C_2H_4 + 3O_2 \rightarrow 2H_2O + 2CO_2$; they were also not given which color spheres represent which element. They were given a hint that the problem was a limiting reactant problem because they were told that O_2 would be in excess. The item was still difficult with an ability of 0.665 separating the C from D students. Further research is necessary to determine if this item was easier because the word "excess" was present in the problem, or if it was because A, B and C students can now understand chemical reactions at the molecular level. What our research does show is that even though molecular level understanding is being discussed in class, only the higher ability students correctly answer the molecular image items.

CHAPTER 7

ANALYSIS OF FALL 2007 – SPRING 2008 ACADEMIC YEAR EXAMINATIONS

IRT Analysis

Just as for the two previous academic years, IRT analysis was performed on the eight fall 2007 – spring 2008 examinations. Both the two and three-parameter IRT models were used to analyze items; the model used was dependent upon the item type. After an ICC was calculated for each item, students were assigned a location on the ability scale, and the ability a student needed in order to receive a specific letter grade was evaluated. The relationship between the IRT calculated student ability and the percentage correct on the examinations for each student is shown in Table 7.1.

Table 7.1: Ability needed for each letter grade for fall 2007 – spring 2008 academic year.

A	B	C	D	F
1.81944	1.22333	0.627231	-0.0311277	< -0.0311277

When comparing the IRT ability needed for a specific letter grade over the four years of this study, the student ability needed for a specific letter grade has slightly increased since before IRT analysis was initiated (Tables 4.1, 5.1, 6.1 and 7.1). This indicates that the general chemistry program is becoming slightly more difficult. In the 2004-2005 academic year, an

ability of 1.48652, as seen in Table 6.1 was needed to be considered an A student. In the 2005-2006 academic year (see Table 5.1) an ability of 1.64177 was necessary to answer 90% of the items correctly, and during the 2006-2007 academic year an ability of 1.70493 was needed (Table 4.1). Finally, in the 2007-2008 academic year, a student ability of 1.81944 corresponded to answering 90% or more of the exam items correctly. A similar increasing ability for each grade level from the 2004-2007 academic years also occurred. As a result of our work, items used on the examinations are more discriminating now than before the institution of IRT analysis, which might cause the ability needed for a specific percent to change slightly. Examination items that do not discriminate well have poorly defined midpoints leading to more uncertain item ability. Since students were assigned ability levels using these poorly defined item abilities, their inherent student abilities are also less certain.

Similar to the 2006 - 2007 academic year, more new test items were included on examinations particularly on the topics that caused the students difficulties in earlier semesters. It was hoped that these new items would provide a new insight in which students are having difficulty with specific items and more importantly why. New items were also added to the JExam database to prevent previous students sharing items with students taking the examination in succeeding years. New items covering the previously discussed difficult topics are described in detail below with their IRT parameters given in Table 7.2.

Table 7.2: IRT parameters for items on the fall 2007 - spring 2008 academic year examinations.

Item Number	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D.F.	Number of Students
1012	0.790 (0.115)	-1.009 (0.335)	0.218 (0.092)	2.7	9.0	1201
1013	1.041 (0.369)	2.413 (0.393)	0.284 (0.038)	8.4	9.0	1201
1032	0.626 (0.122)	0.818 (0.343)	0.186 (0.075)	9.3	9.0	1201
1061	1.277 (0.366)	2.086 (0.240)	0.207 (0.032)	8.4	9.0	1201
1062	0.956 (0.197)	1.553 (0.180)	0.133 (0.045)	5.5	9.0	1201
1063	1.607 (0.182)	0.413 (0.084)	0.122 (0.038)	6.1	9.0	1201
12110	0.969 (0.183)	-1.539 (0.346)	0.209 (0.091)	6.8	7.0	384
12114	1.330 (0.400)	1.583 (0.250)	0.111 (0.038)	2.7	8.0	351
12140	1.034 (0.226)	0.386 (0.225)	0.152 (0.064)	8.0	9.0	351
12492	1.058 (0.230)	1.023 (0.199)	0.125 (0.057)	7.8	7.0	454

Students' Understanding of Ions

On the first examination in the fall of 2007, students were asked item 12140, which is similar to item number 11683 given on the first examination in the fall of 2006. Item 12140 first asked students how many oxygen atoms and hydrogen atoms were present in one formula unit of $Al(OH)_3$. It then asked students how many aluminum ions, oxide ions, hydroxide ions and hydrogen ions were present in one formula unit of $Al(OH)_3$. All six of these questions were free response, but since there are a limited number of logical responses, the three-parameter logistic IRT model was used and successfully converged on the data. Item 12140 had an ability of 0.386, indicating that A, B, C and high D students answered the item correctly whereas low D and F ability students did not. Item, 12140, is much easier than item 11683 given the previous year where students were asked to determine the number of potassium, oxygen, carbon and carbonate ions present in one formula unit of K_2CO_3 (See Table 6.3 for item parameters). In item 11683, students were not asked the number of atoms present in the compound. Including both atoms

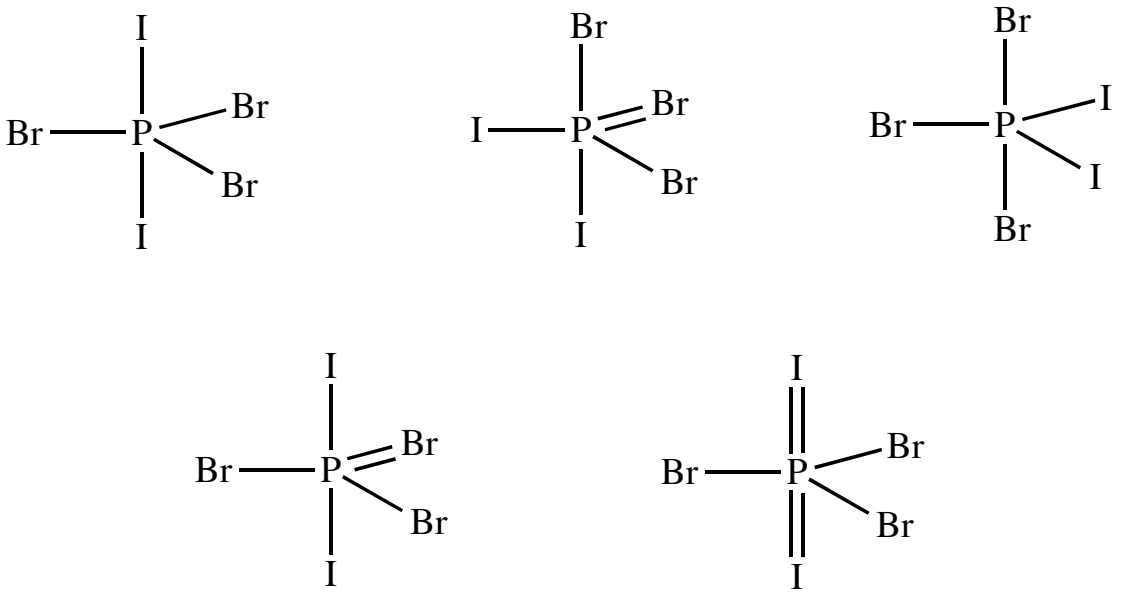
and ions in a single question stimulates the students' thought processes enough to make them distinguish between the two entities.

Previous research has shown that the reason students have difficulty with understanding ions is because they confuse the terms ions, atoms and molecules (*17, 38, 75, 76*). Previous examination items have shown that students at UGA have difficulty understanding ions, and it was believed that for students at UGA this difficulty also stems from them being unable to differentiate between ions, atoms and molecules. To determine the validity of this assumption, on the first examination, students were asked item number 12114 that addresses this specific confusion. Item 12114 read, "Label each of the following as an atom, an ion, a molecule, a formula unit or none of these." The students were instructed to assign the previous descriptors to: MgS, O_2, Mg, SO_3^{2-} and Al^{3+}. This item had an ability of 1.583 indicating that low B, C, D and F students did not answer it correctly. With the present JExam configuration, we are unable to analyze exactly which of the five parts the students could not answer correctly, but from this item we can see that confusion exists for students in identifying atoms, ions, molecules and formula units. This confusion would lead to their difficulty in answering items that asked about ions.

Students' Understanding of Molecular Polarity

During the 2005 and 2006 academic year, IRT analysis proved that deciding if a molecular was polar or nonpolar is difficult for lower ability students. During the 2006 academic year, items 11910-11914 were given on examinations and analyzed (See Table 6.3 for item parameters). These items asked students to look at an imbedded image to determine the intermolecular force present. The analysis of this item indicated that when the students were not

asked to draw the Lewis structures and the dipoles were shown, almost all students could identify an image of hydrogen bonding or dipole-dipole interaction. To determine if the students were struggling with drawing the molecules given or determining if the molecule is polar, item 1013 was given on the first semester final examination. The item read, "A sample of PBr_3I_2 dissolves in water. The best three-dimensional representations of five potential structures are shown below. Which is the most likely structure for this sample?" The five options are shown below.



