10-21-2008

Formal Reasoning and Spatial Ability: A Step towards "Science for All"

Bo Jiang
University of South Florida

Formal Reasoning and Spatial Ability: A Step towards "Science for All"

by

Bo Jiang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Chemistry
College of Arts and Sciences
University of South Florida

Major Professor: Jennifer Lewis, Ph.D.
Abdul Malik, Ph.D.
Robert Potter, Ph.D.
Dana Zeidler, Ph.D.

Date of Approval:
October 21, 2008

Keywords: assessment, evaluation, chemical education, science education,
non-majors science curriculum

Dedication

This work is dedicated to my parents, Gengyin and Shangzhi.

Acknowledgments

First, I am very grateful to Dr. Jennifer Lewis, my major professor, whose patience, intellect, wisdom, and wealth of knowledge contributed greatly to this project. Without her guidance and support this work would not be possible.

Second, I would like to extend my sincerest thanks to Dr. Trace Jordan at New York University for coordinating the national Molecules of Life project and the entire data collection process. I also express my gratitude to all the faculty participants of the Molecules of Life project. I am also thankful to Bob Rumans at the university scanning office for scanning thousands of scantrons into spreadsheets in the data collection process.

I am indebted to my dissertation committee, Drs. Dana Zeidler, Robert Potter, and Abdul Malik for their valuable advice and guidance in preparing this manuscript.

Many thanks to Dr. Scott Lewis for his insights and suggestions on the early part of this work, to Xiaoying Xu for her involvement in part of the GALT data analysis, to Alicia Garcia for data collection in the preparatory chemistry course, and to all other graduate students in the Dr. Lewis group for their support and suggestions on various aspects of the project.

I acknowledge the support from Drs. John Ferron, Robert Dedrick, and Jeffrey Kromrey on different data analysis techniques.

The Molecules of Life project was partially supported by the National Science Foundation awards 0443014 & 0443026. Any opinions, findings, and conclusions or

recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Finally, I thank my parents for their love and encouragement and to whom this work is dedicated.

Table of Contents

List of Tables

List of Figures

Formal Reasoning and Spatial Ability: A Step towards "Science for All"

Bo Jiang

ABSTRACT

This work conducts an evaluation of a non-majors science curriculum named Molecules of Life (MOL) that aims to provide effective science education to undergraduate students who are not majoring in scientific disciplines.

As part of the process of developing an assessment plan for MOL, three related studies were undertaken in order to help us choose assessment instruments for MOL. The first study examined the validity of student evaluations of teaching. The second study investigated the Test of Logical Thinking (TOLT) and Group Assessment of Logical Thinking (GALT), two widely-used instruments for measuring formal reasoning ability. GALT is very similar to TOLT, but contains two additional concrete items. Focusing on the functioning of these two items, we added them into TOLT and created a new test called "TOLT+2". We then compared TOLT with TOLT+2 in terms of reliability, discriminatory power, potential item bias, and predicting students at-risk in a general chemistry course. The two concrete items were found to provide no advantage in these aspects. In the third study, we performed a direct comparison between TOLT and GALT as intact instruments in general chemistry and in preparatory chemistry. GALT showed

no advantage over TOLT for both general and preparatory chemistry in terms of reliability, discriminatory power, potential item bias, and predicting at-risk students. GALT has more frequently occurring, potentially biased items, while TOLT is tenably a less biased test.

Based on the results from the three studies and input from faculty, an assessment plan was developed and refined for the MOL project at two summer workshops that faculty from all eight institutions participated in. Subsequently, a systematic evaluation for MOL was carried out as a fourth study. We found evidence that students learned the enzyme content from the MOL courses at all participating institutions. We also found the MOL curriculum can meaningfully improve students' spatial ability. MOL was able to reduce the gap between high-spatial-ability and low-spatial-ability students at most institutions. Because of the critical link of spatial ability to science learning, this result is very promising for our efforts to move towards "science for all".

Chapter 1: Introduction

Importance of "Science for All"

With the globalization of a knowledge-driven economy (Dwyer, 2008, p. xi), rapid changes of our modern society, and significant advances in science and technology each year, the need for a scientifically literate citizenry and a skilled workforce has never been greater (National Science Board, 2006; National Science Foundation, 2006). The importance of science literacy of the general public goes beyond the "individual self-fulfillment and the immediate national interest of the United States", as "the most serious problems that humans now face are global: unchecked population growth in many parts of the world, acid rain, the shrinking of tropical rain forests and other great sources of species diversity, the pollution of the environment, disease, social strife, the extreme inequities in the distribution of the earth's wealth, the huge investment of human intellect and scarce resources in preparing for and conducting war, the ominous shadow of nuclear holocaust—the list is long, and it is alarming" (Project 2061: American Association for the Advancement of Science, 1990, p. vii). As stated by the National Research Council (NRC), it has become imperative for all U.S. college students to "understand the methods and basic principles of science if they are to succeed" (Fox & Hackerman, 2003, p. 11-12). It has also turned out to be crucial for all college graduates to be "scientifically literate citizens capable of participating in a democracy increasingly influenced by scientific and technological innovations" (Jordan & Lewis, 2008).

As a result, the Division of Undergraduate Education at the National Science Foundation (NSF) states that its mission is to "promote excellence in undergraduate science, technology, engineering, and mathematics (STEM) education for all students" (National Science Foundation Division of Undergraduate Education, 2005). It has become a prevalent objective for colleges and universities across the U.S. to provide "science for all" (Lewis & Lewis, 2008; National Science Foundation, 1996, 2006; Project 2061: American Association for the Advancement of Science, 1990), that is, to provide effective science education for all undergraduate students, including students who are not majoring in science disciplines.

To address this demand, Jordan and Kallenbach at New York University devised a non-majors course named Molecules of Life that introduces students to the modern interface between chemistry, biology, and health as a foundation for exploring the molecular basis of life (Faculty Resource Network at New York University, 2008; Jordan & Lewis, 2008).

The Molecules of Life (MOL) Course

Organized around the motifs of biological molecules and pharmaceuticals, the Molecules of Life course introduces scientific topics in contexts relevant to everyday life to stimulate student interest, such as trans fats; water- and fat-soluble vitamins; sickle cell anemia; DNA and genetic information; enzymes and drug design; the history, success, and side effects of Aspirin; the emergence, ephemeral success, and costly recall of Vioxx (Jordan & Lewis, 2008). This methodology allows non-science students to undergo the excitement of scientific advances in the interdisciplinary field of chemistry, biology,

pharmaceuticals, and health, and to appraise their impact on society (Faculty Resource Network at New York University, 2008). The "Table of Contents" from the first edition of the MOL textbook that Jordan and Kallenbach co-authored (Jordan & Kallenbach, 2008) is listed in Appendix K.

Funding from the National Science Foundation supported the launch of a three-year (2006-2008), nationally coordinated project, titled Molecules of Life: a Partnership to Enhance Undergraduate Science Education for Non-Majors (henceforth referred to as the Molecules of Life or MOL project). A dynamic network partnership was set up between New York University and seven other institutions to adapt and further develop the Molecules of Life course, including Chaminade University (Honolulu, HI), Chicago State University (Chicago, IL), Fairfield University (Fairfield, CT), Nassau Community College (Garden City, NY), Spelman College (Atlanta, GA), University of Puerto Rico at Rio Piedras (Rio Piedras, PR), and Xavier University of Louisiana (New Orleans, LA). Our role at the University of South Florida was to lead the evaluation of this project, while the eight institutions listed above had faculty participants who implemented the MOL course, either offering the entire MOL course, or integrating the enzyme module from MOL into a chemistry or biology course for non-science majors. In other words, whereas we did not implement the MOL course at the University of South Florida (USF), our role at USF was to direct and carry out the evaluation of the MOL project that was implemented at the eight participating institutions listed above.

For all curricular reforms and innovations, assessment is vital for their success (Achacoso & Svinicki, 2005 p. 5-8; Dwyer, 2008 p. 7-8; Lynch, 2000 p. 216-245). *Assessment* can be defined as the process of gathering and interpreting information by

using students' responses to make inferences about students' knowledge, skills, or affective status (Brookhart, 1999 p. 1; Popham, 2000 p. 3-4). In the strict sense, assessment designates the measurements that provide information of student learning, while *evaluation* denotes making judgments based on that information (Brookhart, 1999). Without proper assessment, instructors would not be able to gauge student learning, students would not find out how they are doing and use that information to study more effectively, and we would not be able to get useful feedback to guide and improve science education.

Implementation of the MOL Course at Participating Institutions

The MOL course materials were used at all eight (8) participating institutions by a total of twelve (12) faculty instructors. The last section, Section 6, titled "Enzymes and Drug Design" (also referred to as the enzyme module), includes the last three chapters (see Appendix K) in the MOL textbook and is the only module used at some participating institutions, as described below. There have been three different ways to adopt and implement the MOL course by various instructors at different institutions:

1) Three schools offered the entire MOL course as a science elective for non-science majors, including New York University (in all semesters when MOL was taught, hereafter denoted as "all semesters" for the other schools), Spelman College (all semesters), and Fairfield University (Spring 2007).
2) Two schools integrated the enzyme module into a chemistry course for non-science majors, including Chaminade University (all semesters), and Nassau Community College (Spring 2006, Spring 2007, and Spring 2008).

3) Five schools integrated the enzyme module into a biology course for non-science majors, including University of Puerto Rico (all semesters), Chicago State University (CSU, all semesters), Fairfield University (Spring 2006 and Spring 2008), NCC (Fall 2005), and Xavier University (all semesters).

Outline of This Work

As part of the process of developing an assessment plan for the Molecules of Life (MOL) project, three related studies were undertaken, aimed at helping us choose assessment tools and instruments. The first study examined the tool of SET, namely, student evaluations of teaching. The second study investigated the Test of Logical Thinking (TOLT), an assessment instrument for measuring students' cognitive reasoning ability. As a follow-up to the second study, we carried out a direct comparison of two tests of formal reasoning in the third study. Based on the results from the three studies and input from faculty participants in the MOL project, an assessment plan was developed and refined for the MOL project at two summer workshops that the faculty from all nine institutions participated in. Subsequently, a systematic assessment for the MOL project was carried out as a fourth study. The results from these four studies are presented below. Our presentation of these studies used a series of acronyms, the most common of which for convenience have been collected into a table that is in Appendix A.

Chapter 2: Student Evaluations of Teaching

Student evaluations of teaching (SET, also frequently referred to as *student ratings of instruction* and *student course evaluations* in the literature) are widely used today in colleges and universities in the U.S. as well as in other countries all over the world to improve instructors' awareness and teaching effectiveness. In many cases, SET is used as one of the most important criteria for personnel decisions such as whether or not an instructor gets a promotion, salary increase or tenure. In other situations SET is used for formative and summative assessment of teaching effectiveness. In order for SET to be used as an assessment tool for a curriculum or a course, it is important that it is free of possible biases (Cashin, 1995). *Validity* of SET is thus usually concerned with whether SET is a good measure of teaching effectiveness (Beran & Rokosh, 2008). For the MOL project, we conducted a study to check the validity of SET in multi-section college chemistry courses in order to determine whether we should include SETs as part of the assessment for *Molecules of Life*.

There is a large body of literature that questions the validity of SET. It was shown that factors unrelated to teaching effectiveness may affect SET. These factors include instructor gender (Basow, 1995; Centra & Gaubatz, 2000), ethnicity (Rubin, Ainsworth, Cho et al., 1999), instructor rank and reputation (Griffin, 2001), instructor speaking style or "Dr. Fox effect" (Huemer, 1998; Naftulin, Ware, & Donnelly, 1973), year level and class size (Davies, Hirschberg, Lye et al., 2007; Wigington, Tollefson, & Rodriguez, 1989), work load (Greenwald & Gillmore, 1997b; Marsh, 2001), students' own

personalities and cultural background (Davies et al., 2007; Grimes, Millea, & Woodruff, 2004), students' fee-paying status (Davies et al., 2007), and expected grade (Eiszler, 2002; Griffin, 2004; Grimes et al., 2004).

Grading Leniency Bias

Over the years, much work has been done to examine the possible biases associated with SETs. One of the most frequent debates in the literature over the past decade on whether or not SETs are valid and useful is concerned with the possible "*grading leniency bias"* in SETs, namely, can an instructor receive higher SET rating from his/her students by simply giving higher grades to the students? While some researchers believe that an instructor does receive higher SET by giving higher grades, and that the correlation between grading leniency and SET is statistically significant (Greenwald & Gillmore, 1997a; Griffin, 2004; Olivares, 2001; Seiler & Seiler, 2002), others argue that grading leniency generally does not affect SET much, or that courses are rated lower when they were either too difficult or too easy (Centra, 2003; Marsh & Roche, 2000; Remedios & Lieberman, 2008).

Consequently, there are at least four theories that attempt to explain the relationship between grades and SET:

- The first theory proposes that teaching effectiveness influences both grades and ratings (McKeachie, 1979). The fundamental belief in this theory is that more effective instructors lead to higher student learning and hence higher grades. Within this framework students then give high ratings to the instructor because they learn more. To falsify this theory, SET data from different sections of the same course

taught by the same instructor could be used. Suppose the same instructor teaches all sections of the same course in the same way (e.g. using the same syllabus, textbook, exams, assignments, and the same lecture style…etc) but with different grading leniency, if the instructor receives a higher average SET rating from the section with the higher grading leniency, then this theory would be falsified.

- The second theory argues that students' general academic motivation affects both their grades and the ratings they give to instructors. According to this theory, academically motivated students tend to do better in coursework (thus obtaining better grades) and appreciate the instructors more, accordingly they also tend to give higher SET ratings to their instructors. (Marsh, 1984) Although student motivation is hard to measure and may change from week to week, analysis following a pretest and posttest procedure on students in different sections of the same course may be able to rule out the possibility that students in different sections have different motivation. If a pretest at the beginning of the semester shows no difference in motivation across sections, yet a posttest at the end of the semester shows clear difference in SET rating between different sections, then this theory could be contested.
- The third theory argues that students infer course quality and their own ability from received grades. This is based on social psychological attribution theories that "people tend to accept credit for desired outcomes while denying responsibility for undesired outcomes". (Greenwald & Gillmore, 1997a) This theory predicts that students attribute high grades to self-intelligence or diligence while ascribing low grades to poor instruction (Feldman, 1997; Gigliotti & Buchtel, 1990; Theall, Franklin, & Ludlow, 1990). According to this theory, giving high grades will not

improve the SET rating students give to an instructor, since students will accredit high grades to their own self-intelligence or diligence. Thus to falsify this theory, SET data from different sections of the same course taught by the same instructor could be used. Suppose the same instructor teaches all sections of the same course in the same way but with different grading leniency. If the instructor receives a higher average SET rating from the section with the higher grading leniency, then this theory would be challenged.

- The fourth theory is based on the well-known social psychology notion that praise generates fondness for the praiser (Aronson & Linder, 1965). This theory leads to the prediction of the grading leniency bias. This theory argues that when an instructor gives high grade to students, he/she virtually praises the students, who are then expected to like the instructor more and give high SET ratings to the instructor (Chacko, 1983) (Worthington & Wong, 1979). To test this theory, SET data from different sections of the same course taught by the same instructor could be used. Suppose the same instructor teaches all sections of the same course in the same way but with different grading leniency. If the instructor receives a higher average SET rating from the section with the higher grading leniency, then this theory would be supported.

Although all of the above four theories have been mentioned in the literature, most literature debates in the past decade were focusing on the fourth theory (grading leniency bias), probably because all of the other three theories are related to the fourth theory. The other three theories could all be easily challenged if evidence supports the

fourth theory. Thus this study only focused on the fourth theory above instead of the others. The first research question of this study described later aims at this focus.

Does Curricular Reform have an Effect on SET?

Another factor that might affect SET much is curricular reform, which is hardly explored in the literature. Curriculum reform, especially a pedagogy shift from traditional lecturing to cooperative, inquiry-based learning, is taking place in many colleges and universities all over the world (Maguire & Edmondson, 2001). One of the challenges an instructor most frequently faces is the fear that SET ratings from students may plunge if a course is reformed, as students might not be used to the reformed course. Maguire and Edmondson's work showed that this is not the case for the nursing course they investigated, as most nursing students welcomed the reform and rated the reformed course highly (Maguire & Edmondson, 2001). However, whether similar reform in college chemistry courses will be welcomed by students has not been explored in the literature. It is thus important to examine the effect of curricular reform on SET.

Research Questions

The raison d'être of this study is to examine the validity of SET in college chemistry courses in order to determine whether we should include SETs as part of the assessment for *Molecules of Life*. Hence we attempt to investigate the following two research questions with respect to introductory, college level chemistry courses:

1) Will an instructor receive higher SET rating from students if he/she gives higher grades to the students?

2) What is the effect of curricular reform on SET?

Research Methods

Major Research Techniques

The major techniques utilized in this research were quantitative analyses on the SET data as well as qualitative and quantitative analysis on a related survey that asked students about their expected grades, and the SET score they gave to the instructors. Since the two research questions basically form a correlational study, quantitative, correlational analysis on the SET scores that students gave to instructors was deemed to be most suitable to find the correlations and associations between curricular reform, students' grades, and the SET rating they gave to their instructors. Also, to triangulate the data and results, an online survey completely separate from the SET forms that students filled at the end of semesters was collected on a smaller sample of students to verify how students' perception of grading fairness and the curricular reform affect the SET evaluations they gave to the instructors. The online survey also asked open-ended questions encouraging students' open comments on the course or instructor. Thus there were qualitative as well as quantitative data from the online survey, which were used to triangulate the quantitative data from SET forms.

Sampling and Samples

The research questions limit the population of interest to be college students taking introductory, college level chemistry courses. Hence college students enrolled in two introductory chemistry courses, namely, General Chemistry I and General Chemistry II, were regarded as a representative sample. Thus the course-level SET ratings for all

General Chemistry I and II sections at a large public research university in the southeastern United States for the five-year period from the Spring 2000 semester through the Fall 2004 semester was collected for correlational analyses. Typically there were between 1 and 8 general chemistry sections taught by different instructors in each Spring, Summer, and Fall semester with each section containing 100 to 200 students. Thus for five-year period, 99 different general chemistry sections' grade distribution and SET data were collected, among which 10 sections had missing data, as they only had grade distribution data with no course evaluation data at all. After dropping those 10 sections, the remaining 89 General Chemistry I and General Chemistry II sections, including all available data for all the Spring, Summer, and Fall semesters from 2000 through 2004, were included in the final study. With this many different sections taught by different instructors, it is acceptable to assume that there is enough variance in the SET data. To verify this, a ***power analysis*** was done presupposing an estimated *medium effect size* of 0.5. The reason that a medium effect size is chosen here is because "a medium effect size is conceived as one large enough to be visible to the naked eye" (Cohen, 1988). In other words, a medium effect size is one just large enough that an instructor can notice, hence it is the size that is meaningful when we compare two different sections' SET ratings. A sample size of 86 was found necessary to achieve 80% power of correctly rejecting a false hypothesis. Since our sample size was 89, it was determined that we had sufficient power to perform statistical hypothesis testing.

The General Chemistry I and General Chemistry II courses are typical courses offered to freshmen and sophomore students of various majors, including pre-medicine, engineering, chemistry, biology, pharmacy, psychology, and other arts and sciences

majors. Hence the students in the general chemistry courses are representative of students taking introductory college chemistry courses at large public universities in the United States. Also, the General Chemistry I and II courses are offered every semester in all Spring, Summer and Fall semesters, compared to other chemistry courses that might be only taught once a year. Normally there are about 100 to 200 students in each general chemistry course section, with each section taught by full-time faculty members. For the five-year period from Spring 2000 through Fall 2004, there are more than 10,000 students enrolled in those 99 different sections of the general chemistry courses. Thus the general chemistry courses formed a representative sample suitable for this study. The grade distributions for every General Chemistry I and General Chemistry II course section (separate from the SET data) were used to determine the grading leniency of the instructors. At every Fall semester starting from Fall 2002, there was one section at the university explicitly implemented the peer-led guided inquiry (PLGI) and was thus curricular reformed. Therefore, SET data for both sections with curricular reform and sections without reform were used to compare the effect of curricular reform on SET.

The Instruments: SET Form & Online SALG Survey

The official SET form used at the university of investigation contained eight Likert scale questions. A re-typed copy containing the exact same eight questions as in the official SET form is in Appendix B. The SET data collected using the course evaluation records from university administration contained course-level SET, which included the aggregated average rating on a Likert scale from 1 to 5, as well as the distribution of the ratings (e.g. percent of respondents who gave each of the five possible ratings on a Likert scale of 1 to 5 for each of the 8 questions for each course at each

semester. (A Likert scale measures the extent to which a person agrees or disagrees with the question. The most common scale is 1 to 5. Often the scale will be 1 = strongly disagree, 2 = disagree, 3 = not sure, 4 = agree, and 5 = strongly agree.)

The online survey used in this study was the Student Assessment of Learning Gains (**SALG**) instrument developed by Seymour (Seymour, Wiese, & Hunter, 2000). It is a common faculty experience that "asking students what they 'liked' or 'valued' about their classes, or how they evaluated their teacher's professional performance, offers little information about what students actually gained from the class" (Daffinrud, 1997). In a study on student written comments in SETs conducted at 10 higher education institutions that consist of "3 research universities, 3 liberal arts colleges, 2 community colleges, 1 comprehensive state university, and 1 historically-black college", it was found that "the grand totals for all students' comments evaluating faculty teaching strategies were (for both the reformed, modular courses and the more traditional, non-formed classes) broadly 50% positive and 50% negative" (Seymour et al., 2000). It was also shown that most students answer traditional SET questions that ask about what they liked or disliked about an instructor or course based solely on impression instead of on systematic analysis. Thus it is "more productive to ask students how much they have **gained** from specific aspects of the class than what they liked or disliked about the teacher and his or her pedagogy" (Seymour et al., 2000).

In light of this, Seymour developed the Student Assessment of their Learning Gains (SALG) instrument for instructors of all disciplines who "wish to learn more about how students evaluate various course elements in terms of how much they have gained from them" (Daffinrud, 1997). It was designed to provide instructors "information about

what students feel that they have **gained** from particular aspects of their classes, and from the class overall", and the student feedback from SALG can "guide instructors in modifying their courses to enhance student learning" (Daffinrud, 1997). Initially developed for chemistry instructors interested in finding out the effectiveness of the modular approach in the teaching of chemistry, SALG can be "easily modified to meet the needs of individual faculty in different disciplines and provides an instant statistical analysis of the results", thus "it is argued to be a powerful and useful tool" (Wiese, Seymour, & Hunter, 1999). In addition to Likert scale ratings, the SALG survey also allows each student to write open comments after almost every question to explain their ratings to each Likert scale item.

For this study, the SALG instrument was used in the format of online surveys for different General Chemistry I course sections during both Fall 2003 and Fall 2004 semesters at the university of investigation. The survey questions are shown in Appendix C. Once the results are collected, the students' answers to each question were coded to a Likert scale score from 1 to 5: with "No Help" coded to 1, "A little help" coded to 2, "Moderate help" coded as 3, "Much help" coded to 4, and "Very much help" coded to 5. Items that have "N/A" or blank (i.e. no answer) responses are not included in the data analysis. In the Fall 2003 semester, students from a total of six different General Chemistry I lecture sections participated in the SALG survey. Based on the instructors and on whether or not the section is curricular-reformed, the six sections were labeled as B, E, G, G2, H, and F (PLGI), respectively, where section G and G2 are both taught by instructor G, and **PLGI**, or *Peer-Led Guided Inquiry* (Lewis & Lewis, 2005a, 2008) denotes a section that is curricular-reformed with a combination of guided inquiry

(Farrell, Moog, & Spencer, 1999) and Peer-Led Team Learning, or PLTL (Gosser & Roth, 1998), a form of cooperative learning. In the Fall 2004 semester, students from a total of seven different General Chemistry I lecture sections took part in the SALG survey. The seven sections were labeled as A, A2, A3, C, D, E, and B (PLGI), respectively, where sections A, A2, and A3 were all taught by instructor A, and section B (PLGI) was a curricular-reformed section taught by instructor B.

To obtain meaningful student feedback and a deeper understanding of students' perception of the curricular reform implemented at the university of investigation, student open comments were collected from the SALG surveys during the Fall 2004 semester in both traditional and PLGI-reformed general chemistry course sections.

Conceptualization and Operationalization of Research Concepts

As mentioned earlier, in this study, ***student evaluations of teaching (SET)***, is defined as the course/instructor evaluations that students give. ***Grading leniency bias*** is defined as the supposition that whether an instructor can get higher SET ratings from students if he/she gives higher grades to the students. ***Grade distribution*** is delineated as the percentage of A's and B's in a course. As an example of contrast, instructors who gave an 'A' as the final course grade to 50% or more of the students in the class will be considered as "highly lenient", an instructor who gave an 'A' to between 30% and 50% of the students will be considered as "medium lenient", instructors who gave an 'A' to between 20% and 30% of the students will be considered as "slightly lenient", and instructors who gave an 'A' to less than 20% of the students will be considered as "non-lenient". ***Perceived grading fairness*** is the students' perceptions of how fair the overall grading policy and grading procedures are in a course. Also, for the purpose of this study,

the phrase "***Curricular Reform***" is used to describe the reform on the general chemistry course incurring the explicit and intentional implementation of collaborative, inquiry-based learning. Curricular reform is important as a suddenly reformed course might receive unexpected SET ratings from students as they might not be used to the abruptly reformed course.

For an instructor who teaches a given class at a given semester, each student in that class who chose to complete the SET form and turn it in would give an ***overall SET rating*** to the instructor, i.e. the score for Question #8 that asks for the student's "overall assessment of instructor" in the official SET form (see Appendix B). This score is normally on a Likert scale from 1 to 5, with 1 being "poor", and 5 being "perfect". The average of all the overall SET ratings, i.e. the scores for Question 8 in the official SET form (Appendix B) given by different students of the entire class will be used as the instructor's overall SET score for that class for that given semester. This way a score of 4.8 would be considered very good (or "excellent") while a score of 2.0 would typically be considered bad.

To operationalize the concepts, ***grading leniency*** will be measured as two variables: the percentage of students who get A's and the percentage of students who get B's as their final course grade in the General Chemistry class. Thus an instructor teaching general chemistry who gives an 'A' to 90% of the students will be considered much more "lenient" than an instructor teaching general chemistry who only gives an 'A' to only 20% of the students.

A general chemistry course with ***curricular reform*** can be distinguished from other, non-reformed general chemistry courses by whether or not peer-led guided inquiry

(PLGI) is explicitly and intentionally implemented for that course. PLGI is a new teaching practice combining cooperative learning with guided inquiry and "scaled up for large enrollment classes" (Lewis & Lewis, 2008). Normally PLGI is carried out in a way that divides students into small groups of about four students, with every four to five groups assigned to one peer leader to work in a 50-minute session once a week in separate small classrooms. The peer leader, usually an undergraduate student who has successfully completed the general chemistry course and received an A or B, oversees these groups of four. The peer leader in this setting has the role of a cooperative learning facilitator instead of a lecturer. More details of the PLGI setting and implementation are available in (Lewis & Lewis, 2005a, 2008).

Ethics

All students enrolled in general chemistry at the university of investigation for the pertinent time period were included in the study. The population was roughly 60% women and 40% men, their age range was 18-65, and the range in ethnic background is about 69.8% White (Non-Hispanic), 11.5% Black (Non-Hispanic), 10.1% Hispanic, 5.6% Asian/Pacific Islander, 0.5% American Indian, and 2.5% non-resident alien. Disadvantaged individuals such as prisoners, minors, and the mentally disabled, were not in the participant population.

SET forms and grade distributions for the General Chemistry course offered by the Department of Chemistry at the university of investigation for the five-year period from Spring 2000 through Fall 2004 were collected. SET forms are anonymous and grade distributions provide no data where an individual student can be identified. Also, an online survey was used to gather additional student perceptions about the course during

the Fall 2003 and Fall 2004 semesters. The students was asked to take the survey online on a completely voluntary basis. As customary, the survey responses were not linked to personal identifiers. In other words, the instructors of the generl chemistry course were not able to tell which student wrote down what. Hence the students' rights were protected.

The grade distribution data were used solely for the purpose of this study and unnecessary exposure was avoided, as the data, although being publicly-available records, contain information that might affect an instructor's reputation (especially among students). Also, students' answers to survey questions were kept confidential and were not disclosed to the public or to administrative offices that might judge the instructor's teaching effectiveness. Hence the instructors' rights were protected.

Overall, this study brings about no physical or economical harm to anyone. Ethical guidelines were strictly followed throughout the study to ensure it meets the standards set forth by Federal guidelines, the American Sociological Association Code of Research Ethics, as well as the university Institutional Review Board (IRB) policies.

Results from Quantitative Analysis on SET Data

Reformed vs. Non-reformed Sections

In the data collected, there were 89 General Chemistry I and II sections in total from as early as Spring 2000 to as late as Fall 2004, out of which 3 sections were curricular-reformed, and the remaining 86 sections were non-reformed. Table 2.1 lists the descriptive statistics for the 86 non-reformed and the three reformed sections. The mean, standard deviation (Std.), Skewness and Kurtosis for the grade distribution (including %A as the percentage of students in a section that obtained an A as their final course

grade, and %B as the percentage of students in a section that obtained an B as their final course grade), average SET scores for all 8 questions in the official SET form (labeled as Q1, Q2, Q3, … and Q8), were included (Table 2.1).

Table 2.1 Descriptive Statistics for Reformed and Non-Reformed Sections

	Non-Reformed Sections[a]				Reformed Sections[b]			
Measure	M	SD	Skewness	Kurtosis	M	SD	Skewness	Kurtosis
A%	14.49	13.64	3.03	12.60	13.72	5.52	-0.09	
B%	24.87	8.44	0.09	0.28	23.56	6.64	0.13	
Q1	3.70	0.55	-0.85	0.62	3.70	0.49	-1.64	
Q2	3.30	0.74	-0.66	0.07	3.19	0.61	-1.57	
Q3	3.58	0.63	-0.84	0.44	3.37	0.40	-1.73	N/A[c]
Q4	3.50	0.61	-0.52	-0.02	3.52	0.72	-1.73	
Q5	3.64	0.69	-0.59	-0.35	3.41	1.08	-1.73	
Q6	3.36	0.70	-0.69	0.16	3.19	0.61	-1.57	
Q7	3.38	0.67	-0.62	0.11	3.25	0.64	-1.44	
Q8	3.51	0.72	-0.69	0.10	3.36	0.73	-1.61	

[a]n = 86 (i.e. there were 86 non-reformed sections in our sample).
[b]n = 3 (i.e. there were 3 reformed sections in our sample).
[c]Kurtosis for reformed sections could not be computed as n was too small.

Skewness describes the degree of asymmetry in a frequency distribution. If the mean is larger than the median, then the "tail" of the frequency distribution is towards the positive side, and the skewness would be positive; if mean is smaller than the median, then the "tail" of the frequency distribution is towards the negative side, and the skewness would be negative. Usually if skewness is between -1 and +1, the distribution is approximately symmetric, otherwise the skew is easily visualizable; and if skewness is larger than 2 or more negative than –2, then the distribution has very pronounced skew. *Kurtosis* indicates the degree to which a frequency distribution is peaked with heavy tails. If kurtosis is negative, then the distribution is called *platykurtic*, meaning less outlying values than a normal distribution; if kurtosis is positive, then the distribution is *leptokurtic*, showing a sign of more outlying values than a normal distribution. Usually if kurtosis is between –0.1 and +0.5, it has approximately the same number of outlying

values as a normal distribution; if kurtosis is more negative than -0.1, then it has noticeably less outlying values than a normal distribution; if kurtosis is larger than 0.5, then it has noticeably more outlying values than a normal distribution. The kurtosis for the reformed sections could not be determined as there are only three sections, and a division-by-zero error would occur if kurtosis were calculated using the standard formula. As Table 2.1 shows, the distributions of %B and all eight SET scores (Q1 through Q8) for the non-reformed sections are approximately normal (skewness is between –1 and +1), and most distributions for the three reformed sections have a noticeable negative skewness between –1 and –2, indicating that each of the eight SET scores for the three reformed sections did not form a symmetric normal distribution. However, given the extremely small number of reformed sections (three to be exact), these skewness values were expected and posted no threat to the validity of our comparison between reformed and non-reformed sections.

To compare the grade distribution and SET scores of the reformed vs. non-reformed sections, an independent samples t-test was performed to compare the mean SET scores of the two groups (reformed vs. non-reformed sections). As part of the first steps of the independent samples t-test, the Levene's F-test for equality of variances found no significant difference in variances between the two groups, and then the equal variance independent samples t-test found no significant difference between the two groups for all eight SET scores (on Q1, Q2, …, Q8) as well as the grade distributions of A% and B% (Table 2.3). Table 2.2 lists the average number of students (n), average grade distribution (including %A as the percentage of students in a section that obtained an A as their final course grade, and %B as the percentage of students in a section that

obtained an B as their final course grade), average SET scores for all 8 questions in the official SET form (Q1, Q2, Q3, … and Q8) for the reformed sections, as well as for the non-reformed sections. The differences between the average SET scores for reformed sections and non-reformed sections were found to be very small, ranging from 0.1% to 6.5% (Table 2.2). Since these differences were both small (Table 2.2) and not statistically significant (Table 2.3), we can say that the reformed sections have comparable SET scores as the non-reformed sections. In other words, there is no noticeable difference in SET scores found between the reformed sections and the non-reformed sections. There is no evidence that the PLGI curricular reform affected the SET scores. Note that when comparing the SET scores for reformed vs. non-reformed sections, our null hypothesis was "the SET score for reformed sections equal to the SET score for non-reformed sections". Since there was no direction (i.e. higher or lower than) in the null hypothesis, the two-tailed t-tests (instead of one-tailed t-tests) should be and was used here.

Table 2.2 Average SET Scores of Reformed and Non-Reformed Sections

Section	N	A%	B%	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Reformed	157	13.72	23.56	3.703	3.19	3.38	3.52	3.41	3.19	3.25	3.36
Non-reformed	155	14.57	24.06	3.698	3.30	3.58	3.50	3.64	3.36	3.38	3.51
%Difference[a]	-1.3%	6.2%	2.1%	-0.1%	3.4%	6.1%	-0.8%	6.5%	5.3%	3.9%	4.3%

[a] $\%\text{Difference} = \dfrac{(\textit{nonreformed average - reformed average})}{\textit{reformed average}} \times 100\%$

Table 2.3 t-Test Comparing Reformed with Non-Reformed Sections

	Levene's Test[a]		t-test for Equality of Means (df = 87)				
						95% Conf. Interval of the Difference	
Measure	F	p value	t	p value	Mean Difference	Lower	Upper
Number of Students	.166	.685	-.321	.749	-7.85	-56.48	40.78
A%	.741	.392	-.096	.924	-0.76	-16.54	15.01
B%	.441	.509	-.265	.791	-1.31	-11.11	8.50
Q1	.069	.793	.016	.987	0.01	-0.64	0.65
Q2	.237	.628	-.248	.805	-0.11	-0.97	0.75
Q3	.699	.405	-.561	.576	-0.21	-0.94	0.52
Q4	.103	.749	.076	.939	0.03	-0.68	0.74
Q5	1.426	.236	-.541	.590	-0.22	-1.04	0.59
Q6	.116	.734	-.410	.683	-0.17	-0.99	0.65
Q7	.044	.834	-.326	.746	-0.13	-0.91	0.65
Q8	.013	.911	-.340	.735	-0.14	-0.99	0.70

[a]Levene's test for equality of variances

Grading Leniency Bias

To test the grading leniency bias, a correlation analysis was conducted to investigate whether an instructor will receive higher SET rating from his/her students if he/she gives higher grades to them. The statistical analysis found no significant correlation between the grade distribution (A% and B%) and the SET scores (Q1 through Q8).

As seen in Table 2.4, the correlations between the grade distribution (A% or B%) and the SET scores were very small, ranging from -0.127 to +0.047. Additionally, none of the correlations between grade distribution and the SET scores were statistically significant (Table 2.4). In other words, there is no evidence that an instructor who gave higher grades to his/her students received higher SET ratings from the students.

Table 2.4 Correlations between Grade Distribution and SET Scores

		A%	B%	q1	q2	q3	q4	q5	q6	q7	q8
A%	Pearson Correlation	--									
	p-value (2-tailed)	--									
	n	--									
B%	Pearson Correlation	-.154	--								
	p-value (2-tailed)	.151	--								
	n	89	--								
q1	Pearson Correlation	-.007	.192	--							
	p-value (2-tailed)	.948	.072	--							
	n	89	89	--							
q2	Pearson Correlation	.013	.144	.943*	--						
	p-value (2-tailed)	.901	.179	.000	--						
	n	89	89	89	--						
q3	Pearson Correlation	.047	.155	.965*	.957*	--					
	p-value (2-tailed)	.665	.147	.000	.000	--					
	n	89	89	89	89	--					
q4	Pearson Correlation	-.036	.260*	.918*	.904*	.903*	--				
	p-value (2-tailed)	.735	.014	.000	.000	.000	--				
	n	89	89	89	89	89	--				
q5	Pearson Correlation	-.167	.266*	.789*	.799*	.809*	.860*	--			
	p-value (2-tailed)	.118	.012	.000	.000	.000	.000	--			
	n	89	89	89	89	89	89	--			
q6	Pearson Correlation	.002	.150	.917*	.969*	.938*	.883*	.802*	--		
	p-value (2-tailed)	.988	.161	.000	.000	.000	.000	.000	--		
	n	89	89	89	89	89	89	89	--		
q7	Pearson Correlation	-.019	.179	.948*	.979*	.955*	.924*	.812*	.970*	--	
	p-value (2-tailed)	.857	.093	.000	.000	.000	.000	.000	.000	--	
	n	89	89	89	89	89	89	89	89	--	
q8	Pearson Correlation	-.037	.192	.950*	.980*	.962*	.936*	.872*	.972*	.979*	--
	p-value (2-tailed)	.733	.072	.000	.000	.000	.000	.000	.000	.000	--
	n	89	89	89	89	89	89	89	89	89	--

*Statistically significant at the 0.05 level (2-tailed).

Results from Qualitative Analysis on Students' Comments

Using the HyperRESEARCH software (Version 2.6 for Mac OS X), students' open comments collected from the online SALG surveys in Fall 2004 semester were classified into three major categories based on the subject each comment was referring to: PLGI sessions in general; peer leaders; PLGI homework. These three categories are

summarized one by one below. Since the PLGI sessions only met on each Friday, they are also referred to as "Friday sessions" or "Friday classes".

PLGI Sessions in General

For question Q1D in the SALG ("How much did each of the class activities help your learning: class presentations including lectures in Monday/Wednesday class, class presentations in Friday class, discussion in the Monday/Wednesday class, discussion in the Friday class, group work in the Monday/Wednesday class, group work in the Friday class, hands-on class activities in the Monday/Wednesday class, and hands-on activities in the Friday class?"), 118 different students wrote down comments, 66 of them mentioned Friday Sessions, 51.5% of which were positive, 33.3% were negative, 15.2% of which were mixed messages.

Students who liked the Friday sessions liked it because of its smaller class size, group setting, or more hands-on approach. Some typical positive comments are:

"The Friday discussion groups helped me very much because of the smaller groups."

"The Friday classes were the most help because it felt more like high school and reminded me how much I enjoyed small work groups."

"Lecture on Mon/Wed was helpful to organize the information but the Friday sessions really helped me to understand the information."

"I did not find the lecture to be much help at all, but the Friday's classes were better explained and broken down into simpler terms for us to understand"

"I found the Friday classes to be more helpful because we got to work hands-on and our questions could be worked out on a one to one basis."

"Friday sessions are extremely valuable. Being able to work in smaller groups makes it easier to grasp concepts."

"(The instructor) is a great professor... She really used the Friday classes well. While most professors would have taught the concept to the class as a whole, then done the Friday class to drive the point home, (the instructor) did it the other way around, to our benefit. The Friday before she would teach a concept, we were broken into small groups where we could really focus on clearing up concepts for each individual before we covered them in class on a general level. Then, when we went over the concepts in class, we all had a good idea of what she was talking about and were able to further clarify/ solidify the concepts taught to us. I know that a lot of people skipped the Friday classes, especially toward the end of the semester, but as someone who attended every lecture and every Friday class, I feel that I can very firmly say that the way that (the instructor) set up the Friday class and the lectures was really beneficial. I really feel that the small groups are an awesome way for students to gage how much they know about given concepts and be able to explain those learned concepts to others in a clear way."

"I really enjoyed working with a student just like me on Fridays. I feel that I learned more because it was a more personal setting."

"Lecture classes are too big. Friday groups gave the opportunity to learn the info in small steps."

Some students did not like the Friday sessions because they found the class material covered on Fridays to be too easy or not helpful with their exams. Some typical negative comments are:

"I feel that the materials presented in the Monday/Wednesday classes could have been explained further and the group work on Fridays could have gone a little deeper and focused more on what was on the exams."

"I felt that the Friday classes were not helpful at all. They were more focused on following procedure than trying to help students comprehend subject matter. I feel that time spent in lecture was much more productive and helpful."

"I think that the lecture class was a lot more helpful than the group work on Friday class. I learned more from viewing and taking notes than from discussing."

"Friday class really did not provide any help on test question concepts."

Some other students wrote down mixed messages about Friday sessions. These students found the Friday sessions to be helpful somewhat, but felt that it could be better as there is still much room for improvement. Some typical mixed comments about Friday sessions are:

"The Friday class was really good at teaching us one specific thing, but I am not sure it was worth missing a lecture day."

"I thought the Friday classes helped me to better understand a lot of the material, but I think it would have more beneficial if it was a class where we asked questions about what we did not understand during the Monday/Wednesday classes."

"I feel I learn much better in lecture and reading the text book than in the Friday sessions. The Friday session material would've been better explained in lecture I believe. And the lectures and slideshow were excellent!"

For question Q1E in the SALG, ("how much did the tests, graded activities and assignments help your learning?"), 123 different students wrote down comments. Only

eight of them mentioned Friday sessions, five of which were positive comments, one was a negative comment, and two were mixed messages.

Five students who liked Friday sessions think it's helpful to them. Typical positive comments are:

"The Friday sessions were the biggest help."

"The class and Friday session contents, tests, etc. provided way for understanding concepts through application."

"Friday class lecture and the homework were very helpful."

"The One Key (homework) did not seem to help me that much. I liked Friday classes the best."

One student did not like Friday sessions because of grading dissatisfaction: *"Friday sessions were just a huge 'pain in the neck' and not to mention grade downer."*

Two students wrote down mixed messages because they felt the Friday sessions could be improved by relating to the exams more: *"The Friday classes helped me but a lot of the test questions were more difficult then the practice questions we done!" "Friday class was okay, but not that informative."*

For question Q1G that asks the students to rate the information they were given about three things respectively: class activities each week, how parts of the class-work, reading, or assignments related to each other, and the grading system for the class (Appendix C), 64 students wrote down comments, of which only one mentioned Friday sessions, this is because question Q1G does not ask about Friday sessions at all, that student liked the Friday sessions because the Friday class material was "helpful": *"It was helpful to go over the chemistry material in the Friday sessions."*

For question Q1H that asks about individual support as a learner such as quality of contact with the teacher, with the tutors, with their lab TA and peer leader (Appendix C), 110 students wrote down comments, and 5 students mentioned Friday sessions, three of whom rated the Friday sessions positively, one had mixed comment, and one rated the Friday sessions negatively. The 3 students liked the Friday class because of the smaller group setting was helpful to them: *"… the Friday sessions seemed to help a lot." "I had an easier time relating to people when there was a smaller group like on Fridays and in the lab. During class time things were not very personal." "The Friday classes helped me more than anything else in this course."*

One student had mixed rating for Friday sessions because he/she enjoyed the Friday sessions but did not like the idea that attendance was required for Friday class: *"I enjoyed there being Friday classes but not that they were required."*

One student had a negative comment on Friday sessions: "*Group work (at Friday sessions) did not help me much."*

Peer Leaders

One of the unique characteristics of the Friday class that marks its difference from regular Monday/Wednesday lectures is the group activities led by peer leaders. In consequence, student comments on peer leaders are an important indication of their feedback on Friday sessions. Under question Q1H, 24 students commented on their peer leaders, 18 of which (75%) was positive, 4 of which (16.7%) were negative, and 2 of which (8.3%) were mixed.

Eighteen (18) students liked the idea of peer leaders because peer leaders were available to help and peer leaders related to students better and explained material in a way they can understand better. Some typical positive comments are:

"*Tutors, peer leaders, etc. were almost always available for help."*

"The Peer Leader and the Lab TA and (the instructor) were all very good."

"The Peer Leaders are a very good idea. We learn a lot from them."

"I thought that the Lab TA and my Peer Leader were an essential part for this class. It was easy for me to understand the material by asking them questions. They simplified things and made sure that I understood."

"Small groups with a leader is the best way to ask questions about what you don't know."

"My Lab TA and my Peer Leader were both very helpful, and because the groups were small it was easier to have one on one help."

"The peer leaders on Friday sessions was a good idea since we as students get another approach in how the material is presented"

"I thought that our peer leader had a good command of the subject."

"The Chem TA and the peer leader because they are more our ages and they can relate better"

"My peer leader and other students helped very much, because they explained it in ways that I could easily comprehend."

Four students did not like their peer leaders, probably because some of the peer leaders lacked experience and could not lead the sessions well enough:

"The teacher and peer leader failed to teach material effectively."

"The peer leaders weren't allowed to answer questions, and I couldn't gain contact with or wasn't aware of the other sources."

"Friday peer leader was extremely smart in chemistry, but not teaching or trying to make students understand (couldn't go outside of mind to try and explain stuff)"

'Feel our peer leader was not fit or well prepared to teach a class."

Two students wrote down mixed comments about peer leaders because they felt the peer leader could have been better:

"My Peer Leader helped a little, but not too much."

"My peer leader was okay, but at times she would stand over me and made me feel uncomfortable and incompetent."

In brief, although some students found their peer leaders lacked teaching experience, most students liked their peer leaders because of their high quality of contact with their peer leaders, and because peer leaders related to them better and explained material in a way that they could understand better.

Friday Homework

Some students commented on Friday homework under question Q1D and Q1E. One student thinks the Friday homework was "too easy" and wrote down the following comments under Q1D: *"The Friday Homework was sometimes too easy and pointless."*

For Q1E, 6 students wrote down comments about Friday homework, 2 of which were positive, 4 of which were negative. The two students liked Friday homework because it either helped them learn or helped them to improve grades: *"The homework assignments due on Friday forced you to review the material and learn something…"*

"The online homework assignments, Friday quizzes and the homework that is due on Fridays definitely helped to bring my grade up."

Four students did not think the Friday homework was helpful because the problems in Friday homework were not similar to problems in the tests or because they did not promote understanding of the material:

"The homework each Friday would be more helpful if the tests included problems similar to what was in the homework."

"The homework due for the Friday sessions was a big waste of time though. They were not helpful, and was not explanatory at all."

"I don't think that the work for Friday's class was very useful. The homework did not help me understand the material any better than if I had just read the book."

"The homework due is every Friday does not help at all b/c we are rushed when it is time to review the different answers we came up with."

In summary, a majority of students commented positively about PLGI curricular reform-related items, e.g. comments about peer leaders under Question 1H were 75% positive, and comments under Question 1E about Friday sessions in general were 63% positive. This triangulates the quantitative results in Tables 2.2 & 2.3 and provides further evidence that curricular reform did not negatively affect the SET ratings of the general chemistry course. Educators, instructors, administrators, and curriculum reformers shall therefore be encouraged to implement well-developed curricular reforms such as PLGI.

Conclusions

This study collected quantitative SET data from all 89 General Chemistry I and General Chemistry II sections at a large public research university in the Southeast U.S. from 2000 through 2004 as well as qualitative survey data in the Fall 2003 and Fall 2004 semesters. It was found that there was no evidence of "grading leniency bias": there was no evidence that instructors giving higher grades to students received higher SET ratings from the students. The correlations between the grade distribution and the SET scores of the corresponding course sections were found to be both small and not statistically significant. There was also no evidence found that a peer-led guided inquiry curricular reform affected the SET scores much, as the difference between the SET scores for the reformed sections and non-reformed sections were found to be both small and not statistically significant, and students' comments in the survey were mostly positive about the reform.

However, there may be different possible reasons as to why the grading leniency bias did not show up in our results. First, when students fill the SET forms at the university where our data is collected, they do not know their final course grade yet, as the course evaluations normally happen BEFORE the final exams. This way, even if the students' feelings about their grades may reflect on the SET scores they give to an instructor as grading leniency bias tells us, their perception of what their grades will be may be very inaccurate, which may cause noise and error between the relationship of the grade distribution and the SET scores, which, in turn, may be large enough to mask potential grading leniency bias in our analysis. Secondly, a major part (usually 80%) of the course grade and the grading schema in all the General Chemistry course sections is

determined by the collective of all course instructors instead of by individual instructors. About 80% of the course grade is determined by four multiple-choice exams and a final exam shared across all course sections, which were then machine-graded together using scantrons administered by the common course coordinator who coordinates all sections of the course. Most student may be well aware the situation that the instructor does not have much control over the grading of the course. Therefore, the students might not attribute their grades to their instructor at all, nullifying the grading leniency bias.

Also, our result that curricular reform did not have a significant effect on SET may not be generalizable to other courses or other institutions. The curricular reform implemented in our study was very specific, limited peer-led guided inquiry activities. It may be possible that this specific type of curricular reform does not affect students' opinions about a course instructor, while other types of reforms may dramatically change students' course evaluations. It is also possible that the level at which the curricular reform was implemented in our study is not great enough to significantly affect the SET rating of courses.

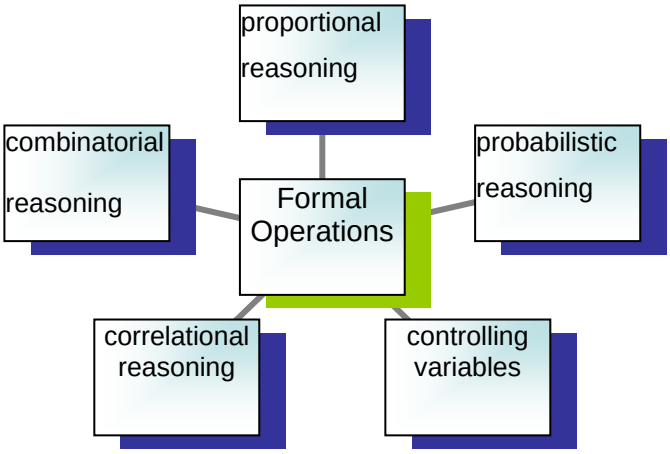
Because of these uncertainties and alternative explanations for the results of our study, and because of the large body of literature that questions the validity of SET in general for assessment purposes (Armstrong, 1998; Billings-Gagliardi, Barrett, & Mazor, 2004; Centra & Gaubatz, 2000; Davies et al., 2007; Eiszler, 2002; Gray & Bergmann, 2003; Griffin, 2004; Grimes et al., 2004; Kogan & Shea, 2007; Lang & Kersting, 2007; Marsh, 2001; Rubin et al., 1999; Trout, 2000; Youmans & Jee, 2007), we decided not to include SET in the assessment tools for the MOL project.

Chapter 3: Two Tests of Formal Reasoning

Introduction: Formal Reasoning Ability

Formal reasoning ability, namely, the ability to reason in the abstract beyond the bounds of specific contexts, has been shown to be essential for student achievement in science and chemistry courses (Cavallo, 1996; Hahn & Polik, 2004; Libby, 1995; Niaz, 1996; Niaz & Robinson, 1992; Nicoll & Francisco, 2001; Noh & Scharmann, 1997; Rubin & Norman, 1992; Uzuntiryaki & Geban, 2005). Formal reasoners have greater comprehension and generalizing skills (Boujaoude, Salloum, & Abd-El-Khalick, 2004; Oliva & Cadiz, 2003). According to Piaget's cognitive development theory (Figure 3.1), the logical or formal operations include theoretical reasoning, combinatorial reasoning, functionality and proportional reasoning, control of variables, and probabilistic reasoning (Good, Mellon, & Kromhout, 1978; Herron, 1975; Satterly, 1987; Williams, Turner, Debreuil et al., 1979; Zeidler, 1985). Piagetian theory expects most students in high school to be able to exhibit these reasoning patterns. However, research studies have shown that as many as 50% of students entering college do not fully have these reasoning abilities (Herron, 1975; Lawson, 1992a; Lawson, Drake, Johnson et al., 2000; McKinnon & Renner, 1971; Shibley, Milakofsky, Bender et al., 2003). Knowledge of students' formal reasoning ability is crucial in assessing their ability to work with and understand the quantitative and abstract nature of chemistry. Students who cannot reason formally have difficulty understanding equations, functional relationships and topics such as entropy, molarity, and concentration.

Figure 3.1 Five Types of Formal Reasoning Operations



According to the American Association for the Advancement of Science, "science, energetically pursued, can provide humanity with the knowledge of the biophysical environment and of social behavior needed to develop effective solutions to its global and local problems; without that knowledge, progress toward a safe world will be unnecessarily handicapped" (Project 2061: American Association for the Advancement of Science, 1990, p. vii). Give this importance of science literacy and the fact that lack of formal reasoning ability hinders students' science learning, it is important for science educators to have a reliable instrument to assess their students' formal reasoning level to guide remedial efforts to improve students' science learning.

There are many instruments used in science education to measure students' formal reasoning ability, including the Arlin Test of Formal Reasoning (Ablard & Tissot, 1998), the Inventory of Piagetian Developmental Tasks or IPDT (Coleman & Gotch, 1998), and the Lawson Classroom Test of Formal Reasoning (Cracolice, Deming, & Ehlert, 2008; Cuicchi, 1992; Lawson, 1978), to name a few. But two instruments in particular have been widely used in the chemical education research literature to measure students'

formal reasoning ability (Jiang, Xu, & Lewis, 2008). The first one is the Test of Logical Thinking (**TOLT**). TOLT is a 10-item/18-question, 40-minute, paper-and-pencil exam originally developed by Tobin and Capie to measure a student's mental capacity in terms of their ability of formal reasoning (Tobin & Capie, 1981). The TOLT test evaluates five reasoning abilities that have relevance to the teaching of science. It provides multiple justifications for the selected answer. The TOLT test contains two items from each of the following: proportional reasoning, probabilistic reasoning, controlling variables, correlational reasoning, and combinatorial reasoning. The official scoring procedures for the TOLT are described in the Data Source section of this paper. The other instrument is the Group Assessment of Logical Thinking (**GALT**). GALT is a 12-item/22-question, 30-minute, paper-and-pencil exam developed by Roadrangka et al (Roadrangka, Yeany, & Padilla, 1983). GALT is very similar to TOLT, as both of them measure the five types of formal reasoning (Figure 1). The major difference is that GALT has two additional concrete items (see Appendix E) that measures students' ability in *concrete thinking*, rooted in the principles of conservation such as conservation of mass and conservation of volume, which are not tested in TOLT.

GALT and TOLT have each been in common use by educators and researchers over the past two decades. For example, Noh and Scharmann found that GALT score was significantly correlated with students' conceptions and problem-solving ability (Noh & Scharmann, 1997); Bunce and Hutchinson found that GALT can be used to identify students at risk of failure in college chemistry (Bunce & Hutchinson, 1993); Poole showed that there were significant relationships between students' GALT scores and microbiology grades (Poole, 1997). Knight et al used TOLT and discovered that there

was no relationship between formal reasoning ability and the connected or separate knowing dimensions of their Knowing Styles Inventory (Knight, Elfenbein, & Martin, 1997). Verzoni and Swan found that TOLT scores were strongly related to conditional reasoning performance, namely, "the ability to reason deductively using inferential rules at progressively higher levels of abstraction" (Verzoni & Swan, 1995). Oliva and Cadiz showed that TOLT scores correlated significantly with mechanics conceptions (Oliva & Cadiz, 1999), and that TOLT score interacts with structural coherence of preconceptions to affect science conceptual change (Oliva & Cadiz, 2003). Boujaoude and coworkers illustrated that TOLT was a significant predictor of performance on conceptual chemistry problems (Boujaoude et al., 2004).

The validity and reliability of TOLT and GALT scores have each been investigated separately. TOLT has even been translated into Spanish (Oliva & Cadiz, 1999; Oliva & Cadiz, 2003) and Greek (Valanides, 1996). *However, no work has been done to explicitly investigate the advantage or disadvantage of the two additional test items on the principle of conservation that GALT contains over and above TOLT.* In a study by Williamson and others, TOLT was chosen because, based on the population of interest, "the shorter TOLT is a better choice" than GALT, as "few if any students would be predicted to lack conservation of matter or conservation of volume" (Williamson, Huffman, & Peck, 2004). But no evidence was given to support that claim. Other researchers prefer GALT simply because they believe including two concrete items on the principle of conservation in GALT enables the test to have more discriminatory power than TOLT (Baird, Shaw, & McLarty, 1996). However, there is no evidence supporting that belief. To be able to make a recommendation for people who are

interested in whether there is an advantage of including the two additional items on principle of conservation in the test, it requires us to compare the psychometric measurements of the test items on a representative sample. It may be useful to compare the TOLT exam with a modified "**TOLT+2**" exam that contains all TOLT items plus the two additional, concrete GALT items that test the principles of conservation. This study attempts to accomplish this examination by investigating the psychometric properties, such as reliability, item difficulty, and underlying factor structure, of the pure "TOLT" test and the "TOLT+2" test.

The research questions this study is concerned with are:

1) *Is there any advantage of adding the two extra concrete items on principles of conservation into TOLT, in terms of reliability and discriminatory power of the test?*
2) *Is it better to use the TOLT or the TOLT+2, in terms of potential bias of test items?*
3) *Which test is better in predicting college students at risk in the general chemistry course?*

Data Source

TOLT and TOLT+2 exams were collected at the beginning of Fall 2005 and Spring 2006 semesters at a large public research university in the southeastern United States. Students received attendance credit for completing the exam as a small portion of the course grade for a general chemistry class. Since the population of interest for this research is college students taking introductory, college-level chemistry courses, it is reasonable to assume that all students enrolled in the General Chemistry I course sections at the university of investigation form a representative sample. Each of these course sections has about 150 to 200 students enrolled, and each section participated in the exam.

The exam was administered in a way such that in each course section, each student was randomly given either "TOLT" or "TOLT+2", but not both, and that the number of students taking "TOLT" was approximately equal to the number of students taking "TOLT+2".

As mentioned earlier, the TOLT contains 10 test items or 18 questions, the first 8 items each being 2 multiple-choice questions, where the first question in the pair asks the student to choose the answer to a problem, and the second question in the pair requires students to choose the reason why that answer was chosen. The student has to answer BOTH questions correctly to be able to get one point for that item. Partial credit was not granted. The remaining two items are one question each, being open-ended combination or permutation problems that require enumeration of lists. Students receive a point of a complete and correct enumeration. No partial credit was given if the enumeration was incomplete or incorrect. Thus, scores on the TOLT test can range from 0 to 10. This grading schema is the official scoring process advised by the original test developer (Tobin & Capie, 1981) and used in the literature (Boujaoude et al., 2004; Knight et al., 1997; Oliva & Cadiz, 2003). In addition to all 10 test items in the TOLT, the TOLT+2 test has two additional concrete items on principle of conservation. The format of these two extra items are the same as the first 8 TOLT items, i.e. each item composed of a pair of two multiple-choice questions and a student has to select the correct choice on the first question and give the right reason on the second question to be able to get credit. Hence scores on the TOLT+2 test could range from 0 to 12.

All "TOLT" and "TOLT+2" tests of students in the general chemistry sections were included in this study. These two groups of students, namely, the TOLT test group

and the "TOLT+2" test group, form the focus of this study. The Fall 2005 sample size was n = 629 for the 10-item TOLT test, and n = 642 for the 12-item "TOLT+2" test. The Spring 2006 sample size was n = 362 for the TOLT, and n = 355 for the TOLT+2. Students' demographic information, as well as their scores in the verbal and quantitative sections of the Scholastic Assessment Test (**SAT**), were also collected with the approval of the institutional review board of the university. Also, using a locally developed questionnaire instrument named "**Day 1 Survey**" (see Appendix D), information concerning the number of semesters of high school chemistry each student took, as well as the highest level of math course each student has completed, was collected on the first day of classes. At the end of each semester, students' scores on a common final exam were collected from the university. The final exam was a third-party instrument developed by the Examinations Institute, Division of Chemical Education of the American Chemical Society (**ACS**), and it has been used as the final exam for general chemistry course at the university of investigation for several years. Due to confidentiality requirements for using the ACS exam, no test item from the ACS exam can be shown here at the present time.

Methods & Analysis

The data of the TOLT and TOLT+2 test scores as well as the Day 1 Survey, SAT scores and ACS exam scores were first input to Microsoft Excel and then converted into SPSS format for analysis using the SPSS software (version 14.0 for Windows).

Student Population in our Sample

There were 1991 students enrolled in the general chemistry course in the Fall 2005 and Spring 2006 semesters. All of them participated in this study. Out of these 1991 students, 1116 (56.5%) of them were female, while 867 (43.5%) of them were male. Also, 198 (9.9%) of them were Asian, 236 (11.9%) were Black, 249 (12.5%) were Hispanic, 1202 (60.4%) were White, and 9 (0.5%) of them were American Indian. This diverse sample of students is typical of the student population taking the general chemistry course at the university of investigation. Table 3.1 compares students who took TOLT to those who took TOLT+2 in academic background, SAT and ACS exam scores. Using the robust equivalence test proposed by Lewis and Lewis (Lewis & Lewis, 2005b), we found the TOLT group and TOLT+2 group were equivalent in academic background, SAT, and ACS exam scores.

Table 3.1 Comparison of Academic Background: TOLT vs. TOLT+2 Group

	Group [a]	No. of Students	Mean	Standard Deviation	Effect Size[a]
Semesters of Completed High School Chemistry	A	877	1.86	0.816	0.012
	B	870	1.85	0.831	
Highest Level of Math[b]	A	874	2.87	0.973	0.050
	B	871	2.82	1.008	
Years in College	A	878	1.69	1.020	0.029
	B	872	1.72	1.056	
Anchor Sum[c]	A	991	6.45	2.655	0.079
	B	997	6.24	2.641	
Percent Score [d] in TOLT/TOLT+2	A	991	.645	0.265	0.051
	B	997	.658	0.240	
SAT Quantitative	A	860	558.40	77.812	0.009
	B	861	557.67	81.326	
SAT Verbal	A	860	544.35	74.115	0.011
	B	861	543.51	78.895	
ACS Exam Score	A	798	21.76	6.875	0.009
	B	802	21.70	6.911	

[a]Group A: students who took TOLT; Group B: students who took TOLT+2; Effect sizes calculated as Cohen's d (Cohen, 1988, p. 20). [b]1 = "haven't taken any math courses as advanced as algebra", 2 = "algebra and/or trigonometry", 3 = "pre-calculus", 4 = "calculus I", 5 = "calculus II". [c]Anchor sum: sum score of the 10 anchor items. [d]Proportion of correctly answered items in TOLT/TOLT+2.

Reliability & Discriminatory Power of TOLT and "TOLT+2"

Reliability of any educational test can be indicated by its *internal consistency*, namely, the extent to which the items on the test are internally consistent with one another, or in other words, the degree to which the items in the test are functioning in a homogeneous fashion (Popham, 2000). One indication of internal consistency is *item-total correlation*, namely, the correlation between the scores on each item and the total score on the test. Another one of the most generalizable methods of estimating the internal consistency of tests is *Coefficient alpha* developed by Cronbach (Crocker & Algina, 1986; Popham, 2000). Coefficient alpha can be any number between 0 and 1.00, with larger values (i.e. values closer to 1.00) indicating better internal consistency. The

widely-accepted social science cutoff value is that alpha of 0.70 or higher is satisfactory for research purposes (Nunnally, 1978).

Coefficient alphas for the TOLT and TOLT+2 were computed. Both tests were found to have relatively high reliability in terms of internal consistency. Table 3.2 lists the item-total correlations and coefficient alphas for TOLT and TOLT+2. The item-total correlations for TOLT ranged from .267 to .513. Coefficient alpha for the TOLT was found to be .756, with the 95% confidence interval between .732 and .778, exhibiting a reasonable level of reliability, as it exceeds the widely-accepted cutoff of .70. The item-total correlations for TOLT+2 ranged from .237 to .536 and Coefficient alpha was found to be .754, with the 95% confidence interval between .731 and .776. This reliability was also very reasonable as it exceeds the cutoff of .70. Therefore, the two tests showed similar level of reliability as measured by their overall internal consistency. However, when the item-total correlations were looked at as a measure of each individual item's contribution to the test reliability, it was found that item 1 had the lowest item-total correlation among all items, and item 2 also had a item-total correlation well below the average of all items (Figure 3.2). This result was an indication that the two extra concrete items, items 1 and 2, did not contribute much to the reliability of the test. Therefore, in terms of reliability, we found no advantage of adding the two extra concrete items into the TOLT test.

The discriminatory power of the two tests can be measured by the difficulty of their items. According to educational measurement convention, *item difficulty* for a test item is defined as the proportion of test takers who answered that item correctly (Crocker & Algina, 1986). Thus the difficulty of test items can range from 0 to 1.00, with higher

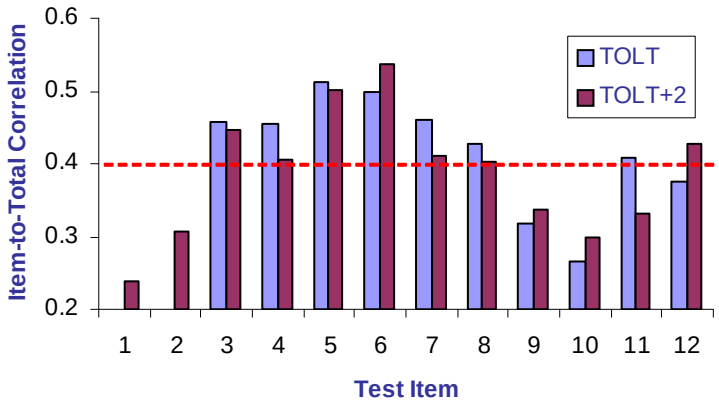
values indicating easier items. The difficulty for each item in the two tests was shown in Table 3.3. As far as items 3 to 12 are concerned, when each item in TOLT+2 is compared to itself in TOLT, it had about the same item difficulty. When items were compared to each other, the majority of items have a difficulty level of .6 to .7 (Table 3.3), indicating that 60% to 70% of students answered each item correctly. However, item 1, one of the two concrete items that only appeared in TOLT+2, had a difficulty of .94, meaning that 94% of test takers answered this item correctly. This item was much easier than all other items in the two tests. This poor discrimination power, as well as the low item-total correlation item 1 in TOLT+2 exhibited that was lower than all other items' (Table 3.2), suggested that there was no advantage of adding the two extra concrete items into TOLT, in terms of discriminatory power of the test.

Table 3.2 Item-Total Correlations and Coefficient Alpha for Each Test

Item	1	2	3	4	5	6	7	8	9	10	11	12	Coefficient alpha	95% Confidence Interval for Coefficient alpha
TOLT*	-	-	.458	.456	.513	.500	.461	.427	.319	.267	.407	.374	.756	.732 ~ .778
TOLT+2	.237	.307	.447	.406	.502	.536	.410	.404	.338	.299	.331	.428	.754	.731 ~ .776

*TOLT does NOT have items 1 and 2, which are the two extra concrete items in TOLT+2.

Figure 3.2 Item-Total Correlations for Each Test Item



* The dotted line is the average item-total correlation of all items

Table 3.3 Item Difficulty for Each Item in Each Test

Item*	1	2	3	4	5	6	7	8	9	10	11	12
TOLT	-	-	.72	.62	.55	.58	.73	.71	.69	.67	.59	.59
TOLT+2	.94	.71	.74	.66	.51	.59	.72	.69	.63	.63	.53	.55

*TOLT does NOT have items 1 and 2, which are the two extra concrete items in TOLT+2.

Construct Validity Measured by Factor Analysis

Similar to all other psychological attributes such as emotional intelligence, creativity, or self-efficacy, formal reasoning ability is a hypothetical concept (a.k.a. construct) that is latent and not directly observable. It is thus important to demonstrate the **construct validity** of TOLT and TOLT+2, i.e. whether each test measures the unobservable construct of formal reasoning ability it purports to measure. One of the widely used approaches to construct validation is **factor analysis**, a multivariate statistical procedure used to investigate the internal structure of a test, such as the number of dimensions (or factors) of the test, correlations between/among dimensions, and the proportion of variance for each observed variable that is explainable by the dimensions (or **factors**) (Crocker & Algina, 1986). The observed variables are usually the scores for

different test items, and they are modeled as linear combinations of the factors, plus "error" terms (also referred to as "residuals"). Confirmatory factor analysis (**CFA**) is a factor analysis that seeks to determine if the number of factors, the relationships of the factors to the observed variables, and the relationships of latent variables to each other, conform to what is expected on the basis of pre-established theory. It can test whether measures created to represent a latent variable really belong together.

To examine the construct validity of the two tests, several confirmatory factor analysis (**CFA**) models were tested using the Mplus software (Muthen & Muthen, 1998-2005) to inspect the factor structures of the two tests. In general, factor analysis requires sample size to be at least 10 times of the number of items (Crocker & Algina, 1986), our sample size of n = 991 for the 10-item TOLT, or n = 997 for the 12-item TOLT +2, far exceeds this criterion, permitting reliable factor analyses.

For Test A (TOLT), a CFA was conducted on a 2nd-order factor model shown in Figure 3.3 (hereby referred to as Model A). Based on formal reasoning theory (Knight et al., 1997), five first-order factors were specified (denoted as *prop, contr, prob, corr,* and *combi* in Figure 3.3), each corresponding to one of the five formal reasoning operations: proportional reasoning, controlling variables, probabilistic reasoning, correlational reasoning, and combinatorial reasoning, respectively. Since there were two different items in the TOLT test for each of the five formal reasoning operations, each factor was specified to load on the two items that measure the formal operation it corresponds to. A second-order factor (denoted as "Formal Reasoning" in Figure 3.3) was then articulated to load on all five first-order factors to explain the correlations between the first-order factors. Using the tetrachoric correlations matrix (Crocker & Algina, 1986) of the 10-item

data, a weighted least squares (WLS) method was employed in Mplus to estimate goodness of fit of the factor model, with the variance on each latent factor initially fixed to 1.0. The estimation of the initial model suggested that model A was an excellent fit of the data, as indicated by the following model fitness indices: χ^2 (n = 991, degrees of freedom: df = 21) = 25.257, chi-square/df ratio = 1.20, Comparative Fit Index (CFI) = .999, Root Mean Square Error of Approximation (RMSEA) = 0.014, Standardized Root Mean Square Residual (SRMR) = .030. The numbers on the arrows in Figure 3.3 were the **factor loadings**, that is, the standardized regression coefficients from the regression of the score of each test item on the set of factors. A factor loading of a factor on a test item is always between 0 and 1, and the closer to 1 it is, the stronger the correlation between the factor and the test item is. All factor loadings in Model A were significant at $p < .05$ level. Normally in confirmatory factor analysis, when chi-square/df ratio < 2.0, CFI > .95, RMSEA < .08, or SRMR < .08, the model is deemed good fit of data (Hu & Bentler, 1999). Hence the above second-order factor model of Test A (TOLT) was an excellent fit, providing strong support for the formal reasoning theory as well as construct validity of the TOLT exam.

For Test B (TOLT+2), two CFA models were tested. The first model (Figure 3.4), hereby referred to as Model B, is similar to Model A above in that it also has the five first-order factors corresponding to the five formal operations. The difference is that an additional sixth factor (denoted as conc in Figure 3.4) was introduced here to tally the two extra concrete items (items 1 and 2) in TOLT+2. A second-order factor (denoted as Formal Reasoning in Figure 3.4) was then articulated to load on all six first-order factors to explain the correlations between the first-order factors. This hypothesized factor

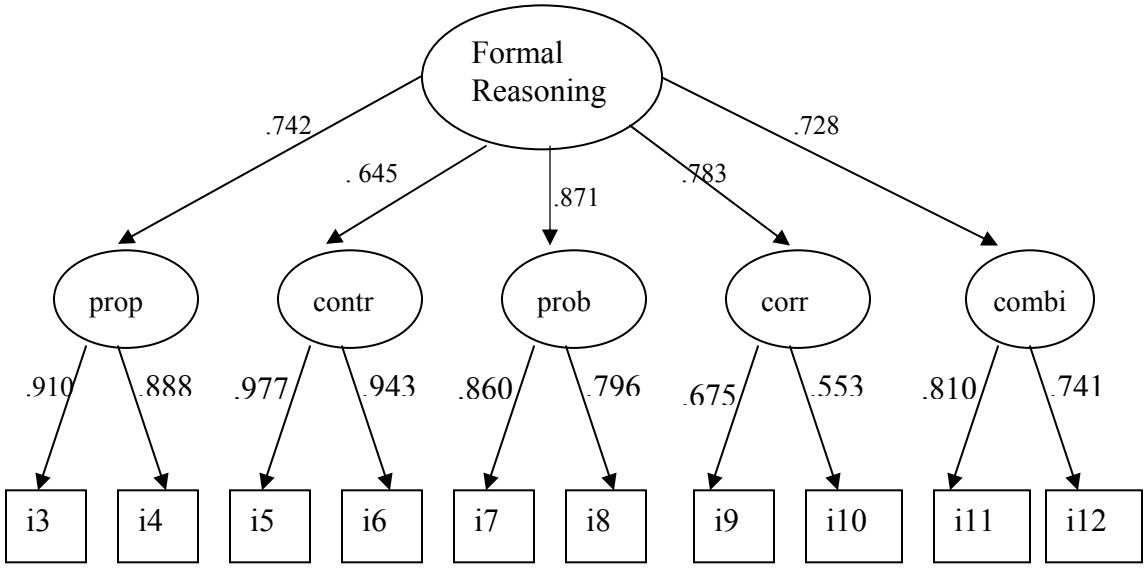
structure was in line with formal reasoning theory (Knight et al., 1997). Using the tetrachoric correlations matrix (Crocker & Algina, 1986) of the 12-item data, the weighted least squares (WLS) method was employed in Mplus to estimate goodness of fit of the factor model, with the variance on each latent factor initially fixed to 1.0. The estimation of the initial model suggested that the model was an excellent fit of the data, as indicated by the following model fitness indices: . χ^2 (n = 997, df = 35) = 42.790, chi-square/df ratio = 1.22, Comparative Fit Index (CFI) = .998, Root Mean Square Error of Approximation (RMSEA) = 0.015, Standardized Root Mean Square Residual (SRMR) = 0.038. The factor loadings in Model B were shown in Figure 3.4. All factor loadings were significant at $p < .05$ level. Since chi-square/df ratio < 2.0, CFI > .95, RMSEA < .08, or SRMR < .08, the above second-order factor model of test B (TOLT +2) was an excellent fit, providing strong support for the formal reasoning theory. On the other hand, Model B, the model for TOLT+2, was no better than Model A, the model for TOLT, since Model A actually had a CFI of closer to 1, a smaller RMSEA, and a smaller SRMR, indicating that Model A had a slightly better model fit than Model B. This result suggests that TOLT+2 did not have any advantage over TOLT in terms of their construct validity.

The second model for the TOLT+2 test (Model C), has the same six first-order factors as Model B (Figure 3.5). However, the sixth factor (denoted as conc) that corresponds to the two extra concrete items in TOLT+2, was specified to be completely uncorrelated with the other five factors here in Model C. Therefore unlike in Model B, the second-order factor in Model C (denoted as Formal Reasoning) was specified to load on the five formal reasoning factors only, but not on the sixth factor. Thus the concrete thinking factor was entirely separate from the rest of the model here. This structure was

to test the notion some science educators have that concrete thinking is a completely different and separate factor from formal reasoning and that adding the two concrete items into TOLT would simply introduce a separate factor. This notion does not conform to formal reasoning theory, as formal reasoning theory states that concrete thinking is a pre-requisite of formal reasoning and that students good at formal reasoning are already good at concrete thinking (Good et al., 1978; Herron, 1975; Shibley et al., 2003), but this notion holds that concrete thinking and formal reasoning are completely different abilities. If this notion was true, then Model C would have a better model fit than Model B. It turned out not to be the case. The model fit indices for Model C were found to be: χ^2 (n = 997, df = 28) = 383.896, chi-square/df ratio = 13.71, CFI = .899, RMSEA = 0.113, SRMR = 0.154. These fit indices obviously did not meet the criteria of chi-square/df ratio < 2.0, CFI > .95, RMSEA < .08, or SRMR < .08. Therefore Model C was a poor fit of our data.

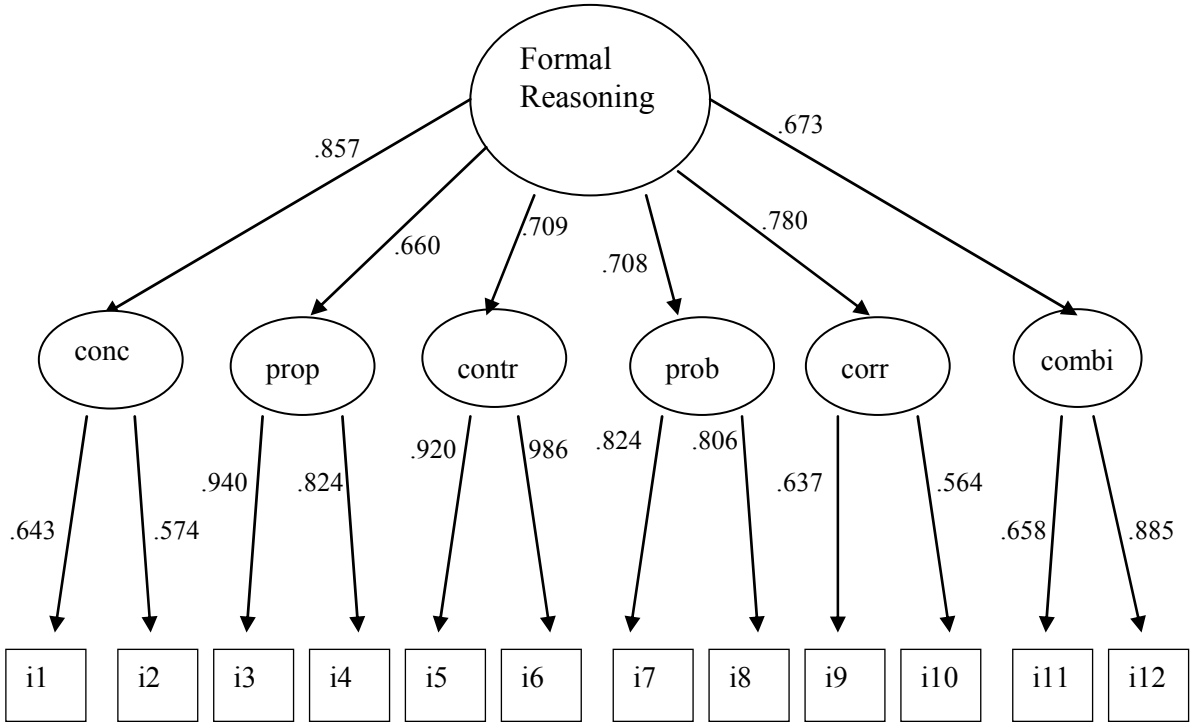
To sum up, when we let the concrete thinking factor to be related to formal reasoning (Model A and B), the model fit was excellent, but when we let concrete thinking factor to be completely unrelated to formal reasoning (Model C), then the model fit was poor. These results are in agreement to what we would expect from Piaget's Formal Reasoning theory, as they indicate that concrete thinking and formal reasoning are indeed related and that the effect of adding two concrete items into TOLT is not simply adding another totally different factor. It also indicates that TOLT+2 showed no obvious advantage over TOLT in terms of their construct validity measured by their underlying factor structures.