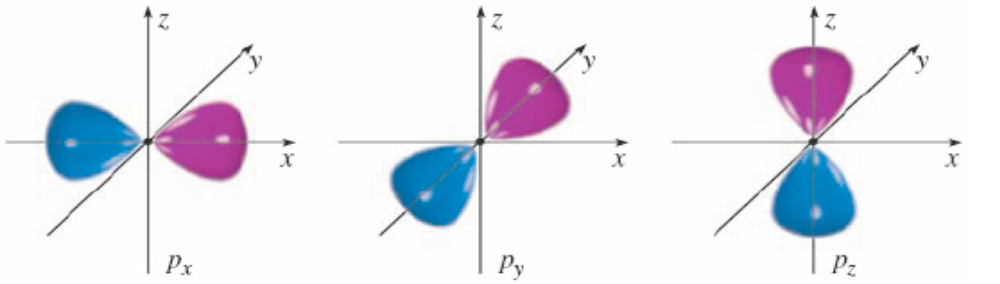
This item had an ability of 2.413 and an ICC with a slope of 1.041 indicating it was extremely difficult for all but the high A students. The students had a 28.4% chance of guessing this item correctly, indicating that most of the students found at least one of the options not "attractive," and avoided choosing it if they guessed the answer. This item indicates again how much the students have difficulty with polarity. If students are given a choice of images and are not given the partial positive and negative parts of the molecule, most students have difficulty determining if the molecule is polar or nonpolar.

To determine if students are struggling with drawing the Lewis structure of the molecule rather than determining the polarity of the molecule, students were asked to determine the molecular and electronic geometry of SO_2 on the same examination (item 1012). The students had little difficulty with this item, indicated by the item's ability of -1.009. A, B, C, D and many F students chose the correct response. It is believed that the difficulty with determining molecular polarity appears to be determining the symmetry of the molecule and the bond diploes not with drawing the Lewis structure.

Students' Understanding of Quantum Numbers

On the final examination of the first semester course, students were asked three items that contained the rules or concepts of quantum numbers. First, students were asked in item 1061, "In the iron atom, how many electrons are in orbitals for which the angular momentum quantum number is 0?" Their options were 2, 8, 6, 10 and 26. Item 1061 discriminated well between high A and A students. Then in item 1062 they were asked, "Which quantum number distinguishes between the three *p* orbitals shown below?

"



The students' options were:

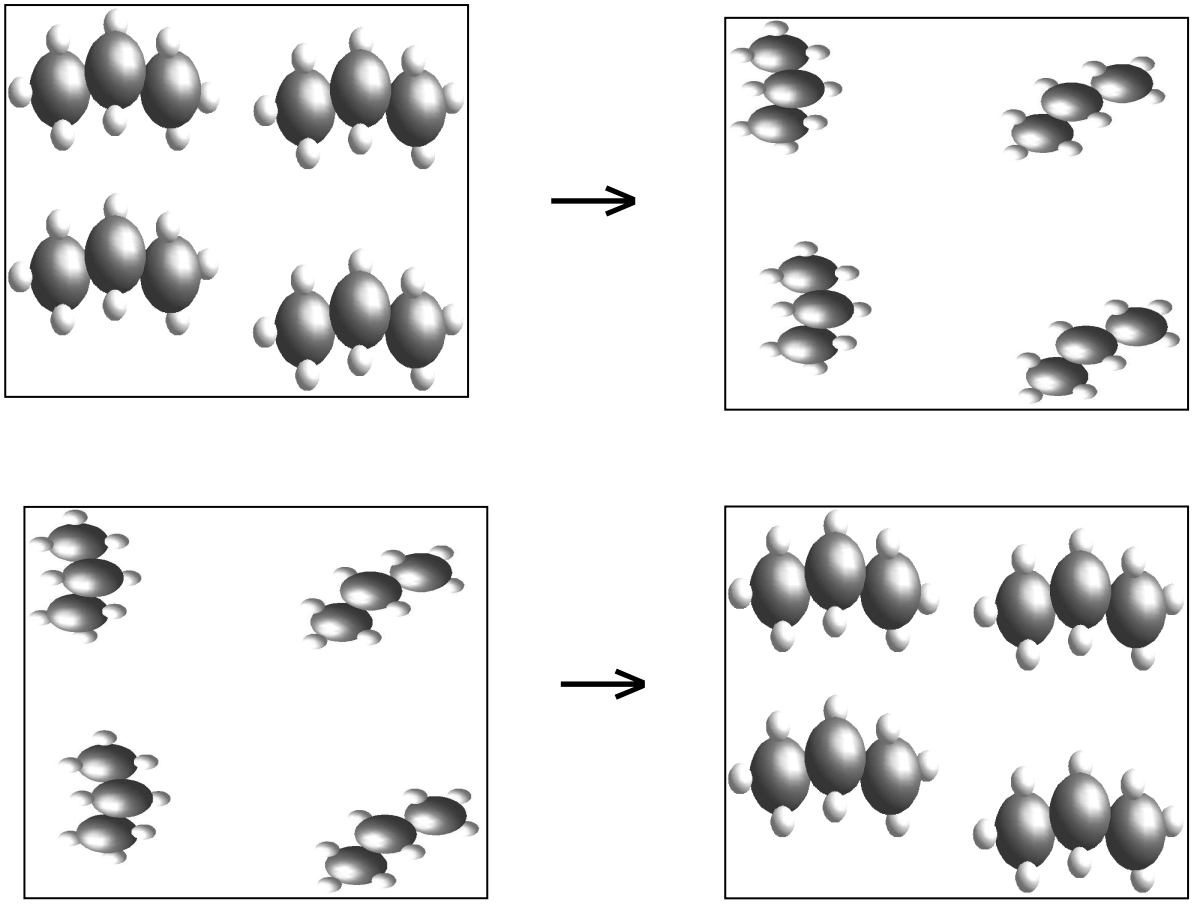
- the principle quantum number
- the magnetic quantum number
- the angular momentum quantum number
- the spin quantum number
- none of the quantum numbers

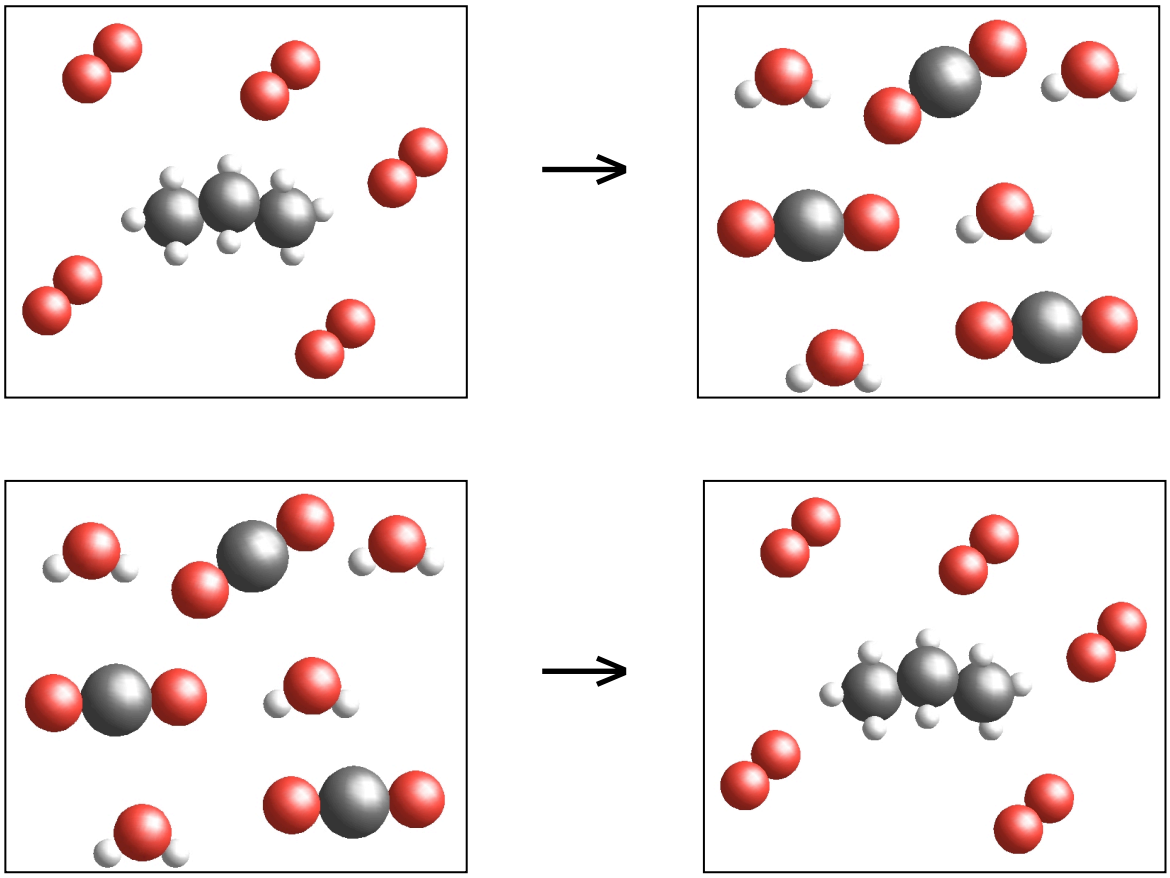
Item 1062 discriminated between B and low B ability students. Lastly in item 1063 the students were asked, "What is the maximum number of electrons that can be described by the following set of quantum numbers: n =5, ℓ =3, m_ℓ = -2." Their options were 48, 40, 10, 2 and 1. Item 1062 is very similar to the first part of item number 11786 given on both the 2006 and 2007 second examinations. Item 11786 asked the students how many electrons can be described by n = 4 and ℓ = 2 quantum numbers? This was difficult for almost everyone in the course (See Table 6.3 for item parameters). However, when students were asked about how many electrons could be described by n =5, ℓ =3, m_ℓ = -2, item 1063, A, B, C and many D students answered the item correctly. To successfully answer the item, students must know how to use the quantum number rules. The difficulty of this item indicates that students learned how to use the rules to answer the previous item after the second examination but before the final examination. Even though there was slight improvement on one item on the final examination, the other two items still show that understanding quantum numbers is difficult for most students. Additional improvement needs to occur with the students' understanding of quantum numbers. Only the high A students could answer all three items on the final examination about quantum numbers.

Students' Understanding of Molecular Image Problems

Understanding chemistry on the molecular level using images previously has been shown to be difficult for almost all students at UGA independent of their chemical knowledge. On the first examination given in the fall of 2005 and the fall of 2006 students were given item number 11237, which asked students to choose from the given images that best represents H_4P_2 boiling. In the fall of 2005, only the highest A students answered the item correctly; in the fall of 2006, only the A and high B students answered the item correctly (See Table 5.2 for item parameters).

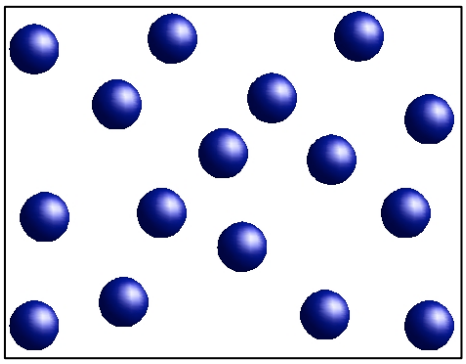
On the first examination given in the fall of 2007 on item 12110 the students were told to first, "Determine whether the statement below describes a physical change or a chemical change. – Propane gas, C_3H_8, condenses to liquid." Then the students had a multiple choice box with only two choices: physical change or chemical change. Next the students were given the series of images shown below and were told that the images "depict chemical or physical changes. The starting image is on the left and the ending image is on the right." The students were asked to select the image set that best described propane gas, C_3H_8 condensing to liquid. The four image options are shown below.



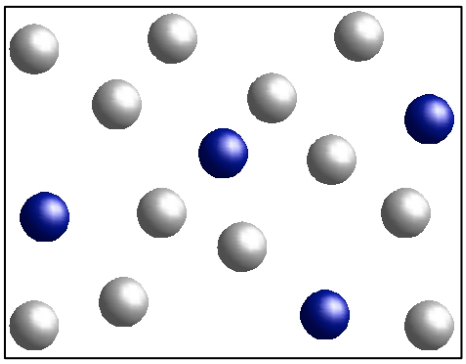


This molecular level understanding item proved to be very easy for students having a difficulty of -1.539. The ICC had a slope of 0.969 indicating that the item discriminated well between lower ability F students. This item is significantly easier than item 11237 that just asked students to choose the image of H_4P_2 boiling. This new item is easier since the students were given hints and prompts in the item to make them conceptually think about why they are choosing a specific image. Two of the four sets of images depicted chemical changes and should not have "distracted" students. The guessing parameter for item 11237 was 0.209 indicating that lower ability F students either could not "guess" the answer correctly, or they were unable to remove these two "distractors" if they did "guess" an answer.

Another new item assessed the students' understanding of molecular level images and the rate of a given reaction. Item 12492 asked students to "Consider the first–order reaction A → B in which A molecules (blue) are converted to B molecules (red). Given the above pictures at t = 0 min and t = 4 min, how many A and how many B molecules will be present at t = 2 minutes?" They were then asked, "How many A and how many B molecules will be present at t = 8 minutes?" Lastly they were asked, "What is the half-life of the reaction?"



t = 0 min

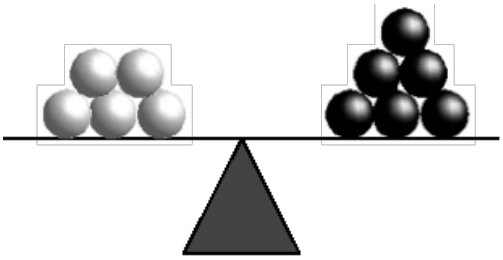


t = 4 min

This item contained free response boxes, but since there are a limited number of logical answers, the three-parameter logistic model was used. This item has a difficulty of 1.023 and therefore it discriminates between B and C students. The difficulty of this problem could lie with the conceptual understanding of first order reactions and half-life, or confusion could have been caused from the molecular level images given. It was most likely a mixture of both that made C, D and F students unable to successfully answer the item. More conceptual items need to be asked about rate laws to determine the students' logic failure.

Students' Understanding of Mole Concepts

It was found in the 2005-2006 academic year that students had difficulty with the mole concept. A new item was written on the first semester final examination to assess the students' understanding of this concept. Item 1032 asked students to look at the following image and determine which element left or right has the greater molar mass.



The students were given the following five options:

- right
- they have the same molar mass
- there is not enough information to answer
- left
- none of the above

The students first have to realize that the scale is balanced, and the scale measures the amount of grams present not the number of moles present. The students then have to recognize that the element on the right has more atoms present than the one on the left. With the amount of grams of the two being equal, there will be more grams per mole of the element on the left compared to the element on the right. The concept of a mole is necessary to understand this item. The item had an ability of 0.818 indicating that A and B students answered the item correctly. An ability of 0.818 would normally separate low C students from high C students, but with the item's ICC having a slope of 0.626, this item did not discriminate well between students in the C range. Some of the lower C students answered the item correctly whereas some of the higher C students did not. When first looking at this item, it would be expected to have a higher ability. An item

very similar to this was given to the students on a practice examination. With seeing a similar problem previously, this probably caused the slope of the ICC for this item to not be as steep. This indicates that some lower ability C students remembered the previous item and other higher ability C students did not. The item does indicate that some C along with D and F students were still struggling with the mole concept after seeing a related problem.

Analysis of Tries on JExam Examinations

Students use the JExam program when taking regular semester examinations at UGA. With this program if a student misses items on the examination, they are given a second and third try to successfully answer the items in the allotted time. To this point, only the first responses on the JExam examinations have been analyzed. It is of interest to calculate the IRT parameters for items on the second and third try of the examinations in order to determine how the discrimination, difficulty and guessing parameters change from the first, second to third try. IRT analysis was performed on try 1, try 2 and try 3 for the first examination given in the fall of 2007. IRT analysis was performed separately for each try and, as in previously discussed IRT analyses, the dichotomous two and three-parameter models were used. If a student answered an entire item successfully on the first try, they were given credit for answering the item correctly in try 1, try 2 and try 3. Furthermore, if a student missed an item on the first try but answered the item correctly on the second try, their response data would show that they answered the item incorrectly on the first try, but correctly on the second and third try.

Since the same items were used for each analysis, the item parameters from each try can directly be compared to each other. This comparison is not meaningful for the lower ability items; most students answer these correctly on the first try and consequently never reach try two

or try three. As expected, it was found that for most items, the ability of the items became progressively lower from the first to second to third try, indicating that they are less difficult as the tries progress. The change in difficulty for an item is much greater when comparing try 1 to try 2 than when comparing try 2 to try 3. For most of the items on the examination, the discriminating power and guessing factors stayed within the error range of each other for the three tries. Examples of items for which the difficulty parameter differed, but the discrimination and guessing parameters did not are shown in Table 7.3.

Table 7.3: IRT analysis of items with different abilities for first, second and third tries on fall 2007 first examinations.