Figure 3.3 Model A: Confirmatory Factor Analysis (CFA) Model for TOLT



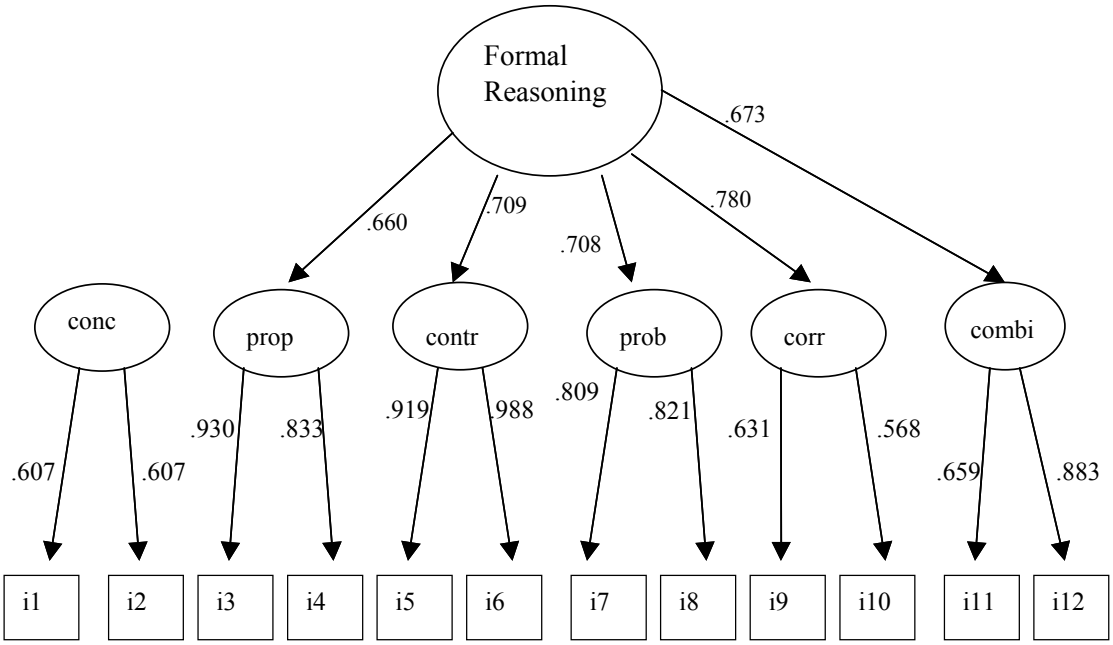
* prop: proportional reasoning, contr: control of variables, prob: probabilistic reasoning, corr: correlational reasoning, combi: combinatorial reasoning.

Figure 3.4 Model B: First CFA model for TOLT+2



* conc: concrete thinking, prop: proportional reasoning, contr: control of variables, prob: probabilistic reasoning, corr: correlational reasoning, combi: combinatorial reasoning.

Figure 3.5 Model C: Second CFA Model for TOLT+2



* conc: concrete thinking, prop: proportional reasoning, contr: control of variables, prob: probabilistic reasoning, corr: correlational reasoning, combi: combinatorial reasoning.

Differential Item Functioning (DIF) Analysis of Items in the Two Tests

The validity of making comparisons of test scores is based on the assumption that the measurement properties of the test are functioning similarly across different demographic groups, e.g., male students and female students. When a test item unfairly favors members of one particular group over another, it is biased. For instance, for male students and female students with the same level of algebra ability, an algebra test item in the context of golfing and football may unfairly favor males over females, as males tend to be more familiar with the contexts of golfing and football. Of course, whether an item favors males or females also depends on the culture they grew up in. A necessary condition for item bias is differential item functioning (**DIF**) (Clauser & Mazor, 1998). DIF occurs when performance on an item for members of two groups differ after they are

matched on the ability measured by the test. DIF is an important issue of test score validity and the most widely used methods to detect DIF is the Mantel-Haenszel (**MH**) statistics . MH statistics is based on the concept of *odds ratio*, the ratio of male students' odds of answering an item correctly over female students' odds of answering the item correctly when their ability are matched to be the same.

To examine whether any item in the TOLT and TOLT+2 tests have DIF, the Mantel-Haenszel statistics were computed using the SPSS software. Δ_{MH}, the measure of the effect size of DIF, i.e., the extent to which male students had an better odds of answering an item correctly than female students with the same level of formal reasoning ability, was calculated based on the Educational Testing Service (ETS) item classification system (Clauser & Mazor, 1998), which takes into consideration of both statistical significance and the practical effect size. Almost all items (except item 2) were classified as level A according to the ETS classification system and they were considered to display little or no DIF. Item 2 was the only one classified as level B using the ETS classification (Table 3.4) and should be deemed exhibiting moderate DIF. Item 2's MH odds ratio was 0.548 (Table 3.4), meaning that on average, the odds for females students to answer Item 2 correctly was only 54.8% of the odds that male students with the same formal reasoning ability had to answering Item 2 correctly. These results indicated that TOLT was better than TOLT+2 from item bias point of view, as TOLT does not contain Item 2, a potentially biased item.

Table 3.4 Mantel-Haenszel (MH) χ^2 and Odds Ratio Estimate for Each Item

Item	1	2	3	4	5	6	7	8	9	10	11	12
MH χ^2 (df = 1)	1.129	12.683[b]	6.783[b]	6.032[b]	.610	.763	9.759[b]	.167	1.277	10.053[b]	.179	3.828
Estimate of common odds ratio[a]	.679	.548	.705	.744	1.108	.891	.669	1.059	1.142	1.426	1.054	1.257
ln(odds ratio)	-.387	-.602	-.350	-.296	.102	-.115	-.401	.057	.132	.355	.053	.229
Std. Error of ln(odds ratio)	.327	.167	.131	.118	.121	.123	.127	.122	.112	.110	.110	.114
p-value (2-sided)	.237	.000	.008	.012	.399	.350	.002	.639	.236	.001	.632	.044
Lower bound of 95% CI[c]	.358	.395	.545	.590	.873	.700	.522	.833	.917	1.149	.849	1.007
Upper bound of 95% CI	1.290	.760	.912	.937	1.405	1.135	.859	1.346	1.421	1.768	1.309	1.571
Δ_{MH}	0.909	1.415	0.823	0.696	-0.240	0.270	0.942	-0.134	-0.310	-0.834	-0.125	-0.538
classification [c]	A	**B**	A	A	A	A	A	A	A	A	A	A

[a]Odds ratio: the ratio of the odds for female students at a certain formal reasoning ability level to answer an item correctly over the odds for male students at the same formal reasoning ability level to answer that item correctly. [b]χ^2 Significant at alpha = .05 level.
[c]95% CI: 95% Confidence interval for the odds ratio.
[d]Educational Testing Service (ETS)'s item classification system: $\Delta MH = -2.35 \ln$ (odds ratio). Items having $|\Delta MH| < 1.0$ or non-significant MH $\chi 2$ are classified as level A (little or no DIF); items having $|\Delta MH| > 1.5$ and significant MH $\chi 2$ are classified as level C (large DIF); items not meeting either criteria are classified as level B (moderate DIF).

Predicting Students At-Risk in General Chemistry: TOLT vs. TOLT+2

Research work by Lewis and Lewis has shown that the TOLT test can be used to predict at-risk students in general chemistry with an accuracy level that is comparable to predictions using SAT scores (Lewis & Lewis, 2007). Our linear regression modeling found that TOLT+2 has no advantage over TOLT in terms of predicting students' actual ACS Exam scores (details available upon request). However, prediction of success is different from predicting which students are at risk of failure in the general chemistry course. It would be interesting to see whether including the two extra items in TOLT+2 brings any advantage in predicting students at risk of nonsuccess in general chemistry. We designate nonsuccess as having a ACS exam score below 20 (i.e. a percent score of

below 50% in the exam), including withdrawals. Since the cutoff score likely affects the accuracy of predictions, a variety of cutoffs were investigated and discussed later.

Since the dependent variable to be predicted here is dichotomous (success vs. nonsuccess), *logistic regression* is a more appropriate technique than linear regression, as linear regression is only appropriate for predicting a continuous response (Legg, Legg, & Greenbowe, 2001). *Logistic regression* is a variant of linear regression as both can be considered as generalized linear models that use a set of independent variables to predict the dependent variable (**DV**). While the DV in linear regression is continuous and assumed to be a linear function of independent variables, the DV in logistic regression is dichotomous (1 vs. 0, e.g. success vs. nonsuccess), and it is not assumed to be a linear function of independent variables (Glass & Hopkins, 1996) . Instead, the logarithm of the odds for the DV to be 1 is assumed to be a linear function of the independent variables. In our logistic regression models, the probability (p) of a student succeeding the general chemistry course can be expressed as

$$p = \frac{1}{1+e^{-z}} \tag{1}$$

where z is the **logit,** defined as the natural logarithm of the odds of success, i.e. $z = \ln(\frac{p}{1-p})$. In logistic regression, the logit z is assumed to be a linear function of the predictors, in our case, TOLT or TOLT+2 scores.

Two logistic regression models were constructed for TOLT and TOLT+2 respectively. When TOLT scores (A) were used to predict logit z, the linear equation was found to be

$$z = -2.027 + 0.321 * \text{TOLT} \tag{2}$$

On the other hand, when TOLT+2 scores (B) were used to predict logit z, the linear equation was found to be

$$z = -2.434 + 0.313 * B \quad (3)$$

where B is the TOLT+2 score. The probability of a student succeeding the general chemistry course could then be expressed using Equation (1) in each model. Figure 3.6 shows the plot of this probability as a function of TOLT or TOLT+2 scores. This plot is similar to what Legg et al obtained in their analysis of success in general chemistry using logistic regression (Legg et al., 2001). However, Legg et al focused only on predicting success, while our model aims at identifying at-risk students. When the probability of success is below 0.5, the student is deemed at-risk. The accuracy of the predictions in logistic regression models can be measured by percent correct predictions, namely, the percentage of actual at-risk students over all students predicted at-risk (Lewis & Lewis, 2007). The TOLT model showed 71.2% correct prediction, while the TOLT+2 model showed 66.7% correct prediction (Table 3.5). Hence there was no considerable difference in the predictive accuracy between the TOLT model and the TOLT+2 model. Thus the TOLT+2 has no sizeable advantage over the TOLT in terms of predicting at-risk students.

These results were based on the cutoff score of below 50% in the ACS exam to be deemed "at-risk". When different cutoffs were used, the percent correct predictions of models varied and they were shown in Figure 3.7. The percent correct predictions for TOLT+2 and TOLT remained close for most cases and they were both between 50% and 70% in general (Figure 3.7), showing no apparent advantage of TOLT+2 over TOLT. Instead, when the cutoff was 30% (a raw score of 12 out of 40) in the ACS exam, the TOLT model had a considerably higher percent correct prediction (69.2%) than the

TOLT+2 model (50.0% correct prediction). Although this result might be due to an artifact of extremely small number of students predicted at-risk at the low ACS exam cut-off of 30%, it was consistent with the general pattern that the TOLT+2 model was no better than the TOLT model. Therefore, regardless of the cutoff for defining "at-risk", the TOLT+2 has no apparent advantage over the TOLT in terms of predicting at-risk students.

Figure 3.6 Predicted Probability of Success from TOLT or TOLT+2

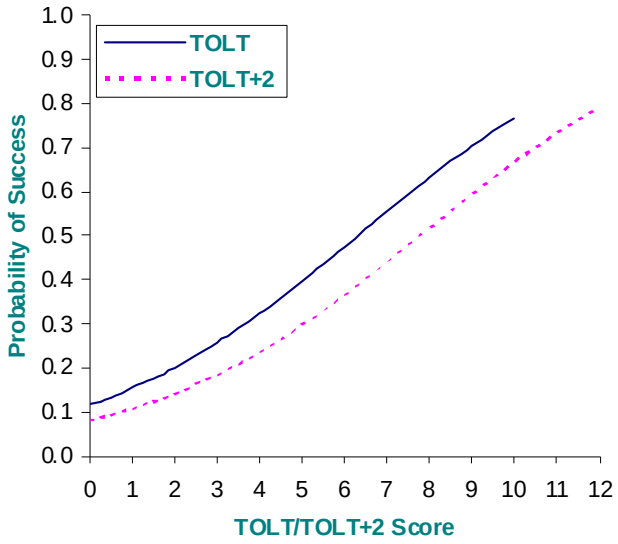
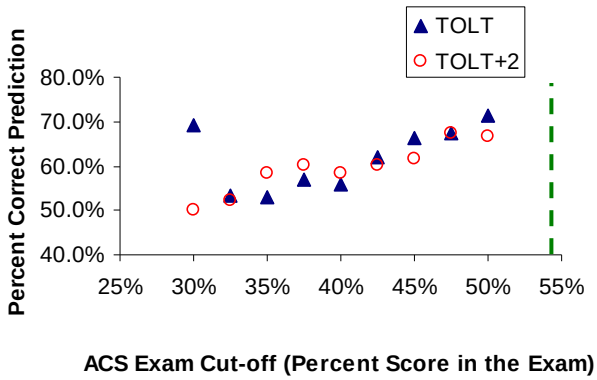


Table 3.5 Predicting At-Risk Students

	Predicted At-Risk	Actual At-Risk	Percent Correct Predictions
TOLT Model	451	321	71.2%
TOLT+2 Model	529	353	66.7%

Figure 3.7 Effect of Changing At-Risk Cutoff



* Percent Correct Prediction cannot be calculated when cutoff was below 30.0%, as the number of predicted at-risk students would be zero (0), and dividing actual at-risk by predicted at-risk would cause division-by-zero errors; the dotted line is the average ACS exam score of all examinees (54.3%, n = 1600)

Missing Data Analysis

The generalizability of findings from quantitative, statistical analyses was based on the assumption that the data used in the analyses were a randomly drawn, representative sample from the population of interest. If some test scores were missing in a non-random fashion, then the data available were not a random sample from the population, which would impact the generalizability of findings based on the available data. Thus it is important to analyze the missing data to find whether there are any non-random patterns in the missing of test scores.

1991 students in total from the Fall 2005 and Spring 2006 semesters took the TOLT/TOLT+2 test, of which 3 students (0.15%) had missing items 11 and 12. No missing data were present for items 1 to 10. Since these 3 missing TOLT/TOLT+2 scores are only 0.15% of our sample, the missing TOLT/TOLT+2 scores were ignorable. Out of the 1991 students, 1721 (86.4%) had SAT scores, while 270 of them (13.6%) had no SAT score. The reason these students had no SAT score was mostly because they took the

American College Testing Assessment (ACT) instead of the SAT. No significant correlations were found between the missing of SAT score and the missing of any TOLT/TOLT+2 item, revealing no pattern such as students having no SAT score also tended to miss certain items on the TOLT/TOLT+2 tests. The correlation between the missing of SAT score and missing of TOLT score was also small and not statistically significant ($r=-.015$, $p>.05$). There was a small but statistically significant, negative correlation found between students' Anchor Sum scores in the TOLT/TOLT+2 test and the missing of SAT score ($r= -.088$, $p<.01$), indicating that students who took SAT tended to do slightly better in TOLT/TOLT+2 tests than those whose did not take SAT. Students who had SAT also tended to have been in college for fewer years and they also have taken more semesters of high school chemistry than those who did not have SAT scores (Table 3.6).

Hence the missing on SAT scores, while not related to missing on TOLT/TOLT+2 scores, was not totally random. This would have a slight impact on analysis involving SAT scores in that the available SAT scores were not a strictly random sample from the population of interest. *However, the main focuses of the research questions in this study are concerned with TOLT and TOLT+2 tests. The non-randomly missing SAT scores would not significantly influence the major conclusions concerning the TOLT and TOLT+2 tests.*

Table 3.6 Comparison of Students Who Had SAT with Those Who Did Not

	Group	N	Mean	Std. Deviation	t-value in t-test (comparing means of SAT group with the missing-SAT group)	Effect Size (Cohen's d)
Anchor Sum	SAT	1718	6.44	2.611	3.741*	0.24
	Missing-SAT	270	5.76	2.814		
% Score in TOLT/TOLT+2	SAT	1718	.66	.250	3.654*	0.23
	Missing-SAT	270	.60	.266		
ACS score	SAT	1405	21.76	6.892	0.494	0.04
	Missing-SAT	195	21.50	6.898		
Semesters of high school chemistry[a]	SAT	1523	2.89	.795	3.770*	0.27
	Missing-SAT	224	2.63	.966		
Highest level of math[b]	SAT	1521	2.83	.971	-1.226	0.09
	Missing-SAT	224	2.93	1.114		
Years in college	SAT	1526	1.53	.861	-14.924*	1.03**
	Missing-SAT	224	2.89	1.325		

[a]1 = "No chemistry in high school", 2 = "1 semester", 3 = "1 full year" 4 = "1-2 full years" 5 = "More than 2 full years". [b]1 = "haven't taken any math courses as advanced as algebra", 2 = "algebra and/or trigonometry", 3 = "pre-calculus", 4 = "calculus I", 5 = "calculus II".
[c]Effect size was large. *Significant at alpha=.05 level.

Out of 1991 students, 1600 of them (80.4%) took the ACS exam. Since the ACS exam was a mandatory final exam for the general chemistry course, the reason the remaining 391 students (19.6%) did not take the ACS exam was mostly because they dropped the course before the end of the semester. No significant correlation were found between the missing of ACS score and missing of TOLT score ($r=-.019$, $p>.05$), while there was a small, but statistically significant correlation between missing of ACS scores and missing of SAT scores ($r=.081$, $p<.01$), suggesting that students who did not take the ACS exam tended to be those who did not take SAT. There was a small but statistically significant negative correlation between students' TOLT scores and the missing of ACS scores ($r=-.111$, $p<.01$), indicating that students who did better in TOLT showed a

slightly higher tendency to remain in the course for the final ACS exam than students who did not do well in the TOLT.

There was also a small but statistically significant negative correlation between students' actual SAT Quantitative scores and the missing of ACS scores ($r=-.145$, $p<.01$), while the correlation between students' SAT Verbal scores and the missing of ACS scores was both small and not statistically significant ($r=-.047$, $p>.05$). This suggests that students who had high SAT Quantitative scores showed a slightly higher tendency to remain in the course for the final ACS exam than students who had low SAT Quantitative scores. Students who took the ACS exam tended to be younger in terms of their years in college, have taken more semesters of high school chemistry, and had a better math background, higher SAT Quantitative scores and TOLT scores, than those who did not take the ACS exam (Table 3.7).

Hence the missing on ACS scores was not random. Students with missing ACS scores tended to be those with poor academic preparations as measured by SAT Quantitative scores, TOLT scores, and high-school math and chemistry background. Thus the results using actual ACS scores, e.g. the logistic regression equations 1) and 2) listed earlier would be generalizable only to students with a certain level of academic background. On the other hand, students missing ACS scores were mostly those who dropped the course before the end of the semester. These students should be considered at-risk and they were indeed considered at-risk in our logistic regression models. In other words, since we already included these students and considered them as actual at-risk in our analysis of the percent correct predictions, the prediction accuracy results from our logistic regression models would not be affected by these missing ACS scores. Therefore

our conclusions about which one of TOLT and TOLT+2 is better to predict at-risk students are valid even with these missing data.

Table 3.7 Comparison of Students Who Took ACS Exam with Those Who Did Not

	Group	n	M	SD	t value in t-test comparing ACS group vs. missing-ACS group	Effect Size (Cohen's d)
Years of College	ACS	1435	1.64	1.000	-5.604*	0.34
	missing-ACS	315	2.03	1.142		
Semesters of High School Chemistry	ACS	1433	2.90	.820	4.790*	0.29
	missing-ACS	314	2.66	.809		
Highest Level of Math	ACS	1430	2.91	.993	6.035*	0.37
	missing-ACS	315	2.54	.924		
SAT Verbal	ACS	1405	545.62	77.477	1.939	0.12
	missing-ACS	316	536.39	71.760		
SAT Quantitative	ACS	1405	563.49	80.074	6.460*	0.37
	missing-ACS	316	533.73	72.563		
Anchor Sum	ACS	1597	6.49	2.622	4.990*	0.28
	missing-ACS	391	5.75	2.680		
% Score in TOLT/TOLT+2	ACS	1597	.6654	.24951	4.974*	0.28
	missing-ACS	391	.5948	.25917		

*Difference between ACS and missing-ACS group significant at alpha=.05 level, all effect sizes were between small and medium.

Conclusions & Implications

There was no advantage of adding the two extra concrete items on principles of conservation into TOLT, in terms of reliability and discriminatory power of the test. Scores from the two tests of formal reasoning ability, "TOLT" and "TOLT+2", have comparable reliability, as scores of both tests exhibited reasonably high internal consistency of about 0.75, demonstrating no apparent advantage of TOLT+2 compared to TOLT. The common items of the two tests, items 3 to 10, have similar discriminatory power measured by item difficulty. But one of the concrete items, item 1, was much easier and showed a much lower item-total correlation than all other items in the two tests,

exhibiting a significant lack of discriminatory power for that GALT item specific to the TOLT+2 test. The result that 94% of students answered item 1 correctly also supported Williamson et al's untested claim that "the shorter TOLT is a better choice [than GALT] because few if any students would be predicted to lack conservation of matter" (Williamson et al., 2004).

One of the two extra items in the TOLT+2 exam, item 2, displayed statistically significant differential item functioning (DIF) with a moderate effect size. It was the only item classified as level B (moderate DIF) using the widely used ETS classification system, while all other items were classified as level A (little or no DIF). Therefore, from potential item bias point of view, the TOLT has advantage over the TOLT+2 since it does not contain item 2.

Also, the TOLT+2 showed no advantage over the TOLT in terms of predicting college students' at risk in the general chemistry course. Each test showed percent correct predictions of about 50% to 70% depending on the cutoff of defining "at-risk". Specifically, 71.2% of the students predicted at-risk by the TOLT model were actually observed to be at-risk, when at-risk was designated as dropping the course or scoring lower than 50% in the final exam. These prediction accuracies were comparable to those reported by Wagner et al (Wagner, Sasser, & DiBiase, 2002), or McFate and Olmsted (McFate & Olmsted, 1999). However, neither Wagner et al nor McFate & Olmsted's assessment could indicate what can be done to assist at-risk students, while our method have the advantage of a clear indication that interventions aimed at improving formal reasoning ability can be implemented, which has a solid research base (Adey & Shayer, 1990; Shayer & Adey, 1992b, 1993; Vass, Schiller, & Nappi, 2000). For example, Vass

and coworkers illustrated that classroom interventions improved students' probabilistic, proportional, and correlational reasoning skills (Vass et al., 2000).

Based on our results, there is no advantage of adding the two extra concrete items on principles of conservation into TOLT, in terms of discriminatory power, reliability, validity, or accuracy in predicting at-risk students. Instead, one of the concrete items significantly lacked discriminatory power, and the other GALT item showed significant differential item functioning, a condition of potential item bias. The TOLT is thus recommended over the TOLT+2 for use in general chemistry teaching to measure college students' formal reasoning abilities and to identify at-risk students.

These findings have practical implications for college chemistry teaching. First of all, when students' background information such as SAT scores are not accessible to instructors , and for students who do not have SAT scores, the TOLT offers an attractive alternative that can be easily used at the beginning of the semester for early identification of students at risk of failing the course. Secondly, given the relative ease of administering the TOLT as a one-time 40-minute test, as well as the fact that the TOLT is readily available to instructors free of charge, the TOLT is a preferred choice over other more costly, more time-consuming tests. More importantly, although several assessment/ placement instruments were reported recently (Legg et al., 2001; McFate & Olmsted, 1999; Wagner et al., 2002) to have percent correct predictions comparable to the TOLT in identifying students at risk in general chemistry, no work has yet been done to examine potential bias of their test items, and our work is the only one known to have investigated DIF, suggesting TOLT as a tenably bias-free test. Finally, for students found to have little or low formal reasoning ability, it is wise to let them participate in interventions such as

those described by Shayer and Adey (Adey & Shayer, 1990; Shayer & Adey, 1992b, 1993) or Vass et al (Vass et al., 2000) to improve their formal reasoning before they take general chemistry. Once their formal reasoning ability has been improved to a certain level, then they will encounter many fewer barriers in learning the abstract concepts in chemistry.

Chapter 4: Direct Comparison of TOLT and GALT as Intact Instruments

Introduction

Chapter 3 focused on the functioning of the two concrete items and compared TOLT and TOLT+2 in a general chemistry course during Fall 2005 and Spring 2006. As a follow-up, we also made a direct comparison between TOLT and GALT as intact instruments in both general chemistry and preparatory chemistry. The preparatory chemistry course at the university of investigation is offered to students with relative weak mathematics and science background, a different population from general chemistry students, e.g. students who either have never taken any chemistry course in high school or have an SAT mathematics score of less than 530.

The same research design and methods described in the first part were applied here. We collected data from general chemistry sections during Fall 2006 and Spring 2007 semesters as well as from the preparatory chemistry course during Fall 2006, Spring 2007 and Fall 2007 semesters. In each course, TOLT and GALT were assigned randomly (e.g. each student was randomly given either TOLT or GALT). Thus students were evenly distributed into the TOLT group and GALT group. Analysis of students' academic background showed no significant difference between the TOLT takers and GALT takers in either course, similar to the equivalence results in Table 3.1.

Reliability and Discriminatory Power of TOLT and GALT

Table 4.1 lists the reliability results of TOLT and GALT for these two courses measured by Coefficient alpha. TOLT was found to have slightly higher reliability than GALT. Also, when item-total correlations were considered as a measure of each individual item's contribution to the discriminatory power and reliability of the test, the GALT items were in general not impressive and had lower item-total correlations than the TOLT items (Table 4.1). One could argue that since TOLT and GALT have different number of test items, their reliabilities are not directly comparable. There are two facts that contradict this argument. First, as discussed in the literature (Bodner, 1980), when the Spearman-Brown prophecy formula is used to standardize the coefficient alpha, the standardized alphas can be directly compared, as the test lengths are already taken into consideration and the alphas are adjusted accordingly during the Spearman-Brown standardization (Bodner, 1980). Therefore, the higher standardized coefficient alpha for the TOLT suggests that the GALT does not have higher level of internal consistency than the TOLT. Secondly, the major difference between TOLT and GALT is the two extra concrete items that the GALT contains over and above TOLT. If we exclude the two concrete items, then the remaining 10 items in the GALT are very similar to the 10 items in the TOLT. One would expect that the standardized coefficient alpha based on the remaining 10 items in the GALT would be comparable to the alpha based on the 10 items in the TOLT. Actually GALT had a standardized alpha of .647 for all the 12 items and .622 for the remaining 10 items for the preparatory chemistry course, which is still no better than the standardized alpha of .669 based on the 10 items in the TOLT. This suggests GALT was no better than the TOLT in terms of test reliability.

Table 4.1 Item-Total Correlations and Coefficient Alpha for Each Test

Item*	1	2	3	4	5	6	7	8	9	10	11	12	Coefficient alpha
For Gen. Chem course													
TOLT	-	-	.404	.368	.453	.465	.408	.372	.317	.253	.332	.363	.716
GALT	.192	.336	.368	.338	.267	.173	.382	.391	.295	.311	.202	.282	.655
For Prep. Chem course													
TOLT	-	-	.457	.350	.409	.441	.371	.317	.280	.226	.217	.245	.669
GALT	.158	.338	.349	.334	.280	.155	.399	.443	.248	.275	.209	.302	.647

*TOLT has 10 items and labeled as items 3-12 here for comparison. TOLT does not have items 1 and 2, the two concrete items in GALT. For general chemistry course, the sample size is 920 for TOLT, 925 for GALT. For preparatory chemistry course, the sample size is 460 for TOLT, 439 for GALT.

Table 4.2 Item Difficulty for Each Item in Each Test

Item*	1	2	3	4	5	6	7	8	9	10	11	12
For Gen. Chem course												
TOLT	-	-	.81	.71	.67	.71	.78	.77	.71	.72	.64	.67
GALT	.92	.76	.73	.61	.79	.71	.86	.86	.60	.25	.87	.70
For Prep. Chem course												
TOLT	-	-	.56	.43	.45	.52	.65	.62	.59	.56	.54	.49
GALT	.83	.60	.52	.47	.66	.69	.76	.74	.36	.12	.84	.58

*TOLT has 10 items and labeled as items 3-12 here for comparison

One possible argument for using GALT was that while it offers no advantage for general chemistry students, its two concrete items might be useful for identifying low-reasoning-ability students in preparatory chemistry. It turned out not to be the case. Table 4.2 shows students in preparatory chemistry tended to score lower for each item in both instruments than students in general chemistry, consistent with the expectation that preparatory chemistry students have lower formal reasoning abilities. Item 1, one of the two concrete items in GALT, was again much easier than all other items, as 92% of general chemistry students and 83% of preparatory chemistry students answered it

correctly (Table 4.2). Not only was this GALT item too easy for general chemistry students, but also it was too easy and had a very low item-total correlation for preparatory chemistry students (Tables 4.1 & 4.2), hence it lacks discriminatory power, similar to results in comparison 1 (Tables 3.2 & 3.3).

Potential Item Bias

In terms of potential item bias, Mantel-Haenszel statistics showed that GALT had more frequently occurring biased items with ETS classification of 'C' (i.e. large DIF) for both general chemistry and preparatory chemistry students (Table 4.3). In this regard, TOLT is better than GALT in that it is tenably a less biased test.

Table 4.3 MH Odds Ratio Estimate for TOLT and GALT Items

TOLT for general chemistry course												
TOLT Item	--	--	3	4	5	6	7	8	9	10	11	12
MH χ^2(df = 1)			.00	2.11	2.65	.245	2.96	.007	.029	1.60	.001	.442
Ln(odds ratio[a])	--	--	.018	-.272	.311	-.113	-.368	-.003	-.045	.225	.017	.129
Δ_{MH}	--	--	-.042	.639	-.731	.266	.865	.007	.106	-.529	-.040	-.303
classification[c]	--	--	A	A	A	A	A	A	A	A	A	A
GALT for general chemistry course												
GALT item	1	2	3	4	5	6	7	8	9	10	11	12
MH χ^2(df = 1)	.246	8.091[b]	3.693	29.99[b]	.478	8.09[b]	.008	.016	1.16	.092	7.85[b]	2.52
Ln(odds ratio)	.175	-.554	-.365	-.908	.149	-.473	.052	.063	.191	.08	.647	.284
Δ_{MH}	-.411	1.302	.858	2.134	-.350	1.112	-.122	-.148	-.449	-.188	-1.520	-.667
classification	A	B	A	C	A	B	A	A	A	A	C	B
TOLT for preparatory chemistry course												
TOLT Item	--	--	3	4	5	6	7	8	9	10	11	12
MH χ^2(df = 1)			2.01	3.76	.019	.001	2.29	.05	.76	1.05	.51	2.71
Ln(odds ratio)	--	--	-.429	-.486	.067	.042	-.482	-.095	.251	.269	.193	-.429
Δ_{MH}	--	--	1.008	1.142	-.157	-.099	1.133	.223	-.590	-.632	-.454	1.008
classification	--	--	B	B	A	A	B	A	A	A	A	B
GALT for preparatory chemistry course												
GALT item	1	2	3	4	5	6	7	8	9	10	11	12
MH χ^2(df = 1)	.96	19.3[b]	2.29	2.71	.01	2.22	.052	2.31	2.40	4.12[b]	.001	.67
Ln(odds ratio)	-.367	-1.265	-.423	-.427	-.006	.402	.125	.516	.418	.817	.049	.222
Δ_{MH}	.862	2.973	.994	1.003	.014	-.945	-.294	-1.213	-.982	-1.920	-.115	-.522
classification	A	C	A	B	A	A	A	B	A	C	A	A

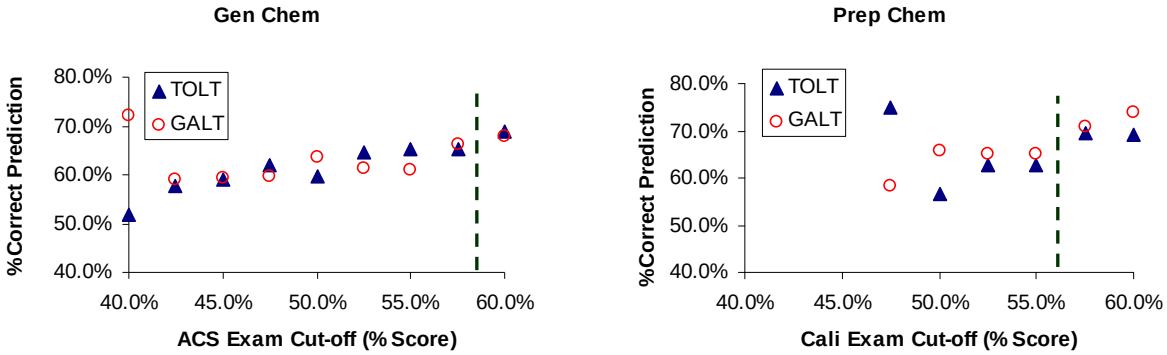
[a]Odds ratio: the ratio of the odds for female students at a certain formal reasoning ability level to answer an item correctly over the odds for male students at the same formal reasoning ability level to answer that item correctly; Δ_{MH} = -2.35 ln (odds ratio). [b]χ^2 significant at .05 level.
[c]Educational Testing Service (ETS)'s item classification system: Level A (little or no DIF): |ΔMH| < 1.0 or non-significant MH χ2; Level C (large DIF): |ΔMH| > 1.5 and significant MH χ2; Level B (moderate DIF): items not meeting either criterion.

Predicting At-Risk Students in General and Preparatory Chemistry

With regard to predicting at-risk students, GALT showed no advantage over TOLT in terms of percent correct predictions. Figure 4.1 shows percent correct predictions for TOLT and GALT from logistic regression models for general chemistry and preparatory chemistry. The ACS exam and the California Chemistry Diagnostic Test (Examinations Institute of the American Chemical Society Division of Chemical Education, 2006) were used as outcome variables for general chemistry and preparatory

chemistry, respectively. The results were based on different cutoff scores from 40 to 60% in the final exam for defining "at-risk". The percent correct predictions for TOLT and GALT models remain close in most cases and are both between 50% and 70% in general (Figure 4.1).

Figure 4.1 Percent Correct Predictions Using Different Cutoffs



*Plot on the left is for general chemistry; plot on the right is for preparatory chemistry. Dotted line in each plot is the mean exam score for all examinees (58.4% for ACS exam, 56.2% for California exam). Percent Correct prediction cannot be calculated when the cutoff is lower than shown, as the number of predicted at-risk students (i.e. students who have a 50% chance of being below the cutoff) becomes very small.

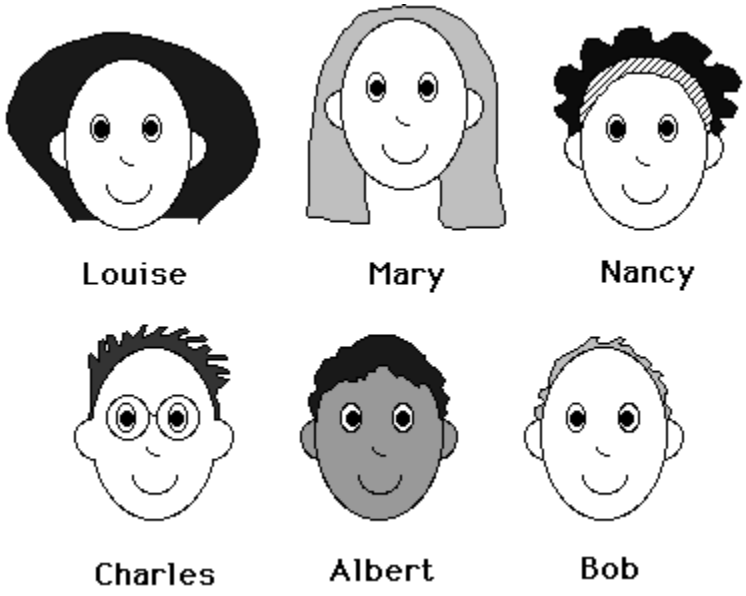
Other Concerns with the GALT

Besides Item 2 exhibiting a large DIF for preparatory chemistry students, Item 11 in the GALT was found to have a large level of DIF in general chemistry (Table 5.4). Item 11 asks student to enumerate all possible pairs of dance partners (Figure 4.2). This item explicitly requests students to "restrict the possible combinations to boys and girls dancing with each other", which lacks cultural sensitivity (Foronda, 2008; Hutnik & Gregory, 2008; Liamputtong, 2008; Rogers, Graham, & Mayes, 2007), as in certain cultures girls are not allowed to dance with boys, such as in Muslim (Brown, 2008; Scrivener, 2003) and Orthodox Jewish (Boroff, 1961; Wolf, 2007) cultures. Students

from such cultures must answer it in a way counter to their cultural norm, and item 11 would in fact be offensive to them. Additionally, the way that boys can only be allowed to dance with girls implies heteronormativity (Desurra & Church, 1994; O'Conor, 1998; Rothing, 2008), which could "marginalize" (Desurra & Church, 1994, p. 23; O'Conor, 1998, p. 66-68) and "stigmatize" students with non-heterosexual identities (Rothing, 2008, p.259).

Figure 4.2 Item 11 from the GALT

Item 11. After dinner, some students decide to go dancing. There are three boys: Albert (A), Bob (B), and Charles (C), and three girls: Louise (L), Mary (M) and Nancy (N).



One possible pair of dance partners is A-L, which means Albert and Louise.

List all other possible pairs of dance partners in the spaces provided on the answer sheet. To reduce the number of possible answers to this question, you can restrict the possible combinations to boys and girls dancing with each other.

Conclusions

From the direct comparison between TOLT and GALT as intact instruments, GALT showed no advantage over TOLT for both general chemistry and preparatory chemistry in terms of reliability, discriminatory power, and potential item bias. GALT also showed no advantage over TOLT in terms of predicting college students to be at-risk in general chemistry and preparatory chemistry. Depending on how at-risk is defined, both instruments showed percent correct predictions between 50% and 70%, with neither model consistently ahead. These prediction accuracies are comparable to those reported in other studies using different instruments (McFate & Olmsted, 1999; Wagner et al., 2002). However, no work has yet been done to examine potential bias of the test items within these instruments, while our work has investigated DIF, suggesting GALT has more frequently occurring biased items, while TOLT is tenably a less biased test. If one wants to use GALT in college chemistry, then Items 2 and 11 in GALT need to be modified, as Item 2 consistently exhibited large level of potential bias against females across general and preparatory chemistry student population, while Item 11 displayed heteronormativity and lack of cultural sensitivity.

Since our sample only includes students in the first-semester general chemistry and preparatory chemistry courses at one large public university, our results do not necessarily apply to other chemistry courses, or to chemistry courses at other types of institutions. Also, formal reasoning is a necessary but not sufficient ability for success in chemistry (Lewis & Lewis, 2007), as it is not the only factor important for learning. A sizeable proportion of students that performed poorly on the ACS exam was not identifiable by TOLT or GALT models. For example, of the 994 students in the TOLT

group in the comparison of TOLT with TOLT+2 (Figure 3.6 & Table 3.5), 515 students finished the course below the ACS cutoff. Of these 515 students, only 321 students (62.3%) were identifiable based on the TOLT model. This suggests a necessity to include other predictors, such as spatial ability (Bodner & Guay, 1997; Yang, Andre, & Greenbowe, 2003) and affective measures like motivation (Cousins, 2007; Hahn & Polik, 2004; Hampton & Reiser, 2004) or self-efficacy (Lawson, Banks, & Logvin, 2007; Poole, 1997), each of which has been shown to be important predictors in science achievement.