Item Number	Try	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D.F.
11677	1	1.396 (0.176)	0.036 (0.080)	0.001 (0.001)	2.8	7.0
	2	1.531 (0.188)	-0.288 (0.080)	0.001 (0.001)	3.8	7.0
	3	1.797 (0.217)	-0.377 (0.071)	0.001 (0.001)	3.1	6.0
11678	1	1.336 (0.183)	-0.212 (0.086)	0.001 (0.001)	7.5	7.0
	2	1.407 (0.174)	-0.577 (0.094)	0.001 (0.001)	3.0	7.0
	3	1.539 (0.184)	-0.660 (0.092)	0.001 (0.001)	9.2	7.0
12091	1	1.694 (0.260)	0.973 (0.113)	0.001 (0.001)	5.7	6.0
	2	1.454 (0.211)	0.314 (0.091)	0.001 (0.001)	12.7	7.0
	3	1.461 (0.205)	0.108 (0.087)	0.001 (0.001)	15.1	7.0
12096	1	1.000 (0.204)	-1.088 (0.317)	0.195 (0.085)	5.1	8.0
	2	0.996 (0.235)	-2.673 (0.540)	0.201 (0.089)	7.0	6.0
	3	1.143 (0.311)	-3.000 (0.623)	0.198 (0.087)	2.2	4.0
12106	1	0.911 (0.288)	0.999 (0.352)	0.234 (0.078)	9.3	9.0
	2	0.766 (0.187)	-2.299 (0.555)	0.210 (0.092)	4.9	8.0
	3	0.896 (0.242)	-3.061 (0.708)	0.202 (0.090)	3.4	6.0
12114	1	1.529 (0.612)	1.694 (0.300)	0.135 (0.038)	6.9	9.0
	2	1.036 (0.291)	0.691 (0.255)	0.182 (0.069)	8.1	8.0
	3	1.077 (0.223)	-0.730 (0.278)	0.192 (0.083)	4.8	8.0

For the items whose ICC's slopes stayed within the same range for the three tries, we conclude that the item discriminates equally well between students on all tries; however, the ability of students that the item discriminates between will change. Surprisingly, the asymptote for many of these items also stayed within the range of each other for the three tries. This indicates that students have the same possibility of guessing the item correctly for the first, second and third try. With purely random guessing, the possibility of guessing an item correct for a multiple choice item should increase from the first to second to third try.

There are two main reasons that describe why the guessing factor in the IRT analysis did not increase for items as the tries progressed; one reason is based upon the students' answering the item, while the other pertains to the way IRT analyzes items. Many of the lower ability students may not have guessed the item correctly on the second and third try because they either ran out of time, or gave up before the examination was over. There is also a possibility that these students assumed their first answer was correct, and decided to keep their answer when submitting the second try. Another reason for the similar guessing factors is that there are fewer students answering the item correctly for the second try in comparison to the first. The students who correctly answered the item at the first attempt are present in the analysis of the second try and are associated with a correct response. With fewer students guessing the item correctly during the second try, the guessing factor for these students is masked by the number of students and the probability of students guessing the item correctly on the first try. Because the difficulty of the item changes for each try, while the guessing and discriminating parameters only slightly change, we conclude that the items which separate students at all ability levels with a low guessing factor are highly discriminating items for all three tries.

On the examination, there were also items present in which the slope of the items were larger for the second try in comparison to the first, or the slope of the ICC was larger for the third try in comparison to the first and second. This indicates that the items separated students at different abilities levels better during the second and third tries than during the first attempt. Table 7.4 contains examples of items where the discrimination parameter increases as the tries increase.

Table 7.4: IRT analysis of items with increased discrimination on first, second and third tries on fall 2007 first examinations.

Item Number	Try	Slope, *a* (Standard Error)	Ability, *b* (Standard Error)	Asymptote, *c* (Standard Error)	Chi Square	D.F.
11576	1	0.777 (0.104)	-1.638 (0.216)	0.001 (0.001)	8.7	9.0
	2	1.140 (0.142)	-1.655 (0.170)	0.001 (0.001)	3.6	7.0
	3	1.260 (0.158)	-1.644 (0.157)	0.001 (0.001)	8.7	7.0
11689	1	1.063 (0.138)	-0.013 (0.094)	0.001 (0.001)	9.2	9.0
	2	1.407 (0.174)	-0.577 (0.094)	0.001 (0.001)	7.6	8.0
	3	1.389 (0.160)	-0.559 (0.086)	0.001 (0.001)	11.4	8.0
12076	1	0.669 (0.154)	-1.377 (0.462)	0.194 (0.086)	7.7	8.0
	2	1.234 (0.244)	-2.420 (0.404)	0.188 (0.085)	3.1	5.0
	3	1.479 (0.370)	-2.621 (0.438)	0.193 (0.087)	0.5	4.0
12104	1	0.455 (0.115)	-0.828 (0.613)	0.214 (0.092)	8.4	9.0
	2	1.020 (0.211)	-2.332 (0.446)	0.190 (0.085)	4.8	7.0
	3	1.263 (0.321)	-2.822 (0.541)	0.193 (0.087)	2.2	3.0

Items, 11576 and 11689 are similar and are algorithmic in nature, whereas items 12076 and 12104 were conceptual problems. Item 11576 is a dimensional analysis problem where the students had to perform many conversions to calculate the correct answer, and 11689 asked students to convert from μL to gal. It is unknown why these items would discriminate better during the second and third try in comparison to the first. Item 12076 is a multiple answer item

that pertains to physical and chemical properties. There is a possibility that students with less chemistry knowledge could answer the item correctly by an educated guess whereas some of the brighter students overanalyzed and were led to one of the distractors. The latter group of students who were uneasy about one of the responses could identify their mistake and answer the question correctly on the second attempt, resulting in that item becoming a more fair test of knowledge *i.e.* more discriminating. Item 12104 is an item that asked students about molecular order in liquid, solid and gas. This item contains three choose boxes, which makes the item much easier on the second and third attempt than on the first, and it asks students about the molecules of liquid H_2O in comparison to the molecules of H_2O in ice. The students must state whether the molecules are: closer together, farther apart or the same distance apart. Many of the higher level ability students would find this item confusing since they also know that H_2O solid floats in H_2O liquid. This extra knowledge would cause higher ability students to be confused about the answers closer together and farther apart; lower ability students would not be confused by over thinking the item. If a brighter student answered the item incorrectly, they were able to answer it correctly on the second try, causing the discrimination parameter to greatly increase.

Try one, two and three for the second examination in general chemistry for the fall 2007 semester were analyzed in the same fashion as the first examination, with similar results found. Most of the items on the examination contained discrimination guessing parameters in the same range, and the ability of the items became easier as the tries progressed. Once again, few items became more discriminating as the tries progressed.

CHAPTER 8

COMPARISON OF STUDENTS' ABILITIES

IRT was used to calculate a location on the ability scale for each student; this location is considered the student's ability. Graphs of student ability distributions were prepared for the 2004, 2005, 2006 and 2007 academic years to compare student performance for each academic year (Figures 8.1, 8.2, 8.3, and 8.4 – 2004, 2005, 2006 and 2007, respectively).

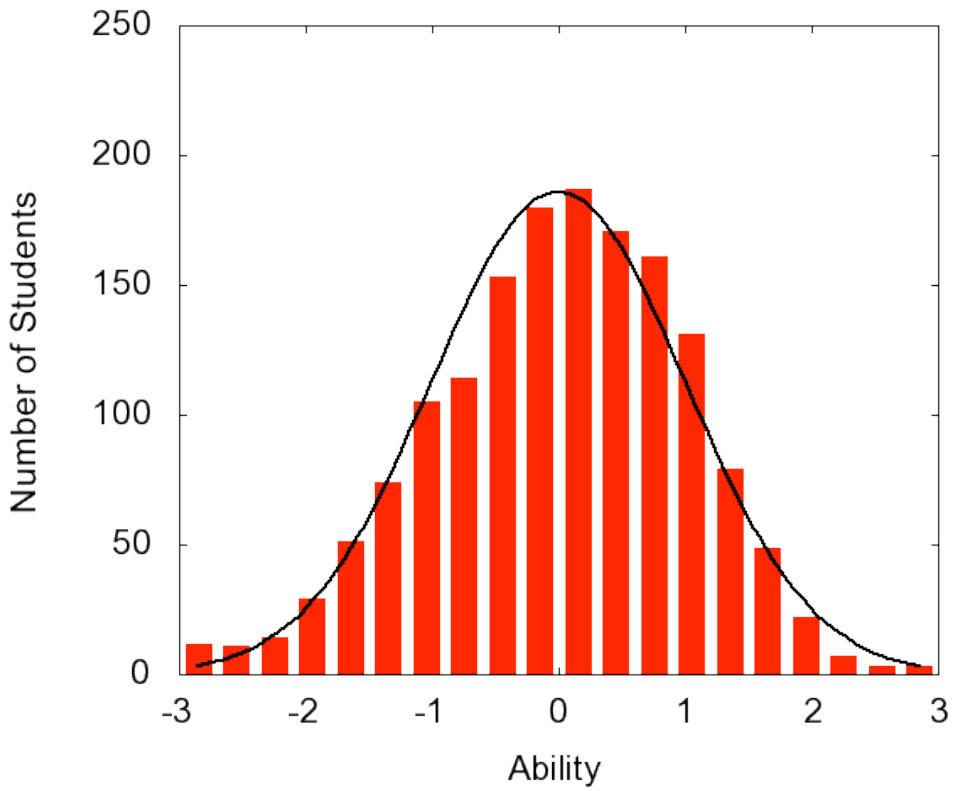


Figure 8.1: Number of students at each ability for the fall 2004 - spring 2005 academic year, with the corresponding normal distribution shown as a black line.

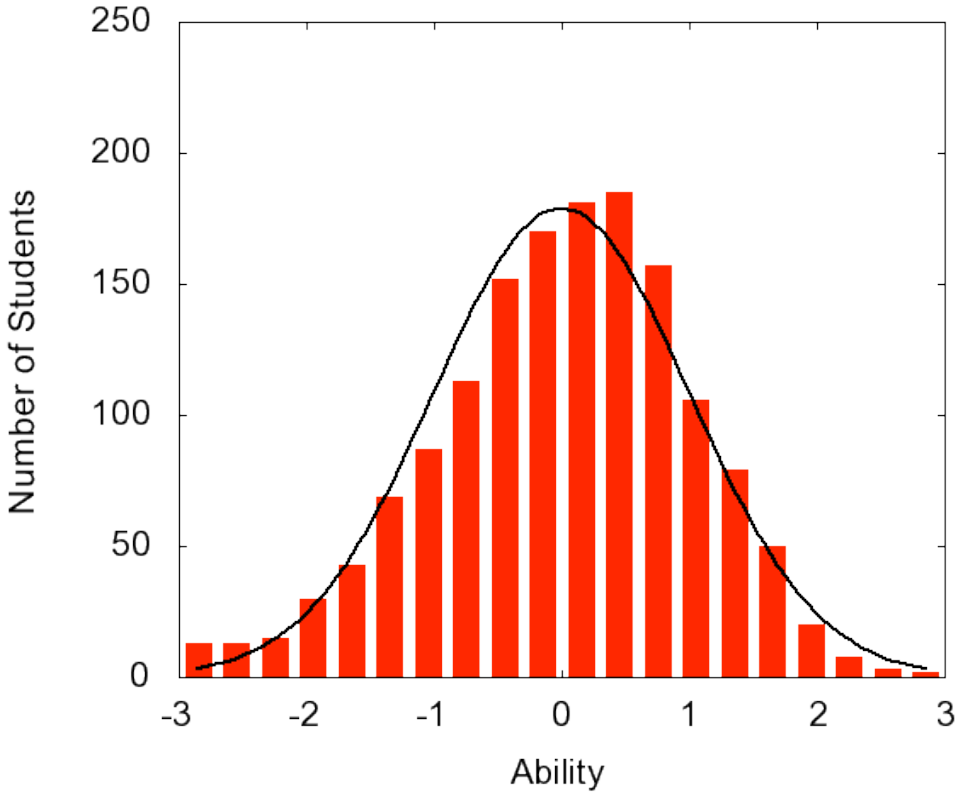


Figure 8.2: Number of students at each ability for the fall 2005 - spring 2006 academic year, with the corresponding normal distribution shown as a black line.

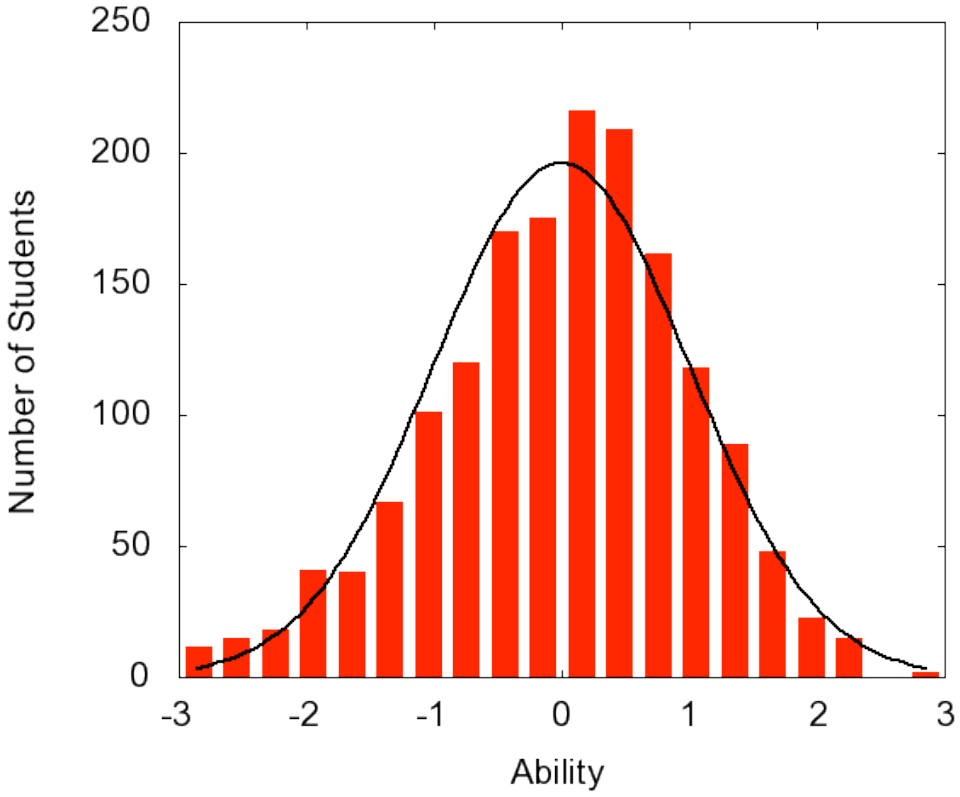


Figure 8.3: Number of students at each ability for the fall 2006 - spring 2007 academic year, with the corresponding normal distribution shown as a black line.

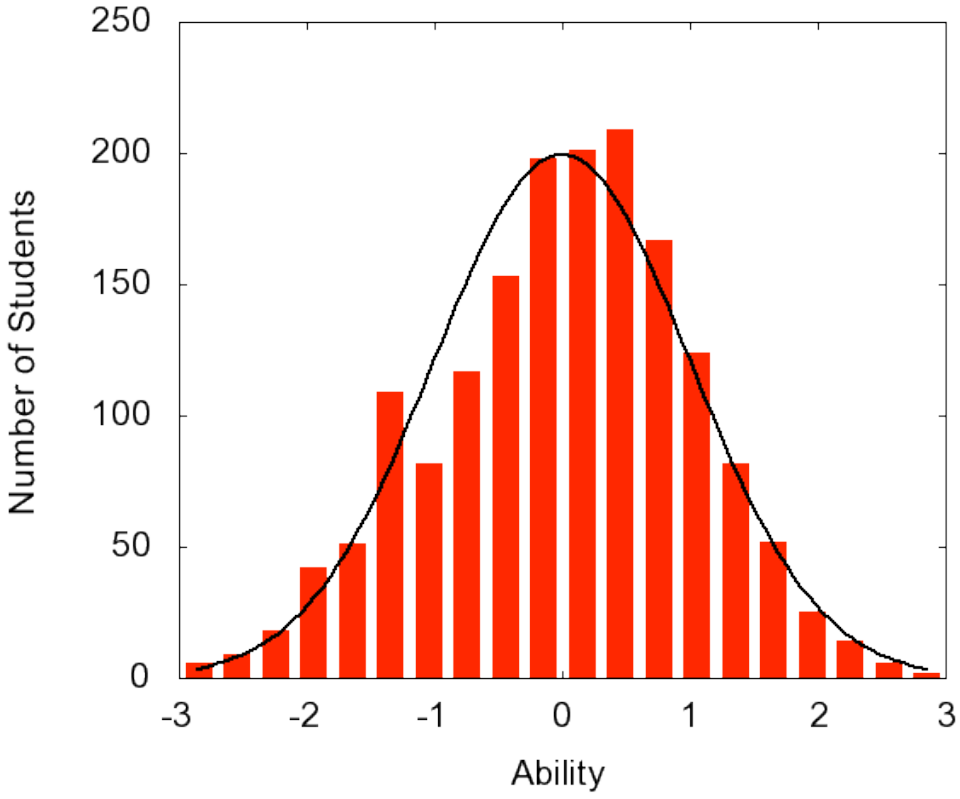


Figure 8.4: Number of students at each ability for the fall 2007 - spring 2008 academic year, with the corresponding normal distribution shown as a black line.

Although it may be tempting to quantitatively compare the distributions depicted in Figures 8.1, 8.2, 8.3 and 8.4, any comparisons must be performed with caution. The student populations for each analysis were assumed to form a normal distribution for MMLE and EAP calculations with the student ability data centered around zero on the abscissae. The standard deviation for the data shown in Figures 8.1, 8.2, 8.3 and 8.4 are all 1.0. Consequently, only qualitative analysis between academic years can be made. Since the ability scale range is from +4 to -4, a random student has a much higher probability of having an average ability near zero than having an ability of +4 or -4. As a result, the observed distribution should resemble a normal distribution.