As discussed earlier, there is an advantage to using a theory-based instrument rather than one that intends to measure prior knowledge of mathematics and chemistry. Entering chemistry students have typically had prior instruction in mathematics and chemistry, but a low score on an instrument containing mathematics and chemistry questions suggests only that this prior instruction was ineffective. What will make the second opportunity to learn basic mathematics and chemistry effective? On the question of an instructional approach for effective remediation, the instrument is silent. On the other hand, a low score on a formal reasoning measure suggests immediately two potential remedies: (1) Ensure that chemistry concepts are presented in a concrete way when they are initially introduced in the general chemistry course (Herron, 1975); (2) Apply specific interventions that have been shown to support the development of formal reasoning ability (Adey & Shayer, 1990; Adey & Shayer, 1994; Cattle & Howie, 2008; Endler & Bond, 2008; Vass et al., 2000). Also, given the need for effective remedial courses, future investigations that evaluate the equity implications of courses aligned with a cognitive development perspective may be beneficial to this important group of at-risk students. As one example, learning cycles have shown benefits for low-formal reasoning

students (Abraham & Renner, 1986), so a remedial course using a learning cycle approach would lend itself well to this type of investigation.

Chapter 5: Evaluation of Molecules of Life

Introduction: Spatial Ability and Science Education

Based on the results from the three related studies (described in Chapters 2, 3, and 4, respectively), the TOLT instead of GALT was included as one of the instruments in the assessment plan for the MOL project. Another instrument used was a measure of spatial ability and the reason it was included in the assessment follows.

Spatial ability is the ability to perceive and mentally manipulate two-dimensional and three-dimensional objects or figures (Ferk, Vrtacnik, & Blejec, 2003; Huk, 2006; Lohman, 1996; Mayer & Sims, 1994; Provo, Lamar, & Newby, 2002; Wu & Shah, 2004; Yang et al., 2003), although it has been defined in slightly different ways (Carroll, 1993; Cronbach & Snow, 1977; Lohman, Pellegrino, Alderton et al., 1987; McGee, 1979; Miyake, Friedman, Rettinger et al., 2001; Smith, 1964). In the multiple intelligences (MI) hypothesis (Hoerr, 2003; Kornhaber, 2004; Shearer, 2004), Gardner considers *spatial intelligence* as one of the seven basic intelligences, which include logical-mathematical, linguistic, musical, spatial, bodily-kinesthetic, intrapersonal, and interpersonal intelligences (Gardner, 1983, 1999; Gardner & Moran, 2006). According to Gardner, "central to spatial intelligence are the capacities to perceive the visual-spatial world accurately, to perform transformations and modifications on one's initial perceptions, and to be able to re-create aspects of one's visual experience even in the absence of relevant physical stimuli" (Gardner, 1983; Gardner & Hatch, 1989).

Two major dimensions or components of spatial ability are usually considered to be relevant to science education: spatial orientation and spatial visualization (Bodner & Guay, 1997; Huk, 2006; McGee, 1979; Shah & Miyake, 2005, p. 124-129). *Spatial orientation,* also referred to as *mental rotation* in the literature, is the ability to remain unconfused by changes in the orientation of objects, and it engages only a mental rotation of the configuration of objects (Bodner & Guay, 1997; Huk, 2006; Shah & Miyake, 2005, p. 124-129). *Spatial visualization* is the ability to mentally restructure or manipulate the components of a figure (Bodner & Guay, 1997; Ferk et al., 2003; McGee, 1979; Wu & Shah, 2004), and it engages "recognizing, retaining, and recalling configurations when the figure or parts of the figure are moved" (Bodner & Guay, 1997). Spatial ability is vital for activities with intense visual, 2-D and 3-D content such as engineering, science, technology, architecture, and medicine.

Research in science education has shown that spatial ability plays a crucial role in problem solving and understanding of scientific concepts. For example, Carter, LaRussa and Bodner illustrated that spatial ability correlated significantly with performance on novel problems, spatially-oriented tasks and tasks that require complex problem-solving skills in general chemistry (Carter, LaRussa, & Bodner, 1987). Yang, Andre, and Greenbowe showed that spatial ability might interact with instructor-guided animation treatment to affect students' understanding of electrochemistry concepts (Yang et al., 2003). Pribyl and Bodner studied the relation between spatial ability and organic chemistry achievement in four organic chemistry courses designed for students with various majors including agriculture, biology, health sciences, pre-med, pre-vet, pharmacy, medicinal chemistry, chemistry, and chemical engineering. They found that

"students with high spatial scores did significantly better on questions which required problem solving skills, such as completing a reaction or outlining a multi-step synthesis, and questions which required students to mentally manipulate two-dimensional representations of a molecule" (Pribyl & Bodner, 1987). Holland showed that chemistry achievement is affected by spatial visualization ability and that chemistry instruction using a greater visual means of instruction increased the achievement of medium and high visualizers, but not low visualizers (Holland, 1995). Provo, Lamar and Newby found that spatial ability affects veterinary students' 3-dimensional knowledge of anatomy of the canine head (Provo et al., 2002). Ferk, Vrtacnik, and Blejec developed a Chemistry Visualization Test (CVT) to assess the correctness of students' "perception of different representations of molecular structure" and their ability to "manipulate these mental images in three dimensions", which included tasks in five categories: 'perception'; 'perception and rotation'; 'perception and reflection'; 'perception, rotation and reflection'; 'perception and mental transfer of information' (Ferk et al., 2003). They found that there was a significant correlation between students' spatial visualization skills and their score on the Chemical Visualization Test (Ferk et al., 2003). Supasorn and coworkers found that high-spatial-ability students were better able to answer high cognitive thinking questions in organic extraction than low-spatial-ability students (Supasorn, Suits, Jones et al., 2008).

In addition to spatial ability, another cognitive construct extensively studied in the last few decades is *formal reasoning.* As mentioned earlier in Chapter 3, formal reasoning is the ability to reason in the abstract beyond the bounds of specific contexts; it has been found to have significant relationships with principled moral reasoning (Zeidler,

1985); and it has also be shown to be essential for students' successful achievement in science (Cavallo, 1996; Cuicchi, 1992; Giuliano, 1997; Griffin, 1997; Holland, 1995; Lawson, 1992a, 1992b; Lawson et al., 2007; Lawson et al., 2000; Lekhavat, 1996; Niaz, 1996; Niaz & Robinson, 1992; Noh & Scharmann, 1997; Rubin & Norman, 1992; Uzuntiryaki & Geban, 2005).

Spatial ability and formal reasoning are two fundamental cognitive constructs important for science teaching and learning. In the past two decades, much research has been done concerning the separate effects of spatial ability and of formal reasoning on science achievement, while very little work has been done to compare the relationships between spatial ability and formal reasoning. For example, it is possible that the considerable correlation between spatial ability and chemistry problem-solving skills is rooted in a "general cognitive factor" (Wu & Shah, 2004), which could be formal reasoning. Wu and Shah (2004) suggested that before spatial ability is studied as a prominent predictor of chemistry problem solving, the role of the general cognitive factor needs to be elucidated. Therefore, it will be interesting to investigate what the relation is between students' formal reasoning and their spatial ability.

In fields such as physics, engineering, and instructional technology, there are plenty of published research work showing that certain interventions were able to significantly improve spatial ability (Study, 2006). For example, Pallrand and Seeber showed that spatial intervention and taking introductory physics can improve college students' visual-spatial abilities (Pallrand & Seeber, 1984); Lord demonstrated that women have the capacity to improve their spatial abilities and often catch up to men's level by participating in meaningful visuospatial interventions (Lord, 1987); Kwon

illustrated that a web-based virtual reality (VR) graphics program was effective in improving the spatial visualization skills of ninth-grade students (Kwon, 2003); Piburn, Reynolds, and coworkers showed that "spatial ability can be improved through instruction, that learning of geological content will improve as a result, and that differences in performance between the genders can be eliminated" (Piburn, Reynolds, McAuliffe et al., 2005); Rafi and coworkers demonstrated that college students' spatial ability can be improved using a web-based virtual environment for 5 weeks (Rafi, Anuar, Samad et al., 2005), or through a five-week computer-mediated engineering drawing instruction (Rafi, Samsudin, & Ismail, 2006).

As mentioned earlier, in the field of science education, there are many published papers showing that spatial ability has significant correlations with students' science achievement (Carter et al., 1987; Ferk et al., 2003; Holland, 1995; Pribyl & Bodner, 1987; Yang et al., 2003). However, little or no work has been done in the field of science education on possible interventions or curriculum that increase students' spatial ability. Because of the critical linkage between spatial ability and science learning, and because the MOL course contains an large amount of spatial-visual content, e.g. molecules to cells, 2-D and 3-D DNA structure, genetic information, and protein architecture, it would be interesting to see whether or not students' spatial ability improves from encountering the visuospatial content and training in the MOL course. Therefore, it would be very useful to include a measure of students' spatial ability as well as the potential improvement of students' spatial ability into the assessment for the MOL course. Hence any spatial ability test we use would need to be given twice during the

semester of the MOL course: a pretest in the beginning of the semester, and a posttest at the end of the semester.

One of the most widely used instruments to measure college students' spatial ability is the Purdue Visualization of Rotations (**ROT**) test (Bodner & Guay, 1997; Carter et al., 1987; Pribyl & Bodner, 1987; Schoenfeld-Tacher, 2000; Study, 2006; Wu & Shah, 2004). The ROT test was initially developed at Purdue University and it measures students' ability in both spatial visualization and mental rotation. It contains 20 multiple-choice items, with 1 point for each item. Scores for the ROT test can range from 0 to 20. Each item requires mental operations on the mental representation a three-dimensional object being represented by two-dimensional drawings. It also contains "questions in which the object is rotated around more than one axis" (Bodner & Guay, 1997). Study by Bodner and Guay showed ROT had reasonable construct validity and that its reliability coefficients measured by internal consistency ranged from 0.78 to 0.85 for their samples of science/engineering, health science, and biology students with each group's sample size ranging from 127 to 1648 (Bodner & Guay, 1997), which were reasonable reliabilities according to the commonly-used criterion of 0.70 (Nunnally, 1978).

Based on the results from the three related studies (described in Chapters 2 through 4) and faculty input at the summer workshops, the finalized assessment plan for the MOL project included five instruments: 1) a student survey as listed in Appendix F, given at the beginning of the semester to collect students' demographics and their prior academic background, e.g. high school and college math, chemistry, biology courses taken; 2) the TOLT, given at the beginning of the semester, to measure students' formal reasoning ability at their entrance of the MOL course; 3) The ROT, given twice during

the semester: one before the enzyme module as a pretest of students' spatial ability, and the other after the enzyme module as a posttest of students' spatial ability; 4) a content knowledge pretest (hereby referred to as the *enzyme pretest)*, given before the enzyme module, to measure students' knowledge prior to the enzyme module; 5) a content knowledge posttest (hereby referred to as the *enzyme posttest*), given at the end of the semester as part of the final exam, to measure students' content knowledge after the enzyme module (Jordan & Lewis, 2008).

The finalized enzyme pretest has 20 multiple-choice questions and it focuses on six (6) learning goals: understand the concept of activation energy and reaction energetics; describe the overall pathway of a chemical reaction from reactants through the transition state to the final products; explain how a catalyst affects the rate of a chemical reaction; describe the role of enzymes as catalysts for biological processes (including the life cycle of HIV); explain the molecular principles of enzyme inhibition and apply them to HIV protease inhibitor drugs; describe the stages by which a new pharmaceutical is developed, tested, and approved. There are three to four questions for each learning goal.

The finalized enzyme posttest has 43 multiple-choice items as well as two open-ended questions, with four to five items on each of a set of learning goals. Eighteen of the multiple choice questions are from the pretest. The learning goals include all six goals specified in the pretest above, as well as additional goals such as: describe the overall pathway of a chemical reaction from reactants through transitional state to the final products; understand and visualize three-dimensional molecular structures; discuss substrate specificity and contrast the "lock and key" model with the "induced fit" theory of substrate binding; describe enzymatic function and discuss factors contributing to the

catalytic efficiency of enzymes; understand the molecular principles of enzyme inhibition; explain how specific drugs function on a molecular level (e.g. HIV protease inhibitors, Aspirin, COX-2 inhibitors); describe the strategies by which a new pharmaceutical is developed, tested, and approved. Faculty from all eight (8) participating schools developed these learning goals and served as expert panel for establishing content validity of the enzyme pretest and the enzyme posttest. Due to test score reliability concerns, only the multiple-choice items (i.e. the first 43 items) from the posttest was used in analyses. Since each multiple-choice item is worth 1 point, scores on the Enzyme Posttest can range from 0 to 43.

Research Questions

The research questions investigated in the assessment for the MOL project include:

1) Did MOL reach a diverse group of students?
2) Did students learn the enzyme content in the MOL course?
3) What is the relation between students' formal reasoning and their spatial ability?
4) Can the MOL course meaningfully improve students' spatial ability? Or, in other words, can it reduce the gap between high and low spatial ability students?

Student Demographics at Participating Schools

There were 905 students in the data set (including all schools and all semesters' data received as of October 21, 2008). Table 5.1 lists the number of students who participated in this study from each institution at each semester. Note that not every student took the Student Survey (Appendix F). For example, 16 students from NYU in Spring 2008 participated in the MOL study, but only 14 of them took the Student Survey. 60 students from UPR in Fall 2007 participated in the MOL study, but only 57 of them took the Student Survey. The detailed demographic information (e.g. sex and race) was only available for those students who took the survey, and students who did not take the survey are listed as "Unspecified" for both their sex and their race in Table 5.2.

Table 5.1 Number of Students at Each School Each Semester (Total n = 905)

Semester	NYU	UPR	Chaminade	Chicago	Fairfield	NCC	Spelman	Xavier
Fall05	80	58	0	0	0	38	0	0
Spring06	78	0	8	22	25	15	13	52
Fall06	80	56	0	0	0	0	0	0
Spring07	0	0	16	23	23	16	13	34
Fall07	79	60	0	0	0	0	14	0
Spring08	16	0	0	0	35	22	0	29
Total by School	333	174	24	45	83	91	40	115
%	37%	19%	3%	5%	9%	10%	4%	13%

Table 5.2 Demographics: Number of Students by Sex and Ethnicity

		NYU (n = 333)	UPR (n =174)	Other (n = 398)	All Schools Overall (total n = 905)
Sex	Male	123 (37%)	41 (24%)	118 (30%)	282 (31%)
	Female	200 (60%)	112 (64%)	189 (47%)	501 (55%)
	Unspecified	10 (3%)	21 (12%)	91 (23%)	122 (13%)
Race	American Indian/ Native Alaskan	0	4 (2%)	0	4 (0.4%)
	Native Hawaiian/ Pacific Islander	3 (1%)	3 (2%)	8 (2%)	14 (2%)
	Asian	84 (25%)	4 (2%)	12 (3%)	100 (11%)
	Black	15 (5%)	35 (20%)	160 (36%)	210 (23%)
	White	215 (65%)	103 (59%)	114 (29%)	432 (48%)
	Unspecified	16 (5%)	25 (14%)	104 (26%)	145 (16%)
Hispanic /Latino	Yes	29 (9%)	151 (87%)	24 (6%)	204 (23%)
	No	290 (87%)	1 (1%)	278 (70%)	569 (63%)
	Unspecified	14 (4%)	22 (13%)	96 (24%)	132 (15%)

Demographics of students from all 8 schools were listed in Table 5.2. NYU and UPR students are the two largest groups, with NYU making up 38% of the total sample and UPR making up 20% of the sample. Each of the other 6 schools makes up 10% or less of the total sample. On the whole, about 55% of all students were female, 31% of all students were male, while 14% of them did not specify their sex. 0.5% of students were American Indian/Native Alaskan, 2% of them were Native Hawaiian/other Pacific Islander, 11% were Asian, 21% were Black, 49% were White, while 16% of them did not specify their race. Also, about 23% of all students considered themselves Hispanic/Latino (Table 5.2). Because the MOL project was able to reach a significant number of female and minority students underrepresented in science (American Indian/Native Alaskan, Black, and Hispanic/Latino students) (Micari & Drane, 2007), it is a good step for our effort to move toward "science for all".

Outline of This Chapter

This chapter will focus on the assessment results from two schools – NYU and UPR-Rio Piedras. These two schools were selected for special attention because of their large classroom enrollment and also their pronounced educational differences; the other six schools either had a sample size that is too small (Table 5.1), or had incomplete data with so much missing data that their assessment results will be presented but not be a focus. NYU is a private institution with high selectivity and tuition costs: it's the only one among all eight (8) MOL participating schools to be ranked as one of the "Top 50, Tier 1 National Universities" by U.S. News & World Report (U.S. News & World Report, 2008); its Fall 2007 acceptance rate was only 36.7 %, and tuition and fees for 2008-2009

is $37,372 (U.S. News & World Report, 2008). Science courses for non-majors are taught in English as part of a college-based general education curriculum (the Morse Academic Plan). NYU students in this study took a semester-long *Molecules* course with one instructor that used the entire MOL curriculum. The typical class size was 80 students, which split into groups of 20 students for the laboratory sessions. For comparison, UPR-Rio Piedras is the largest of the eleven campuses within the university system of the Commonwealth of Puerto Rico. Non-majors science courses are taught in Spanish within the College of General Studies and are usually taken by students in their first year of undergraduate study. Two instructors modified their introductory biology class to enable integration of the *Enzymes and Drug Design* module from MOL. The curriculum materials for this module – draft chapters and laboratory exercises – were translated into Spanish for use by the UPR students. The enzyme module was placed as the third of five course units and was preceded by an introduction to biological investigation plus an overview of the chemical characteristics of living organisms. The course at UPR involved two 1.5-hour lecture periods along with one two-hour laboratory period weekly, and it devoted approximately three weeks, i.e. six lecture periods (1.5 hours each) in conjunction with three laboratory periods (two hours each) to teaching the enzyme module.

Descriptive Statistics of the Assessments

At NYU, the full MOL course was implemented for five semesters, including Fall 2005, Spring 2006, Fall 2006, Fall 2007, and Spring 2008 (Table 5.1). For the first four semesters, there were about 79 students enrolled at each semester and these four

semesters' MOL course was taught in a large lecture hall by the same instructor. For the fifth semester (Spring 2008), however, there were only 16 students enrolled in the MOL course and a different instructor taught it. Owing to the small class size, the course was taught in a very different format in Spring 2008 that is more similar to a seminar/discussion. To control the extraneous confounding variables of instructor, class size, and class format, we only included the first four semesters' data in our analyses for NYU, as the fifth semester had a different instructor, different class format, and a much smaller class size.

Tables 5.3 and 5.4 list the descriptive statistics (mean, standard deviation, minimum, maximum scores, skewness, and kurtosis) and the reliability for the assessments for NYU and UPR, respectively. Most assessments had a reasonable reliability with a raw Cronbach's alpha close to or above 0.70, the widely accepted cutoff value for adequate reliability for research purposes in social sciences (Nunnally, 1978). As discussed above, the pre/post content assessment test was re-designed during the project to improve validity and reliability. The revised test was incorporated into courses taught later in the project and was therefore used by a smaller number of students in comparison to the TOLT and ROT.

Table 5.3 Descriptive Statistics and Reliability of the Assessments at NYU

Test	n	M	SD	Minimum	Maximum	Skewness	Kurtosis	Reliability (Cronbach's alpha)
TOLT	304	8.28	1.89	1	10	-1.28	1.11	0.675
ROT Pretest	313	12.87	3.58	4	20	-0.13	-0.66	0.734
ROT Posttest	273	14.03	3.53	2	20	-0.53	0.17	0.735
Enzyme Pretest*	74	16.11	2.96	2	20	-2.04	6.43	0.713
Enzyme Posttest	75	37.75	3.25	25	43	-1.55	4.73	0.665

*Only the Fall 2007 students took the finalized version of the enzyme posttest with all 43 questions, therefore only the Fall 07 NYU sample was included in calculating the descriptive statistics & reliabilities of the enzyme pre- and posttest.

Table 5.4 Descriptive Statistics and Reliability of the Assessments at UPR

Test	n	M	SD	Minimum	Maximum	Skewness	Kurtosis	Reliability (Cronbach's alpha)
TOLT	153	2.60	2.26	0	10	1.12	0.80	0.721
ROT Pretest	136	10.46	4.12	2	19	0.08	-0.58	0.763
ROT Posttest	134	11.00	4.57	0	20	-0.23	-0.51	0.824
Enzyme Pretest*	91	10.30	3.14	3	17	0.01	-0.56	0.588
Enzyme Posttest	92	21.51	5.63	6	35	0.08	-0.17	0.724

*Only the Fall 2006 & Fall 2007 students took the finalized version of the enzyme pre- and posttest. Therefore only the Fall 06 & Fall 07 UPR sample was included in calculating the descriptive statistics & reliabilities of the enzyme pre- and posttest.

A comparison of the descriptive statistics in Tables 5.3 and 5.4 reveals certain key differences between the two student populations in our study. The average TOLT score for NYU students was much higher than for the UPR students. This result can be understood in the context of the student survey results, which showed that NYU students generally had a more extensive preparation in mathematics. (The detailed results from the student survey, including demographics and academic background for NYU and UPR students, will be presented later in Table 5.13.) In addition, a much higher percentage of the UPR students were in their first year of university study. The results from the Purdue ROT showed a slightly higher average pretest score for the NYU students, with both groups demonstrating a small average gain in the posttest scores. The pretest and posttest scores for the content assessment showed that NYU students had a more extensive knowledge of the topics in the *Enzymes and Drug Design* module. This result is not surprising since the NYU students were taking a semester-long course in *Molecules of Life*, with a strong curriculum focus on biological molecules, whereas the UPR students were studying a single module embedded within an introductory biology course. These variations in student backgrounds, experiences, and abilities provide an interesting

context in which to examine whether the MOL curriculum can be effective in different educational environments.

MOL Assessment at NYU

Assessment Results for NYU students

As mentioned earlier, there are eighteen (18) questions shared by the enzyme pretest and posttest. Therefore, students' performance on these eighteen anchor items provides a measure on whether students learned the enzyme concepts. It was found that on average, students' total score on the eighteen anchor items increased from 14.4 to 17.23, with an average gain of 2.84, which was large and statistically significant (Table 5.7). This improvement provided evidence that the students indeed learned the enzyme content in the MOL course. Also, analysis of item difficulty and item-total correlations was performed. *Item difficulty* was defined as the proportion of students who answered that item correctly. An item difficulty of greater than 0.9 means an item is too "easy" for students in the sample. Table 5.5 lists the item difficulty and item-total correlation for the Enzyme Pretest items for NYU students. Some Enzyme Pretest items, including items 1, 5, 8, 10, and 11, had item difficulty of greater than 0.9. These items were too "easy" for NYU students even at Pretest. The most difficult item at Pretest was item 12, as only 42.5% students got it correct. But at Posttest, 82.4% of students got this same item correct (Table 5.5). The same trend held true for all eighteen anchor items for both NYU and UPR, namely, a much higher percentage of students got each anchor item correct at Posttest than at Pretest, providing another perspective of evidence that students learned the enzyme content during the semester.

Table 5.5 Enzyme Pretest and 18 Anchor Items: Difficulty for NYU students

Item # in Pretest	Item Difficulty* in Pretest	SD in Pretest	Item-Total Correlation in Pretest	Item # in Posttest	Item Difficulty in Posttest
enzyPreQ1	**0.932****	0.253	0.112	enzyPostQ1	**1.00**
enzyPreQ2	0.676	0.471	0.183.	enzyPostQ16	**0.933**
enzyPreQ3	0.662	0.476	0.214	enzyPostQ3	**0.920**
enzyPreQ4	0.824	0.383	0.201	enzyPostQ15	**0.946**
enzyPreQ5	**0.946**	0.228	0.642	enzyPostQ17	**0.987**
enzyPreQ6	0.851	0.358	0.108	enzyPostQ18	**0.973**
enzyPreQ7	0.770	0.424	0.415	enzyPostQ2	**0.920**
enzyPreQ8	**0.919**	0.275	0.267	enzyPostQ13	**1.00**
enzyPreQ9	0.797	0.405	0.482	enzyPostQ14	**0.987**
enzyPreQ10	**0.919**	0.275	0.468	--	--
enzyPreQ11	**0.959**	0.199	0.424	enzyPostQ19	**1.00**
enzyPreQ12	0.425	0.498	**-0.076**	enzyPostQ5	0.824
enzyPreQ13	0.851	0.358	0.272	--	--
enzyPreQ14	0.822	0.385	0.488	enzyPostQ4	**0.973**
enzyPreQ15	0.568	0.499	0.282	enzyPostQ9	**0.947**
enzyPreQ16	0.878	0.329	0.270	enzyPostQ8	**0.987**
enzyPreQ17	0.797	0.405	0.395	enzyPostQ33	**1.00**
enzyPreQ18	0.892	0.313	0.392	enzyPostQ42	**0.947**
enzyPreQ19	0.770	0.424	0.161	enzyPostQ35	**0.973**
enzyPreQ20	0.877	0.331	0.446	enzyPostQ41	**0.973**

*Item difficulty also equals the proportion of students who answered that item correctly; n = 74 for the first 19 items, while n = 73 for the last item (enzyPreQ20).
**Items difficulties in bold were greater than .9 and denotes items being too "easy" at that time

For the Enzyme Posttest, the item difficulty, standard deviations, and item-total correlations for NYU students are listed in Table 5.6. 27 out of 43 items on the enzyme posttest had item difficulty of above 0.9, indicating that these times were too "easy" for NYU students. Items 1, 13, 19, and 33 had item difficulty of 1.00 with zero variability as all NYU students answered them correctly. Items 4, 7, 11, 14, 28, 30, 35, 38, and 43 had negative item-total correlations, although these correlations were all small and close to zero in effect size, which did not have any large negative impact on the overall reliability

of the test. Similar results were found for UPR students, which are presented in the next sections. The enzyme pre and posttests are designed as *criterion-referenced test (Popham, 2000, p. 30-34)*, i.e. test results are to be interpreted *absolutely* with a clearly-defined assessment domain (in our case, the enzyme learning goals). Therefore, limited variability and item-total correlations, and low reliability are normal and expected. That the items matched instructional objectives and students did so well in the posttest further illustrated that students learned the enzyme content. (The other type of test is *norm-referenced test*, in which, test results are to be interpreted *relatively* on how each student's performance compares to the performance of other students or norm groups, but no absolute information can be obtained about what a student can or can't do in a defined assessment domain.)

Table 5.6 Enzyme Posttest at NYU: Item Difficulty and Item-Total Correlations

Item	n	Item Difficulty*	SD	Item-Total Correlation
enzyPostQ1	75	**1.00****	0.00	N/A***
enzyPostQ2	75	**0.920**	0.273	.46
enzyPostQ3	75	**0.920**	0.273	.16
enzyPostQ4	75	**0.973**	0.162	-.03
enzyPostQ5	74	0.824	0.383	.28
enzyPostQ6	68	**0.985**	0.121	.004
enzyPostQ7	75	0.840	0.369	-.01
enzyPostQ8	75	**0.987**	0.115	.03
enzyPostQ9	75	**0.947**	0.226	.59
enzyPostQ10	75	0.507	0.503	.19
enzyPostQ11	75	**0.987**	0.115	-.002
enzyPostQ12	74	0.784	0.414	.13
enzyPostQ13	75	1.00	0.00	N/A
enzyPostQ14	75	**0.987**	0.115	-.001
enzyPostQ15	74	**0.946**	0.228	.23
enzyPostQ16	75	**0.933**	0.251	.57
enzyPostQ17	75	**0.987**	0.115	.11
enzyPostQ18	75	**0.973**	0.162	.02
enzyPostQ19	75	**1.00**	0.00	N/A
enzyPostQ20	75	0.880	0.327	.14
enzyPostQ21	75	**0.987**	0.115	.45
enzyPostQ22	75	**0.933**	0.251	.35
enzyPostQ23	75	0.867	0.342	.32
enzyPostQ24	75	0.840	0.369	.05
enzyPostQ25	75	0.787	0.412	.27
enzyPostQ26	74	0.878	0.329	.14
enzyPostQ27	75	0.867	0.342	.41
enzyPostQ28	75	**0.960**	0.197	-.06
enzyPostQ29	74	0.689	0.466	.34
enzyPostQ30	74	0.527	0.503	-.02
enzyPostQ31	74	0.459	0.502	.16
enzyPostQ32	75	0.413	0.496	.15
enzyPostQ33	75	**1.00**	0.00	N/A
enzyPostQ34	75	**0.800**	**0.403**	**.35**
enzyPostQ35	75	**0.973**	0.162	-.05
enzyPostQ36	75	**0.907**	0.293	.37
enzyPostQ37	74	**0.973**	0.163	.38
enzyPostQ38	75	**0.987**	0.115	-.04
enzyPostQ39	75	**0.960**	0.197	.53
enzyPostQ40	75	**0.920**	0.273	.14
enzyPostQ41	74	**0.973**	0.163	.29
enzyPostQ42	75	**0.947**	0.226	.10
enzyPostQ43	75	**0.907**	0.293	-.08

* Item difficulty also equals the proportion of students who answered that item correctly.

** Items difficulties in bold are greater than .9, indicating those items were too "easy" for the sample of students. *** N/A: Not Applicable, items 1, 13, 19, and 33 had zero variability as all students answered them correctly, thus correlation with total score couldn't be computed

Table 5.8 lists the mean, standard deviation and t value and p value for the ROT_Gain score for each semester as well as for the aggregate data, respectively. *ROT_Gain* score is defined as the ROT Posttest score subtracting the ROT Pretest score. For the aggregate data (all semesters overall), the mean ROT Gain from pretest to posttest was 1.199, and it was significantly different from 0 ($p < .001$). However, the effect size of this gain measured by Cohen's d was small (0.33, i.e. on average, students' ROT Posttest score was about 0.33 standard deviations higher than the ROT Pretest score), indicating that in general, there was only a slight improvement in spatial ability from pretest to posttest. (According to Cohen's rule of thumb for t-tests, an effect size below 0.5 is small, an effect size between 0.5 and 0.8 is medium, and an effect size of 0.8 or above is large for t-test comparisons.)

Table 5.9 lists the correlations between TOLT, ROT Pretest, and ROT_Gain scores for each semester and for the aggregate data, respectively. For the aggregate data, the correlation between TOLT and ROT Pretest score was 0.18 and it was statistically significant. According to Cohen, correlations smaller than .3 would be classified as a small effect size, correlations between .3 and .5 would be classified as a medium effect size, and correlations of .5 or higher would be classified as large effect size (Cohen, 1988). So, the correlation between TOLT and ROT Pretest (0.18) was small. The smallness of this correlation was apparent in the scatterplot (Figure 5.1). On one hand, this result seems to provide some support for Gardner's "multiple intelligences" hypothesis to some extent, in which Gardner considers spatial ability and logical reasoning as separate and unrelated skills. On the other hand, this result may be simply an objet d'art of "ceiling effect" in which the lack of variability in these NYU students'

TOLT scores attenuated the correlation between these students' TOLT and ROT Pretest scores, as there was a "ceiling" of high TOLT scores for the NYU students clearly visible in Figure 5.1.

Table 5.7 NYU Enzyme Content Assessment

n	Pretest mean (SD)	Posttest mean (SD)	Anchor_Gain (SD)	Effect size	t -value (p value)
73	14.40 (2.681)	17.23 (1.208)	2.84 (2.625)	1.36	t = 9.23 (p < .0001)

Table 5.8 ROT Gain Score for NYU Students

Semester	n	M	SD	t value	p value	Effect size (Cohen's d)
Fall 2005	67	1.761	2.481	5.81	<0.001	0.52
Spring 2006	67	1.134	2.735	3.39	0.0012	0.32
Fall 2006	69	1.014	3.265	2.58	0.0120	0.27
Fall 2007	68	0.897	3.12	2.37	0.0206	0.23
Aggregate (all semesters)	271	1.199	2.923	6.75	<0.001	0.33

Table 5.9 Correlations between TOLT, ROT Pretest and ROT_Gain for NYU

Semester	Correlation between TOLT and ROT Pretest (n, p value)	Correlation between TOLT and ROT_Gain	Correlation between ROT Pretest and ROT_Gain
Fall 2005	0.43** (n = 73, p < .01)	-0.004 (n = 64, p > .05)	-0.13 (n = 67, p > .05)
Spring 2006	0.14 (n = 76, p > .05)	0.08 (n = 66, p > .05)	-0.55** (n = 67, p < .01)
Fall 2006	0.16 (n = 77, p > .05)	-0.03 (n = 68, p > .05)	-0.54** (n = 69, p < .01)
Fall 2007	0.11 (n = 74, p > .05)	0.16 (n = 66 , p > .05)	-0.41** (n = 68, p < .01)
Aggregate (all semesters)	0.18** (n = 300, p < .01)	0.07 (n = 264, p > .05)	-0.44** (n = 271, p < .01)

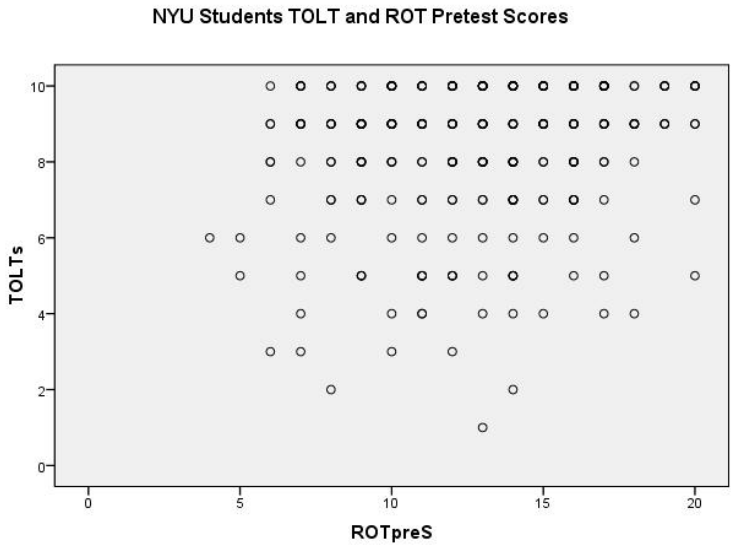
*Pair-wise exclusion of missing values were used for all correlations. **significant at .05 level.

Table 5.10 Low vs. High Spatial Ability Group in ROT_Gain for NYU

Spatial ability group	ROT pretest score (n, SD)	ROT posttest score (n, SD)	ROT _Gain (n, SD)	t -test comparing ROT_Gain to 0 (n, p value)	t-test comparing two groups in their ROT_Gain (Cohen's d, p value)
Low	10.13 (167, 2.256)	12.20 (147, 3.255)	2.14 * (147, 2.848)	t = 9.12 (n = 147, p < .01)	t = 6.17 (d = 0.75, p < .01)
High	16.00 (146, 1.789)	16.21 (124, 2.496)	0.08 (124, 2.609)	t = 0.34 (n = 124, p > .05)	

* ROT_Gain significantly higher than 0

Figure 5.1 Scatterplot of NYU Students' TOLT and ROT Pretest Scores



Another interesting result was that the correlation between ROT Pretest and ROT_Gain score was negative (-0.44, see Table 5.9) and this correlation was statistically significant with medium-to-large effect size. This suggests that students who did poorly on the ROT Pretest tended to have a larger ROT_Gain score, which makes sense, since these students would have a larger room for improvement in their spatial ability than those other students who had good spatial ability to begin with.

To further look into this difference in ROT_Gain between low ability students and high ability students, the sample was divided into two groups: students with a ROT Pretest score above the median score of 13.00 were classified as "High" spatial ability group, while those with a ROT Pretest score at or below the median score were classified as "Low" spatial ability group. The average ROT_Gain for the low spatial ability group was 2.14, a large and statistically significant improvement in spatial ability, while the average ROT_Gain for the high spatial ability group was only 0.08, which was small and

not statistically significant (t =0.34, p>.05, Table 5.10). An independent samples t-test found that there was a statistically significant difference between the two groups in their ROT_Gains (t = 6.17, p < .0001). The effect size of this t-test comparison was d = 0.75 (Table 5.10), indicating a medium-to-large difference between the low group and the high group in their average ROT_Gain. The low group had a large and statistically significant improvement in their spatial ability from ROT pretest to posttest, while the high group's ROT score essentially did not improve from pretest to posttest. The average gap between the low group and high group at the ROT pretest was 5.87 (10.13 for the low group vs. 16.00 for the high group, Table 4), but the gap dropped considerably to 4.01 at the posttest (12.20 for the low group vs. 16.21 for the high group, see Table 4). This result suggests that the Molecules of Life course at NYU was successful in improving the spatial ability of students who began with low spatial skills, and to some degree, the MOL course was able to reduce the gap in spatial ability between low spatial ability students and high spatial ability students.

Missing Data Analysis

There were 317 students in total in our NYU sample, 4% of them (13 students) missed TOLT, 1% (4 students) missed ROT Pretest, and 14% (44 students) missed ROT Posttest. Since less than 5% of the sample missed TOLT or ROT Pretest, effect of those missing values is ignorable. For the missing ROT Posttest scores, students who took ROT Posttest were compared to those who did not take ROT Posttest. Little difference was found between these two groups in their TOLT and ROT Pretest scores (Table 5.11). An equivalence test recommended in (Lewis & Lewis, 2005b) was also performed. Due to small sample size, the equivalence test failed to rule out non-equivalence. But since the

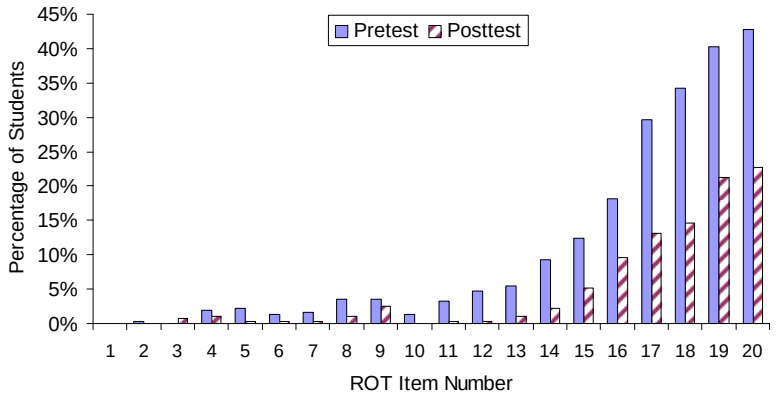
effect size for the differences between these two groups were so small (0.04 and 0.07 for TOLT and ROT Pretest, respectively), it was reasonable to say there was little or no difference between students who took the ROT Posttest and those who did not. Thus, had the missing ROT Posttest scores not been missing, they would not affect our results.

Table 5.11 NYU Students Who Took ROT Posttest vs. Those Who Did Not

Took ROT Posttest?	TOLT (n, SD)	ROT Pretest (n, SD)
Yes (n=273)	8.29 (266, 1.91)	12.83 (271, 3.66)
No (n=44)	8.21 (38, 1.82)	13.09 (42, 3.02)
Effect size (Cohen's d)	0.04	0.07
t value & p value	0.24 (p>.1)	-0.44 (p>.1)

Also, the "missingness" of test scores (TOLT, ROT Pretest, and ROT Posttest) appeared to be random, i.e. missing one test was not related to missing of another test, and missing one test was not related to scores on other tests, as all corresponding correlations were small.