From a visual comparison of Figures 8.1 and 8.2 it can be seen that even though roughly the same ability was needed to receive a specific grade in the class, the general student (the largest peak) increased from the 2004 – 2005 academic year to the 2005 - 2006 academic year. This is an increase in the mode of the student's ability from before IRT was used to the time IRT analyses began to influence the examinations. The number of students taking the general chemistry examinations greatly increased from the 2005 - 2006 academic year to the 2006 – 2007 academic year (Figures 8.2 and 8.3). During the same time frame, the number of students with an ability greater than 2.5 decreased. This is not because there were fewer bright students in the 2006 - 2007 compared to the 2005 - 2006 year, but because the exams were composed with fewer really difficult questions. While writing the 2005 -2006 year examinations, it was known that a student needs an ability around 1.5 to make a letter grade of an A in the class. At most, only one question having an ability greater than 2.5 was asked on the examinations. Since there were few questions with an ability of 2.5 or larger, students could not be assigned these high abilities. The student abilities are dependent upon the ability of the questions that they are asked, and in IRT, students are not assigned abilities higher than the ability of the items they answered correctly.

When comparing the number of students at each ability during the 2006 – 2007 academic year (Figure 8.3), the number of students having abilities of 0.15 and 0.45 are significantly larger than the number of students with abilities of -0.15 and -0.45. This indicates that the students having ability levels around 0.0 are the ones being positively affected by our informing the instructors teaching the class what was difficult and easy for the students.

The ability distribution for the 2007 – 2008 academic year, Figure 8.4 shows a larger than expected number of students with an ability of -1.35. This occurred because many items on the

examinations discriminated between F students causing quite a few items on the examinations with abilities around -1.35. With many items having this ability, there is a high probability of students being assigned an ability at this location on the ability scale if they answered these items correctly but were unable to answer the more difficult items correctly. If students are only able to correctly answer items with an ability of -1.35 and are assigned this ability, they would not have passed the examinations. It is likely that most of the students with the ability of -1.35 and below dropped the course due to their poor examination grades. Their response data was still analyzed using IRT with the remainder of the students to afford an accurate analysis of topics on the early examinations.

Notwithstanding the difficulties in comparing the plots presented here, the general skewing of the distributions towards higher ability levels on the later examinations is a pleasing trend that is indicative of increasing student knowledge year-to-year. For a better perspective of students' increase in abilities coming as a result of the IRT-inspired modification of teaching methods, the number of students at each ability level in the fall 2004 academic year was added to the number of students at each ability in the fall of 2005 and plotted (Figure 8.5). In the 2004 academic year, examinations were written without IRT knowledge, whereas in the 2005 academic year they were written with item parameters in mind. Modification in the classroom did not occur until the 2006 academic year and continued during the 2007 academic year. The 2006 and 2007 academic years number of students at each ability were also added together and graphed (Figure 8.6).

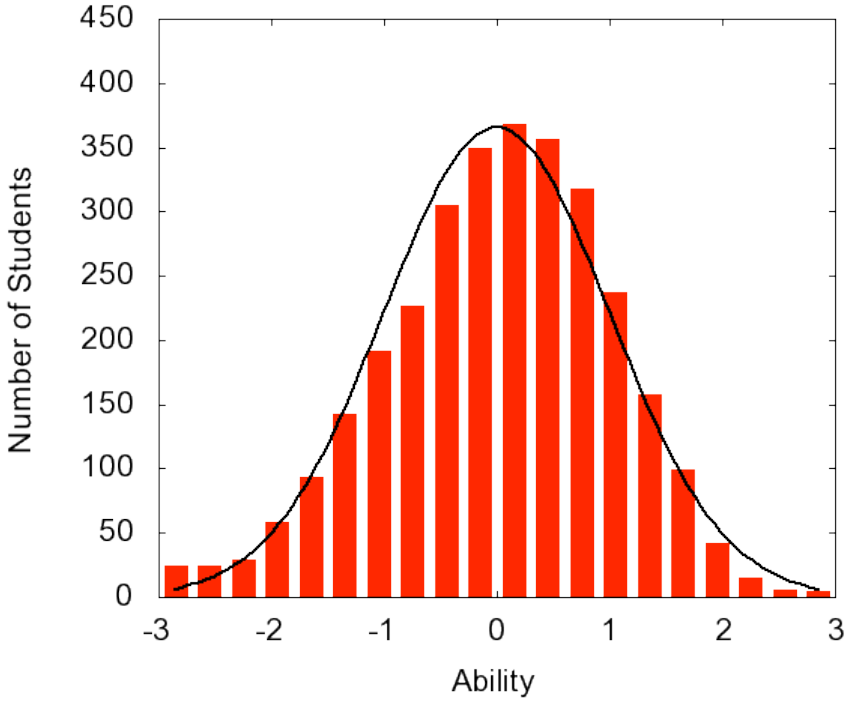


Figure 8.5: Number of students at each ability before modification occurred in the classroom.

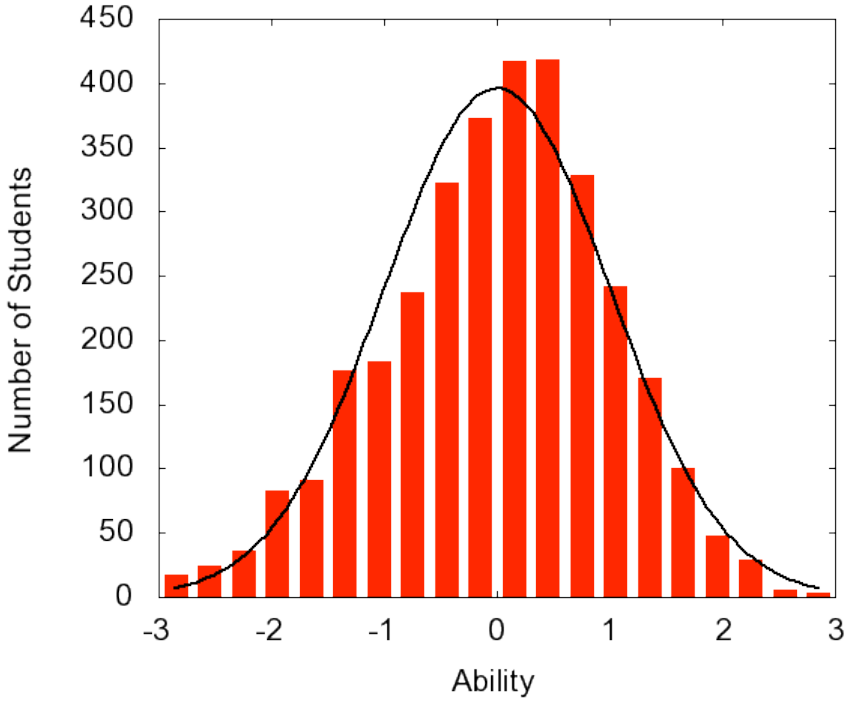


Figure 8.6: Number of students at each ability after modification occurred in the classroom.

These graphs show a more noticeable improvement in the students' abilities than the individual year data plots compared previously. More students took the class in the 2006 and 2007 academic years than the 2004 and 2005 academic years, so the number of students at each ability should slightly increase in Figure 8.6, compared to those in Figure 8.5. There was a much larger increase in the number of students at abilities 0.15 and 0.45, which indicates that overall the students' abilities improved from before modification in the classroom took place to after. There is also a significant increase in the number of students having abilities of -1.35 and -1.95; this increase in lower ability students preserves the 0.0 mean for the data in Figure 8.6. As with the individual analyses, we conclude that an improvement is observed as a result of IRT guided changes to teaching.

To make the comparison of Figures 8.5 and 8.6 more transparent, the number of students at each ability was transformed to the percentage of students at each ability for each group of years. Given that the resulting data are on the same scale, a direct comparison can be made from before to after modification in the classroom took place. The percentage of students at each ability before and after a difference to classroom teaching occurred are juxtaposed for an easier visual comparison in Figure 8.7.

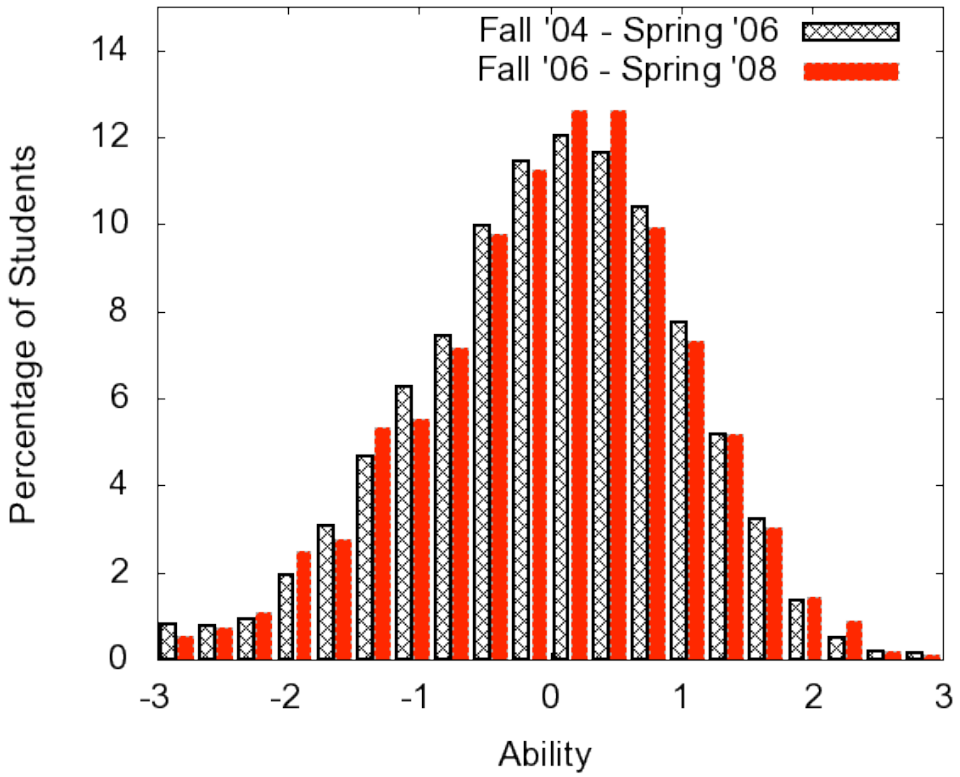


Figure 8.7: Percentage of students at each ability before and after modification occurred in the classroom.

As seen in Figure 8.7, there is an increase in the percentage of students with an ability of 2.25 and higher after classroom modification. Before modification, 0.85% of students had an ability of 2.25 and higher, after modification, this rose to 1.18%. When the instructors changed how, and which, topics they spent more time on in the classroom, they were able to positively affect the brightest students in the classroom, shifting their abilities slightly higher. Likewise, 52.56% of students had an ability of 0.0 or higher before modification, which rose to 53.33% after modification. The populations of the ability levels between 0 and -1, *i.e.* below average knowledge, were depleted by the change to the teaching methods. Before modification 35.19% of students had a calculated ability between 0 and -1, after modification in the classroom this dropped to 33.72%. This skew in the distribution, towards a higher ability, proves that there was

a slight increase in the student's overall abilities after the instructors changed the teaching in the classroom. When analyzing the differences in these two distributions, the emphasis should be placed on the ability range where the bulk of students lie and not the outliers at the extremes. Even though the increase or decrease in percentage of students with the abilities discussed above is small, with the number of students so large, this small change is statistically valid.

CHAPTER 9

CONCLUSIONS

Item response theory analysis is more complex and rigorous than classical test theory; however, the amount of information learned about students, difficult topics and the items asked significantly surpasses the information that is obtained from CTT. IRT analysis of general chemistry examinations was used to develop several improved and more effective teaching methods and tools. For example, future examinations can be written to a given difficulty, *i.e.* the average score can be accurately predicted at the time of writing. Consequently, the common practice of "curving" examinations is no longer necessary, while the fairness and integrity of the testing is maintained. Examinations are now written containing equivalent items for each student, which results in unique examinations of consistent difficulty for the many students being examined. Before the IRT analysis was performed, the items that constitute a given question were designed to be equivalent; however, we have shown a number of illuminating examples where some items had a significantly different difficulty level to their 'equivalent' counterparts. Remedying these imbalances, after identifying them with IRT, causes grades to be much more accurately and fairly assigned to students in general chemistry while maintaining the freedom to administer different examinations to different students, which helps to eradicate cheating.

Our research has proven that UGA's general chemistry program has been fairly consistent in regards to the amount of chemistry knowledge (student ability) needed for a student to receive a specific letter grade in the course. From the 2004 to 2007 academic years, data

analysis has shown that a slight increase in chemical knowledge was necessary to receive the same letter grade from one year to the next. This increase in necessary knowledge indicates that the program has become slightly more difficult over recent years and that the overall performance of the students has improved.

General chemistry topics that are difficult for UGA students were also ascertained using IRT. Previous research has shown these topics to be universally difficult for all students, not just for students at UGA. However, previous research has not informed instructors which of these topics were pervasively difficult throughout the student population. Our research indicated for which grade level students the topics were difficult. For example, we found that topics such as molecular level understanding and the understanding of quantum numbers are difficult for all UGA students, except the brightest A level students. Other topics, such as understanding the difference between an ion, atom or molecule and calorimetry are difficult for only C, D and F level students. Before this research was conducted, difficult topics were discussed in the literature, and it was assumed that these topics were difficult for all students taking general chemistry. We proved that this is not the case at UGA.

This research can also be used to help students when they come in for help one-on-one. With the ability of each student known, this ability can be used to calculate the probability of the student answering each item correctly. This will inform the instructor which items on the examination that the student was close to being able to answer correctly, and the instructor can focus on the topics that the students were close to grasping. This is extremely important, as it offers a highly pragmatic route to improving students' grades by highlighting those weaknesses that can be most easily addressed.

Furthermore, given the agreement between the list of difficult topics determined in our studies and those from national studies in the literature, we note that IRT should be applied to data from other institutions to determine whether the same groups (ability levels) of students have difficulties with the same topics outside of UGA. This knowledge could be extremely useful in constructing specific guidelines that can be universally distributed to instructors, making them aware of the difficulties that specific groups have and how to remedy these.

Our research also elucidated the core chemistry topics that initially are difficult for students and thus require emphasis in instruction at the beginning of the course. For example, understanding the structure of ionic compounds (particularly polyatomic ions) is necessary to fully understand the number of ions dissolved in solutions, which impacts the following topics: acids and bases, strong, weak and non-electrolytes, freezing point depression, boiling point elevation and vapor pressure. The understanding of bond polarity vs. molecular polarity is also difficult for many students, and is required to understand intermolecular forces. The understanding of such core topics is absolutely crucial for success in the remainder of the course or students may become overwhelmed as a result of falling behind early on. This knowledge will be highly influential in reducing the attrition rate of the general chemistry courses.

A number of topics that are easy for all students have also been identified, such as unit conversion, balancing equations and electron configuration. Our understanding of which students do not grasp certain topics has helped instructors plan their lessons accordingly to emphasize specific topics. Additionally, with knowledge of which topics are easy for all students, instructors have been asked to cover them less in class, which facilitates greater student understanding without sacrificing course content. While IRT has proven to be very informative about topic difficulty, future analysis and research, such as interviews, are necessary to

determine the reason why students are struggling with these topics. Common misconceptions often underlie a student's problems as opposed to an inability to grasp a concept; these cannot be identified through statistics alone and must be eradicated before the student is capable of understanding a topic.

Through our analyses, we have been able to determine a difficulty scale of the many topics in general chemistry. This ordering is independent of who and where the examinations are taken. We have proven that the concept of quantum numbers is difficult for all students at UGA, and this concept is more difficult than understanding electron configuration. At a more prestigious university, the concept of quantum numbers might not be difficult for all students, but it should still be a more difficult topic for students than electron configuration. At a less prestigious university, some of the topics that were found to be easier at UGA might be difficult for most of the students, but that same topic should still be easier than topics that we found to be more difficult.

Just as this analysis can be used in a variety of ways to improve general chemistry teaching, other programs can also use IRT to analyze the consistency of their examinations and improve teaching methods. We have shown that, after IRT was used to write examinations, abilities of specific students increased. Additionally, after the difficult topics were determined and the instructors were informed, another group of students' abilities increased. With knowledge of which grade level students struggle with each topic, we can identify the students that are on the verge of grasping the material and help them fully understand the topic. Holding help sessions to discuss a topic that has been found to be difficult for the D and F students or teaching the topic in a more innovative way could lead to a better understanding of these topics. After partitioning the topics into more and less difficult categories, two or more distinct help

sessions could be held in order to not deter the brightest students who would be bored if discussing an easier topic; conversely, the lower ability students would profit most from concentrating on the topics that elude them but are within reach.

Research investigating why these topics are difficult for students needs to occur on a large scale. This will help instructors identify misconceptions or lack of understanding that specific students have. Currently only small-scale studies of why topics are difficult have taken place. Statistical validity of difficulty across academic years also needs to be researched. Item Response Theory should be used to determine if the students improved over a given time frame. Once this is accomplished, it can be used to determine if an experimental teaching technique tried in class was successful, which would be extremely useful with future chemical education research and students' learning. Current IRT research should be used to look at the diversity and the impact of demographic on understanding of particular topics. If particular groups of students struggle with certain topics, identifying those topics is a crucial first step to achieving equality in teaching for all students.

REFERENCES

(1) Emons, W. H. M.; Meijer, R. R.; Denollet, J. *Journal of Psychosomatic Research* **2007**, *63*, 27-39.

(2) Bhakta, B.; Tennant, A.; Horton, M.; Lawton, G.; Andrich, D. *BMC Medical Education* **2005**, *5*.

(3) Scherbaum, C. A.; Finlinson, S.; Barden, K.; Tamanini, K. *The Leadership Quarterly* **2006**, *17*, 366-386.

(4) Cohen, A. S.; Bottge, B. A.; Wells, C. S. *Council for Exceptional Children* **2001**, *68*, 23-44.