Figure 5.2 Percentage of NYU Students Missing Each ROT item



At the item level, for each TOLT item, less than 1% of the students who took the TOLT test missed it. However, this is not the case for ROT pre and posttest items. Figure 5.2 shows the percentage of test-takers who missed (i.e. did not answer) each ROT item in the pre and posttest. Note the pretest percentages are out of the 313 students who took

the ROT pretest, and the posttest percentages are out of the 273 students who took the ROT posttest. There were two apparent trends in Figure 5.2. First, there were few students missing the beginning items but many more students missing items toward the end of the test, e.g. at pretest there were less than 5% of students missing items 1 to 12, respectively, while more than 25% students missing items 17 to 20 each. This pattern is not surprising as the ROT test was a 20-item-10-minute test and we expect students to have problems finishing all 20 items in 10 minutes. Secondly, when pretest and posttest are compared, there were many more students missing items in the pretest than the same items in the posttest, e.g. there were 12.5% students missing item 15 in the pretest, but only 5.1% missing the same item in the posttest. One possible explanation could be students did not take the Pretest seriously, but our reliability analysis showed that both pretest and posttest were quite reliable (Table 5.1), suggesting most students did take the tests seriously. Another explanation could be the *test-retest* effect, i.e. students might have remembered the ROT items from the first time (pretest), so they were able to answer more items correctly at the second time (posttest). But the literature suggests when there's a gap of five (5) weeks or more between pretest and posttest, any improvement is not attributed to test-retest effect (Rafi et al., 2005; Rafi et al., 2006). Because the ROT pretest was given near the beginning of the semester, while the posttest was given near the end of the semester, there were at least seven (7) weeks between pre and posttest, making the test-retest effect highly unlikely. Additionally, there were less than 5% of students missing each of the first 12 ROT items even at pretest. If students' scores on these 12 items improve significantly from pretest to posttest, then the observed performance gain on the ROT is not simply due to students answering more items at

posttest, but mostly because of a real improvement in spatial ability. When we inspect the low-spatial-ability group's score on the first 12 items, we found their score on these 12 items improved significantly at posttest (Table 5.12). Because of the gap of at least seven weeks between pre and posttest that nullifies the test-retest effect, as well as the significant improvement of low-spatial-ability group's score on the first 12 items, the low-spatial-ability group's performance gain on the ROT overall (fewer missing items and higher scores) is most likely an indication of these students' improvement in spatial ability, not the test-retest effect.

Table 5.12 NYU Low-Ability Students' Score on the First 12 ROT Items

Score at Pretest (n, SD)	Score at Posttest (n, SD)	Gain (n, SD)	t -test comparing gain to zero (n, p value)	Effect size (Cohen's d)
8.088 (113, 1.976)	8.690 (113, 2,342)	0.602* (113, 2.262)	t = 2.83 (n = 113, p < .01)	0.28

* Gain significantly higher than 0

Validity of Measured Gains in Spatial Ability

Molecules of Life was taught at NYU as a semester-long course by the same instructor in the four semesters (Fall 05, Spring 06, Fall 06, and Fall 07). Numbers of students at different semesters in the course were about the same. Therefore the confounding variables of school, class size, and instructor effect were controlled.

Another threat to the validity of the gain in spatial ability is the potential regression effect (also known as *regression to the mean* in the literature), namely, scores very different from the population mean on a initial measurement will tend to be closer to the mean on a subsequent measurement (Linden, 2007; Weeks, 2007). In our case, it was possible the large spatial-ability gain observed for the low ability group might be simply due to the regression effect at the posttest. There are two contractions to this threat. First,

no substantial regression effect was found in the literature concerning spatial ability improvements. Even for studies using non-random assignment of control and treatment groups (Piburn et al., 2005; Study, 2006), no sizeable regression effect was reported on spatial ability gains. Secondly, we used the following formula recommended in Weeks (2007) to estimate the expected posttest score based on regression effect: $r_{xy}(X-\mu)+\mu$, where X represents the pretest score, μ is the population mean estimate, r_{xy} is the pretest-posttest correlation. Based on our entire sample, μ=12.87, r_{xy}=.671. So, for the low-ability group with mean pretest score of 10.13 (Table 5.10, the expected posttest score would be .671*(10.13-12.87)+12.87=11.03. The spatial-ability gain from the regression effect would be 11.03-10.13=0.9. Since our observed average gain for the low-ability group was 2.14, considerably greater than 0.9, our gain would be most likely due to the MOL course, not an objet d'art of the regression effect. Also, for the high-ability group, if the MOL course had no effect on students' spatial ability and the regression effect is the only source of potential spatial-ability gains, the expected posttest score for them would be .671*(16.00-12.87)+12.87=14.97. But our observed posttest score for the high-ability group was 16.21 (Table 5.10, much higher than 14.97. These results suggest the regression effect was not a major factor for either high- or low-ability group, and the observed spatial-ability gain was most likely the effect of the MOL course.

What Contributed to the Improvement of Spatial Ability?

In a faculty survey developed with Dr. Trace Jordan from NYU (see Appendix J for the full survey instrument), we asked the faculty participants at all participating institutions about the specifics of the MOL course implemented at each institution,

especially about the contents and/or activities that the instructors believe contributed to the potential improvement of students' spatial ability.

Besides taking part in the development of the faculty survey, Dr. Jordan is also the instructor who taught the MOL course for the four semesters we discussed earlier, and he believes that both the course content and class/lab activities contributed to the meaningful improvement of students' spatial ability at NYU. An important part of the MOL course was the enzyme module placed at the end of the course. The enzyme module has three (3) chapters, and they contain several topics that involve 3-D visualization of molecular structures: Chapter 1 (Reactions & Catalysts) compares different molecular structures of the reactants, transition state, and products in a chemical reaction, and students study the example of a substitution reaction, which requires the visualization of 3-D structures; in Chapter 2 (Enzymes as Biological Catalysts), understanding the function of the HIV protease enzyme involves the visualization of a tetrahedral transition state, the lock-and-key model is used to show how the enzyme active site is complementary to the 3-D structure of the substrate, and the positioning of amino acid sidechains in the active site is illustrated to allow for the discrimination between chiral isomers of the same compound; Chapter 3 (Enzymes & Drug Design) demonstrates how HIV protease inhibitor drugs achieve their effect by mimicking the 3-D geometry of the transition state for the enzyme-catalyzed reaction, and it requires carefully studying the structure of a complex molecular and identifying the tetrahedral region. Also, well before the enzyme module, the instructor spend a lot of time on 3-D structures in the early parts of the course, so students already have a strong foundation by the time the enzyme module begins. Some relevant topics that the instructor teaches

before the module include: predicting the 3-D structures of simple molecules from electron pair repulsion; molecular conformation of alkanes (e.g., ethane) and alkenes (e.g. ethene); structures of molecules with functional groups (e.g., amines); chirality of amino acids (L- and D- structures).

In addition to these contents that involve spatial ability, the MOL course at NYU used a variety of class and lab activities that involved visual-spatial thinking. These activities include:

- Drawing molecular structures – students were given two in-class drawing exercises. The first was for simple molecules (e.g., NH_3) and the second was for the 3-D structure of a hydrocarbon (propane). Students also drew molecular structures in two laboratory sessions (see below).
- Building models using model kits – students had two laboratory exercises (1 hr 30 min) where they build molecular models using kits.
- Using computer graphics software – the model-building lab projects also included the use of the CHIME software for molecular visualization. CHIME allows students to rotate the molecule on the screen, zoom in for a closer look, and show different representations (e.g., ball-and-stick, spacefill, etc.) A typical laboratory exercise would be the following: (a) students study the 3-D structure of a molecular using the interactive CHIME software; (b) they use the kit to build a model of the molecule; (c) they draw the 3-D structure of the model in their lab manuals using the standard chemical methods for spatial representation.

- Other types of activity – the instructor also used large-sized demonstration models of molecular structures during his lecture presentations of these topics.

From observing students work in the lecture and labs, the instructor believes that a combination of activities is most likely to lead to improved spatial ability (computer graphics, model kits, and drawings). Like any skill, 3-D visualization and representation of molecular structure is best achieved through regular practice (Williamson & José, 2008). Of the methods used, the instructor suggests that having students practice their drawing of molecular structures in the lecture and/or lab is a key aspect of 3-D training.

MOL Assessment at UPR

Difference between NYU and UPR Students

As mentioned earlier, the classes and student population at NYU and UPR differ significantly in terms of course content, demographics, and language of instruction. A summary of the student demographics and academic background for the two schools is summarized in Table 5.13. In comparing NYU with UPR students in our sample, a much higher proportion of UPR students are freshmen (1st year in college), a much lower proportion of UPR students took 5 or more semesters of high school math, a much higher proportion of UPR students took no (i.e. zero semester of) college-level science course, and a much higher proportion of UPR students took no college-level math course (see Table 5.13). In other words, the UPR students tend to have a lower level of science and mathematics background in general.

Figure 5.3 Scatterplot of UPR Students' TOLT and ROT Pretest Scores

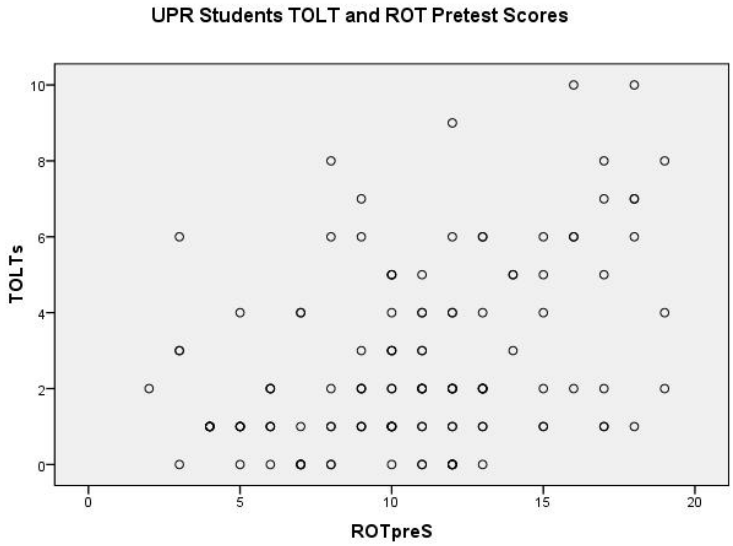


Table 5.13 Demographics and Academic Background of NYU and UPR Students

		NYU (n = 333)	UPR (n =174)
Sex	Male	123 (37%)	41 (24%)
	Female	200 (60%)	112 (64%)
	Unspecified	10 (3%)	21 (12%)
Race	American Indian/ Native Alaskan	0	4 (2%)
	Native Hawaiian/ other Pacific Islander	3 (1%)	3 (2%)
	Asian	84 (25%)	4 (2%)
	Black	15 (5%)	35 (20%)
	White	215 (65%)	103 (59%)
	Unspecified	16 (5%)	25 (14%)
Hispanic /Latino	Yes	29 (9%)	151 (87%)
	No	290 (87%)	1 (1%)
	Unspecified	14 (4%)	22 (13%)
Years in College	1	66 (21%)	147 (96%)
	2	134 (43%)	3 (2%)
	3	79 (25%)	1 (.7%)
	4	32 (10%)	1 (.7%)
	5 or more	1 (.3%)	1 (.7%)
Semesters of High School Chemistry	0	8 (3%)	11 (7%)
	1	28 (9%)	6 (4%)
	2	221 (71%)	130 (85%)
	3 to 4	48 (15%)	5 (3%)
	5 or more	7 (2%)	1 (.7%)
Semesters of High School Biology	0	12 (4%)	0
	1	19 (6%)	4 (3%)
	2	227 (73%)	131 (86%)
	3 to 4	49 (16%)	18 (12%)
	5 or more	5 (2%)	0
Semesters of High School Math	0	1 (.3%)	29 (19%)
	1	1 (.3%)	9 (6%)
	2	0	11 (7%)
	3 to 4	6 (2%)	13 (9%)
	5 or more	304 (97%)	91 (59%)
Semesters of College Level Science Courses Taken	0	80 (26%)	101 (66%)
	1	191 (61%)	19 (12%)
	2	28 (9%)	26 (17%)
	3 to 4	11 (4%)	3 (2%)
	5 or more	2 (.6%)	4 (3%)
Semesters of College Level Math Courses Taken	0	121 (39%)	98 (64%)
	1	107 (34%)	51 (33%)
	2	52 (17%)	3 (2%)
	3 to 4	22 (7%)	1 (.7%)
	5 or more	10 (3%)	0 (0%)

Assessment Results for UPR students

Table 5.14 lists the item difficulty and item-total correlations for the enzyme pretest for UPR students. For most of the eighteen (18) anchor items shared by the enzyme pretest and posttest, it was noticeable that a higher proportion of students answered these anchor items correctly at posttest than at pretest. This result was similar to the NYU result elucidated earlier, imparting evidence that students learned the enzyme content during the semester.

Table 5.14 Enzyme Pretest and 18 Anchor Items: Difficulty for UPR students

Item Number in Pretest	n	Item Difficulty* in Pretest	SD in Pretest	Item-Total Correlation in Pretest	Item # in Posttest	Item Difficulty in Posttest
enzyPreQ1	91	.747	0.437	.129	enzyPostQ1	**.902**
enzyPreQ2	90	.367	0.485	.228	enzyPostQ16	.511
enzyPreQ3	89	.169	0.376	.015	enzyPostQ3	.391
enzyPreQ4	90	.711	0.456	.142	enzyPostQ15	.793
enzyPreQ5	91	.802	0.401	.262	enzyPostQ17	**.912**
enzyPreQ6	91	.725	0.449	.148	enzyPostQ18	.663
enzyPreQ7	91	.516	0.502	.389	enzyPostQ2	.663
enzyPreQ8	90	.278	0.450	.350	enzyPostQ13	.714
enzyPreQ9	89	.371	0.486	.367	enzyPostQ14	.703
enzyPreQ10	88	.523	0.502	.219	--	--
enzyPreQ11	91	.780	0.416	.198	enzyPostQ19	**.912**
enzyPreQ12	90	.322	0.470	.168	enzyPostQ5	.286
enzyPreQ13	90	.567	0.498	.069	--	--
enzyPreQ14	89	.663	0.475	.227	enzyPostQ4	.783
enzyPreQ15	90	.356	0.481	.259	enzyPostQ9	.739
enzyPreQ16	90	.522	0.502	.141	enzyPostQ8	.560
enzyPreQ17	89	.483	0.503	.100	enzyPostQ33	.693
enzyPreQ18	87	.854	0.355	.259	enzyPostQ42	.851
enzyPreQ19	87	.460	0.501	.218	enzyPostQ35	.591
enzyPreQ20	88	.216	0.414	.025	enzyPostQ41	.276

*Item difficulty also equals the proportion of students who answered that item correctly.
**Items difficulties in bold were greater than .9 and denotes items being too "easy" at that time.

Table 5.15 Enzyme Posttest at UPR: Item Difficulty and Item-Total Correlations

Item	n	Item Difficulty*	SD	Item-Total Correlation
enzyPostQ1	92	**.902****	0.299	.05
enzyPostQ2	92	.663	0.475	.15
enzyPostQ3	92	.391	0.491	.19
enzyPostQ4	92	.783	0.415	.10
enzyPostQ5	91	.286	0.454	.07
enzyPostQ6	92	.489	0.503	.38
enzyPostQ7	92	.196	0.399	.15
enzyPostQ8	91	.560	0.499	.38
enzyPostQ9	92	.739	0.442	.27
enzyPostQ10	92	.380	0.488	.27
enzyPostQ11	91	.791	0.409	.19
enzyPostQ12	90	.422	0.497	.12
enzyPostQ13	91	.714	0.454	.29
enzyPostQ14	91	.703	0.459	.23
enzyPostQ15	92	.793	0.407	-.03
enzyPostQ16	90	.511	0.503	.40
enzyPostQ17	91	**.912**	0.285	.31
enzyPostQ18	92	.663	0.475	.32
enzyPostQ19	91	**.912**	0.285	.20
enzyPostQ20	89	.382	0.489	.29
enzyPostQ21	91	.659	0.477	.24
enzyPostQ22	92	.641	0.482	.23
enzyPostQ23	89	.169	0.376	.23
enzyPostQ24	89	.213	0.412	-.001
enzyPostQ25	91	.220	0.416	.08
enzyPostQ26	91	.637	0.483	.18
enzyPostQ27	91	.549	0.500	.15
enzyPostQ28	89	.573	0.497	.22
enzyPostQ29	90	.189	0.394	.28
enzyPostQ30	91	.209	0.409	-.22
enzyPostQ31	89	.191	0.395	-.003
enzyPostQ32	89	.337	0.475	.14
enzyPostQ33	88	.693	0.464	.20
enzyPostQ34	87	.172	0.380	.19
enzyPostQ35	88	.591	0.494	.26
enzyPostQ36	86	.407	0.494	.35
enzyPostQ37	85	.306	0.464	.12
enzyPostQ38	86	.488	0.503	.36
enzyPostQ39	86	.523	0.502	.32
enzyPostQ40	86	.453	0.501	.27
enzyPostQ41	87	.276	0.450	.35
enzyPostQ42	87	.851	0.359	.16
enzyPostQ43	80	.463	0.502	.48

*Item difficulty equals the proportion of students who answered that item correctly. **Item difficulties in bold are greater than .9, indicating they were too "easy" for the sample of students.

For the enzyme posttest, the item difficulty, standard deviations, and item-total correlations for UPR students are listed in Table 5.15. Similar to NYU results explicate explicated earlier, some items (Items 15, 24, 30 and 31) exhibited slightly negative item-total-correlations, although these correlations were all small and close to zero in effect size, which did not have any large negative impact on the overall reliability of the test. Again, since the enzyme posttest is a criterion-referenced test, limited variability and item-total correlations, and low reliability are normal and expected. That the items matched instructional objectives and students did well in most items in the enzyme posttest further illustrated that students learned the enzyme content.

Moreover, UPR students' total scores on the eighteen anchor items (shared by the enzyme pretest and posttest) increased from 9.3 to 11.84 on average, with a mean gain score of 2.55, which was large and statistically significant (Table 5.16). This improvement was similar to the results for NYU students and provided yet another solid evidence that the students indeed learned the enzyme content in the MOL course.

Table 5.17 lists the mean, standard deviation and t value and p value for the ROT_Gain score for each semester as well as for the aggregate data, respectively. For the aggregate data (all semesters overall), the mean ROT Gain from pretest to posttest was 0.948, and it was significantly different from 0 ($p < .001$). However, the effect size of this gain measured by Cohen's d was small (0.22, i.e. on average, students' ROT Posttest score was about 0.33 standard deviations higher than the ROT Pretest score), indicating that in general, there was only a slight improvement in spatial ability from pretest to posttest.

Table 5.18 lists the correlations between TOLT, ROT Pretest, and ROT_Gain scores for each semester and for the aggregate data, respectively. For the aggregate data, the correlation between TOLT and ROT Pretest score was 0.41 and it was statistically significant. According to Cohen's notation (Cohen, 1988), this correlation was medium. This correlation was noticeable in the scatterplot (Figure 5.3). This result is different from the NYU results and it does not provide support for Gardner's "multiple intelligences" hypothesis, in which Gardner considers spatial ability and logical reasoning as separate and unrelated skills.

Another interesting result was that the correlation between ROT Pretest and ROT_Gain score was negative (-0.39, Table 5.18) and this correlation was statistically significant with medium effect size. This suggests that students who did poorly on the ROT Pretest tended to have a larger ROT_Gain score, which makes sense, since these students would have a larger room for improvement in their spatial ability than those other students who had good spatial ability to begin with.

To further look into this difference in ROT_Gain between low ability students and high ability students, the sample was divided into two groups in a way similar to the analysis for NYU data: students with a ROT Pretest score above the UPR median score of 10.5 were classified as "High" spatial ability group, while those with a ROT Pretest score at or below the median score were classified as "Low" spatial ability group. The average ROT_Gain for the low spatial ability group was 2.02, a large and statistically significant improvement in spatial ability, while the average ROT_Gain for the high spatial ability group was -0.05, which was negative and not statistically significant ($t = 0.34$, $p > .05$, Table 5.19). An independent samples t-test found that there was a

statistically significant difference between the two groups in their ROT_Gains ($t = 3.70$, $p < .001$). The effect size of this t-test comparison was $d = 0.69$, indicating a medium-to-large difference between the low group and the high group in their average ROT_Gain. The low group had a large and statistically significant improvement in their spatial ability from ROT pretest to posttest, while the high group's ROT score essentially did not improve from pretest to posttest. The average gap between the low group and high group at the ROT pretest was 6.58 (7.12 for the low group vs. 13.70 for the high group), but the gap dropped considerably to 4.98 at the posttest (8.89 for the low group vs. 13.87 for the high group). This result suggests that the Molecules of Life course at UPR was successful in improving the spatial ability of students who began with low spatial skills, and to some degree, the MOL course was able to reduce the gap in spatial ability between low spatial ability students and high spatial ability students.

Table 5.16 UPR Enzyme Content Assessment

n	Pretest mean (SD)	Posttest mean (SD)	Anchor_Gain (SD)	Effect size	t -test (p value)
77	9.30 (2.681)	11.84 (2.857)	2.55 (2.775)	0.92	t = 8.05 (p < .0001)

Table 5.17 ROT Gain Score for UPR Students

Semester	n	Mean	SD	t value	p value	Effect size (Cohen's d)
Fall 2005	38	0.263	2.993	0.54	>0.05	0.06
Fall 2006	35	1.029	2.728	2.23	0.032	0.26
Fall 2007	43	1.488	3.514	2.78	0.008	0.36
Aggregate (all semesters)	116	0.948	3.14	3.25	0.001	0.22

Table 5.18 Correlations between TOLT, ROT Pretest and ROT_Gain for UPR

Semester	Correlation between TOLT and ROT Pretest (n, p value)	Correlation between TOLT and ROT_Gain	Correlation between ROT Pretest and ROT_Gain
Fall 2005	0.39** (n = 44, p =.0083)	0.40** (n = 38, p = .013)	-0.09 (n = 38, p > .05)
Fall 2006	0.17 (n = 38, p > .05)	0.12 (n = 33, p > .05)	-0.29** (n = 35, p > .05)
Fall 2007	0.56** (n = 48, p < .001)	-0.29 (n = 39 , p > .05)	-0.64** (n = 43, p < .001)
Aggregate	0.41** (n = 130, p < .001)	0.02 (n = 110, p > .05)	-0.39** (n = 116, p < .001)

*Pair-wise exclusion of missing values were used for all correlations. **Significant at .05 level.

Table 5.19 Low vs. High Spatial Ability Group in ROT_Gain for UPR

Spatial ability group	ROT pretest score (n, SD)	ROT posttest score (n, SD)	ROT _Gain (n, SD)	t -test comparing ROT_Gain to zero (n, p value)	t-test comparing two groups in their ROT_Gain (Cohen's d, p value)
Low	7.12 (67, 2.409)	8.89 (56, 3.601)	2.02* (56, 3.371)	t = 4.48 (n = 56, p < .001)	t = 3.70 (d = 0.69, p < .001)
High	13.70 (69, 2.528)	13.87 (60, 3.332)	-0.05 (60,2.554)	t = -0.15 (n = 60, p > .05)	

*ROT_Gain significantly higher than 0

Missing Data Analysis

There were 174 UPR students in total in our sample, 12.1% of them (21 students) missed TOLT, 21.8% of them (38 students) missed ROT pretest, and 23.0% of them (40 students) missed ROT posttest. Unlike the NYU data, the "missingness" of test scores for UPR (TOLT, ROT pretest, and ROT posttest) appeared to be non-random. In fact, missing each test was strongly correlated with missing other tests (Table 5.21). For instance, the correlation between missing the TOLT and missing the ROT pretest was .44, a medium and statistically significant correlation, demonstrating that UPR students who missed the TOLT also had a strong tendency to miss the ROT pretest. Even the smallest correlation listed in Table 5.21, the correlation between missing the TOLT and missing the ROT posttest was statistically significant. This non-randomness of missing test scores was further revealed when we compared UPR students who took each test and those who

missed it. For example, we found an almost medium effect size difference in TOLT scores between students who took ROT posttest and those who did not (Table 5.20). An equivalence test recommended in (Lewis & Lewis, 2005b) was performed and it failed to rule out non-equivalence between these two groups. Hence these two groups may well be not equivalent.

Table 5.20 UPR Students Who Took ROT Posttest vs. Those Who Did Not

Took ROT Posttest?	TOLT (n, SD)	ROT Pretest (n, SD)
Yes (n=134)	2.75 (123, 2.35)	10.52 (116, 4.33)
No (n=40)	2.00 (30, 1.74)	10.10 (20, 2.55)
Effect size (d)	0.33	0.10
t-test & p value	1.96 (p>.05)	0.60 (p>.1)

Table 5.21 Correlations between Missing Different Tests for UPR

"Missingness" of Test	Missing TOLT	Missing ROT Pretest	Missing ROT Posttest
Missing TOLT	--		
Missing ROT Pretest	.44*	--	
Missing ROT Posttest	.22*	.37*	--

* n= 174; all correlations were found statistically significant with $p < .001$.

These results suggest that the missing data for UPR students were not random and there was indeed a pattern in those UPR students who have missing data. The occurrence of such pattern begets a limitation in the generalizability of the UPR results involving TOLT, ROT pretest, and ROT posttest scores. Nevertheless, most of our research questions, as stated earlier in this Chapter, are not concerned with the generalizability of the results from any individual institution. For instance, our research question 1 is: "did MOL reach a diverse group of students", and research question 4 is: "can MOL meaningfully improve students' spatial ability". We would still conclude that "MOL did reach a diverse group of students" regardless of the non-random missing data, since the students' demographics at the eight participating institutions undeniably showed that

MOL did reach a significant number of women and minority students. Also, since the MOL course at UPR did meaningfully improve students' spatial ability by reducing the gap between high ability students and low ability students whose data were available, we can still draw the conclusion that "MOL **can** meaningfully improve students' spatial ability, at least for those students who took the ROT pre and posttest", notwithstanding the non-randomly missing ROT pre- and posttest scores. Accordingly, the non-randomly missing data from UPR would not have a major impact on our conclusions vis-à-vis our research questions.

At the item level, for eight out of the ten TOLT items, less than 1% of the 153 UPR students who took the TOLT missed each item. Even for the remaining two TOLT items, only 2.6% of the 153 UPR TOLT test-takers missed TOLT Item 2, and only 1.3% of these test-takers missed Item 8. These percentages were too small to have any large impact on our results. This is not the case for ROT pre and posttest items. Figure 5.4 the percentage of students shows the percentage of UPR test-takers who missed (i.e. did not answer) each ROT item in the pre and posttest. Note the pretest percentages are out of the 136 UPR students who took the ROT pretest, and the posttest percentages are out of the 134 UPR students who took the ROT posttest. Similar to the NYU results in Figure 5.2, there were two trends in Figure 5.4. First, there were few students missing the beginning items but more students missing items toward the end of the test, e.g. at pretest there were less than 2% of students missing items 1 to 12, respectively, while more than 4% of UPR students missing items 17 to 20 each. Secondly, when pretest and posttest are compared, there were generally more students missing items in the pretest than missing the same items in the posttest, e.g. there were 5.9% of UPR test-takers missing item 18 in the

pretest, but only 1.5% missing the same item in the posttest. Additionally, there were no more than 1.5% of UPR test-takers missing each of the first 12 ROT items even at pretest, and the UPR low-spatial-ability group's score on these first 12 items improved significantly from pretest to posttest (Table 5.22). Due to the rationales similar to the ones presented for the NYU missing data analysis, we believe that the UPR low-spatial-ability group's performance gain on the ROT overall (fewer missing items and higher scores on the posttest) is most likely an indication of these students' improvement in spatial ability, not the test-retest effect.

Figure 5.4 Percentage of UPR Students Missing Each ROT item

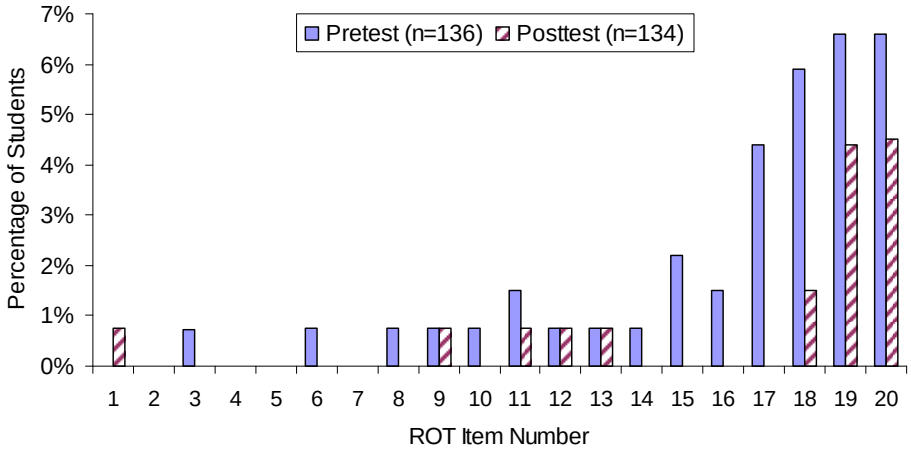


Table 5.22 UPR Low-Ability Students' Score on the First 12 ROT Items

Pretest score (n, SD)	Posttest score (n, SD)	Gain (n, SD)	t-test comparing gain to zero (n, p value)	Effect size (Cohen's d)
5.055 (55, 1.948)	6.255 (55, 2.605)	1.200* (55, 2.453)	t = 3.63 (55, p < .01)	0.52

* Gain significantly higher than 0.

Validity of Measured Gains in Spatial Ability

As described in Chapter 1, UPR implemented the MOL course by integrating the enzyme module into a biology course for non-science majors. It was co-taught as a

semester-long course by the same two full-time professors in all three semesters in which the course was offered (Fall 2005, Fall 2006, and Fall 2007). Numbers of students at different semesters in the course were about the same. Therefore the confounding variables of school, class size, and instructor effect were controlled.

Another threat to the validity of the spatial ability gain is the potential regression effect mentioned earlier. If we used the formula recommended in Weeks (2007) to estimate the expected posttest score based on regression effect: $r_{xy}(X - \mu) + \mu$, where X represents the pretest score, μ is the population mean estimate, r_{xy} is the pretest-posttest correlation. Based on our entire UPR sample, $\mu = 10.46$ (Table 5.4), $r_{xy} = .733$. So, for the low-ability group with mean pretest score of 7.12 (Table 5.19), the expected posttest score would be .733*(7.12 - 10.46) + 10.46 = 8.01. The spatial-ability gain from the regression effect would be 8.01 - 7.12 = 0.89. Since our observed average gain for the low-ability group was 2.02, considerably greater than 0.89, our gain would be most likely due to the MOL course, not an objet d'art of the regression effect. Also, for the high-ability group, if the MOL course had no effect on students' spatial ability and the regression effect is the only source of potential spatial-ability gains, the expected posttest score for them would be .733*(13.70 - 10.46) + 10.46 = 12.83. But our observed posttest score for the high-ability group was 13.87 (Table 5.19), much higher than 12.83. Similar to the justifications discussed earlier for NYU, these UPR results suggest the regression effect was not a major factor for either high- or low-ability group, and the observed spatial-ability gain at UPR was also most likely the effect of the MOL course.

What Contributed to the Improvement of Spatial Ability?

The MOL course at UPR devoted approximately six lecture periods (1.5 hours each) and three laboratory periods (two hours each) to teaching the enzyme module. The enzyme module has three chapters (see Appendix K): Chapter 1. Reactions & Catalysts (Comparing catalysts for N_2 fixation – Haber process & N_2-fixing bacteria; Chemical reactions – activation energy & transition state); Chapter 2. Enzymes as Biological Catalysts (Role of enzymes in HIV infection; HIV protease as a model enzyme; Principles of enzyme function, e.g. lock & key, induced fit); Chapter 3. Enzymes & Drug Design (Designing an effective anti-HIV drug – chemical & biological principles; How do HIV-protease inhibitors work).

As mentioned earlier, in a faculty survey (Appendix J) developed with Dr. Jordan from NYU, we asked the faculty participants at all participating institutions about the specifics of the MOL course implemented at each institution, especially about the contents and/or activities that the instructors believe contributed to the potential improvement of students' spatial ability. For the question "which topics within the enzyme module chapters do you believe contributed to helping students develop spatial ability" in the survey, the two instructors at UPR wrote the following answer: "*Chapter 1: the part on the chemical reactions, specially the section on Reaction Pathways. We devoted time in explaining and having the students visualize the changes that occur from the reactant to the transition state to the products. Chapter 2: the part on HIV protease as a model enzyme and that of Stages of enzyme reactions, especially the part on how the enzyme recognizes its substrate. The analogy of the baseball player making the catch with his mitt really helped the students visualize the relation between the enzyme and its*

substrate. Chapter 3: the part of principles of enzyme inhibition and how an inhibitor competes with the substrate for the enzyme active site was very important for our students to understand that spatial and structural conformation of the molecules involved in the chemical reaction are decisive in the type of interaction they undergo. The section on designing an effective HIV protease inhibitor also contributed to the students understanding why a transition state analog is more effective as a possible drug to treat the condition."

When asked "were there any content topics from other parts of the course (i.e., NOT the enzyme module) that you believe contributed to improving your students' spatial ability", the two instructor answered "*in our Biological Science course, after the Introductory Unit we enter into the unit of Chemical Characteristics of Living Organisms, were we talk about the chemical composition of matter and formation of molecules (chemical bonding, etc.). We discuss the chemical characteristics of the water molecule and also that of acids and bases. We think that the discussion of this information just before teaching the Enzyme & Drug Design module help the students with their visualization skills*."

In addition to these contents that involve spatial ability, the MOL course at UPR used a variety of class and lab activities that involved visual-spatial thinking. These activities include:

- Drawing molecular structures – students were asked in both the class and the laboratory to draw molecular structures.

- Building models using model kits – "The students were asked to build molecules using model kits. A whole laboratory section was devoted to this activity which the students enjoy very much."
- Other types of activity – "*as part of the activities related to the module we showed the students the movie Lorenzo's Oil. The idea was to have them understand how molecules (in this case, enzymes) and their structures (which are dictated by the genetic information) may affect our lives. It was also our intention to expose the students to a real life situation in which learning about chemistry made the big difference.*"

When asked "which activities and/or experiments do you think were especially helpful for improving your students' spatial ability", the instructors at UPR wrote: "we think that the laboratory activity of building molecular models was very effective. We were able to observe the students using their spatial skills (sometimes limited) to arrange the balls (atoms) and sticks (bonds) in a logical way. Sometimes they noticed that the bonds they were creating were not a possible option and started to discuss other possibilities between them. This interaction was very positive in terms of developing their visual skills."

Similar to the instructor at NYU, the instructors at UPR also believe that a combination of activities is most likely to lead to improved spatial ability, and these activities include drawing molecular structures, building molecular models using model kits, using computer graphics software, and using analogies to help students visualize. Similar to the instructor at NYU, the instructors at UPR believe that three dimensional

visualization and representation of molecular structure is best achieved through regular practice (Williamson & José, 2008).

MOL Assessment Results for Students at Xavier and Other Schools

After NYU and UPR, Xavier had the 3rd largest sample size with data available to answer at lease one of our research questions. Xavier did not use the revised version of enzyme tests and its enzyme content assessment cannot be analyzed to answer our first research question. For our third research question concerning whether students' spatial ability can improve, the assessment results for Xavier students are similar to those for NYU and UPR. Table 5.23 lists the mean, standard deviation and t value and p value for the ROT_Gain score for each semester as well as for the aggregate data, respectively. For the aggregate data (all semesters overall), the mean ROT Gain from pretest to posttest was 0.929, and it was significantly different from 0 ($p < .05$). However, the effect size of this gain measured by Cohen's d was small (0.27), indicating that in general, there was only a slight improvement in spatial ability from pretest to posttest.

The scatterplot of Xavier Students' TOLT and ROT pretest scores (Figure 5.5) shows that there is no ceiling effect mentioned earlier, unlike the NYU data shown in Figure 5.1. But the correlations between spatial ability and formal reasoning for Xavier students were similar to those for NYU students: small and not significant (Table 5.24), providing some support for Gardner's multiple intelligence hypothesis that spatial ability and logical reasoning are separate and unrelated skills. This was different from the UPR results.

When we compared the low vs. high spatial ability group in terms of their ROT Gain scores, the Xavier results were different from NYU and UPR. This time, no

difference was found in the average ROT_Gain for Low spatial ability group vs. for High ability group; for both groups, their ROT_Gain was small and not significant (Table 5.25). One possible reason for this may be that NYU did the entire MOL course, which contains a lot of spatial thinking, while Xavier and UPR did the enzyme module only within a biology course. But the other parts of the course would still have an effect on spatial ability. For example, the other parts of the biology course at UPR involved copious visuospatial activities, such as drawing molecular structures in both the class and the lab, and building models using model kits, which led to the large and significant gain in spatial ability for the low ability group at UPR, while the other parts of the biology course at Xavier probably did not entail a lot of spatial thinking, which could explain why the low ability group at Xavier did not improve their spatial ability.

Figure 5.5 Scatterplot of Xavier Students' TOLT and ROT Pretest Scores

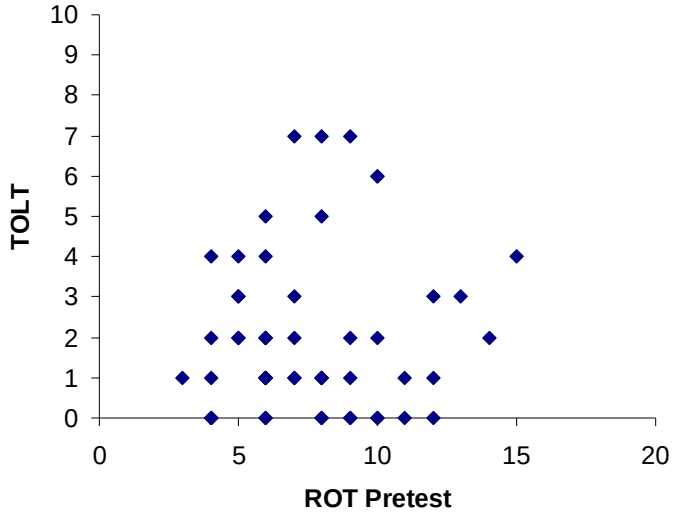


Table 5.23 ROT Gain Score for Xavier Students

Semester	n	Mean	SD	t value	p value	Effect size (Cohen's d)
Spring 06	32	0.844	3.428	1.39	>.05	0.23
Spring 07	24	1.042	2.545	2.01	>.05	0.31
Aggregate (all semesters)	56	0.929	3.056	2.27	.027	0.27

Table 5.24 Correlations between TOLT, ROT pretest, and ROT_Gain for Xavier

Semester	Correlation between TOLT and ROT Pretest (n, p value)	Correlation between TOLT and ROT_Gain	Correlation between ROT Pretest and ROT_Gain
Spring 06	0.12 (n = 29, p > .05)	0.19 (n = 28, p > .05)	-0.12 (n = 32, p > .05)
Spring 07	0.06 (n = 25, p > .05)	0.05 (n = 23 , p > .05)	-0.20 (n = 24, p > .05))
Aggregate (both semesters)	.093 (n = 54, p > .05)	0.15 (n = 51, p > .05)	-0.14 (n = 56, p > .05)

Table 5.25 Low vs. High Spatial Ability Group in ROT_Gain for Xavier

Spatial ability group	Pretest score (n, SD)	Posttest score (n, SD)	ROT_Gain (n, SD)	t -test comparing ROT_Gain to 0 (d, n, p value)	t-test comparing two groups in their ROT_Gain (Cohen's d, p value)
Low	4.96 (28, 1.261)	5.77 (26, 3.229)	0.77 (26, 3.179)	t = 1.23 (d=0.33, n=26, p>.05)	t = -0.36 (d = 0.10, p > .05)
High	9.63 (32, 2.121)	10.63 (30, 2.953)	1.07 (30,2.993)	t = 1.95 (d=0.39, n=60, p>.05)	

For the other 5 schools' data, we found that in general, the full MOL course (such as Fairfield Spring 07) had a larger effect on improving students' spatial ability than just integrating the enzyme module into another course (such as Fairfield Spring 06, Table 5.26). But also, the other parts of the course could also have a significant effect on spatial ability (such as Chaminade Spring 07, and Chicago State Spring 07). The results for spatial ability improvement at most other schools were similar to NYU and UPR. For example, for students at Chicago State University (CSU), the average gap between the low ability group and high ability group at the ROT Pretest was 6.70 (3.86 for the low

ability group vs. 10.56 for the high ability group, Table 5.27), but the gap dropped considerably to 5.53 at the Posttest (6.14 for the low ability group vs. 11.67 for the high ability group). This result suggests that the MOL course at CSU was successful in improving the spatial ability of students who began with low spatial skills, and to some degree, the MOL course was able to reduce the gap in spatial ability between low spatial ability students and high spatial ability students. The results from the other schools and semesters where ROT Gain data were available (including Fairfield Spring 06, Fairfield Spring 07, and Chaminade Spring 07) were similar.

Table 5.26 ROT Gain Score for Students at Other Schools

School/Semester	n	Mean ROT Gain	SD	t value	p value	Cohen's d
Fairfield Spring 06 (Biology*)	18	0.56	2.75	0.86	>.05	0.14
Fairfield Spring 07 (Full MOL course)	18	2.56	1.92	5.66	<.0001*	0.80
Chaminade Spring 07 (Chemistry)	14	2.79	4.54	2.29	.0391	0.75
Chicago Spring 07 (Biology)	16	1.63	2.53	2.57	.0212	0.39

* Implementation of the MOL course: 1) Biology: integrating the enzyme module into a biology course for non-majors; 2) Chemistry: integrating the enzyme module into a chemistry course for non-majors; 3) Full MOL course: offering the entire MOL course as a non-majors science elective.

Table 5.27 ROT_Gain: Low vs. High Ability Group at CSU

Spatial ability group	Pretest score (n, SD)	Posttest score (n, SD)	ROT_Gain (n, SD)	t -test comparing ROT_Gain to zero (Cohen's d, n, p value)	t-test comparing two groups in their ROT_Gain (Cohen's d, p value)
Low	3.86 (7, 1.215)	6.14 (7, 1.864)	2.29 (7, 2.812)	t = 2.15 (d= 1.45, n=7, p > .05)	t = 0.37 (d = 0.46, p > .05)
High	10.56 (9, 3.812)	11.67 (9, 3.000)	1.11 (9, 2.315)	t = 1.44 (d= 0.32, n=9, p > .05)	

Chapter 6: Conclusions and Discussion

Summary of Our Four Studies

The primary focus of this work was to evaluate the effectiveness of a non-majors science course named Molecules of Life (MOL) that aimed to provide useful science education to undergraduate students who are not majoring in scientific disciplines, the raison d'être of the MOL partnership established between eight participating institutions. To tackle this focus, we needed to first develop an assessment plan for MOL. Thus three related studies were carried out to help us choose assessment instruments for MOL, as part of the process of developing an effective assessment plan for MOL.

The first study examined the validity of student evaluations of teaching (SET), specifically looking at whether there was grading leniency bias in SET and whether a peer-led guided inquiry (PLGI) reform negatively affected SET. Although no grading leniency bias was found in our data collected for that study, the reason for that result was likely not because there was no grading leniency bias, but because students did not attribute their grades to their instructor due to the setting in which SET ratings were normally given at the university of investigation. Also, the result that the PLGI reform did not have a negative effect on SET may not be generalizable to other types of reforms, other courses, or other institutions. Due to these uncertainties and a large body of literature questioning the validity of SET for assessment purposes, we determined not to use SET as one of the assessment instruments for MOL.

The second and third study investigated the Test of Logical Thinking (TOLT) and Group Assessment of Logical Thinking (GALT), two widely-used instruments for measuring formal reasoning ability. The second study focused on the functioning of the two additional concrete items that GALT contains over and above TOLT for general chemistry students, while the third study was a direct comparison between TOLT and GALT as intact instruments in both general chemistry and preparatory chemistry. We found that GALT showed no advantage over TOLT for both general and preparatory chemistry in terms of reliability, discriminatory power, potential item bias, and predicting at-risk students. GALT has more frequently occurring, potentially biased items, while TOLT is tenably a less biased test. TOLT is thus recommended over GALT for use in college chemistry teaching to measure college students' formal reasoning abilities and to identify at-risk students. These findings have valuable practical implications for college chemistry teachers. First, when students' SAT scores are not available or not accessible to us as chemistry instructors, as not all students take SAT and not all colleges require SAT scores for admission, TOLT offers an attractive option for us that we can easily utilize at the beginning of the semester for early identification of students at risk of failing the course. Secondly, TOLT is available to us free of charge since it was published (Tobin & Capie, 1981) and it is easy to administer as a 40-minute test, making it a more preferred choice over other more costly, more time-consuming instruments. Also, out of all the instruments reported to have correct percent predictions comparable to the TOLT in identifying students at risk in general chemistry (Legg et al., 2001; McFate & Olmsted, 1999; Wagner et al., 2002), none has been examined of the potential bias of their test items, while our work is the only one known to have investigated differential item functioning and verified TOLT as a tenably little-bias test. This analysis on potential item bias is one of the unique contributions that this work brings to

the field of chemical education research. Furthermore, there is an advantage for chemistry instructors to use a theory-based instrument such as the TOLT rather than an empirical one that intends to measure prior knowledge of mathematics and/or chemistry. Entering chemistry students have typically had prior instruction in mathematics and chemistry, but a low score on an empirical instrument containing mathematics and/or chemistry questions suggests only that this prior instruction was ineffective. There is no theoretical underpinning that ensures that giving students the second opportunity to learn basic mathematics and chemistry will be effective. On the question of an instructional approach for effective remediation, the empirical instrument is silent. On the other hand, a low score on a theory-based formal reasoning measure immediately suggests two potential remedies: (1) Ensure that chemistry concepts are presented in a concrete way when they are initially introduced in the general chemistry course (Herron, 1975); (2) Apply specific interventions that have been shown to support the development of formal reasoning ability (Adey & Shayer, 1990; Adey & Shayer, 1994; Cattle & Howie, 2008; Endler & Bond, 2008; Shayer & Adey, 1992a, 1992b, 1993; Vass et al., 2000). Compared to other instruments reported in the literature to be able to predict at-risk students with comparable prediction accuracies (McFate & Olmsted, 1999; Wagner et al., 2002), our approach using the TOLT with a solid theory-base and clear indication of effective means that instructors can take advantage of to improve their students' chemistry learning is another unique contribution that this work brings to the field of chemical education.

Anchored in the above results from the three related studies, an assessment plan was developed for the Molecules of Life (MOL) project and a systematic evaluation for the MOL course was carried out as a fourth study, as described earlier in Chapter 5.

A Caveat about the Statistical Analysis Results

A caveat about the statistical analysis results in these four studies is that all of the parametric statistical analyses assume that the observations are independent, e.g. one student's scores are not correlated to any other student's scores, as referred to as the independence assumption (Stevens, 1999). A small amount of violation of this independence assumption causes the actual type I error rate (i.e. actual α level) to be several times greater than the nominal α level of .05 set for each statistical analysis in this work. In that case, we may think we are falsely rejecting a true null hypothesis 5% of the time (nominal α), but in fact the false rejection rate may be much higher (actual α). Since the MOL course at each institution was not taught individually to one student at a time, students in the same classroom may have interacted with each other through discussion, group work, or even through the positive or negative classroom atmosphere that the good or troublemaking students caused. Therefore, one student's achievement was possibly influenced by other students', i.e. the observations may have influenced each other, and the independence assumption may have been violated. However, the violation of independence assumption mostly affects type I error rate (α level, which determines statistical significance), but not the effect size, which determines practical significance in pretest vs. posttest comparisons or in two-group comparisons. Since we derived our results mostly from the effect sizes and not from the statistical significance tests, we believe our results are valid even in cases where the assumption of independence might have been violated and the actual type I error rate might have become higher than the nominal rate of .05, because a large-effect-size difference would still be practically significant although not statistically significant, even if we made a type I error. Therefore, whether or not the assumption of independence was violated, it would not change our conclusions much.

Conclusions and Discussion

Did MOL reach a diverse group of students?

As depicted in Chapter 1, the MOL course draws on a context-based approach in introducing scientific topics to simulate student interest. According to Bennett, Lubben, & Hogarth (2007), "*context-based approaches* are approaches adopted in science teaching where contexts and applications of science are used as the starting point for the development of scientific ideas. This contrasts with more traditional approaches that cover scientific ideas first, before looking at applications" (p. 348).

Although there are myriads of studies providing evidences that encourage the use of context-based approaches to promote effectual science learning in primary and secondary education (Ben-Zvi, 1999; Bennett & Lubben, 2006; King, Bellocchi, & Ritchie, 2008; Ramsden, 1997; Rubba, McGuyer, & Wahlund, 1991; Tsai, 2000; Wierstra & Wubbels, 1994; Winther & Volk, 1994; Yager & Weld, 1999), very few studies examined context-based approaches at the college level. Perchance context-based approaches work well for middle- and high-school students, but do they work at the college level too? Research has shown as many as 40% to 50% of students entering college do not yet have fully formal operational reasoning ability (Herron, 1975; Lawson et al., 2000; Shibley et al., 2003). If we want our college students to become scientifically literate citizens when they graduate, the importance of successful college science education for all students, including non-science majors, cannot be overemphasized. Context-based approaches may be one of the effective vehicles to achieve this goal. But out of the few studies that explored context-based approaches at the college level (Gutwill-Wise, 2001; Schwartz, 2006), none delved into whether their context-based

approach reached a diverse group of students, especially Black and Hispanic/Latino students who have been traditionally estranged by science. Maybe these reported context-based approaches worked well for their samples of study, but what contribution did they make to the noble goal of "science for all", if they did not even reach a number of underrepresented minority students?

Contrary to the existing lines of published work, this study included a look into whether a context-based approach at the college level, purposely the MOL course, reached a diverse group of students. As a matter of fact, based on the evaluation results for the MOL project, the first conclusion we can draw is that the MOL course did reach a diverse group of students, including a significant number of women, and minority students underrepresented in science (American Indian/Native Alaskan, Black, and Hispanic/Latino students) (Micari & Drane, 2007). In our modern society, the need for science literacy and for us to provide science education for all citizens, including the public and students not majoring in scientific disciplines, has become imperative. According to the American Association for the Advancement of Science, "the life-enhancing potential of science and technology cannot be realized unless the public in general comes to understand science, mathematics, and technology and to acquire scientific habits of mind. Without a science-literate population, the outlook for a better world is not promising" (Project 2061: American Association for the Advancement of Science, 1990, p. viii). Also, "the world has changed in such a way that science literacy has become necessary for everyone, not just a privileged few; science education will have to change to make that possible. We are all responsible for the current deplorable state of affairs in education, and it will take all of us to reform it" (Project 2061: American

Association for the Advancement of Science, 1990, p. ix). Thus, "science for all" is an important goal of education.

The MOL course contributes to this goal in at least three ways. First, as a non-majors science course, the MOL course contributes to the science literacy of future citizens (or the general public) of our society by introducing students who are not majoring in scientific disciplines to the interface between chemistry, biology, pharmaceuticals, and health. Secondly, The MOL course reached a large number of women and minority students who are underrepresented in science and engineering fields. This result alone was a contribution to the equity in science education. Thirdly, feedback from faculty participants at all eight participating institutions in the MOL project showed that the MOL course was very successful at stimulating students' interest in science. A number of the women and minority students who took the MOL course, while being non-science majors at the time when they enrolled for the MOL course, may become interested in pursuing a science career after their successful experience in the MOL course. This possibility would be a contribution to increasing women and underrepresented minorities to pursue scientific careers.

Did Students in the MOL Courses Learn the Enzyme Content?

In a literature review into the research evidence on the effects of context-based approaches to science teaching, Bennett et al (2007) points out, "there is a noticeable absence of studies [of context-based approaches] on students from ethnic minority groups" (p. 368). Indeed, despite the large body of work on context-based approaches, what remains in question is whether context-based approaches will be effective for diverse groups of students, especially for students in other cultures whose first language

is not English. Our study fills in this literature gap by involving eight very different institutions in our study, including a highly selective private school (NYU) that enrolls students from throughout the United States and abroad, a large public university within the local university system of the Commonwealth of Puerto Rico (UPR), a small private Catholic and Jesuit University (Fairfield University), a historically black liberal arts college for women (Spelman College), to name a few. These very different institutions included incredibly diverse groups of students. A case in point is the UPR student population: 87% of the 174 UPR students in our study identify themselves as Hispanic/Latino, and a vast majority of them speak Spanish as their first language. The MOL course was translated and taught in Spanish at UPR. Students at UPR and the other seven, very different institutions were all found to have learned the science concepts well in the MOL curriculum. This finding fills in the existing literature gap by demonstrating that a context-based approach indeed can be effective for diverse groups of students, including students in a different culture who speak a different language.

In fact, our second major conclusion is that the MOL curriculum was successful in fostering science content learning for diverse groups of students and that students learned the enzyme content in the course at the eight very different participating schools. The content assessment results on the eighteen (18) anchor items shared by the enzyme pretest and posttest showed that students' scores on these eighteen (18) anchor items improved significantly from pretest to posttest at the eight participating schools. These improvements had large effect sizes and were statistically significant for most schools' data. Furthermore, even on the non-anchor items in the enzyme posttest, students did well too at most schools. Students' scores on the enzyme posttest revealed that they grasped

most of the learning goals of the enzyme module listed in Appendix G. These learning goals were developed and established during the two summer workshops at NYU that faculty members from all eight schools participated in, which happened before the MOL courses were taught at each participating school. Students' grip of these pre-determined learning goals indicates that students at these schools learned the content knowledge in the core scientific topics contained within the module on Enzymes and Drug Design.

Formal Reasoning and Spatial Ability

As indicated earlier in Chapter 5, the existing literature suggests that formal reasoning (Boujaoude et al., 2004; Cattle & Howie, 2008; Endler & Bond, 2008; Lawson et al., 2007; Taylor & Jones, 2008) and spatial ability (Huk, 2006; Lee, 2007; Wu & Shah, 2004) are two fundamental cognitive constructs important for science teaching and learning. This study found that formal reasoning and spatial ability were significantly correlated with science content learning in the MOL course (measured by the enzyme posttest) for students at most schools. This result lends credence to the existing published research in the literature that shows the importance of formal reasoning and spatial ability for students' science learning.

In addition to the studies mentioned above, much research has been done in the past two decades vis-à-vis the separate effects of spatial ability (Carter et al., 1987; Ferk et al., 2003; Holland, 1995; Pribyl & Bodner, 1987; Provo et al., 2002; Supasorn et al., 2008; Yang et al., 2003) and of formal reasoning (Cavallo, 1996; Hahn & Polik, 2004; Libby, 1995; Niaz, 1996; Niaz & Robinson, 1992; Nicoll & Francisco, 2001; Noh & Scharmann, 1997; Rubin & Norman, 1992; Uzuntiryaki & Geban, 2005) on students'

science achievement, whereas no work has yet been conducted to investigate the potential relationship between students' formal reasoning and their spatial ability.

This study was the first one to examine the latent relationship between formal reasoning and spatial ability, two essential cognitive constructs important for science education. Although we obtained different results for students from different institutions, we stress that this work is a good, first step in coming to understand the relationship between formal reasoning and spatial ability.

In point of fact, our third major conclusion is: whether formal reasoning and spatial ability are related is still an open question. At six participating schools including NYU, Chicago State, Fairfield, NCC, Spelman, and Xavier, the correlation between their students' spatial ability and formal reasoning ability, as well as the correlation between spatial ability improvement and formal reasoning ability, were small (i.e. less than .3), indicating that formal reasoning and spatial ability were unrelated, separate skills for students at these six schools. Results from these six schools seem to provide some level of support for the multiple intelligences hypothesis (Gardner, 1983). However, at the other two participating schools (UPR and Chaminade), medium correlations were found between their students' spatial ability and formal reasoning ability, and the correlations were statistically significant, suggesting a significant relationship between spatial ability and formal reasoning for students at UPR and Chaminade. The small correlation between formal reasoning and spatial ability for NYU students could be due to the ceiling effect mentioned earlier, namely, the lack of variability (i.e. greater homogeneity) of test scores for NYU students, particularly their high scores on the TOLT, attenuated the correlation between these NYU students' TOLT and ROT pretest scores. However, the variability of

test scores at institutions such as Chicago State, Fairfield, NCC, Spelman, and Xavier was great enough that if there was a relationship between formal reasoning and spatial ability for students at these institutions, then we would expect such relationship to show up as medium or even large correlations for students at these schools. The fact that our observed correlations for students at these six very different schools (including NYU, Chicago State, Fairfield, NCC, Spelman, and Xavier) were all small suggests that formal reasoning and spatial ability are most likely not correlated for students at these six institutions. On the other hand, however, formal reasoning and spatial ability were found to be significantly related with medium correlation of greater than .3 for students at the other two schools (UPR and Chaminade). Both UPR and Chaminade are island schools with UPR located on the island of Puerto Rico and Chaminade located on an island in Hawaii. One can speculate that there might be an "island effect", namely, students growing up on an island and going to college on the same island may experience a cognitive growth pattern different from other student populations in that these island students' formal reasoning and spatial ability could be co-dependent on each other, i.e. their formal reasoning ability and spatial ability may be correlated with each other more than normally expected. However, we could find no published research supporting this conjecture. Since we do not have data in our current study to test this hypothesis of "island effect", we deem it an interesting future research area to examine this hypothesis. Taking into consideration of our results from all eight participating institutions, we cannot make any definite assertion concerning the relationship between formal reasoning and spatial ability. In other words, the different results from different schools indicate

that whether formal reasoning and spatial ability are related (or not) is indeed still an open question.

Can MOL Meaningfully Improve Students' Spatial Ability?

As mentioned earlier, there are copious published studies showing evidences that support the use of context-based approaches to promote expedient science learning in primary and secondary education (Ben-Zvi, 1999; Bennett & Lubben, 2006; King et al., 2008; Ramsden, 1997; Rubba et al., 1991; Tsai, 2000; Wierstra & Wubbels, 1994; Winther & Volk, 1994; Yager & Weld, 1999), or even at the college level (Gutwill-Wise, 2001; Schwartz, 2006). Nonetheless, these existing studies only did comparisons of generic students' learning in context-based approaches versus the learning in conventional approaches. A crucial question remains: will a context-based approach be effective for low-ability groups of students as well? As Lewis & Lewis (2008) pointed out, "even when student scores improve on average, there is a distinct possibility that certain groups of students may not be benefiting at all, or worse be put at a disadvantage, but this phenomenon is simply masked by the improvements from other groups. In other words, when reform evaluations consider only the generic student, without attention to which groups of students might preferentially benefit and which might be disadvantaged, overall effectiveness can be a misleading measure" (p. 2). Conceivably the context-based approaches usually improve the **average** scores of **generic** students in their attitudes towards science as well as in their understanding of science concepts (Bennett, Lubben, & Hogarth, 2007), but what about the low-ability students who have been historically left behind by scientific disciplines? Whereas several recent studies have scrutinized the equity issue and the effect of other curricular reforms, particularly small-group

cooperative learning approaches, on low-ability students (Lewis & Lewis, 2008; Micari & Drane, 2007), there has been little advancement in creating and evaluating context-based approaches that can be easily made use of at the college level by science instructors for low-ability students who have typically been falling behind in scientific courses.

On the contrary to the abovementioned research papers, our work examined the effectiveness of MOL, a context-based curriculum, on the **low-ability** students at dissimilar institutions, including NYU and UPR, among others. In truth, our fourth and most notable conclusion is that a context-based approach such as the MOL course can meaningfully improve students' spatial ability and that it can reduce the gap between high-ability and low-ability students. The first part of this result is consistent with and adds to current published research work in other fields showing that certain interventions are able to significantly improve students' spatial ability. Some of these fields include physics (Pallrand & Seeber, 1984), geology (Piburn et al., 2005), secondary education (Kwon, 2003; Rafi et al., 2005), and information technology (Rafi et al., 2006), but our study is the first work of its kind in the field of chemical education to show that a context-based curriculum such as MOL can enhance students' spatial ability at the college level. More importantly, not only did the MOL course improve students' spatial ability, but also the improvement was meaningful, as it reduced the gap between high-spatial-ability and low-spatial-ability students at most participating schools such as NYU, UPR, Chaminade, Chicago State, and Fairfield.

This meaningfulness of students' spatial ability gains in our study serves well the purpose of equity, and it extends the literature beyond typical, universal comparisons of generic students' learning in context-based curricula with learning in conventional

courses. The only other study we are aware of that looked at whether low-ability students benefit from a context-based approach is (Yager & Weld, 1999), in which it was found that low-ability students in classes using a context-based approach improved their science learning in concept, application, process, creativity, attitude, and world view domains significantly more than their low-ability peers taking conventional classes. However, Yager and Weld did not define their "low-ability" group in a clear manner, as they did not delineate exactly which students were considered "low ability" in their study. Moreover, when their "low-ability" students were compared with "high-ability" students within the same classes using their context-based approach, the large gap between "high-ability" students and "low-ability" students remained at posttest. In other words, their approach did not reduce the gap between "high-ability" and "low-ability" students (Yager & Weld, 1999, p. 187). In contrast, our study is the first one to show that a context-based approach can not only improve student aptitude, but also **meaningfully** improve it by reducing the gap between high-spatial-ability and low-spatial-ability students.

Because of the critical link of spatial ability to science learning, this result is very promising for our efforts to move towards "science for all", an important goal of science education, as described in many authoritative documents and NSF mission statements (National Science Foundation, 1996, 2006; Project 2061: American Association for the Advancement of Science, 1990), as well as in various research conference presentations (Jiang & Lewis, 2008; Jordan, Kallenbach, Lewis et al., 2008; Jordan & Lewis, 2008) and published research work (Bianchini & Cavazos, 2007; Lewis & Lewis, 2008).

Broadening the work by Williamson & José (Williamson & José, 2008), where working with 3-D model kits and computer graphics software was found to be able to

improve students' spatial ability during two three-week sessions, our results indicate that these activities (building molecular models using model kits and utilizing computer graphics software to help students visualize two and three dimensional structures), along with practice in drawing two and three dimensional molecular structures (e.g. drawing wedge-dashed-wedge-line structures with three dimensional perspectives), seem to contribute to better spatial ability. But Williamson and José did not discuss the test-retest effect or the "regression to the mean effect" (a.k.a. regression effect) that we conferred earlier in Chapter 5, hence the observed gains in their students' spatial ability might be simply due to the test-retest effect or the regression effect. Our finding with regard to the meaningful improvement of students' spatial ability is more robust to these alternative explanations, owing to the reasons discussed earlier. An implication of our finding for classroom teachers is that instructors of chemistry should be encouraged to include the aforementioned visuospatial activities in classes and labs if they are concerned with improving students' spatial ability to enhance students' conceptual understanding of difficult concepts.

Additionally, the MOL course used certain socioscientific issues in some cases as the context that served as the starting point for the development of scientific concepts. *Socioscientific issues* designate "social dilemmas with conceptual, procedural, or technological associations with science" (Sadler & Zeidler, 2005a, p. 112). For example, in a section titled "How do we know that drugs are safe and effective?" in Chapter 1 in the draft textbook for the full MOL course, the rapid but transitory success and costly recall of the analgesic drug Vioxx is introduced. Then it is discussed that "*the example of Vioxx serves to illustrate the complex challenges – scientific, social, financial, and*

political – that must be overcome in order to provide the public with a successful new drug. For example, what is the appropriate balance between rigorous testing of drug candidates (which is both expensive and time-consuming) versus making treatments available more rapidly to people who need them? Should life-saving medications, such as drugs to fight HIV, be handled differently than treatments with cosmetic function like Botox? Has the explosion of direct-consumer advertising proved beneficial by providing patients with more information about pharmaceuticals, or does it foster misplaced enthusiasm for expensive new drugs like Vioxx that in some cases are no better than older treatments?" (Jordan & Kallenbach, 2008, p. 18) The MOL course was successful in using socioscientific issues (SSI) like this one to arouse student interest and promote learning gains in both scientific content and spatial ability, according to our assessment results and faculty feedback. While some recent studies (Sadler & Zeidler, 2005b; Walker & Zeidler, 2007) seemed to suggest a positive association between scientific content knowledge and *informal reasoning*, to wit, the "generation and evaluation of positions in response to complex issues [or contentious problems] that lack clear-cut solutions" (Sadler, 2004, p. 514), our work is among the first reports to illustrate that a context-based college science curriculum for non-majors can effectively advance student learning gains in both their content knowledge and cognitive ability, which is an **important** and **necessary** first step not only for empowering students to become responsible citizens able to "carefully consider SSI and make reflective decisions regarding those issues" (Zeidler, Sadler, Simmons et al., 2005, p. 372), but also for equipping students to "participate thoughtfully with fellow citizens in building and protecting a society that is open, decent,

and vital" (Project 2061: American Association for the Advancement of Science, 1990, p. xiii).

Other Conclusions

A fifth conclusion we can draw is that in general, the full MOL course has larger effects on improving students' spatial ability than just integrating the enzyme module into another course. But also, other parts of the course can have an effect on spatial ability too. As mentioned earlier, implementations at some institutions (such as UPR and Chaminade) only used the enzyme module instead of the full MOL course and integrated the enzyme module into a biology or chemistry course for non-science majors. Instructors teaching Molecules of Life at these institutions had the freedom to make many choices in terms of contents, activities, or even teaching styles used in the non-majors chemistry or biology course encompassing the enzyme module. For example, in addition to letting students build molecular models using model kits and asking them to draw 2-D and 3-D molecular structures in both the classroom and the laboratory, the UPR instructors showed students the movie "Lorenzo's Oil" to help students understand how enzymes and their structures affect their lives and to expose students to real-life situations. At Chaminade, there were a small number of students (eight during Spring 2006, and sixteen during Spring 2007) taking the non-majors chemistry course that incorporated the enzyme module. During the lecture in the small classrooms, the instructor introduced students into drawing wedge and dash bonds, included short exercises for students to construct simple molecules to illustrate that enantiomers are non-superimposable, asked students to use the ChemDraw and ChemDraw 3-D computer graphics software, and had students perform simple simulations of lock and key (enzyme-substrate) concepts by drawing on paper circles of

matching colors to models and by illustrating that only one enantiomer will match the colors. These different activities at UPR and Chaminade most likely had an effect on their students' interest, motivation, as well as spatial ability.

Finally, we found that collecting data from many different courses at many different institutions is very challenging. There were various administrative problems and inconsistent, laissez faire handling of assessments by a number of staff members and teaching assistants at different institutions that resulted in incomplete data with too much missing information. For example, at Chaminade, the student survey as shown in Appendix F was not used in the Spring 2006 semester, and during the Spring 2007 semester, the test administer did not record students' responses on the last two items of the TOLT test, resulting in missing TOLT scores for all Chaminade students in Spring 2007; at Chicago State University, the enzyme posttest was not given but the enzyme pretest was given twice in the Spring 2007 semester, and, the enzyme pretest, enzyme posttest, and ROT posttest were not given in the Spring 2006 semester at Chicago State, making it not possible to perform content assessment for either Spring 2006 or Spring 2007, nor to conduct any analysis on students' potential improvement in spatial ability for Spring 2006; at Fairfield, TOLT item responses were not recorded for the Spring 2007 semester and only total TOLT scores were available without any item scores, disabling any item analysis or reliability analysis on the TOLT test scores for the Fairfield students in Spring 2007; at NCC during the Spring 2007 semester, the student survey was not used and no ROT pretest nor ROT posttest was given, leading to the lack of important data on student demographics, academic background, and spatial ability measure for these NCC students; at Spelman during the Fall 2007 semester, most students did not answer the first

two questions on the ROT pretest and ROT posttest, which is a very abnormal pattern. Yet the test administers could provide no explanation at all on why this phenomenon happened. At Xavier University during the Spring 2007 semester, only total scores of the TOLT, ROT pretest and ROT posttest were available while individual item responses were not recorded, making it impossible to conduct any item analysis or even reliability analysis of these assessments for Xavier.

These problems reveal to us that it is very difficult and challenging to conduct large scale educational studies and to collect data from many different courses involving many different instructors, staff members, and teaching assistants at different institutions. An important implication of this challenge is that any future educational research collecting assessment data from multiple institutions or locations must have a well-developed plan to deal with potential logistic and administrative problems such as those mentioned above, and, additionally, it needs to not only have a uniform assessment plan but also make sure that different institutions strictly follow the uniform assessment plan with highest level of consistency across different locations.

Future Research

A limitation of this work is that we did not have data to directly examine how and why the MOL course improved students' spatial ability. The instructors' answers to the faculty survey from NYU, UPR, and the other participating institutions showed us that the instructors believe a combination of classroom and laboratory activities in the MOL course, such as those described earlier, contributed to their students' growth in spatial ability. But we did not have data or evidence to directly support these instructors' beliefs.

Since numerical data alone often "obscure many of the nuances" (Cousins, 2007, p.728) and qualitative analysis is usually necessary to "uncover the story behind the numerical data" (Cousins, 2007, p. 716), we believe a future study involving qualitative methods, such as think-aloud probing and interviewing of students is necessitated to find out how exactly the MOL course improves students' spatial ability.

Given the importance of both formal reasoning and spatial ability for science learning, we believe that the potential relationship (or lack thereof) between these two fundamental cognitive skills merits further investigation. Future studies in this area should collect reliable data with large-enough sample sizes from a variety of institutions with diverse student populations to see if there is any consistent relationship (or lack thereof) between formal reasoning and spatial ability for different student populations. If it is found that there are consistently medium or even large correlations between these two skills for a variety of student populations at different institutions, then it would support the proposition that the correlation between spatial ability and chemistry problem-solving skills is based on "a more general cognitive factor" (Wu & Shah, 2004, p. 472), which may be formal reasoning ability.

Another type of worthwhile future work is to employ a true experimental design with random assignment of large sample of students into control groups and treatment groups to investigate whether the enzyme module can lead to learning gains in spatial ability and science learning for women and underrepresented minority students. The biology course at UPR devoted approximately three weeks (i.e. six 1.5-hour lecture periods along with three two-hour lab periods) to covering the enzyme module, and we found it was successful in improving its mostly Spanish-speaking Latino/Hispanic

students' spatial ability. But a limitation is that we did not have a randomized control group to compare with the treatment group to enhance the external validity of our results. Future research shall preferably utilize a true experimental design, that is, the "gold standard" of randomized control trial commonly regarded as offering the strongest substantiation of "what works" (Bennett et al., 2007, p. 365). Previous studies revealed that many students' poor science feat are due to lack of proper understanding of scientific concepts (Cuicchi, 1992; Griffin, 1997; Mulford & Robinson, 2002; Noh & Scharmann, 1997; Oliva & Cadiz, 1999; Oliva & Cadiz, 2003; Supasorn et al., 2008; Uzuntiryaki & Geban, 2005; Yezierski & Birk, 2006). If students' depleted spatial ability is one of the major causes for their lack of adequate conceptual understanding, then educators and researchers should come up with ways to boost up students' spatial ability. The MOL course, or even the enzyme module from the MOL course, can be promising candidates for good approaches that enhance students' spatial ability.

Concluding Remarks

In summary, our study demonstrated that the Molecules of Life curriculum was successful in promoting learning gains in both spatial ability and scientific content for vastly different student populations at very diverse institutions such as NYU and UPR, among others. The Molecules of Life course is an effective step for our efforts to move towards "science for all". As a result, the Molecules of Life curriculum is strongly recommended for institutions and instructors to use to provide valuable science education for college students not majoring in scientific disciplines.

References Cited

Ablard, K. E., & Tissot, S. L. (1998). Young Students' Readiness for Advanced Math: Precocious Abstract Reasoning. *Journal for the Education of the Gifted,* 21(2), 206-223.

Abraham, M. R., & Renner, J. W. (1986). The Sequence of Learning Cycle Activities in High School Chemistry. *Journal of Research in Science Teaching,* 23(2), 121-143.

Achacoso, M. V., & Svinicki, M. D. (Eds.). (2005). *Alternative Strategies for Evaluating Student Learning* (Vol. 2004, Issue 100). San Francisco, CA: Wiley Periodicals, Inc.

Adey, P. S., & Shayer, M. (1990). Accelerating the Development of Formal Thinking in Middle and High School Students. *Journal of Research in Science Teaching,* 27(3), 267-285.

Adey, P. S., & Shayer, M. (1994). *Really Raising Standards: Cognitive Intervention and Academic Achievement* (First ed.). New York, NY: Routledge.

Armstrong, J. S. (1998). Are Student Ratings of Instruction Useful? *American Psychologist,* 53(11), 1223-1224.

Aronson, E., & Linder, D. E. (1965). Gain and Loss of Esteem as Determinants of Interpersonal Attractiveness. *Journal of Experimental Social Psychology,* 1, 156-171.

Baird, W. E., Shaw, E. L., & McLarty, P. (1996). Predicting Success in Selected Events of the Science Olympiad. *School Science and Mathematics,* 96(2), 85-93.

Basow, S. A. (1995). Student Evaluations of College Professors: When Gender Matters. *Journal of Educational Psychology,* 87(4), 656-665.

Ben-Zvi, R. (1999). Non-Science Oriented Students and the Second Law of Thermodynamics. *International Journal of Science Education,* 21(12), 1251-1267.

Bennett, J., & Lubben, F. (2006). Context-Based Chemistry: The Salters Approach. *International Journal of Science Education,* 28(9), 999-1015.

Bennett, J., Lubben, F., & Hogarth, S. (2007). Bringing Science to Life: A Synthesis of the Research Evidence on the Effects of Context-Based and STS Approaches to Science Teaching. *Science Education,* 91(3), 347-370.

Beran, T. N., & Rokosh, J. L. (2008). Instructors' Perspectives on the Utility of Student Ratings of Instruction. *Instructional Science,* 36.

Bianchini, J. A., & Cavazos, L. M. (2007). Learning from Students, Inquiry into Practice, and Participation in Professional Communities: Beginning Teachers' Uneven Progress toward Equitable Science Teaching. *Journal of Research in Science Teaching,* 44(4), 586-612.

Billings-Gagliardi, S., Barrett, S. V., & Mazor, K. M. (2004). Interpreting Course Evaluation Results: Insights from Thinkaloud Interviews with Medical Students. *Medical Education,* 38(10), 1061-1070.

Bodner, G. M. (1980). Statistical Analysis of Multiple-Choice Exams. *Journal of Chemical Education,* 57(3), 188-190.

Bodner, G. M., & Guay, R. B. (1997). The Purdue Visualization of Rotations Test. *The Chemical Educator,* 2(4), 138-154.

Boroff, D. (1961). Jewish Teen-Age Culture. *Annals of the American Academy of Political and Social Science,* 338, 79-90.

Boujaoude, S., Salloum, S., & Abd-El-Khalick, F. (2004). Relationships between Selective Cognitive Variables and Students' Ability to Solve Chemistry Problems. *International Journal of Science Education,* 26(1), 63-84.

Brookhart, S. M. (1999). *The Art and Science of Classroom Assessment: The Missing Part of Pedagogy* (Vol. 27, No. 1). Washington, DC: The George Washington University Graduate School of Education and Human Development.

Brown, P. L. (2008). For the Muslim Prom Queen, There Are No Kings Allowed [Electronic Version]. *iViews.com*. Retrieved September 24, 2008 from http://www.iviews.com/articles/Articles.asp?ref=NT0807-3621.

Bunce, D. M., & Hutchinson, K. D. (1993). The Use of the GALT (Group Assessment of Logical Thinking) as a Predictor of Academic Success in College Chemistry (Sym). *Journal of Chemical Education,* 70(3), 183-187.

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York: Cambridge University Press.

Carter, C. S., LaRussa, M. A., & Bodner, G. M. (1987). A Study of Two Measures of Spatial Ability as Predictors of Success in Different Levels of General Chemistry. *Journal of Research in Science Teaching,* 24(7), 645-657.

Cashin, W. E. (1995). *Student Ratings of Teaching: The Research Revisited (Idea Paper No.32)*, Kansas State University, Center for Faculty Evaluation and Development.

Cattle, J., & Howie, D. (2008). An Evaluation of a School Programme for the Development of Thinking Skills through the CASE@KS1 Approach. *International Journal of Science Education,* 30(2), 185-202.

Cavallo, A. M. L. (1996). Meaningful Learning, Reasoning Ability, and Students' Understanding and Problem Solving of Topics in Genetics. *Journal of Research in Science Teaching,* 33(6), 625 - 656.

Centra, J. A. (2003). Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work? *Research in Higher Education,* 44(5), 495-518.

Centra, J. A., & Gaubatz, N. B. (2000). Is There Gender Bias in Student Evaluations of Teaching? *The Journal of Higher Education,* 70, 17-33.

Chacko, T. I. (1983). Student Ratings of Instruction: A Function of Grading Standards. *Educational Research Quarterly,* 8(2), 19-25.

Clauser, B. E., & Mazor, K. (1998). Using Statistical Procedures to Identify Differentially Functioning Test Items. *Educational Measurement: Issues and Practice,* 17, 31-44.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Second ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Coleman, S. L., & Gotch, A. J. (1998). Spatial Perception Skills of Chemistry Students. *Journal of Chemical Education,* 75(2), 206.

Cousins, A. (2007). Gender Inclusivity in Secondary Chemistry: A Study of Male and Female Participation in Secondary School Chemistry. *International Journal of Science Education,* 29(6), 711-730.

Cracolice, M. S., Deming, J. C., & Ehlert, B. (2008). Concept Learning Versus Problem Solving: A Cognitive Difference. *Journal of Chemical Education,* 85(6), 873-878.

Crocker, L. M., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York, NY: CBS College Publishing: Holt, Rinehart and Winston.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*. New York: Irvington Publishers.

Cuicchi, P. M. (1992). *The Effect of Formal Reasoning Ability and Grouping by Formal Reasoning Ability in Cooperative Study Groups Upon the Alleviation of Misconceptions in High School Physics.* Ed.D thesis, Mississippi State University.

Daffinrud, S. (1997). Student Assessment of Learning Gains -- Instrument Description. Retrieved August 15, 2008, from http://www.wcer.wisc.edu/salgains/instructor/SALGains.asp

Davies, M., Hirschberg, J., Lye, J., Johnston, C., & Mcdonald, I. (2007). Systematic Influences on Teaching Evaluations: The Case for Caution. *Australian Economic Papers,* 46(1), 18-38.

Desurra, C. J., & Church, K. A. (1994). *Unlocking the Classroom Closet: Privileging the Marginalized Voices of Gay/Lesbian College Students*. Paper presented at the 80th Annual Meeting of the Speech Communication Association (New Orleans, LA, November 19-22, 1994).

Dwyer, C. A. (Ed.). (2008). *The Future of Assessment*. New York, NY: Lawrence Erlbaum Associates.

Eiszler, C. F. (2002). College Students' Evaluations of Teaching and Grade Inflation. *Research in Higher Education,* 43(4), 483-501.

Endler, L. C., & Bond, T. G. (2008). Changing Science Outcomes: Cognitive Acceleration in a US Setting. *Research in Science Education,* 38(2), 149-166.

Examinations Institute of the American Chemical Society Division of Chemical Education. (2006). California Chemistry Diagnostic Test 2006.

Faculty Resource Network at New York University. (2008). *The Molecules of Life: A National Dissemination Conference*. Atlanta, GA.

Farrell, J. J., Moog, R. S., & Spencer, J. N. (1999). A Guided-Inquiry General Chemistry Course. *Journal of Chemical Education,* 76(4), 570-574.

Feldman, K. A. (1997). Identifying Exemplary Teachers and Teaching: Evidence from Student Ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective Teaching in Higher Education: Research and Practice* (pp. 368-395). New York: Agathon Press.

Ferk, V., Vrtacnik, M., & Blejec, A. (2003). Students' Understanding of Molecular Structure Representations. *International Journal of Science Education,* 25(10), 1227-1245.