(5) Lawson, D. M. *Journal of Manipulative and Physiological Therapeutics* **2006**, *29*, 393-397.

(6) Wagner, T. A.; Harvey, R. J. *Psychological Assessment* **2006**, *18*, 100-105.

(7) Puhan, G.; Gierl, M. J. *Journal of Cross-Cultural Psychology* **2006**, *37*, 136-154.

(8) de Ayala, R. J. *The Theory and Practice of Item Response Theory*; The Guilford Press: New York, 2009.

(9) Zimowski, M.; Muraki, E.; Mislevy, R.; Bock, D. Bilog-MG 3: Scientific Software International, Inc,2002.

(10) Caughran, J. A.; Martin, J. G.; Atwood, C. H. In *CONFCHEM* 2001.

(11) Atwood, C. H.; Martin, J. G.; Caughran, J. A. In *221st ACS National Meeting* San Diego, CA, 2001.

(12) Atwood, C. H.; Lautenschlager, G. J.; Marsh, R. L.; Martin, J. G.; Caughran, J. A. *Initial results from computerized assessment of large freshman chemistry classes at the University of Georgia*, 2003; Vol. 225.

(13) Morris, G. A.; Branum-Martin, L.; Harshman, N.; Baker, S. D.; Mazur, E.; Dutta, S.; Mzoughi, T.; McCauley, V. *American Journal of Physics* **2006**, *74*, 449-453.

(14) Yu, C. H.; Popp, S. E. O. In *Practical Assessment, Research & Evaluation* 2005; Vol. 10.

(15) Cook, L. L.; Eignor, D. R. *International Journal of Educational Research* **1989**, *13*, 161-173.

(16) Bodner, G. M. *Journal of Chemical Education* **1991**, *68*, 385-388.

(17) Nicoll, G. *International Journal of Science Education* **2001**, *23*, 707-730.

(18) Kruse, R. A.; Roehrig, G. H. *Journal of Chemical Education* **2005**, *82*, 1246-1250.

(19) Rich, S. G. *Journal of Chemical Education* **1925**, 142-145.

(20) Özmen, H. *Journal of Science Education and Technology* **2004**, *13*, 147-159.

(21) Ingram, E.; Lehman, E.; Love, A. C.; Polacek, K. M. *Journal of College Science Teaching* **2004**, *34*, 39-43.

(22) Uzuntiryaki, E.; Geban, Ö. *Instructional Science* **2005**, *33*, 311-339.

(23) Tien, L. T.; Teichert, M. A.; Rickey, D. *Journal of Chemical Education* **2007**, *84*, 175-181.

(24) Nakhleh, M. B. *Journal of Chemical Education* **1992**, *69*, 191-196.

(25) Robinson, W. R. *Journal of Chemical Education* **1998**, *75*, 1074.

(26) Schmidt, H.-J. *Science Education* **1997**, *81*, 123-135.

(27) Wirtz, M. C.; Kaufmann, J.; Hawley, G. *Journal of Chemical Education* **2006**, *83*, 595-898.

(28) Sanger, M. J.; Phelps, A. J. *Journal of Chemical Education* **2007**, *84*, 870-883.

(29) Nurrenbern, S. C.; Pickering, M. J. *Journal of Chemical Education* **1987**, *64*, 508-510.

(30) Devetak, I.; Urbancic, M.; Grm, K. S. W.; Krnel, D.; Glazer, S. A. *Acta. Chim. Slov.* **2004**, *51*, 799-814.

(31) Ardac, D. *Journal of Chemical Education* **2002**, *79*, 510.

(32) Pinarbasi, T.; Canpolat, N. *Journal of Chemical Education* **2003**, *80*, 1328-1332.

(33) Birk, J. P.; Kurtz, M. J. *Journal of Chemical Education* **1999**, *76*, 124-128.

(34) Tarhan, L.; Ayar-Kayali, H.; Urek, R. O.; Acar, B. *Research in Science Education* **2008**, *38*, 285-300.

(35) Peterson, R. F.; Treagust, D. F. *Journal of Chemical Education* **1989**, *66*, 459-460.

(36) Teichert, M. A.; Stacy, A. M. *Journal of Research Science Teaching* **2002**, *39*, 464-496.

(37) Khan, S. *Science Education* **2007**, *91*, 877-905.

(38) Harrison, A. G.; Treagust, D. F. *Science Education* **2000**, *84*, 352-381.

(39) Teichert, M. A.; Tien, L. T.; Anthony, S.; Rickey, D. *International Journal of Science Education* **2008**, *30*, 1095-1114.

(40) Chandrasegaran, A. L.; Treagust, D. F.; Mocerino, M. *Research in Science Education* **2008**, *38*, 237-248.

(41) Nahum, T.; Mamlock-Naaman, R.; Hofstein, A.; Krajcik, J. *Science Education* **2007**, *91*, 579-603.

(42) Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; MESA Press: Chicago, 1960.

(43) Rasch, G. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, 1961; Vol. 4, p 321-333.

(44) Wilson, M. *Methods of Psychological Research Online* **2003**, *8*, 1-22.

(45) Embretson, S. E.; Reise, S. P. *Item Response Theory for Psychologists*; Lawrence Erlbaum Associates Inc., 2000.

(46) Baker, F. B. *The Basics of Item Response Theory*; 2nd ed.; ERIC Clearinghouse on Assessment and Evaluation, 2001.

(47) The correlation data analysis for this paper was generated using SAS/STAT software, Version 9.1.3 of the SAS System for Windows. Copyright © 2004 SAS Institute Inc. SAS and all other SAS Institute Inc. Product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA:

(48) Toit, M. d. *IRT from SSI: Bilog-Mg, Multilog, Parscale, Testfact*; Scientific Software International, Inc.: Lincolnwood, IL, 2003.

(49) Bock, R. D.; Gibbons, R.; Schilling, S. G.; Muraki, E.; Wilson, D. T.; Wood, R. Testfact 4: Scientific Software International, Lincolnwood, IL, 2003.

(50) Muraki, E.; Bock, D. Parscale 4: Scientific Software International, Lincolnwood, IL, 2003.

(51) Thissen, D.; Chen, W.-H.; Bock, D. Multilog 7: Scientific Software International, Lincolnwood, IL, 2003.

(52) Glas, C. A. W.; Linden, w. J. v. d. *Applied Psychological Measurement* **2003**, *27*, 247-261.

(53) Ingram, E. L.; Nelson, C. E. *The American Biology Teacher* **2006**, *68*, 275-279.

(54) Mulford, D. R.; Robinson, W. R. *Journal of Chemical Education* **2002**, *79*, 739-744.

(55) Smith, K. J.; Metz, P. A. *Journal of Chemical Education* **1996**, *73*, 233-235.

(56) Taber, K. S. *International Journal of Science Education* **1998**, *20*, 597-608.

(57) Furió, C.; Calatayud, L. *Journal of Chemical Education* **1996**, *73*, 36-41.

(58) Taagepera, M.; Arasasingham, R.; Potter, F.; Soroudi, A.; Lam, G. *Journal of Chemical Education* **2002**, *79*, 756-762.

(59) Cokelez, A.; Dumon, A.; Taber, K. S. *International Journal of Science Education* **2008**, *30*, 807-836.

(60) Ludwig, O. G. *Journal of Chemical Education* **2001**, *78*, 634.

(61) Vaarik, A.; Taagepera, M.; Tamm, L. *Journal of Baltic Science Education* **2008**, *7*, 27-36.

(62) Cervellati, R.; Perugini, D. *Journal of Chemical Education* **1981**, *58*, 568-569.

(63) Gabel, D. *Journal of Chemical Education* **1999**, *76*, 548-554.

(64) Ebenezer, J. V.; Erickson, G. L. *Science Education* **1996**, *80*, 181-201.

(65) Raviolo, A. *Journal of Chemical Education* **2001**, *78*, 629-631.

(66) Sanger, M. J. *Journal of Chemical Education* **2005**, *82*, 131-134.

(67) Krieger, C. R. *Journal of Chemical Education* **1997**, *74*, 306-309.

(68) Yin, M. *Journal of Chemical Education* **1345**, *78*, 1345-1347.

(69) O Farrell, F. J. In *Chemistry in Action* 2000.

(70) Gorin, G. *Journal of Chemical Education* **1994**, *71*, 114-116.

(71) Greenbowe, T. J. *International Journal of Science Education* **2003**, *25*, 779-800.

(72) Cassels, J. R. T.; Johnstone, A. H. *Journal of Chemical Education* **1982**, *61*, 613-615.

(73) Johnstone, A. H. *J. Comp. Assist. Learn.* **1991**, *7*, 701-703.

(74) Evans, J. D. *School Science Review* **1974**, *55*, 585-590.

(75) Stains, M.; Talanquer, V. *Journal of Chemical Education* **2007**, *84*, 880-883.

(76) Robinson, W. R. *Journal of Chemical Education* **1999**, *76*, 297-298.