Foronda, C. L. (2008). A Concept Analysis of Cultural Sensitivity. *Journal of Transcultural Nursing,* 19(3), 207-212.

Fox, M. A., & Hackerman, N. (Eds.). (2003). *Evaluating and Improving Undergraduate Teaching in Science, Technology, Engineering, and Mathematics*. Washington, DC: The National Academies Press.

Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences*. New York: BasicBooks.

Gardner, H. (1999). *Intelligence Reframed*. New York: BasicBooks.

Gardner, H., & Hatch, T. (1989). Multiple Intelligences Go to School: Educational Implications of the Theory of Multiple Intelligences. *Educational Researcher,* 18(8), 4-9.

Gardner, H., & Moran, S. (2006). The Science of Multiple Intelligences Theory: A Response to Lynn Waterhouse. *Educational Psychologist,* 41(4), 227–232.

Gigliotti, R. J., & Buchtel, F. S. (1990). Attributional Bias and Course Evaluations. *Journal of Educational Psychology,* 82, 341-351.

Giuliano, F. J. (1997). *The Relationships among Cognitive Variables and Students' Problem-Solving Strategies in an Interactive Chemistry Classroom.* Ph.D. thesis, Syracuse University, Syracuse, NY.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical Methods in Education and Psychology* (Third ed.). Needham Heights, MA: Allyn and Bacon.

Good, R., Mellon, E. K., & Kromhout, R. A. (1978). The Work of Jean Piaget. *Journal of Chemical Education,* 55(11), 688-693.

Gosser, D. K., & Roth, V. (1998). The Workshop Chemistry Project: Peer-Led Team-Learning. *Journal of Chemical Education,* 75(2), 185-187.

Gray, M., & Bergmann, B. R. (2003). Student Teaching Evaluations. *Academe,* 89(5), 44-46.

Greenwald, A. G., & Gillmore, G. M. (1997a). Grading Leniency Is a Removable Contaminant of Student Ratings. *American Psychologist,* 52(11), 1209-1217.

Greenwald, A. G., & Gillmore, G. M. (1997b). No Pain, No Gain? The Importance of Measuring Course Workload in Student Ratings of Instruction. *Journal of Educational Psychology,* 89, 743–751.
Griffin, B. W. (2001). Instructor Reputation and Student Ratings of Instruction. *Contemporary Educational Psychology,* 26(4), 534-552.
Griffin, B. W. (2004). Grading Leniency, Grade Discrepancy, and Student Ratings of Instruction. *Contemporary Educational Psychology,* 29(4), 410-425.
Griffin, L. L. (1997). *Relationships among Selected Physical Science Misconceptions Held by Preservice Elementary Teachers and Four Variables: Formal Reasoning Ability, Working Memory Capacity, Verbal Intelligence, and Field Dependence/Independence.* Ed.D thesis, The University of Mississippi, Mississippi.
Grimes, P. W., Millea, M. J., & Woodruff, T. W. (2004). Grades - Who's to Blame? Student Evaluation of Teaching and Locus of Control. *Journal of Economic Education,* 35(2), 129-147.
Gutwill-Wise, J. (2001). The Impact of Active and Context-Based Learning in Introductory Chemistry Courses: An Early Evaluation of the Modular Approach. *Journal of Chemical Education,* 78(5), 684–690.
Hahn, K. E., & Polik, W. F. (2004). Factors Influencing Success in Physical Chemistry. *Journal of Chemical Education,* 81(4), 567-572.
Hampton, S. E., & Reiser, R. A. (2004). Effects of a Theory-Based Feedback and Consultation Process on Instruction and Learning in College Classrooms. *RESEARCH IN HIGHER EDUCATION,* 45(5), 497-527.
Herron, J. D. (1975). Piaget for Chemists: Explaining What "Good" Students Cannot Understand. *Journal of Chemical Education,* 52(3), 146-150.
Hoerr, T. R. (2003). It's No Fad: Fifteen Years of Implementing Multiple Intelligences. *Educational Horizons,* 81(2), 92-94.
Holland, C. T. (1995). *The Effects of Formal Reasoning Ability, Spatial Ability, and Type of Instruction on Chemistry Achievement.* Ph.D. thesis, University of Florida, Gainesville, FL.
Hu, L. T., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indices in a Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling,* 6, 1-55.
Huemer, M. (1998). Student Evaluations: A Critical Review. Retrieved August 15, 2008, from http://home.sprynet.com/~owl1/sef.htm
Huk, T. (2006). Who Benefits from Learning with 3d Models? The Case of Spatial Ability. *Journal of Computer Assisted Learning,* 22(6), 392-404.
Hutnik, N., & Gregory, J. (2008). Cultural Sensitivity Training: Description and Evaluation of a Workshop. *Nurse Education Today,* 28, 171-178.
Jiang, B., & Lewis, J. (2008). *Project Assessment for the Molecules of Life.* Paper presented at the Molecules of Life: A National Dissemination Conference, 14-15 March 2008, Atlanta, GA.
Jiang, B., Xu, X., & Lewis, J. E. (2008). *Two Tests of Formal Reasoning and Their Applications in College Chemistry.* Paper presented at the 20th Biennial Conference on Chemical Education (BCCE), 27-31 July 2008, the Division of Chemical Education of the American Chemical Society, Bloomington, Indiana.
Jordan, T., & Kallenbach, N. (2008). *Molecules of Life: A Chemical Approach* (1st ed.). Unpublished Draft Textbook for the MOL Course. New York, NY: New York University.
Jordan, T., Kallenbach, N., Lewis, J., & Jiang, B. (2008). *Molecules of Life: Exploring Chemical Principles in a Biological Context.* Paper presented at the 20th Biennial Conference on

Chemical Education (BCCE), 27-31 July 2008, the Division of Chemical Education of the American Chemical Society, Bloomington, Indiana.
Jordan, T., & Lewis, J. E. (2008). *Molecules of Life: A Partnership to Enhance Undergraduate Science Education.* Paper presented at the Molecules of Life: A National Dissemination Conference, 14-15 March 2008, Atlanta, GA.
King, D., Bellocchi, A., & Ritchie, S. M. (2008). Making Connections: Learning and Teaching Chemistry in Context. *Research in Science Education,* 38(3), 365-384.
Knight, K. H., Elfenbein, M. H., & Martin, M. B. (1997). Relationship of Connected and Separate Knowing to the Learning Styles of Kolb, Formal Reasoning, and Intelligence. *Sex Roles,* 37(5-6), 401-414.
Kogan, J. R., & Shea, J. A. (2007). Course Evaluation in Medical Education. *Teaching and Teacher Education,* 23(3), 251-264.
Kornhaber, M. L. (2004). Multiple Intelligences: From the Ivory Tower to the Dusty Classroom—but Why? *Teachers College Record,* 106(1), 67–76.
Kwon, O. N. (2003, August 2003). *Fostering Spatial Visualization Ability through Web-Based Virtual-Reality Program and Paper-Based Program.* Paper presented at the Web and Communication Technologies and Internet-Related Social Issues – HSI 2003, Seoul, South Korea.
Lang, J. W. B., & Kersting, M. (2007). Regular Feedback from Student Ratings of Instruction: Do College Teachers Improve Their Ratings in the Long Run? *Instructional Science,* 35(3), 187-205.
Lawson, A. E. (1978). The Development and Validation of a Classroom Test of Formal Reasoning. *Journal of Research in Science Teaching,* 15(1), 11-24.
Lawson, A. E. (1992a). The Development of Reasoning among College Biology Students—a Review of Research. *Journal of College Science Teaching,* 21(6), 338-344.
Lawson, A. E. (1992b). What Do Tests of Formal Reasoning Actually Measure? *Journal of Research in Science Teaching,* 29(9), 965–983.
Lawson, A. E., Banks, D. L., & Logvin, M. (2007). Self-Efficacy, Reasoning Ability, and Achievement in College Biology. *Journal of Research in Science Teaching,* 44(5), 706-724.
Lawson, A. E., Drake, N., Johnson, J., Kwon, Y.-J., & Scarpone, C. (2000). How Good Are Students at Testing Alternative Explanations of Unseen Entities? *The American Biology Teacher,* 62(4), 249–255.
Lee, H. (2007). Instructional Design of Web-Based Simulations for Learners with Different Levels of Spatial Ability. *Instructional Science,* 35, 467-479.
Legg, M. J., Legg, J. C., & Greenbowe, T. J. (2001). Analysis of Success in General Chemistry Based on Diagnostic Testing Using Logistic Regression. *Journal of Chemical Education,* 78(8), 1117-1121.
Lekhavat, P. (1996). *The Impact of Adjunct Questions Emphasizing the Particulate Nature of Matter on Students' Understanding of Chemical Concepts Presented in Multimedia Lessons.* Ph.D. thesis, University of Northern Colorado, Colorado.
Lewis, S. E., & Lewis, J. E. (2005a). Departing from Lectures: An Evaluation of a Peer-Led Guided Inquiry Alternative. *Journal of Chemical Education,* 82(1), 135-139.
Lewis, S. E., & Lewis, J. E. (2005b). The Same or Not the Same: Equivalence as an Issue in Educational Research. *Journal of Chemical Education,* 82(9), 1408-1412.
Lewis, S. E., & Lewis, J. E. (2007). Predicting at-Risk Students in General Chemistry: Comparing Formal Thought to a General Achievement Measure. *Chemistry Education Research and Practice,* 8(1), 32-51.

Lewis, S. E., & Lewis, J. E. (2008). Seeking Effectiveness and Equity in a Large College Chemistry Course: An HLM Investigation of Peer-Led Guided Inquiry. *Journal of Research in Science Teaching,* 9999(9999), 1-18.

Liamputtong, P. (2008). Doing Research in a Cross-Cultural Context: Methodological and Ethical Challenges. In *Doing Cross-Cultural Research: Ethical and Methodological Perspectives* (pp. 3-20): Springer Netherlands.

Libby, R. D. (1995). Piaget and Organic Chemistry: Teaching Introductory Organic Chemistry through Learning Cycles. *Journal of Chemical Education,* 72(7), 626-631.

Linden, A. (2007). Estimating the Effect of in Health Management Regression to the Mean Programs. *Disease Management & Health Outcomes,* 15(1), 7-12.

Lohman, D. F. (1996). Spatial Ability and G. In I. Dennis & P. Tapsfield (Eds.), *Human Abilities: Their Nature and Assessment* (pp. 97-116). Hillsdale, NJ: Erlbaum.

Lohman, D. F., Pellegrino, J. W., Alderton, D. L., & Regian, J. W. (1987). Dimensions and Components of Individual Differences in Spatial Abilities. In S. H. Irvine & S. E. Newstead (Eds.), *Intelligence and Cognition: Contemporary Frames of Reference* (pp. 253-312). Dordrecht, The Netherlands: Martinis Nijhoff.

Lord, T. R. (1987). A Look at Spatial Abilities in Undergraduate Women Science Majors. *Journal of Research in Science Teaching,* 24(8), 757-767.

Lynch, S. J. (2000). *Equity and Science Education Reform*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Maguire, S., & Edmondson, S. (2001). Student Evaluation and Assessment of Group Projects. *Journal of Geography in Higher Education,* 25(2), 209-217.

Marsh, H. W. (1984). Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases, and Utility. *Journal of Educational Psychology,* 76, 707-754.

Marsh, H. W. (2001). Distinguishing between Good (Useful) and Bad Workloads on Students' Evaluations of Teaching. *American Educational Research Journal,* 38(1), 183-212.

Marsh, H. W., & Roche, L. A. (2000). Effects of Grading Leniency and Low Workload on Students' Evaluations of Teaching: Popular Myth, Bias, Validity, or Innocent Bystanders? *Journal of Educational Psychology,* 92(1), 202-228.

Mayer, R. E., & Sims, V. K. (1994). From Whom Is a Picture Worth a Thousand Words? Extensions of a Dual-Coding Theory of Multimedia Learning. *Journal of Educational Psychology,* 86(3), 389-401.

McFate, C., & Olmsted, J. A. I. (1999). Assessing Student Preparation through Placement Tests. *Journal of Chemical Education,* 76(4), 562-565.

McGee, M. G. (1979). Human Spatial Abilities: Psychometric Studies and Environmental, Genetic, Hormonal, and Neurological Influences. *Psychological Bulletin* 86(5), 889-918.

McKeachie, W. J. (1979). Student Ratings of Faculty: A Reprise. *Academe,* 65, 384-397.

McKinnon, J. W., & Renner, J. W. (1971). Are Colleges Concerned with Intellectual Development? *American Journal of Physics,* 39(9), 1047-1052.

Micari, M., & Drane, D. (2007). Promoting Success: Possible Factors Behind Achievement of Underrepresented Students in a Peer-Led Small-Group Stem Workshop Program. *Journal of Women and Minorities in Science and Engineering,* 13, 295-315.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How Are Visuospatial Working Memory, Executive Functioning, and Spatial Abilities Related? A Latent-Variable Analysis. *Journal of Experimental Psychology: General* 130(4), 621-640.

Mulford, D. R., & Robinson, W. R. (2002). An Inventory for Alternate Conceptions among First-Semester General Chemistry Students. *Journal of Chemical Education,* 79(6), 739-744.

Muthen, L. K., & Muthen, B. O. (1998-2005). *Mplus User's Guide* (Third ed.). Los Angeles, CA: Muthen & Muthen.

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor Fox Lecture: A Paradigm of Educational Seduction. *Journal of Medical Education,* 48(1973), 630-635.
National Science Board. (2006). America's Pressing Challenge — Building a Stronger Foundation [Electronic Version]. Retrieved October 1, 2008 from http://www.nsf.gov/statistics/nsb0602/.
National Science Foundation. (1996). *Shaping the Future: New Expectations for Undergraduate Education in Science, Mathematics, Engineering, and Technology* (No. NSF 96-139). Arlington, VA: National Science Foundation.
National Science Foundation. (2006). *Investing in America's Future Strategic Plan 2006-2011* (No. NSF 06-48). Arlington, VA: National Science Foundation.
National Science Foundation Division of Undergraduate Education. (2005). Mission Statement [Electronic Version]. Retrieved September 27, 2008 from http://www.nsf.gov/ehr/due/about.jsp.
Niaz, M. (1996). Reasoning Strategies of Students in Solving Chemistry Problems as a Function of Developmental Level, Functional M-Capacity and Disembedding Ability. *International Journal of Science Education,* 18(5), 525-541.
Niaz, M., & Robinson, W. R. (1992). Manipulation of Logical Structure of Chemistry Problems and Its Effect on Student Performance. *Journal of Research in Science Teaching,* 29(3), 211-226.
Nicoll, G., & Francisco, J. S. (2001). An Investigation of the Factors Influencing Student Performance in Physical Chemistry. *Journal of Chemical Education,* 78(1), 99-102.
Noh, T., & Scharmann, L. C. (1997). Instructional Influence of a Molecular-Level Pictorial Presentation of Matter on Students' Conceptions and Problem-Solving Ability. *Journal of Research in Science Teaching,* 34(2), 199–217.
Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.
O'Conor, A. D. (1998). *The Cultural Logic of Gender in College: Heterosexism, Homophobia and Sexism in Campus Peer Groups.* Ph.D. thesis, University of Colorado at Boulder.
Oliva, J. M., & Cadiz, C. P. (1999). Structural Patterns in Students' Conceptions in Mechanics. *International Journal of Science Education,* 21(9), 903-920.
Oliva, J. M., & Cadiz, C. P. (2003). The Structural Coherence of Students' Conceptions in Mechanics and Conceptual Change. *International Journal of Science Education,* 25(5), 539-561.
Olivares, O. J. (2001). Student Interest, Grading Leniency, and Teacher Ratings: A Conceptual Analysis. *Contemporary Educational Psychology,* 26, 382–399.
Pallrand, G. J., & Seeber, F. (1984). Spatial Ability and Achievement in Introductory Physics. *Journal of Research in Science Teaching,* 21(5), 507 - 516.
Piburn, M. D., Reynolds, S. J., McAuliffe, C., Leedy, D. E., Birk, J. P., & Johnson, J. K. (2005). The Role of Visualization in Learning from Computer-Based Images. *International Journal of Science Education,* 27(5), 513-527.
Poole, B. A. M. (1997). *An Exploration of the Perceptions, Developmental Reasoning Levels, Differences in Learning Processes, and Academic Achievement Levels of Students in Introductory College Microbiology.* Ph.D. thesis, The University of Southern Mississippi, Mississippi.
Popham, W. J. (2000). *Modern Educational Measurement: Practical Guidelines for Educational Leaders* (Third ed.). Needham, MA: Allyn & Bacon.
Pribyl, J. R., & Bodner, G. M. (1987). Spatial Ability and Its Role in Organic Chemistry: A Study of Four Organic Courses. *Journal of Research in Science Teaching,* 24(3), 229-240.
Project 2061: American Association for the Advancement of Science. (1990). *Science for All Americans*. New York: Oxford University Press.

Provo, J., Lamar, C., & Newby, T. (2002). Using a Cross Section to Train Veterinary Students to Visualize Anatomical Structures in Three Dimensions. *Journal of Research in Science Teaching,* 39(1), 10-34.

Rafi, A., Anuar, K., Samad, A., Hayati, M., & Mahadzir, M. (2005). Improving Spatial Ability Using a Web-Based Virtual Environment (WbVE). *Automation in Construction,* 14(6), 707-715.

Rafi, A., Samsudin, K. A., & Ismail, A. (2006). On Improving Spatial Ability through Computer-Mediated Engineering Drawing Instruction. *Educational Technology & Society,* 9(3), 149-159.

Ramsden, J. M. (1997). How Does a Context-Based Approach Influence Understanding of Key Chemical Ideas at 16+? *International Journal of Science Education,* 19(6), 697-710.

Remedios, R., & Lieberman, D. A. (2008). I Liked Your Course Because You Taught Me Well: The Influence of Grades, Workload, Expectations and Goals on Students' Evaluations of Teaching. *British Educational Research Journal,* 34(1), 91-115.

Roadrangka, V., Yeany, R. H., & Padilla, M. J. (1983). *The Construction and Validation of the Group Assessment of Logical Thinking (GALT).* Paper presented at the Paper presented at the annual meeting of the National Association for Research in Science Teaching, Dallas, TX.

Rogers, P. C., Graham, C. R., & Mayes, C. T. (2007). Cultural Competence and Instructional Design: Exploration Research into the Delivery of Online Instruction Cross-Culturally. *Educational Technology Research and Development,* 55(2), 197-217.

Rothing, A. (2008). Homotolerance and Heteronormativity in Norwegian Classrooms. *Gender and Education,* 20(3), 253–266.

Rubba, P. A., McGuyer, M., & Wahlund, T. M. (1991). The Effects of Infusing STS Vignettes into the Genetics Unit of Biology on Learner Outcomes in STS and Genetics: A Report of Two Investigations. *Journal of Research in Science Teaching,* 28(6), 537–552.

Rubin, D. L., Ainsworth, S., Cho, E., Turk, D., & Winn, L. (1999). Are Greek Letter Social Organizations a Factor in Undergraduates Perceptions of International Instructors? *International Journal of Intercultural Relations,* 23(1), 1-12.

Rubin, R. L., & Norman, J. T. (1992). Systematic Modeling Versus the Learning Cycle - Comparative Effects on Integrated Science Process Skill Achievement. *Journal of Research in Science Teaching,* 29(7), 715-727.

Sadler, T. D. (2004). Informal Reasoning Regarding Socioscientific Issues: A Critical Review of Research. *Journal of Research in Science Teaching,* 41(5), 513-536.

Sadler, T. D., & Zeidler, D. L. (2005a). Patterns of Informal Reasoning in the Context of Socioscientific Decision Making. *Journal of Research in Science Teaching,* 42(1), 112 - 138.

Sadler, T. D., & Zeidler, D. L. (2005b). The Significance of Content Knowledge for Informal Reasoning Regarding Socioscientific Issues: Applying Genetics Knowledge to Genetic Engineering Issues *Science Education,* 89(1), 71-93.

Satterly, D. (Ed.). (1987). *Piaget and Education.* Oxford: Oxford University Press.

Schoenfeld-Tacher, R. M. (2000). *Relation of Student Characteristics to Learning of Basic Biochemistry Concepts from a Multimedia Goal-Based Scenario.* Ph.D. thesis, University of Northern Colorado, Greeley, Colorado.

Schwartz, A. T. (2006). Contextualized Chemistry Education: The American Experience. *International Journal of Science Education,* 28(9), 977-998.

Scrivener, L. (2003, June 29). A Happily Dateless Prom. *Toronto Star*.

Seiler, V. L., & Seiler, M. J. (2002). Professors Who Make the Grade: Factors That Affect Students' Grades of Professors. *Review of Business,* 23(2), 39-44.

Seymour, E., Wiese, D. J., & Hunter, A.-B. (2000, March 27). *Creating a Better Mousetrap: On-Line Student Assessment of Their Learning Gains.* Paper presented at the National Meeting of the American Chemical Society Symposium, San Francisco, CA.

Shah, P., & Miyake, A. (Eds.). (2005). *The Cambridge Handbook of Visuospatial Thinking*. New York, NY: Cambridge University Press.

Shayer, M., & Adey, P. S. (1992a). Accelerating the Development of Formal Thinking in Middle and High School Students Ii: Post-Project Effects on Science Achievement. *Journal of Research in Science Teaching,* 29(1), 81-92.

Shayer, M., & Adey, P. S. (1992b). Accelerating the Development of Formal Thinking in Middle and High School Students Iii: Testing the Permancy of Effects. *Journal of Research in Science Teaching,* 29(10), 1101-1115.

Shayer, M., & Adey, P. S. (1993). Accelerating the Development of Formal Thinking in Middle and High School Students Iv: Three Years after a Two-Year Intervention. *Journal of Research in Science Teaching,* 30(4), 351-366.

Shearer, B. (2004). Multiple Intelligences Theory after 20 Years. *Teachers College Record,* 106(1), 2-16.

Shibley, I. A., Milakofsky, L., Bender, D. S., & Patterson, H. O. (2003). College Chemistry and Piaget: An Analysis of Gender Difference, Cognitive Abilities, and Achievement Measures Seventeen Years Apart. *Journal of Chemical Education,* 80(5), 569-573.

Smith, I. M. (1964). *Spatial Ability: Its Educational and Social Significance*. London: University of London Press.

Stevens, J. P. (1999). *Intermediate Statistics: A Modern Approach* (Second ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Study, N. E. (2006). Assessing and Improving the Below Average Visualization Abilities of a Group of Minority Engineering and Technology Students. *Journal of Women and Minorities in Science and Engineering,* 12, 367-380.

Supasorn, S., Suits, J. P., Jones, L. L., & Vibuljan, S. (2008). Impact of a Pre-Laboratory Organic-Extraction Simulation on Comprehension and Attitudes of Undergraduate Chemistry Students. *Chemistry Education Research and Practice,* 9(2), 169 - 181.

Taylor, A., & Jones, G. (2008). Proportional Reasoning Ability and Concepts of Scale: Surface Area to Volume Relationships in Science. *International Journal of Science Education,* 99999, 1-17.

Theall, M., Franklin, J., & Ludlow, L. (1990). Attributions and Retributions: Student Ratings and the Perceived Causes of Performance. *Instructional Evaluation,* 11, 12-17.

Tobin, K. G., & Capie, W. (1981). The Development and Validation of a Group Test of Logical Thinking. *Educational and Psychological Measurement,* 41(2), 4l3-423.

Trout, P. (2000). Flunking the Test: The Dismal Record of Student Evaluations. *Academe,* 86(4), 58-61.

Tsai, C.-C. (2000). The Effects of STS-Oriented Instructions on Female Tenth Graders' Cognitive Structure Outcomes and the Role of Student Scientific Epistemological Beliefs. *International Journal of Science Education,* 22(10), 1099–1115.

U.S. News & World Report. (2008). America's Best Colleges 2009. Retrieved October 5, 2008, from http://www.usnews.com/usnews/edu/college/rankings/brief/t1natudoc_brief.php

Uzuntiryaki, E., & Geban, O. (2005). Effect of Conceptual Change Approach Accompanied with Concept Mapping on Understanding of Solution Concepts. *Instructional Science,* 33(4), 311-339.

Valanides, N. C. (1996). Formal Reasoning and Science Teaching. *School Science and Mathematics,* 96(2), 99-106.

Vass, E., Schiller, D., & Nappi, A. J. (2000). The Effects of Instructional Intervention on Improving Proportional, Probabilistic, and Correlational Reasoning Skills among

Undergraduate Education Majors. *Journal of Research in Science Teaching,* 37(9), 981-995.
Verzoni, K., & Swan, K. (1995). On the Nature and Development of Conditional Reasoning in Early Adolescence. *Applied Cognitive Psychology,* 9(3), 213-234.
Wagner, E. P., Sasser, H., & DiBiase, W. J. (2002). Predicting Students at Risk in General Chemistry Using Pre-Semester Assessments and Demographic Information. *Journal of Chemical Education,* 79(6), 749-755.
Walker, K. A., & Zeidler, D. L. (2007). Promoting Discourse About Socioscientific Issues through Scaffolded Inquiry. *International Journal of Science Education,* 29(11), 1387 - 1410.
Weeks, D. L. (2007). The Regression Effect as a Neglected Source of Bias in Nonrandomized Intervention Trials and Systematic Reviews of Observational Studies. *Evaluation & the Health Professions,* 30(3), 254-265.
Wierstra, R. F. A., & Wubbels, T. (1994). Student Perception and Appraisal of the Learning Environment: Core Concepts in the Evaluation of the PLON Physics Curriculum. *Studies in Educational Evaluation,* 20(4), 437-455.
Wiese, D., Seymour, E., & Hunter, A. B. (1999). *Report on a Panel Testing of the Student Assessment of Their Learning Gains Instrument by Faculty Using Modular Methods to Teach Undergraduate Chemistry.* (No. Report to the Exxon Education Foundation). Boulder, CO: Bureau of Sociological Research, University of Colorado.
Wigington, H., Tollefson, N., & Rodriguez, E. (1989). Student's Ratings of Instructors Revisited: Interactions among Class and Instructor Variables. *Research in Higher Education,* 30(3), 331-344.
Williams, H., Turner, C. W., Debreuil, L., Fast, J., & Berestiansky, J. (1979). Formal Operational Reasoning by Chemistry Students. *Journal of Chemical Education,* 56(9), 599-600.
Williamson, V., Huffman, J., & Peck, L. (2004). Testing Students' Use of the Particulate Theory. *Journal of Chemical Education,* 81(6), 891-896.
Williamson, V. M., & José, T. J. (2008). The Effects of a Two-Year Molecular Visualization Experience on Teachers' Attitudes, Content Knowledge, and Spatial Ability. *Journal of Chemical Education,* 85(5), 718-723.
Winther, A. A., & Volk, T. L. (1994). Comparing Achievement of Inner-City High School Students in Traditional Versus STS-Based Chemistry Courses. *Journal of Chemical Education,* 71(6), 501-505.
Wolf, A. (2007). Tradition and Tolerance: A Look into Orthodox Judaism in the Northwest [Electronic Version]. Retrieved September 24, 2008 from http://www.oregonhum.org/pdf/YS-2007-Wolf.pdf.
Worthington, A. G., & Wong, P. T. P. (1979). Effects of Earned and Assigned Grades on Student Evaluations of an Instructor. *Journal of Educational Psychology,* 71, 764-775.
Wu, H.-K., & Shah, P. (2004). Exploring Visuospatial Thinking in Chemistry Learning. *Science Education,* 88(3), 465 - 492.
Yager, R. E., & Weld, J. D. (1999). Scope, Sequence and Coordination: The Iowa Project, a National Reform Effort in the USA. *International Journal of Science Education,* 21(2), 169 - 194.
Yang, E., Andre, T., & Greenbowe, T. J. (2003). Spatial Ability and the Impact of Visualization/Animation on Learning Electrochemistry. *International Journal of Science Education,* 25(3), 329-349.
Yezierski, E. J., & Birk, J. P. (2006). Misconceptions About the Particulate Nature of Matter. *Journal of Chemical Education,* 83(6), 954-960.

Youmans, R. J., & Jee, B. D. (2007). Fudging the Numbers: Distributing Chocolate Influences Student Evaluations of an Undergraduate Course. *Teaching of Psychology,* 34(4), 245-247.

Zeidler, D. L. (1985). Hierarchical Relationships among Formal Cognitive Structures and Their Relationship to Principled Moral Reasoning. *Journal of Research in Science Teaching,* 22(5), 461-471.

Zeidler, D. L., Sadler, T. D., Simmons, M. L., & Howes, E. V. (2005). Beyond STS: A Research-Based Framework for Socioscientific Issues Education. *Science Education,* 89(3), 357-377.

Appendices

Appendix A: Commonly Used Acronyms

Acronym	**Name**
MOL	Molecules of Life
SET	Student Evaluations of Teaching (a.k.a. Student Course Evaluations)
PLGI	Peer-Led Guided Inquiry
SALG	Student Assessment of Learning Gains
TOLT	Test of Logical Thinking
GALT	Group Assessment of Logical Thinking
TOLT+2	A test containing all TOLT items plus the two additional, concrete items that the GALT contains over and above the TOLT
ACS	American Chemical Society
ACS Exam	American Chemical Society First Semester General Chemistry (Special) Examination
SAT	Scholastic Assessment Test
CFA	Confirmatory Factor Analysis
DIF	Differential Item Functioning
2-D	Two Dimensional
3-D	Three Dimensional
ROT	Purdue Visualization of Rotations Test
ROT_GAIN	ROT Posttest score subtracting the ROT Pretest score
NYU	New York University
UPR	University of Puerto Rico at Rio Piedras
CCI	Chemistry Concepts Inventory

Appendix B: Questions in the Official Course Evaluation Forms

A re-typed copy of the official SET form is below (all the questions are exactly as they were in the actual SET form):

Please select ratings according to the following scale:

5 – Excellent 4 – Very Good 3 – Good 2 – Fair 1 – Poor

	5	4	3	2	1
1. Description of course objectives and assignments					
2. Communication of ideas and information					
3. Expression of expectations for performance in this class					
4. Availability to assist students in or out of class					
5. Respect and concern for students					
6. Stimulation of interest in the course					
7. Facilitation of learning					
8. Overall assessment of instructor					

Appendix C: SALG Surveys Used for Fall 2003 and Fall 2004 Semesters

Fall 2003 – Regular Survey

Instructions:
Check one value for each question on each scale. If the question is not applicable, check 'NA'. You may add a comment for any item in the text box at the end of the survey.
Q1: How much did each of the following aspects of the class help your learning?
NA No help A little help Moderate help Much help Very much help
A. The way in which the material was approached
B. How the class activities, labs, reading, and assignments fit together
C. The pace at which we worked
D. The class activities NA No help A little help Moderate help Much help Very much help
1. Class presentations (including lectures)
2. Discussion in class
3. Group work in class
4. Hands-on class activities
E. Tests, graded activities and assignments NA No help A little help Moderate help Much help Very much help
1. Opportunities for in-class review
2. The number and spacing of tests
3. The fairness of test content
4. The mental stretch required of us
5. The grading system used
6. The feedback we received
F. Resources NA No help A little help Moderate help Much help Very much help
1. The text
2. Other reading materials
3. Posted lecture notes
4. Use made of the WWW in this class
G. The information we were given about NA No help A little help Moderate help Much help Very much help
1. Class activities for each week
2. How parts of the classwork, reading, or assignments relate to each other
3. The grading system for the class
H. Individual support as a learner NA No help A little help Moderate help Much help Very much help
1. The quality of contact with the teacher
2. The quality of contact with the tutors in Chemistry 106
3. Working with other students outside of the designated class times
4. The quality of contact with your Chemistry Lab TA
K. The way this class was taught overall

Fall 2003 – PLGI Survey

Instructions:
Check one value for each question on each scale. If the question is not applicable, check 'NA'. You may add a comment for any item in the text box at the end of the survey.
Q1: How much did each of the following aspects of the class help your learning?
NA No help A little help Moderate help Much help Very much help
A. The way in which the material was approached
B. How the class activities, labs, reading, and assignments fit together
C. The pace at which we worked
D. The class activities NA No help A little help Moderate help Much help Very much help
1. Class presentations including lectures (Monday / Wednesday class)

2. Class presentations (Friday class)
3. Discussion in the Monday / Wednesday class
4. Discussion in the Friday class
5. Group work in the Monday / Wednesday class
6. Group work in the Friday class
7. Hands-on class activities in the Monday/Wednesday class
8. Hands-on class activities in the Friday class

E. Tests, graded activities and assignments NA No help A little help Moderate help Much help Very much help

1. Opportunities for in-class review
2. The number and spacing of tests
3. The fairness of test content
4. The mental stretch required of us
5. The grading system used
6. The feedback we received
7. The quizzes in the Friday session
8. Homework due for each Friday session

F. Resources NA No help A little help Moderate help Much help Very much help

1. The chemistry textbook (McMurry and Fay)
2. The workbook for the Friday sessions (Chemistry: A Guided Inquiry)
3. Other reading materials
4. Posted lecture notes
5. Use made of the WWW in this class

G. The information we were given about NA No help A little help Moderate help Much help Very much help

1. Class activities for each week
2. How parts of the classwork, reading, or assignments related to each other
3. The grading system for the class

H. Individual support as a learner NA No help A little help Moderate help Much help Very much help

1. The quality of contact with the teacher
2. The quality of contact with the tutors in Chemistry 106
3. Working with other students outside of the designated class times
4. The quality of contact with your Chemistry Lab TA
5. The quality of contact with your Peer Leader (during the Friday sessions)

K. The way this class was taught overall

Fall 2004 – Regular Survey

Instructions:
Check one value for each question on each scale. If the question is not applicable, check 'NA'. You may add a comment for any item in the text box at the end of the survey.

Q1: How much did each of the following aspects of the class help your learning?

NA No help A little help Moderate help Much help Very much help

A. The way in which the material was approached
B. How the class activities, labs, reading, and assignments fit together
C. The pace at which we worked
D. The class activities NA No help A little help Moderate help Much help Very much help

1. Class presentations (including lectures)
2. Discussion in class
3. Group work in class
4. Hands-on class activities

Please explain your ratings in the space provided.

E. Tests, graded activities and assignments NA No help A little help Moderate help Much help Very much help

1. Opportunities for in-class review
2. The number and spacing of tests

3. The fairness of test content
4. The mental stretch required of us
5. The grading system used
6. The feedback we received
7. The on-line "One Key" homework
Please explain your ratings in the space provided.
F. Resources NA No help A little help Moderate help Much help Very much help
1. The text
2. Other reading materials
3. Posted lecture notes
4. Use made of the WWW in this class
Please explain your ratings in the space provided.
G. The information we were given about NA No help A little help Moderate help Much help Very much help
1. Class activities for each week
2. How parts of the classwork, reading, or assignments relate to each other
3. The grading system for the class
Please explain your ratings in the space provided.
H. Individual support as a learner NA No help A little help Moderate help Much help Very much help
1. The quality of contact with the teacher
2. The quality of contact with the tutors in Chemistry 106
3. Working with other students outside of the designated class times
4. The quality of contact with your Chemistry Lab TA
Please explain your ratings in the space provided.
K. The way this class was taught overall

Fall 2004 – PLGI Survey

Instructions:
Check one value for each question on each scale. If the question is not applicable, check 'NA'. You may add a comment for any item in the text box at the end of the survey.
Q1: How much did each of the following aspects of the class help your learning?
NA No help A little help Moderate help Much help Very much help
A. The way in which the material was approached
B. How the class activities, labs, reading, and assignments fit together
C. The pace at which we worked
D. The class activities NA No help A little help Moderate help Much help Very much help
1. Class presentations including lectures (Monday / Wednesday class)
2. Class presentations (Friday class)
3. Discussion in the Monday / Wednesday class
4. Discussion in the Friday class
5. Group work in the Monday / Wednesday class
6. Group work in the Friday class
7. Hands-on class activities in the Monday/Wednesday class
8. Hands-on class activities in the Friday class
Please explain your ratings in the space provided.
E. Tests, graded activities and assignments NA No help A little help Moderate help Much help Very much help
1. Opportunities for in-class review
2. The number and spacing of tests
3. The fairness of test content
4. The mental stretch required of us
5. The grading system used
6. The feedback we received
7. The on-line "One Key" homework assignments
8. The quizzes in the Friday session

9. Homework due for each Friday session
Please explain your ratings in the space provided.

F. Resources NA No help A little help Moderate help Much help Very much help

1. The chemistry textbook (McMurry and Fay)
2. The workbook for the Friday sessions (Chemistry: A Guided Inquiry)
3. Other reading materials
4. Posted lecture notes
5. Use made of the WWW in this class

Please explain your ratings in the space provided.

G. The information we were given about NA No help A little help Moderate help Much help Very much help

1. Class activities for each week
2. How parts of the classwork, reading, or assignments related to each other
3. The grading system for the class

Please explain your ratings in the space provided.

H. Individual support as a learner NA No help A little help Moderate help Much help Very much help

1. The quality of contact with the teacher
2. The quality of contact with the tutors in Chemistry 106
3. Working with other students outside of the designated class times
4. The quality of contact with your Chemistry Lab TA
5. The quality of contact with your Peer Leader (during the Friday sessions)

Please explain your ratings in the space provided.

K. The way this class was taught overall

Appendix D: Day 1 Survey Used in the Second Study

Please fill out your name, University Identification Number (U#) and your answers to the following questions on a scantron bubble sheet. The scantron will serve as the attendance check. Your answers to the 15 questions below are appreciated as we work to improve this course.

1. How many years (including this one) have you attended a college or university?
a) 1st year **b)** 2nd year **c)** 3rd year **d)** 4th year **e)** more than 4 years

2. Are you a transfer student from another college or university? **a)** Yes **b)** No

3. What is your major or intended major?
a) Chemistry **b)** Pre-med or allied-health **c)** Engineering **d)** Other science **e)** Non-science

4. How much chemistry did you have in high school?
a) No chemistry in high school **b)** 1 semester **c)** 1 full year **d)** 1-2 full years
e) More than 2 full years

5. Which best describes the highest level of math you've completed?
a) I have not taken any math courses as advanced as algebra
b) algebra and/or trigonometry (MAC 1105)
c) pre-calculus (MAC 1140)
d) calculus I (MAC 2241, 2281 or 2311)
e) calculus II (MAC 2242, 2282 or 2312)

6. Which best describes the math course you are taking now?
a) I am not currently taking a math course
b) algebra and/or trigonometry (MAC 1105)
c) pre-calculus (MAC 1140)
d) calculus I or calculus II (MAC 2241, 2242, 2281, 2282, 2311 or 2312)
e) other

7. Have you taken Chemistry for Today (CHM 2021 or equivalent)? **a)** Yes **b)** No

8. Do you currently plan to take General Chemistry II (CHM 2046)? **a)** Yes **b)** No

9. With regard to General Chemistry I (CHM 2045 or equivalent), which best describes you:
a) I am retaking General Chemistry I **b)** I am enrolled in General Chemistry I for the 1st time

10. With regard to General Chemistry I Lab (CHM 2045L or equivalent), which best describes you:
a) I am currently enrolled in the General Chemistry I Lab
b) I am planning to take General Chemistry I Lab
c) I have already completed General Chemistry I Lab
d) I have no plans to take General Chemistry I Lab

11. What grade do you expect to earn in General Chemistry I (CHM 2045)?
a) A **b)** B **c)** C **d)** D **e)** F

12. Are you: **a)** Male **b)** Female

13. Are you a U.S. citizen? **a)** Yes **b)** No

14. Race/National Origin that best describes you (categories taken from USF admissions application):
a) American Indian and Native Alaskan **b)** Native Hawaiian or other Pacific Islander
c) Asian **d)** Black **e)** White

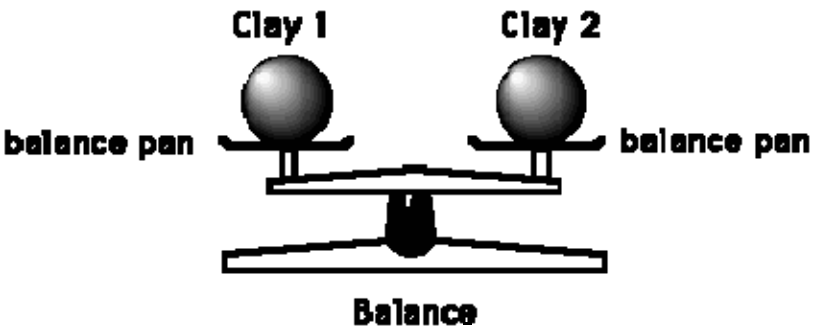
15. Do you consider yourself Hispanic or Latino? **a)** Yes **b)** No

Appendix E: Two Concrete Items GALT Contains Over and Above TOLT

The following shows the two additional concrete items. Item 1 is about "Piece of Clay" (questions 1 and 2), and Item 2 is about "Metal Weights" (questions 3 and 4).

Piece of Clay

1. Tom has two balls of clay. They are the same size and shape. When he places them on the balance, they weigh the same.



The balls of clay are removed from the balance pans. Clay 2 is flattened like a pancake.

Which of these statements is true?

a) The pancake-shaped clay weighs more.

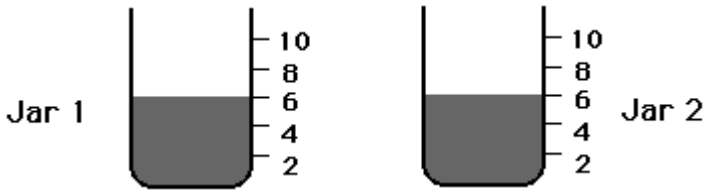
b) The two pieces weigh the same.
c) The ball weighs more.

2. What was the reason for your answer to question 1?

a) You did not add or take away any clay.
b) When clay 2 was flattened like a pancake, it had greater area.
c) When something is flattened, it loses weight.
d) Because of its density, the round ball had more clay in it.

Metal Weights

3. Linn has two jars. They are the same size and shape. Each is filled with the same amount of water.



She also has two metal weights of the same volume. One weight is light. The other is heavy.

She lowers the light weight into jar 1. The water level in the jar rises and looks like this:

If the heavy weight is lowered into Jar 2, what will happen?

a) The water will rise to a higher level than in jar 1.
b) The water will rise to a lower level than in jar 1.
c) The water will rise to the same level as in jar 1.

4. What was the reason for your answer to question 3?

a) The weights are the same size so they will take up equal amounts of space.
b) The heavier the metal weight, the higher the water will rise.
c) The heavy metal weight has more pressure, therefore the water will rise.
d) The heavier the metal weight, the lower the water will rise.

Appendix F: Student Survey Used in the Molecules of Life Courses

Please fill out your name, student identification number, and your answers to the following questions on a scantron bubble sheet. Your answers to the 16 questions below are appreciated as we work to improve this course.

1. How many years (including this one) have you attended a college or university?
A) 1 year **B)** 2 years **C)** 3 years **D)** 4 years **E)** more than 4 years

2. Are you a transfer student from another college or university?
A) Yes **B)** No

3. What is your major or intended major?
A) arts or humanities **B)** business or social science **C)** education **D)** health professions **E)** science or engineering

4. How much chemistry did you have in high school?
A) No chemistry in high school **B)** 1 semester **C)** 1 full year **D)** 1 -- 2 full years **E)** More than 2 full years

5. How much biology did you have in high school?
A) No biology in high school **B)** 1 semester **C)** 1 full year **D)** 1 -- 2 full years **E)** More than 2 full years

6. How many science courses have you taken at the college level (other than this course)?
A) No science courses in college **B)** 1 semester **C)** 1 full year **D)** 1 -- 2 full years **E)** More than 2 full years

7. How much math did you have in high school?
A) No math in high school **B)** 1 semester **C)** 1 full year **D)** 1 -- 2 full years **E)** More than 2 full years

8. How many math courses have you taken at the college level ?
A) No math courses in college **B)** 1 semester **C)** 1 full year **D)** 1 -- 2 full years **E)** More than 2 full years

9. How many hours per week do you plan to spend studying for this course outside of class time?
A) less than 1 hour **B)** 1-2 hours **C)** 3-4 hours **D)** 5-6 hours **E)** more than 6 hours

10. Considering both on-campus and off-campus paid employment, how many hours do you work per week?
A) I have no paid employment, **B)** 1-10 hours **C)** 11-20 hours **D)** 21-40 hours **E)** more than 40 hours

11. With regard to this course, which best describes you?
A) I am retaking this course. **B)** I am enrolled in this course for the first time.

12. What grade do you expect to earn in this course?
A) A **B)** B **C)** C **D)** D **E)** F

13. Are you:

A) Male **B)** Female

14. Are you a U.S. citizen?

A) Yes **B)** No

15. What Race/National Origin best describes you? (categories are taken from National Science Foundation):

A) American Indian or Alaskan Native
B) Native Hawaiian or other Pacific Islander
C) Asian
D) Black
E) White

16. Do you consider yourself Hispanic or Latino?

A) Yes **B)** No

Appendix G: Learning Goals for the Enzyme Module

Chapter 1: Chemical Reactions and Catalysis

After studying this chapter students should be able to:

1.1 Understand the concept of activation energy and reaction energetics.

1.2 Describe the overall pathway of a chemical reaction from reactants through the transition state to the final products.

1.3 Explain how a catalyst affects the rate of a chemical reaction.

Chapter 2 – Enzymes as Biological Catalysts

After studying this chapter students should be able to:

2.1 Describe the role of enzymes as catalysts for biological processes (including the life cycle of HIV).

2.2 Describe the structure and mechanism of HIV protease as an example of enzyme function.

2.3 Explain how enzymes bind specific substrates by comparing the "lock and key" model with the "induced fit" model.

2.4 Visualize and interpret three-dimensional molecular structures.

Chapter 3 – Enzymes and Drug Design

After studying this chapter students should be able to:

3.1 Explain the molecular principles of enzyme inhibition and apply them to HIV protease inhibitor drugs.

3.2 Compare the function of *Aspirin* and *Vioxx* as enzyme inhibitors.

3.3 Describe the stages by which a new pharmaceutical is developed, tested, and approved.

Six learning goals were used for questions in the content pre-test:

1.1, 1.2, 1.3, 2.1, 3.1, 3.3

All learning goals were used for questions in the post-test.

Appendix H: Enzymes and Drug Design Pretest (a.k.a. the Enzyme Pretest)

Question 1: Before cooking foods on a charcoal barbecue, you must first start the fire with a match or a lighter. Why is this necessary?

(a) The oxygen molecules in the air are not chemically reactive.
(b) The reaction requires an input of energy to get it started.
(c) Burning charcoal does not release heat.
(d) The charcoal always contains moisture that needs to evaporate.

Question 2: New cars in the U.S. are equipped with a catalytic converter. This device reduces the amount of poisonous carbon monoxide that is released into the air from the tailpipe. How does a catalytic converter work?

(a) It removes carbon monoxide from the exhaust gas by absorbing it like a sponge.
(b) It prevents the production of carbon monoxide from burning gasoline.
(c) It generates heat to ensure complete combustion of the gasoline.
(d) It speeds up the chemical conversion of carbon monoxide into carbon dioxide.

Question 3: An automobile engine obtains energy by burning gasoline in the presence of oxygen. What other component of the engine is necessary for this chemical reaction to occur?

(a) crankshaft
(b) spark plug
(c) piston
(d) tailpipe

Question 4: Consider a chemical reaction that takes 1 second to complete under normal conditions. Suppose that you are able to increase the rate of the reaction by a factor of 1 million. How many times could the chemical reaction be completed during one minute?

(a) 60
(b) 60 thousand
(c) 60 million
(d) 60 billion

Question 5: What is the general function of a biological enzyme?

(a) It preserves the structural integrity of the cell.
(b) It stores the cell's genetic information.
(c) It serves to transport oxygen from the lungs to other regions of the body.
(d) It acts as a biological catalyst by speeding up chemical reactions.

Question 6: What type of biological molecule typically functions as an enzyme?

(a) protein
(b) fat
(c) DNA
(d) sugar

Question 7: Suppose that you mix together two chemical compounds in a beaker. At room temperature you observe no changes. But when you heat the mixture to 80^{o}C, you observe a color change that indicates a chemical reaction has occurred. Why is the outcome different at the two temperatures?

(a) The heat changes the chemical structure of the molecules in the mixture.
(b) The heat changes the type of chemical reaction that takes place.
(c) The heat gives the molecules sufficient energy to react with each other.
(d) The heat breaks apart the molecules into their atomic components.

Question 8: What effect does a catalyst have on a chemical reaction?

(a) It increases the rate of the reaction.
(b) It changes the relative energies of reactants and products.
(c) It alters the type of chemical products generated by the reaction.
(d) It increases the amount of heat released by the reaction.

Question 9: How does a catalyst achieve its effect?

(a) By lowering the energy of the reactants
(b) By lowering the energy barrier for the reaction.
(c) By lowering the reaction rate.
(d) By lowering the energy of the products.

Question 10: Many people cannot tolerate eating dairy products. This condition is known as lactose intolerance since it arises from a type of sugar called lactose that is found in milk. What causes this condition?

(a) The enzyme for breaking down lactose is missing or reduced.
(b) The immune system triggers an allergic reaction to lactose.
(c) Bacteria in the intestines of affected people convert the lactose into a toxic product.
(d) It arises from excess acid that is generated in the stomach.

Question 11: Souring of milk is causes by bacteria that utilize sugars to generate acids. It is a common observation that milk kept in a refrigerator does not sour as rapidly as milk left on a kitchen table. What is the reason?

(a) The cold temperatures kill the bacteria.
(b) Enzymatic chemical reactions in the bacteria occur more slowly at colder temperatures.
(c) It is not possible to form acids via chemical reactions at colder temperatures.
(d) Milk gets thicker at lower temperatures and stops the bacteria from being mobile.

Question 12: Burning natural gas generates heat energy, which can be used to heat homes. This type of chemical reaction is called combustion. Where does the heat energy come from?

(a) The products of combustion are less energetically stable than the reactant molecules.
(b) The reactant molecules move more quickly than the product molecules.

(c) The combustion reaction absorbs energy from its surrounding environment.
(d) The products of combustion are more energetically stable than the reactant molecules.

Question 13: The name of an enzyme often tells you something about its function. One enzyme involved in HIV replication is called HIV proteinase (which is sometimes shortened to HIV protease). Based on its name, what is the likely function of this enzyme?

(a) It cuts up HIV's DNA molecules.
(b) It cuts up HIV's sugar molecules.
(c) It cuts up HIV's protein molecules.
(d) It cuts up HIV's RNA molecules.

Question 14: Which of the following statements is an accurate description of an **exothermic** chemical reaction?

(a) The energy of the products is higher than the energy of the reactants.
(b) The reaction absorbs heat energy from its surroundings.
(c) There is no energy change during the reaction.
(d) The reaction releases heat energy to the surroundings.

Question 15: What is the **transition state** of a chemical reaction?

(a) A molecular structure that is intermediate between reactants and products.
(b) A state that determines whether a reaction absorbs or releases heat.
(c) A state that contains one of the transition metals in the periodic table.
(d) A change of state from a liquid to a gas.

Question 16: Chemical reactions often have an "energy barrier" that exists between the reactants and the products. What is the relationship between the height of the energy barrier and the rate of a chemical reaction?

(a) A higher energy barrier corresponds to a faster chemical reaction.
(b) A lower energy barrier corresponds to a faster chemical reaction.
(c) The energy barrier does not affect the rate of the chemical reaction.
(d) A lower energy barrier corresponds to a slower chemical reaction.

Question 17: Why are some drug molecules called **competitive inhibitors**?

(a) They compete with the products to be released from the enzyme.
(b) They compete with the cell for essential nutrients.
(c) They compete with the reactant molecules that bind to the enzyme.
(d) These drugs are competitive in the marketplace.

Question 18: All pharmaceuticals are tested by clinical trials on patients. Suppose that a new painkiller drug – called *Cureall* - is being tested to see if it is **more effective** in reducing pain than an older medication like *Aspirin*. How would you set up a clinical trial to answer this question?

(a) Give *Cureall* to one patient group and a placebo to another patient group.
(b) Give *Aspirin* to one patient group and a placebo to another patient group.

(c) Give *Aspirin* to one patient group and *Cureall* to another patient group.
(d) Give *Cureall* to all patients.

Question 19: Why is it an advantage to have a drug that binds very tightly to an enzyme?

(a) The drug can be used in lower doses.
(b) The drug is cheaper to manufacture.
(c) The drug can be converted more easily to products.
(d) The drug can be used in higher doses.

Question 20: The LD_{50} value of a drug is a measure of toxicity and indicates the dose required to kill 50% of test animals. Suppose you are comparing the LD_{50} values of two drugs that are tested on mice. Drug A has an LD_{50} of 1 milligram (mg) and Drug B has an LD_{50} of 20 mg. Which drug is more toxic to mice?

(a) Drug A is more toxic than Drug B.
(b) Drug B is more toxic than Drug A.
(c) The two drugs are equally toxic.
(d) It's not possible to tell using these LD_{50} values.

Appendix I: Enzymes and Drug Design Posttest (a.k.a. the Enzyme Posttest)

SECTION 1: CHEMICAL REACTIONS AND CATALYSIS

Question 1: Before cooking foods on a charcoal barbecue, you must first start the fire with a match or a lighter. Why is this necessary?

(a) The oxygen molecules in the air are not chemically reactive
(b) The reaction requires an input of energy to get it started
(c) Burning charcoal does not release heat
(d) The charcoal always contains moisture that needs to evaporate

Question 2: Suppose that you mix together two chemical compounds in a beaker. At room temperature you observe no changes. But when you heat the mixture to 80^{o}C, you observe a color change that indicates a chemical reaction has occurred. Why is the outcome different at the two temperatures?

(a) The heat changes the chemical structure of the molecules in the mixture
(b) The heat changes the type of chemical reaction that takes place
(c) The heat gives the molecules sufficient energy to react with each other
(d) The heat breaks apart the molecules into their atomic components

Question 3: An automobile engine obtains energy by burning gasoline in the presence of oxygen. What other component of the engine is necessary for this chemical reaction to occur?

(a) crankshaft
(b) spark plug
(c) piston
(d) tailpipe

Question 4: Which of the following statements is an accurate description of an **exothermic** chemical reaction?

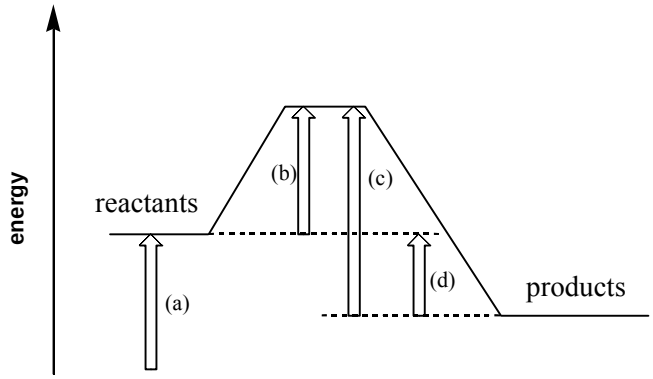
(a) The energy of the products is higher than the energy of the reactants
(b) The reaction absorbs heat energy from its surroundings
(c) There is no energy change during the reaction
(d) The reaction releases heat energy to the surroundings

Question 5: Burning natural gas generates heat energy, which can be used to heat homes. This type of chemical reaction is called combustion. Where does the heat energy come from?

(a) The products of combustion are less energetically stable than the reactant molecules
(b) The reactant molecules move more quickly than the product molecules
(c) The combustion reaction absorbs energy from its surrounding environment

(d) The products of combustion are more energetically stable than the reactant molecules

Question 6: The diagram below shows the energy profile of a chemical reaction. Which of the labeled arrows shows the **activation energy** for the reaction?



(a) **(b)** (c) (d)

Question 7: In the same diagram, which labeled arrow shows the **reaction energy**?

(a) (b) (c) **(d)**

Question 8: Chemical reactions often have an "energy barrier" that exists between the reactants and the products. What is the relationship between the height of the energy barrier and the rate of a chemical reaction?

(a) A higher energy barrier corresponds to a faster chemical reaction
(b) A lower energy barrier corresponds to a faster chemical reaction
(c) The energy barrier does not affect the rate of the chemical reaction
(d) A lower energy barrier corresponds to a slower chemical reaction

Question 9: What is the **transition state** of a chemical reaction?

(a) A molecular structure that is intermediate between reactants and products
(b) A state that determines whether a reaction absorbs or releases heat
(c) A state that contains one of the transition metals in the periodic table
(d) A change of state from a liquid to a gas

Question 10: Consider a chemical reaction with a transition state that is very unstable (in energetic terms). Which of the following characteristics would you predict for this reaction?

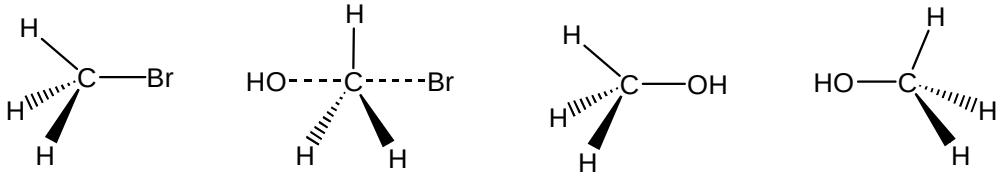
(a) The reaction has a very slow rate
(b) The reaction is exothermic
(c) The reaction has a very fast rate

(d) The reaction is endothermic

Question 11: The equation below describes a substitution reaction:

$$\underset{\text{hydroxide ion}}{OH^-} + \underset{\text{methyl bromide}}{CH_3Br} \rightarrow \underset{\text{methanol}}{CH_3OH} + \underset{\text{bromide ion}}{Br^-}$$

Which of the following structures depicts the **transition state** for the reaction?

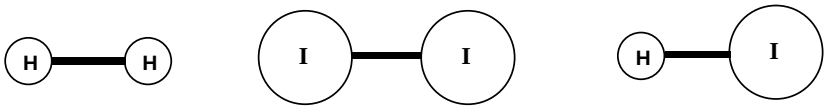


(a) (b) (c) (d)

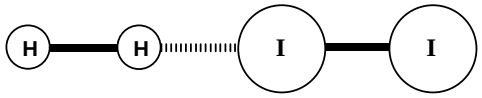
Question 12: Consider a chemical reaction between molecular hydrogen (H_2) and molecular iodine (I_2) to produce hydrogen iodide H I.

$$\mathbf{H_2 + I_2 \rightarrow 2\,H\,I}$$

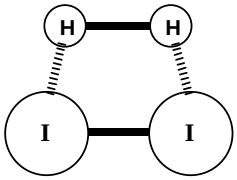
The H_2, I_2, and HI molecules are shown below (Note: Iodine is drawn larger that hydrogen because its atomic radius is greater).



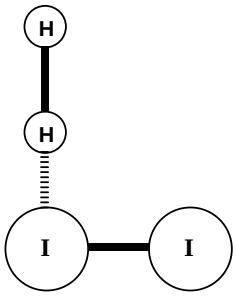
What structure do you predict for the transition state of the chemical reaction between H_2 and I_2?



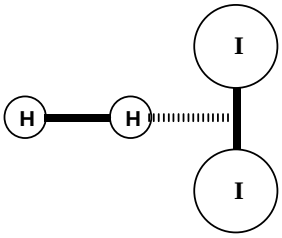
(a)



(b)



(c)



(d)

Question 13: What effect does a catalyst have on a chemical reaction?

(a) It increases the rate of the reaction.
(b) It changes the relative energies of reactants and products.
(c) It alters the type of chemical products generated by the reaction.
(d) It increases the amount of heat released by the reaction.

Question 14: How does a catalyst achieve its effect?

(a) By lowering the energy of the reactants
(b) By lowering the energy barrier for the reaction.
(c) By lowering the reaction rate.
(d) By lowering the energy of the products.

Question 15: Consider a chemical reaction that takes 1 second to complete under normal conditions. Suppose that you are able to increase the rate of the reaction by a factor of 1 million. How many times could the chemical reaction be completed during one minute?

(a) 60
(b) 60 thousand
(c) 60 million
(d) 60 billion

Question 16: New cars in the U.S. are equipped with a catalytic converter. This device reduces the amount of poisonous carbon monoxide that is released into the air from the tailpipe. How does a catalytic converter work?

(a) It removes carbon monoxide from the exhaust gas by absorbing it like a sponge.
(b) It prevents the production of carbon monoxide from burning gasoline.
(c) It generates heat to ensure complete combustion of the gasoline.
(d) It speeds up the chemical conversion of carbon monoxide into carbon dioxide.

SECTION 2: ENZYMES AS BIOLOGICAL CATALYSTS

Question 17: What is the general function of a biological enzyme?

(a) It preserves the structural integrity of the cell
(b) It stores the cell's genetic information
(c) It serves to transport oxygen from the lungs to other regions of the body
(d) It acts as a biological catalyst by speeding up chemical reactions

Question 18: What type of biological molecule typically functions as an enzyme?

(a) protein
(b) fat
(c) DNA
(d) sugar

Question 19: Souring of milk is causes by bacteria that utilize sugars to generate acids. It is a common observation that milk kept in a refrigerator does not sour as rapidly as milk left on a kitchen table. What is the reason?

(a) The cold temperatures kill the bacteria.
(b) Enzymatic chemical reactions in the bacteria occur more slowly at colder temperatures.
(c) It is not possible to form acids via chemical reactions at colder temperatures.
(d) Milk gets thicker at lower temperatures and stops the bacteria from being mobile.

Question 20: HIV's genetic information is stored in the form of RNA. What enzyme is responsible for copying this RNA into DNA within an HIV-infected cell?

(a) HIV protease
(b) HIV integrase
(c) HIV transcriptase
(d) reverse transcriptase

Question 21: Why is the HIV protease enzyme essential for HIV to replicate itself within an infected cell?

(a) It enables HIV to escape attack by the body's immune system.

(b) It enables the integration of HIV's genetic information into the DNA of the infected cell.
(c) It cuts a large polypeptide chain into smaller chains so the virus can assemble.
(d) It enables HIV to bind to the surface of the immune cells that it infects.

Question 22: What type of chemical reaction does the HIV protease enzyme catalyze in order to break a peptide bond?

(a) hydrolysis
(b) polymerization
(c) condensation
(d) substitution

Question 23: In the first step of the HIV protease mechanism, an aspartic acid sidechain removes a proton from a nearby water molecule. Why is this necessary?

(a) H_2O molecules cannot react with the peptide bond.
(b) The resulting ^{-}OH ion is more reactive than H_2O.
(c) The aspartic acid sidechain is not stable unless it has a proton attached to it.
(d) The binding pocket of HIV protease is hydrophobic and cannot accommodate H_2O.

Question 24: For the chemical reaction catalyzed by HIV protease, how does the structure of the transition state compare with the structure of the substrate?

(a) The substrate and transition state are both planar.
(b) The substrate is planar and the transition state is trigonal bipyramidal.
(c) The substrate and transition state are both tetrahedral.
(d) The substrate is planar and the transition state is tetrahedral.

Question 25: In order to achieve its catalytic effect, which molecular structure in the reaction pathway does the HIV protease enzyme bind most tightly?

(a) the substrate(s)
(b) the transition state
(c) the products(s)
(d) the binding is equal for all of them

Question 26: For any catalyst – including enzymes – it is important to quickly release the products of the reaction. Why?

(a) The products must be released to allow the catalytic cycle to begin again.
(b) The chemical energy of the products is very high.
(c) The products of the reaction are damaging to the enzyme.
(d) Bound products react with the starting materials to reverse the chemical reaction.

Question 27: What type of interaction is described by the "lock-and-key" model of enzyme function?

(a) The substrate is the lock and the enzyme is the key.
(b) The enzyme is the lock and the substrate is the key.
(c) The enzyme is the lock and an active site amino acid sidechain is the key.
(d) Two substrates in the active site act as the lock and the key

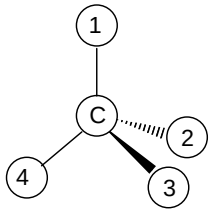
Question 28: How does the "induced fit" model for enzyme function differ from the "lock-and-key" model?

(a) The induced fit model applies only to large enzymes with multiple subunits.
(b) The induced fit model does not utilize complementary molecular interactions.
(c) The two models are equivalent but emphasize different aspects of the enzyme reaction.
(d) The induced fit model involves a structural change of the enzyme during binding.

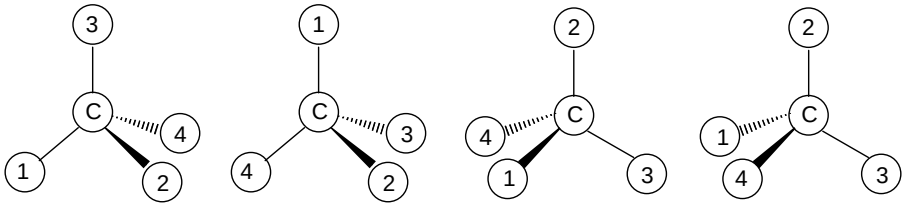
Question 29: Based on the principle of complementary chemical interactions, what type of amino acid sidechain in the enzyme active site would form the most favorable interaction with an $-NH_3^+$ group on a substrate molecule?

(a) A nonpolar sidechain containing a $-CH_3$ group.
(b) A polar sidechain containing an –OH or NH_2 group
(c) A charged sidechain containing a $-COO^-$ group.
(d) A charged sidechain containing a $-NH_3^+$ group.

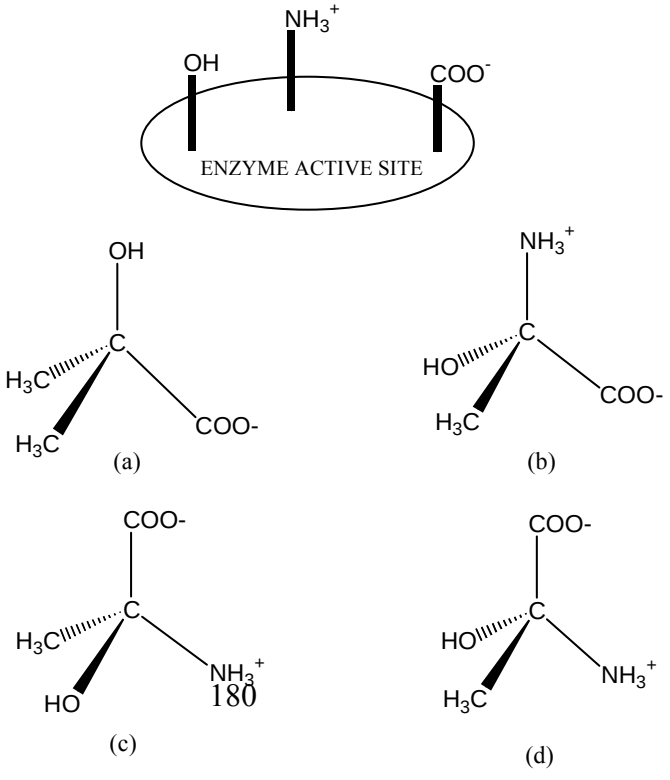
Question 30: The function of biological molecules or drug molecules typically depends on their **specific three-dimensional structure**. The 3-D structure of a hypothetical molecule is shown below. This molecule contains a central carbon at (C) that is bonded to four different chemical groups (represented by 1, 2, 3, and 4).



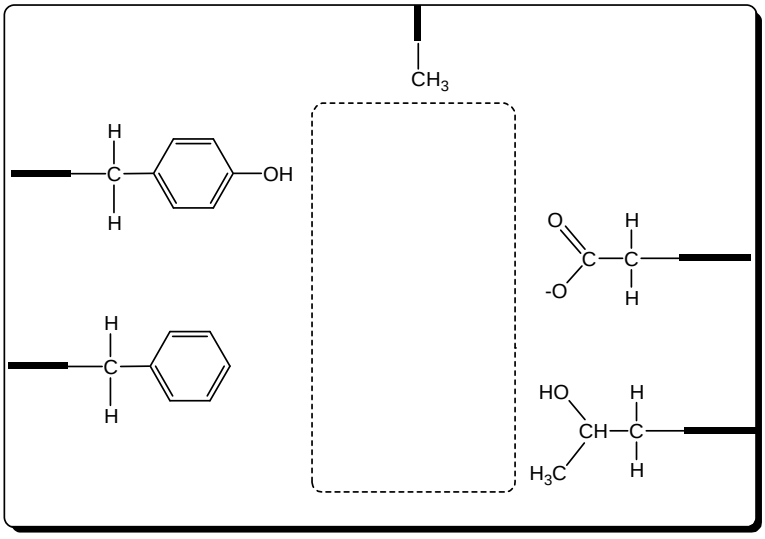
Suppose that this molecule is rotated in three-dimensional space. Which of the drawings below corresponds to the same molecule?



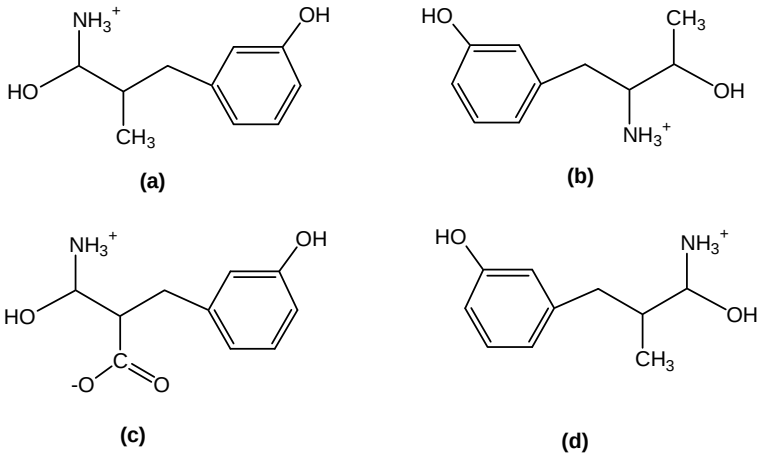
Question 31: Enzymes bind to substrate molecules using complementary chemical interactions. The active site of an enzyme is shown schematically below, with three functional groups arranged in a particular geometry. Which of the substrates shows the tightest binding to the active site by maximizing the number of complementary interactions? *(Note: You may need to rotate the molecules in space to get the best fit).*



Question 32: The diagram below shows the active site of an enzyme. The space for binding the substrate molecule is shown as a dashed box. Several amino acid sidechains are shown that can interact with the substrate.



Based on complementary chemical interactions, which of the substrate molecules below would achieve the best binding to the amino acid sidechains in the enzyme active site? You can assume that the molecule remains planar and you may need to rotate the structure in space.



SECTION 3: ENZYMES AND DRUG DESIGN

Question 33: Why are some drug molecules called **competitive inhibitors**?

(a) They compete with the products to be released from the enzyme
(b) They compete with the cell for essential nutrients.
(c) They compete with the reactant molecules that bind to the enzyme

(d) These drugs are competitive in the marketplace.

Question 34: What type of drug molecule would function as an effective enzyme inhibitor by binding most tightly to the enzyme active site?

(a) A drug molecule that resembles the substrate
(b) A drug molecule that resemble the transition state
(c) A drug molecule that resembles the product.
(d) A drug molecule that resembles the enzyme

Question 35: Why is it an advantage to have a drug that binds very tightly to an enzyme?

(a) The drug can be used in lower doses
(b) The drug is cheaper to manufacture
(c) The drug can be converted more easily to products.
(d) The drug can be used in higher doses

Question 36: HIV protease inhibitors are a successful class of pharmaceuticals. What molecular feature enables these drugs to function as a potent inhibitor of the HIV protease enzyme?

(a) They contain regions that resemble sidechains in the peptide substrate for HIV protease.
(b) They have a structural region that mimics the planar peptide bond.
(c) They contain a tetrahedral structure similar to the transition state for the reaction.
(d) They resemble the product of the reaction that cleaves the peptide bond.

Question 37: HIV protease inhibitors often become less effective over time even though the patient continues to take them. What causes this effect?

(a) The patient's immune system changes over time to become resistant to the drug
(b) The HIV population evolves to increase the proportion of drug resistant strains
(c) The drug molecules are broken down over time within the patient's body
(d) Each virus particle changes its genetic make-up in response to exposure to the drug

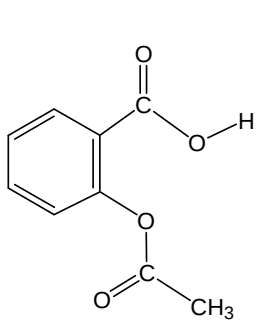
Question 38: The **cyclo-oxygenase enzyme** is the target of painkiller drugs such as *Aspirin*. How is this enzyme involved in the creation of painful sensations?

(a) It facilitates production of prostaglandins that stimulate inflammation.
(b) It increases the transmission of pain impulses within the brain.
(c) It inhibits the production of endorphins that modulate the body's response to pain.
(d) It synthesizes molecules that produce a natural anesthetic in our brains.

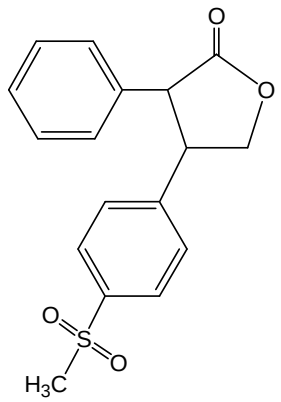
Question 39: Drugs like *Vioxx* and *Celebrex* were developed to reduce the unpleasant side-effects of older drugs like Aspirin. How do these new drugs function at a molecular level?

(a) They selectively inhibit the COX-1 enzyme that produces protective prostaglandins.
(b) They target all COX enzyme variants and inhibit both COX-1 and COX-2.
(c) They selectively inhibit the COX-2 enzyme that produces inflammatory prostaglandins.
(d) They reduce the formation of blot clots.

Question 40: The molecular structures of *Aspirin* and *Vioxx* are shown below. What feature accounts for the selective action of *Vioxx* as compared to *Aspirin*.



aspirin



vioxx

(a) *Vioxx* contains a cyclic ring of 6 carbon atoms whereas *Aspirin* does not.
(b) *Vioxx* is larger than *Aspirin* and fits into the bigger active site of the COX-2 enzyme.
(c) *Aspirin* does not bind to the COX-2 enzyme and therefore does not inhibit its function.
(d) A bulky amino acid sidechain in the COX-2 active site allows *Vioxx* to bind.

Question 41: The LD_{50} value of a drug is a measure of toxicity and indicates the dose required to kill 50% of test animals. Suppose you are comparing the LD_{50} values of two drugs that are tested on mice. Drug A has an LD_{50} of 1 milligram (mg) and Drug B has an LD_{50} of 20 mg. Which drug is more toxic to mice?

(a) Drug A is more toxic than Drug B
(b) Drug B is more toxic than Drug A
(c) The two drugs are equally toxic
(d) It's not possible to tell using these LD_{50} values

Question 42: All pharmaceuticals are tested by clinical trials on patients. Suppose that a new painkiller drug – called *Cureall* - is being tested to see if it is **more effective** in reducing pain than an older medication like *Aspirin*. How would you set up a clinical trial to answer this question?

(a) Give *Cureall* to one patient group and a placebo to another patient group
(b) Give *Aspirin* to one patient group and a placebo to another patient group
(c) Give *Aspirin* to one patient group and *Cureall* to another patient group
(d) Give *Cureall* to all patients.

Question 43: Suppose that a pharmaceutical company is proposing *Cureall* for approval by the Food and Drug Administration (FDA). They present the FDA with data from clinical trials that compared two groups of patients. One patient group took *Cureall* and the other group took *Aspirin*. Which of the following questions could not be answered from these data?

(a) Does *Cureall* reduce pain more effectively than *Aspirin*?
(b) Does *Cureall* produce fewer problems with stomach irritation than *Aspirin*?
(c) Does *Cureall* show no additional benefit with pain reduction compared to *Aspirin*?
(d) Does *Cureall* increase the risk of heart attack and stroke?

SECTION 4: INTEGRATIVE QUESTIONS

Please provide written answers to the following two questions.

Question 44: An outerspace probe has discovered a new type of organism that can derive its energy needs from the following chemical reaction:

$$Cl_2 + CH_4 \rightarrow CH_3Cl + HCl$$

Performing this reaction in the laboratory requires high temperatures and a metal catalyst. But the newly-discovered organism is able to utilize the reaction at room temperature. Discuss how this could be possible using the concepts covered in the course.

Question 45: The typical enzyme is a large biological molecule, often with hundreds or thousands of amino acid components. But the active site that carries out the catalytic function is much smaller and involves perhaps 10 amino acids. Propose two reasons to explain why enzymes are so large.

Appendix J: Faculty Survey for Molecules of Life (MOL)

This survey is designed to ask for information about how you taught the topic of **3-D molecular structure** in the context of the *Molecules of Life* curriculum materials. We want to learn about what methods you used and what you believe was effective in helping students improve their 3-D visualization skills, as measured by the Purdue Rotation pre/post tests. Please type your results in the spaces below and provide as much **detail** as possible. Thank you very much for your assistance.

Question 1: Course Information

How did you teach the *Molecules of Life* course materials? Was it a **semester-long course** or did you use the *Enzyme & Drug Design* **module** as part of a chemistry or biology course? If you taught the module, how many class periods were devoted to the MOL materials?

Question 2: Course Content

(a) Which topics within the **module chapters** do you believe contributed to helping students develop spatial ability? A sample of the relevant chapter sections is given below.

- ***Chapter 1. Reactions & Catalysts*** *(Comparing catalysts for N_2 fixation – Haber process & N_2-fixing bacteria; Chemical reactions – activation energy & transition state)*

- ***Chapter 2. Enzymes as Biological Catalysts*** *(Role of enzymes in HIV infection; HIV protease as a model enzyme; Principles of enzyme function, e.g. lock & key, induced fit)*

- ***Chapter 3. Enzymes & Drug Design*** *(Designing an effective anti-HIV drug – chemical & biological principles; How do HIV-protease inhibitors work?*

(b) Were there any content topics from **other parts of the course** (i.e., NOT the module) that you believe contributed to improving your students' spatial ability?

Question 3: Class & Lab Activities

Did you ask your student to perform any of the following activities in your **classroom or in a lab session?** Please answer **YES** or **NO** for each one and indicate whether you used it in the class or a lab. Any details about the activities you performed would be helpful.

- Drawing molecular structures

- Building models using model kits

- Using computer graphics software

- Did you have student perform any other type of activity? If so, please describe it.

Question 4: Effect of the Activities

Which activities and/or experiments do you think were especially helpful for improving your students' spatial ability? Please explain.

Question 5: Further Comments?

Do you have any further comments and insights on what might have worked, or did not work, in terms of potential improvement of your students' spatial ability? We would be grateful for any additional information.

Appendix K: The MOL Textbook: Table of Contents

(Note: Section 6 below is referred to as the "enzyme module".)

About the Author

Bo Jiang received a Bachelor of Science degree in Chemistry from Peking University in Beijing, China in June 1999. After that he attended Rensselaer Polytechnic Institute in Troy, NY and obtained a Master of Science degree in Computational Chemistry in May 2002. He then attended the University of South Florida (USF) and received a Master of Computer Science degree in May 2004. He entered the Ph.D. program in Chemistry Education at USF in August 2004.

Bo was awarded an "Investing in the Future" Fellowship at USF for 2004-2006. He held numerous graduate assistantships and taught many undergraduate laboratory courses at USF, including General Chemistry I & II, and Organic Chemistry II Lab. He has presented at several academic conferences, including the Molecules of Life National Dissemination Conference, and the 20th Biennial Conference on Chemical Education sponsored by the Division of Chemical Education of the American Chemical Society (ACS).